IBM SPSS Modeler Text Analytics 17.1 - Guida dell'utente

Nota
Prima di utilizzare queste informazioni e il relativo prodotto, leggere le informazioni disponibili in “Informazioni particolari” a pagina 237.

Informazioni sul prodotto

Questa edizione si applica alla versione 17.1, release 0, livello di modifica 0 di IBM SPSS Modeler Text Analytics e a tutti i successivi release e modifiche fino a eventuali disposizioni contrarie indicate in nuove edizioni.

Indice

Prefazione

IBM® SPSS Modeler Text Analytics offre potenti funzionalità di analisi testuale, che utilizzano tecnologie linguistiche avanzate e di Natural Language Processing (NLP) per elaborare rapidamente una grande varietà di dati di testo non strutturati e, da questo testo, estrarre e organizzare i concetti chiave. Inoltre, IBM SPSS Modeler Text Analytics può raggruppare tali concetti in categorie.

Circa l'80% dei dati di un'organizzazione vengono conservati sotto forma di documenti di testo, ad esempio, report, pagine Web, posta elettronica e note di call center. Il testo è un fattore chiave per consentire a un'organizzazione di acquisire una migliore comprensione del comportamento dei propri clienti. Un sistema che incorpora la tecnologia NLP è in grado di estrarre in modo intelligente termini e persino frasi composte. Inoltre, la conoscenza della lingua sottostante consente la classificazione dei termini in gruppi affini, quali prodotti, organizzazioni o persone utilizzando il significato e il contesto del termine. È possibile determinare la rilevanza delle informazioni in base alle proprie esigenze. È quindi possibile combinare i concetti e le categorie estratti con dati strutturati esistenti, per esempio dati demografici, e applicarli alla modellazione utilizzando la suite completa degli strumenti di data mining di IBM SPSS Modeler per prendere decisioni migliori e più mirate.

I sistemi linguistici sono sensibili alla conoscenza: più informazioni sono contenute nei dizionari, più elevata è la qualità dei risultati. IBM SPSS Modeler Text Analytics viene distribuito con una serie di risorse linguistiche, come dizionari per i termini e i sinonimi, le librerie e i modelli. Questo prodotto consente inoltre di sviluppare e affinare queste risorse linguistiche in base al proprio contesto. L'adattamento delle risorse linguistiche è spesso un processo iterativo ed è necessario per il recupero accurato e la categorizzazione dei concetti. Sono inclusi inoltre modelli, librerie e dizionari personalizzati per domini specifici, quali CRM e genomica.

Informazioni su IBM Business Analytics

Il software IBM Business Analytics fornisce informazioni complete, coerenti e accurate a cui i responsabili delle decisioni possono affidarsi per ottimizzare le prestazioni di business. Un ampio portafoglio di applicazioni di business intelligence, analisi predittiva, gestione delle prestazioni e delle strategie finanziarie e analisi offre una panoramica chiara, istantanea e interattiva delle prestazioni attuali e la possibilità di prevedere i risultati futuri. Utilizzato in combinazione con potenti soluzioni di settore, prassi consolidate e servizi professionali, questo software consente alle aziende di tutte le dimensioni di ottimizzare la produttività, automatizzare le decisioni senza problemi e fornire risultati migliori.

Come parte di questo portafoglio, il software IBM SPSS Predictive Analytics consente alle aziende di prevedere gli eventi futuri e di agire tempestivamente in modo da migliorare i risultati delle attività di business. Le aziende, gli enti governativi e le università di tutto il mondo si affidano alla tecnologia IBM SPSS perché rappresenta un vantaggio concorrenziale in termini di attrazione, retention e aumento dei clienti, riducendo al tempo stesso le frodi e limitando i rischi. Incorporando il software IBM SPSS nelle attività quotidiane, le aziende diventano imprese in grado di effettuare previsioni e di gestire e automatizzare le decisioni, per raggiungere gli obiettivi di business e vantaggi tangibili sulla concorrenza. Per ulteriori informazioni o per contattare un rappresentante, visitare il sito http://www.ibm.com/spss.

Supporto tecnico

Il supporto tecnico è a disposizione dei clienti che dispongono di un contratto di manutenzione. I clienti possono contattare il supporto tecnico per richiedere assistenza per l'utilizzo di IBM Corp. o per l'installazione di uno degli ambienti hardware supportati. Per contattare il supporto tecnico, visitare il sito Web IBM Corp. all'indirizzo http://www.ibm.com/support. Ricordare che durante la richiesta di assistenza sarà necessario fornire i dati di identificazione personali, i dati relativi alla propria società e il numero del accordo di manutenzione.

Capitolo 1. Informazioni su IBM SPSS Modeler Text Analytics

IBM SPSS Modeler Text Analytics offre potenti funzionalità di analisi testuale, che utilizzano tecnologie linguistiche avanzate e di Natural Language Processing (NLP) per elaborare rapidamente una grande varietà di dati di testo non strutturati e, da questo testo, estrarre e organizzare i concetti chiave. Inoltre, IBM SPSS Modeler Text Analytics può raggruppare tali concetti in categorie.

Circa l'80% dei dati di un'organizzazione vengono conservati sotto forma di documenti di testo, ad esempio, report, pagine Web, posta elettronica e note di call center. Il testo è un fattore chiave per consentire a un'organizzazione di acquisire una migliore comprensione del comportamento dei propri clienti. Un sistema che incorpora la tecnologia NLP è in grado di estrarre in modo intelligente termini e persino frasi composte. Inoltre, la conoscenza della lingua sottostante consente la classificazione dei termini in gruppi affini, quali prodotti, organizzazioni o persone utilizzando il significato e il contesto del termine. È possibile determinare la rilevanza delle informazioni in base alle proprie esigenze. È quindi possibile combinare i concetti e le categorie estratti con dati strutturati esistenti, per esempio dati demografici, e applicarli alla modellazione utilizzando la suite completa degli strumenti di data mining di IBM SPSS Modeler per prendere decisioni migliori e più mirate.

I sistemi linguistici sono sensibili alla conoscenza: più informazioni sono contenute nei dizionari, più elevata è la qualità dei risultati. IBM SPSS Modeler Text Analytics viene distribuito con una serie di risorse linguistiche, come dizionari per i termini e i sinonimi, le librerie e i modelli. Questo prodotto consente inoltre di sviluppare e affinare queste risorse linguistiche in base al proprio contesto. L'adattamento delle risorse linguistiche è spesso un processo iterativo ed è necessario per il recupero accurato e la categorizzazione dei concetti. Sono inclusi inoltre modelli, librerie e dizionari personalizzati per domini specifici, quali CRM e genomica.

Distribuzione. È possibile distribuire flussi di estrazione testo utilizzando IBM SPSS Modeler Solution Publisher per il calcolo del punteggio in tempo reale dei dati non strutturati. La possibilità di distribuire questi flussi garantisce implementazioni efficaci di estrazione testo a circuito chiuso. Ad esempio, l'organizzazione può analizzare ora note da chiamanti in entrata o in uscita applicando i modelli predittivi per aumentare la precisione dei propri messaggi di marketing in tempo reale.

Nota: per eseguire IBM SPSS Modeler Text Analytics con IBM SPSS Modeler Solution Publisher, aggiungere la directory `<install_directory>/ext/bin/spss.TMWBServer` alla variabile d'ambiente `$LD_LIBRARY_PATH`.

Traduzione automatica delle lingue supportate. IBM SPSS Modeler Text Analytics, in collaborazione con SaaS (SDL Software as a Service), è possibile tradurre il testo da un elenco di lingue supportate, incluse arabo, cinese e persiano, in inglese. È possibile eseguire l'analisi di testo del testo tradotto e distribuire questi risultati a persone che non comprendono il contenuto della lingua di origine. Poiché i risultati di estrazione testo vengono collegati automaticamente al testo di lingua straniera corrispondente, l'azienda è in grado di focalizzare le risorse di madrelingua più necessarie solo sui risultati più significativi dell'analisi. SDL offre una traduzione automatica di lingua utilizzando algoritmi statistici scaturiti da anni e anni di ricerche avanzate nel campo della traduzione.

Aggiornamento a IBM SPSS Modeler Text Analytics Versione 17.1

Aggiornamento da una versione precedente di PASW Text Analytics o Text Mining for Clementine.

Prima di installare IBM SPSS Modeler Text Analytics versione 17.1 è necessario salvare ed esportare i file TAP, i modelli e le librerie della versione attuale che si desidera utilizzare nella nuova versione. Si consiglia di salvare questi file in una directory che non sarà cancellata né sovrascritta durante l'installazione dell'ultima versione.

Dopo aver installato l'ultima versione di IBM SPSS Modeler Text Analytics è possibile caricare i file TAP salvati, aggiungere le eventuali librerie salvate o importare e caricare i modelli salvati per poterli usare nell'ultima versione.

Importante: Se si disinstalla la versione corrente senza salvare ed esportare prima i file necessari, l'eventuale lavoro svolto con i file TAP, i modelli, e le librerie pubbliche nella precedente versione andrà perso e non potrà essere utilizzato in IBM SPSS Modeler Text Analytics versione 17.1.

Informazioni sull'estrazione testo

Oggi una crescente quantità di informazioni viene mantenuta in formati non strutturati e semistrutturati, come e-mail del cliente, note di call center, risposte a sondaggi, flussi di notizie, moduli Web eccetera. Questa abbondanza di informazioni pone un problema per molte organizzazioni che si chiedono "Come possiamo raccogliere, esplorare ed utilizzare queste informazioni?"

L'*estrazione testo* è il processo di analisi delle raccolte di materiali di testo che consentono di catturare concetti chiave e temi e scoprire le relazioni nascoste e le tendenze senza che si conosca l'esatta parole o i termini che gli autori hanno utilizzato per esprimere questi concetti. Sebbene siano molto diverse, il processo di estrazione testo viene talvolta confuso con il recupero di informazioni. Laddove il recupero accurato e l'archiviazione delle informazioni è una sfida enorme, l'estrazione e la gestione dei contenuti qualitativi, la terminologia e le relazioni contenute all'interno delle informazioni sono processi fondamentali e critici.

Estrazione testo e data mining

Per ogni articolo del testo, l'estrazione testo basata sulla lingua restituisce un indice di concetti, nonché di informazioni relative a questi concetti. Queste informazioni distillate e strutturate possono essere combinate con altre origini di dati per affrontare questioni quali:

- Quale concetti si verificano insieme?
- A cos'altro sono collegati?
- Quali categorie di alto livello possono essere create dalle informazioni estratte?
- Cosa possono predeterminare i concetti o le categorie?
- Come i concetti o le categorie determinano i comportamenti?

La combinazione di estrazione testo con il data mining offre una maggiore percezione di quanta ne offrono i dati strutturati o non strutturati. Questo processo prevede le seguenti fasi:

1. **Identificazione del testo da estrarre.** Preparare il testo per l'estrazione. Se il testo esiste in più file, salvare i file in un'ubicazione singola. Per i database, determinare il campo contenente il testo.
2. **Estrarre il testo e i dati strutturati.** Applicare gli algoritmi di estrazione testo al testo di origine.
3. **Creare modelli di concetti e di categorie.** Identificare i concetti chiave e/o creare le categorie. Il numero di concetti restituiti dai dati non strutturati è generalmente molto grande. Identificare i migliori concetti e categorie per il calcolo del punteggio.
4. **Analizzare i dati strutturati.** Utilizzare le tecniche di data mining tradizionali, come il raggruppamento, classificazione e modelli predittivi, per rilevare le relazioni tra i concetti. Unire i concetti estratti con altri dati strutturati per prevedere il comportamento futuro in base ai concetti.

Analisi e categorizzazione del testo

L'analisi del testo, una forma di analisi qualitativa, è l'estrazione di informazioni utili dal testo che consente di raggruppare idee o concetti chiave contenuti nel testo in un numero appropriato di categorie. L'analisi del testo possono essere eseguite su tutti i tipi e le lunghezze di testo, sebbene l'approccio all'analisi può variare.

I record brevi o documenti vengono categorizzati più facilmente, poiché non sono molto complessi e di solito contengono meno parole e risposte ambigue. Ad esempio, se con domande brevi e aperte, si chiede alle persone di rispondere sulle loro tre attività di tempo libero favorite, ci si potrebbe aspettare di vedere molte risposte brevi, ad esempio *spiaggia*, *visita ai parchi pubblici* o *non fare nulla*. Risposte più lunghe, aperte, d'altro canto, possono essere abbastanza complesse e molto lunghe, soprattutto se gli intervistati sono istruiti, motivati e hanno tempo sufficiente per completare il questionario. Se si chiede alla gente di parlare delle loro convinzioni politiche in un sondaggio o stabilire un blog politico, potrebbero risultare lunghe osservazioni su ogni sorta di temi e posizioni.

La possibilità di estrarre concetti chiave e creare categorie profonde da queste origini di testo più lunghe in un periodo di tempo molto breve, è un vantaggio chiave dell'uso di IBM SPSS Modeler Text Analytics. Questo vantaggio è dato dalla combinazione di tecniche linguistiche e statistiche che produce risultati più affidabili per ciascuna fase del processo di analisi del testo.

Elaborazione linguistica e NLP

Il problema principale con la gestione di tutti questi dati di testo non strutturati è che non esistono norme standard per la scrittura di testo che un computer può comprendere. La lingua, e quindi il significato, varia per ogni documento e ogni blocco di testo. L'unico modo per richiamare correttamente e organizzare tali dati non strutturati, è analizzare la lingua e quindi scoprire il suo significato. Esistono diversi approcci automatizzati per l'estrazione di concetti di informazioni non strutturate. Tali approcci possono essere suddivisi in due tipi, linguistici e non linguistici.

Alcune aziende hanno cercato di impiegare soluzioni non linguistiche automatizzate basate su statistiche e reti neurali. Utilizzando la tecnologia del computer, queste soluzioni possono eseguire la scansione dei concetti chiave e categorizzarli più rapidamente rispetto a semplici lettori. Purtroppo, la precisione di tali soluzioni è molto basso. La maggior parte dei sistemi basati su statistiche conteggiano semplicemente il numero di volte in cui le parole ricorrono e calcolano le approssimazioni statistiche ai concetti correlati. Essi producono molti risultati non pertinenti o inutili e perdono i risultati che devono invece trovare, quelli rilevanti.

Per compensare la loro limitata precisione, alcune soluzioni incorporano regole non linguistiche complesse che aiutano a distinguere tra risultati rilevanti e irrilevanti. Ciò viene indicato come *estrazione di testo basato su regole*.

L'estrazione di testo basata sulla linguistica, d'altro canto, applica i principi di elaborazione lingua naturale (NLP) -l'analisi computerizzata delle lingue umane- all'analisi delle parole, frasi e della sintassi o struttura del testo. Un sistema che incorpora la tecnologia NLP è in grado di estrarre in modo intelligente concetti, compreso frasi composte. Inoltre, la conoscenza della lingua sottostante consente la classificazione dei concetti in gruppi affini, quali prodotti, organizzazioni o persone utilizzando il significato e il contesto.

L'estrazione di testo basato sulla linguistica rileva il significato del testo come succede per la mente umana, riconoscendo una varietà di forme di parole con significati simili e analizzando la struttura della frase per fornire un quadro per comprendere il testo. Questo approccio offre la rapidità ed efficacia dei sistemi basati sulle statistiche, ma offre un grado molto più elevato di precisione e richiede un intervento umano molto minore.

Per illustrare la differenza tra approcci basati sulle statistiche e quelli basati sulla linguistica durante il processo di estrazione con tutti i testi in lingua escluso giapponese, considerare in che modo ciascuno risponderebbe ad una interrogazione sulla `riproduzione dei documenti`. Entrambe le soluzioni basate sulle statistiche e basate su linguistica dovranno espandere la parola `riproduzione` per includere i sinonimi, ad esempio `copia` e `duplicazione`. In caso contrario, le informazioni relative verranno trascurate. Ma se una soluzione basata sulle statistiche tenta di eseguire questo tipo di ricerca di sinonimi per altri termini con lo stesso significato è probabile che includano anche il termine `nascita`, generando un certo numero di risultati non pertinenti. La comprensione linguistica riduce l'ambiguità del testo, rendendo, per definizione, l'estrazione testo basata sulla linguistica l'approccio più affidabile.

L'uso delle tecniche basate sulla linguistica tramite l'analisi di opinione rende possibile estrarre più espressioni significative. L'analisi e la cattura di emozioni riduce l'ambiguità del testo, rendendo, per definizione, l'estrazione testo basata sulla linguistica l'approccio più affidabile.

Comprendere come il funzionamento del processo di estrazione può aiutare l'utente a prendere decisioni chiave durante l'ottimizzazione delle risorse linguistiche (librerie, tipi, sinonimi e altro). Le fasi nel processo di estrazione includono:

- Conversione dei dati di origine in un formato standard
- Identificazione di termini candidati
- Identificazione delle classi di equivalenza e l'integrazione dei sinonimi
- Assegnazione di un tipo
- Indicizzazione e, quando richiesto, corrispondenza di modello con un analizzatore secondario

Passo 1. Conversione dei dati di origine in un formato standard

In questo primo passo, l'importazione di dati viene convertita in un formato uniforme che può essere utilizzato per ulteriori analisi. Questa conversione viene eseguita internamente e non modifica i dati originali.

Passo 2. Identificazione di termini candidati

È importante comprendere il ruolo delle risorse linguistiche nell'identificazione dei termini candidati durante l'estrazione. Le risorse linguistiche vengono utilizzate ogni volta che viene eseguita un'estrazione. Essi sono presenti sotto forma di modelli, librerie e risorse compilate. Le librerie includono elenchi di parole, relazioni e altre informazioni utilizzate per specificare o regolare l'estrazione. Le risorse compilate non possono essere visualizzate o modificate. Tuttavia, le restanti risorse possono essere modificate in Editor di modelli o, se ci si trova in una sessione di workbench interattivo, in Editor risorse.

Le risorse compilate sono componenti interni, principali, del motore di estrazione all'interno di IBM SPSS Modeler Text Analytics. Tali risorse includono un dizionario generale che contiene un elenco di moduli di base con un codice parte del discorso (nome, verbo, aggettivo e così via).

Oltre a queste risorse compilate, molte librerie sono fornite con il prodotto e possono essere utilizzate per completare le definizioni di tipi e concetti nelle risorse compilate, nonché per offrire sinonimi. Queste librerie e tutte quelle personalizzate create sono costituite da vari dizionari. Essi comprendono dizionari di tipo, dizionari dei sinonimi e dizionari di esclusione.

Una volta che i dati sono stati importati e convertiti, il motore di estrazione inizierà a identificare termini candidati per l'estrazione. I termini candidati sono parole o gruppi di parole che vengono utilizzati per identificare i concetti nel testo. Durante l'elaborazione del testo, le parole singole (**termini univoci**) e le parole composte (**termini multipli**) vengono identificati utilizzando gli estrattori di modello di parte del discorso. Quindi, le parole chiave di opinione candidate sono identificate mediante l'analisi del collegamento del testo di opinione.

Nota: i termini nel dizionario generale compilato summenzionati rappresentano un elenco di tutte le parole che sono probabilmente non interessanti o linguisticamente ambigue come termini univoci. Queste parole sono escluse dall'estrazione quando si identificano i termini univoci. Tuttavia, vengono rivalutate quando si stanno determinando parti del discorso o si stanno osservando parole composte candidate più estese (termini multipli).

Passo 3. Identificazione delle classi di equivalenza e integrazione dei sinonimi

Una volta identificati termini univoci e multipli, il software utilizza un dizionario di normalizzazione per identificare le classi di equivalenza. Una classe di equivalenza è un forma di base di una frase o una sola

forma di due varianti della stessa frase. Per determinare quale concetto utilizzare per la classe di equivalenza il motore di estrazione applica le seguenti regole nell'ordine elencato:

- Il formato specificato dall'utente in una libreria.
- La forma più frequente, come definito dalle risorse precompilate.

Passo 4. Assegnazione di tipo

Quindi, i tipi vengono assegnati a concetti estratti. Un tipo è un raggruppamento semantico di concetti. In questo passo vengono utilizzate risorse compilate e librerie. I tipi includono elementi come i concetti di livello superiore, le parole positive e negative, i nomi, i luoghi, le organizzazioni e molto altro ancora. Consultare la sezione “Dizionari di tipo” a pagina 189 per ulteriori informazioni.

Notare che le risorse della lingua giapponese hanno una serie distinta di tipi.

I sistemi linguistici sono sensibili alla conoscenza - più informazioni sono contenute nei dizionari, più elevata è la qualità dei risultati. La modifica del contenuto del dizionario, ad esempio definizioni di sinonimi, può semplificare le informazioni risultanti. Si tratta spesso di un processo iterativo, necessario per il recupero accurato dei concetti. NLP è un elemento principale di IBM SPSS Modeler Text Analytics.

Come funziona il processo di estrazione

Durante l'estrazione di concetti chiave e idee dalle risposte, IBM SPSS Modeler Text Analytics si basa su analisi del testo basate su linguistica. Questo approccio offre velocità e redditività nei sistemi basati su statistiche. Ma fornisce altresì un più alto livello di accuratezza, con intervento utente molto contenuto. L'analisi di testo basata su linguistica si fonda sul campo di studi conosciuto come elaborazione di lingua o anche come linguistica informatica.

Importante: Per il testo in lingua giapponese, il processo di estrazione segue una serie differente di procedure.

Comprendere come il funzionamento del processo di estrazione può aiutare l'utente a prendere decisioni chiave durante l'ottimizzazione delle risorse linguistiche (librerie, tipi, sinonimi e altro). Le fasi nel processo di estrazione includono:

- Conversione dei dati di origine in un formato standard
- Identificazione di termini candidati
- Identificazione delle classi di equivalenza e l'integrazione dei sinonimi
- Assegnazione di un tipo
- Indicizzazione
- Corrispondenza tra modelli ed estrazione di eventi

Passo 1. Conversione dei dati di origine in un formato standard

In questo primo passo, l'importazione di dati viene convertita in un formato uniforme che può essere utilizzato per ulteriori analisi. Questa conversione viene eseguita internamente e non modifica i dati originali.

Passo 2. Identificazione di termini candidati

È importante comprendere il ruolo delle risorse linguistiche nell'identificazione dei termini candidati durante l'estrazione. Le risorse linguistiche vengono utilizzate ogni volta che viene eseguita un'estrazione. Essi sono presenti sotto forma di modelli, librerie e risorse compilate. Le librerie includono elenchi di parole, relazioni e altre informazioni utilizzate per specificare o regolare l'estrazione. Le risorse compilate non possono essere visualizzate o modificate. Tuttavia, le restanti risorse (modelli) possono essere modificate in Editor di modelli o, se ci si trova in una sessione di workbench interattivo, in Editor risorse.

Le risorse compilate sono componenti interni principali del motore di estrazione all'interno di IBM SPSS Modeler Text Analytics. Tali risorse includono un dizionario generale che contiene un elenco di moduli di base con un codice parte del discorso (nome, verbo, aggettivo, avverbio, participio, coordinatore, determinativo o preposizione). Le risorse includono anche tipi incorporati, riservati, utilizzati per assegnare molti termini estratti ai seguenti tipi: `<Ubicazione>`, `<Organizzazione>` o `<Persona>`. Consultare la sezione “Tipi incorporati” a pagina 190 per ulteriori informazioni.

Oltre a queste risorse compilate, molte librerie sono fornite con il prodotto e possono essere utilizzate come complemento alle definizioni di tipi e concetti nelle risorse compilate, nonché per offrire altri tipi e sinonimi. Queste librerie e tutte quelle personalizzate create sono costituite da vari dizionari. Queste includono i dizionari di tipo, i dizionari di sostituzione (sinonimi ed elementi facoltativi) e i dizionari di esclusione. Consultare la sezione Capitolo 16, “Gestione delle librerie”, a pagina 179 per ulteriori informazioni.

Una volta che i dati sono stati importati e convertiti, il motore di estrazione inizierà a identificare termini candidati per l'estrazione. I termini candidati sono parole o gruppi di parole che vengono utilizzati per identificare i concetti nel testo. Durante l'elaborazione del testo, le parole singole (*termini univoci*) che non sono nelle risorse compilate vengono considerati come estrazioni di termine candidato. Le parole composte candidate (*termini multipli*) vengono identificate utilizzando gli estrattori di modello parte del discorso. Ad esempio, il termine multiplo `auto sportiva` che segue il modello parte del discorso "nome aggettivo", contiene due componenti. Il termine multiplo `auto sportiva veloce` che segue il modello parte del discorso "nome aggettivo aggettivo", contiene tre componenti.

Nota: i termini nel dizionario generale compilato summenzionati rappresentano un elenco di tutte le parole che sono probabilmente non interessanti o linguisticamente ambigue come termini univoci. Queste parole sono escluse dall'estrazione quando si identificano i termini univoci. Tuttavia, vengono rivalutate quando si stanno determinando parti del discorso o si stanno osservando parole composte candidate più estese (termini multipli).

Infine, viene utilizzato un algoritmo speciale per gestire le stringhe di lettera maiuscola, come i titoli professionali, in modo tale che questi particolari modelli possono essere estratti.

Passo 3. Identificazione delle classi di equivalenza e integrazione dei sinonimi

Una volta identificati i termini univoci e multipli candidati, il software utilizza una serie di algoritmi per confrontare le classi di equivalenza. Una classe di equivalenza è una forma di base di una frase o un singolo modulo di due varianti della stessa frase. Lo scopo di assegnare frasi alle classi di equivalenza è garantire che, ad esempio, `presidente dell'azienda` `presidente azienda` non vengono trattati come concetti separati. Per determinare quale concetto utilizzare per la classe di equivalenza, e cioè se viene utilizzato come termine principale `presidente dell'azienda` o `presidente azienda`, il motore di estrazione applica le seguenti regole nell'ordine elencato:

- Il formato specificato dall'utente in una libreria.
- La forma più frequente nel corpo del testo completo.
- La forma breve nel corpo completo di testo (che normalmente corrisponde al modulo di base).

Passo 4. Assegnazione di tipo

Quindi, i tipi vengono assegnati a concetti estratti. Un tipo è un raggruppamento semantico di concetti. In questo passo vengono utilizzate risorse compilate e librerie. I tipi includono elementi come i concetti di livello superiore, le parole positive e negative, i nomi, i luoghi, le organizzazioni e molto altro ancora. Tipi aggiuntivi possono essere definiti dall'utente. Consultare la sezione “Dizionari di tipo” a pagina 189 per ulteriori informazioni.

Passo 5. Indicizzazione

L'intera serie di record o documenti viene indicizzato stabilendo un puntatore tra una posizione del testo e il termine rappresentativo per ogni classe di equivalenza. Ciò presuppone che tutte le istanze del modulo con inflessioni di un concetto candidato vengano indicizzate come modulo di base candidato. La frequenza globale viene calcolata per ogni modulo di base.

Passo 6. Corrispondenza tra modelli ed estrazione di eventi

IBM SPSS Modeler Text Analytics può rilevare non solo tipi e concetti ma anche le relazioni tra di essi. Molti algoritmi e librerie sono disponibili con questo prodotto e forniscono la capacità di estrazione dei modelli di relazione tra tipi e concetti. Essi sono particolarmente utili quando si tenta di rilevare le opinioni specifico (ad esempio, reazioni al prodotto) o le relazioni tra persone o oggetti (ad esempio, i collegamenti tra gruppi politici o genomi).

Come funziona la categorizzazione

Quando vengono creati modelli di categoria in IBM SPSS Modeler Text Analytics, sono disponibili diverse tecniche per la creazione di categorie. Poiché ogni dataset è univoco, il numero di tecniche e l'ordine in cui si desidera applicarle possono cambiare. Poiché l'interpretazione dei risultati può essere diversa da un utente all'altro, potrebbe essere necessario sperimentare le diverse tecniche per capire quale di queste genera i migliori risultati per i dati di testo. In IBM SPSS Modeler Text Analytics, è possibile creare modelli di categoria in una sessione del workbench in cui è possibile esaminare e regolare le proprie categorie.

In questa guida, per **creazione di categoria** si intende la creazione delle definizioni e classificazioni mediante l'uso di uno o più tecniche integrate mentre la **categorizzazione** fa riferimento al calcolo del punteggio o etichettatura, processo in cui identificativi univoci (nome/ID/valore) vengono assegnati alle definizioni di categoria per ogni record o documento.

Durante la creazione della categoria, i concetti e i tipi che sono stati estratti vengono utilizzati come blocchi di creazione per le categorie. Quando si generano categorie, i record o documenti vengono automaticamente assegnati a categorie se contengono testo che corrisponda ad un elemento di definizione di una categoria.

IBM SPSS Modeler Text Analytics fornisce diverse tecniche di creazione automatica di categorie che facilitano la categorizzazione veloce di documenti o record.

Tecniche di raggruppamento

Ciascuna delle tecniche disponibili si adatta adeguatamente ad alcuni tipi di dati e situazioni, ma spesso è utile combinare le tecniche nella stessa analisi e acquisire la gamma completa di documenti o record. Ciò significa che è possibile visualizzare un concetto in più categorie o trovare categorie ridondanti.

Derivazione principale di concetto. Questa tecnica crea categorie prendendo un concetto e cercando altri concetti che sono in relazione con esso analizzando se uno dei componenti del concetto sono morfologicamente correlati. Questa tecnica è molto utile per identificare i concetti di parola composta sinonimo, poiché i concetti in ciascuna categoria generata sono sinonimi o strettamente correlati nel significato. Essa lavora con dati di lunghezza variabile e genera un numero inferiore di categorie compatte. Ad esempio, la nozione di `opportunità di avanzare` potrebbe essere raggruppata con i concetti di `opportunità per l'avanzamento` e `opportunità di avanzamento`. Consultare la sezione “Derivazione principale di concetto” a pagina 117 per ulteriori informazioni. Questa opzione non è disponibile per il testo giapponese.

Rete semantica. Questa tecnica inizia individuando i possibili sensi di ciascun concetto dal suo ampio indice di relazioni di parole e poi crea le categorie raggruppando concetti correlati. Questa tecnica è migliore quando i concetti sono noti alla rete semantica e non sono troppo ambigui. La tecnica è meno utile quando il testo contiene terminologia specialistica o gergo sconosciuto alla rete. In un esempio, il

concetto mela della nonna potrebbe essere raggruppato con mela gala e mela poiché sono discendenti del primo concetto. In un altro esempio, il concetto animale potrebbe essere raggruppato con gatto e canguro poiché sono iponimi di animale. Questa tecnica è disponibile solo per il testo inglese in questo rilascio del prodotto. Consultare la sezione “Reti semantiche” a pagina 119 per ulteriori informazioni.

Inclusione di concetto. Questa tecnica crea categorie raggruppando concetti di più termini (parole composte) basata su se i concetti contengono parole che sono sottoserie o superserie di una parola nell'altra. Ad esempio, il concetto sedile viene raggruppato con sedile di sicurezza, cinture di sicurezza e fibbia di cintura di sicurezza. Consultare la sezione “Inclusione concetti” a pagina 118 per ulteriori informazioni.

Ricorrenza. Questa tecnica crea categorie da ricorrenze trovate nel testo. L'idea è che quando i concetti vengono spesso trovati insieme in documenti e record, tale ricorrenza riflette una relazione sottostante che ha probabilmente un valore nelle definizioni di categoria. Quando le parole ricorrono in modo significativo, viene creata una regola di ricorrenza e può essere utilizzata come descrittore di categoria per una sottocategoria nuova. Ad esempio, se molti record contengono le parole prezzo e disponibilità, questi concetti possono essere raggruppati in una regola di ricorrenza, (prezzo & disponibile) e assegnati, ad esempio, ad una sottocategoria della categoria prezzo. Consultare la sezione “Regole di ricorrenza” a pagina 120 per ulteriori informazioni.

Numero minimo di documenti Per determinare quanto interessanti possono essere le ricorrenze, definire il numero minimo di documenti o record che devono contenere una data ricorrenza per essere utilizzati come descrittori in una categoria.

IBM SPSS Modeler Text Analytics Nodi

Oltre ai numerosi nodi standard forniti con IBM SPSS Modeler, è possibile anche gestire i nodi di estrazione testo per integrare il potere di analisi del testo nei propri flussi. IBM SPSS Modeler Text Analytics offre diversi nodi di estrazione testo per farlo. Tali nodi sono memorizzati in IBM SPSS Modeler Text Analytics scheda della tavolozza dei nodi.

Sono inclusi i nodi seguenti:

- Il **nodo origine di elenco file** genera un elenco di nomi di documento come input del processo di estrazione testo. È utile quando il testo risiede in documenti esterni anziché in un database o altro file strutturato. Il nodo restituisce un unico campo con un record per ogni documento o cartella elencati, che possono essere selezionati come input in un successivo nodo di estrazione testo. Consultare la sezione “Nodo Elenco file” a pagina 11 per ulteriori informazioni.
- Il **nodo di origine di flusso web** rende possibile leggere nel testo di flusso web, per esempio blog o flussi di notizie in formato RSS o HTML e di utilizzare questi dati nel processo di estrazione testo. Il nodo restituisce uno o più campi per record trovato nel flusso che possono essere selezionati come input in un successivo nodo di estrazione testo. Consultare la sezione “Nodo di flusso Web” a pagina 13 per ulteriori informazioni.
- Il **nodo di estrazione testo** utilizza metodi linguistici per estrarre concetti chiave dal testo, consente di creare categorie con tali concetti e altri dati e offre la possibilità di identificare relazioni e associazioni tra i concetti in base a modelli conosciuti (analisi del collegamento del testo). Il nodo può essere utilizzato per esaminare i contenuti dei dati di testo o per produrre un modello di concetto o un modello di categoria. I concetti e le categorie possono essere quindi combinati con dati strutturati esistenti, per esempio dati demografici, e applicate a modelli. Consultare la sezione “Nodo di modellazione di estrazione testo” a pagina 20 per ulteriori informazioni.
- Il **nodo di analisi collegamenti del testo** estrae concetti e identifica anche le relazioni tra concetti in base a modelli conosciuti all'interno del testo. L'estrazione di modello può essere utilizzato per rilevare le relazioni tra i concetti, così come pareri o eventuali qualificativi collegati a questi concetti. Il nodo analisi collegamenti del testo offre un modo più diretto per identificare ed estrarre i modelli dal testo e quindi aggiungere i risultati del modello di dati nel flusso. Ma è possibile anche eseguire TLA

utilizzando una sessione workbench interattiva nel nodo di modellazione di estrazione testo. Consultare la sezione “Nodo Analisi di collegamento del testo” a pagina 49 per ulteriori informazioni.

- Il **nodo di traduzione** può essere utilizzato per tradurre il testo dalle lingue supportate, come l'arabo, il cinese, e persiano, in inglese o altre lingue per fini di modellazione. Ciò consente di eseguire il mining di documenti in lingue a doppio byte che non sarebbero altrimenti supportate e permette agli analisti di estrarre concetti da tali documenti anche se non conoscono la lingua in questione. È possibile richiamare la stessa funzionalità da qualsiasi altro nodo di modellazione testo, ma l'uso di un nodo Traduci separato consente di memorizzare nella cache e riutilizzare una traduzione in più nodi. Consultare la sezione “Nodo di traduzione” a pagina 57 per ulteriori informazioni.
- Quando si esegue l'estrazione testo da documenti esterni, è possibile utilizzare il **nodo di output di estrazione testo** per generare una pagina HTML contenente collegamenti ai documenti da cui sono stati estratti i concetti. Consultare la sezione “Nodo Visualizzatore file” a pagina 61 per ulteriori informazioni.

Applicazioni

In generale, chi abitualmente deve rivedere grandi volumi di documenti per identificare gli elementi chiave per ulteriori esplorazioni può trarre vantaggio da IBM SPSS Modeler Text Analytics.

Alcune applicazioni specifiche includono:

- **Ricerca scientifica e medica.** Esplora materiali di ricerca secondaria, come notifiche di brevetti, articoli di giornale e pubblicazioni di protocolli. Identifica le associazioni che erano precedentemente sconosciute (come un medico associati ad un prodotto particolare), presentando possibilità di ulteriori esplorazioni. Minimizza il tempo impiegato nel processo di rilevamento dei medicinali. Utilizzare come aiuto nella ricerca genomica.
- **Ricerca di investimenti.** Esamina i report quotidiani di analista, articoli di notizie e comunicati stampa aziendali per identificare i punti chiave delle strategie di mercato. L'analisi del trend di tali informazioni rivela le opportunità o problemi emergenti per un'impresa o industria in un determinato periodo di tempo.
- **Rilevamento di comportamenti fraudolenti.** Utilizzato per le frodi bancarie o nel campo della salute pubblica per rilevare anomalie e scoprire segnali di allarme in grandi quantità di testo.
- **Analisi di mercato.** Utilizzare nelle attività di ricerca di mercato per identificare gli argomenti chiave nelle risposte ai sondaggi aperti.
- **Analisi di blog e di flusso web.** Esplorare e creare modelli utilizzando le idee chiave trovate nei flussi di notizie, blog, ecc.
- **CRM.** Crea modelli utilizzando dati provenienti da tutti i punti di contatto cliente, come posta elettronica, transazioni e sondaggi.

Capitolo 2. Lettura in testo di origine

I dati per l'estrazione testo possono trovarsi in uno qualsiasi dei formati standard utilizzato da IBM SPSS Modeler, inclusi i database o altri formati "rettangolari" che rappresentano i dati in righe e colonne o in formati di documento, come Microsoft Word, Adobe PDF o HTML, non conformi a questa struttura.

- Per leggere nel testo da documenti non conformi alla struttura dei dati standard compreso Microsoft Word, Microsoft Excel e Microsoft PowerPoint oltre che Adobe PDF, XML, HTML e altri, il nodo di elenco file può essere utilizzato per generare un elenco di documenti o cartelle come input per il processo di estrazione testo. Consultare la sezione “Nodo Elenco file” per ulteriori informazioni.
- Per leggere nel testo da flussi web, per esempio di blog o flussi di notizie in formato RSS o HTML, il nodo di flusso Web può essere utilizzato per formattare i dati di flusso Web per l'input nel processo di estrazione testo. Consultare la sezione “Nodo di flusso Web” a pagina 13 per ulteriori informazioni.
- Per leggere nel testo da uno dei formati di dati standard utilizzato da IBM SPSS Modeler, come un database con uno o più campi di testo per i commenti del cliente, possono essere utilizzati tutti i nodi di origine standard nativi di IBM SPSS Modeler. Consultare la documentazione del nodo di IBM SPSS Modeler per ulteriori informazioni.

Nodo Elenco file

Per leggere nel testo da documenti non strutturati salvati in formati come Microsoft Word, Microsoft Excel e Microsoft PowerPoint oltre che Adobe PDF, XML, HTML e altri, il nodo di elenco file può essere utilizzato per generare un elenco di documenti o cartelle come input per il processo di estrazione testo. Questa operazione è necessaria perché i documenti di testo non strutturati non possono essere rappresentati da campi e record a righe e colonne nella stessa maniera di altri dati utilizzati da IBM SPSS Modeler. Questo nodo può essere trovato nella tavolozza di estrazione testo.

Il nodo Elenco file funziona come un nodo origine; tuttavia, oltre che leggere e inserire automaticamente i dati effettivi dai file origine, è possibile utilizzare alternativamente il nodo per leggere i nomi dei documenti e directory sotto la root specificata e produrli come un elenco. Quando viene utilizzato per leggere i nomi documento o i nomi di directory, l'output è un solo campo con un record per ogni file elencato che può essere selezionato come input per un successivo nodo di estrazione testo o di analisi di collegamento del testo.

È possibile rilevare questo nodo sulla scheda IBM SPSS Modeler Text Analytics della tavolozza dei nodi nella parte inferiore della finestra IBM SPSS Modeler. Per ulteriori informazioni, consultare la sezione “IBM SPSS Modeler Text Analytics Nodi” a pagina 8.

Importante: Tutti i nomi di directory e i nomi file contenenti caratteri che non sono inclusi nella codifica locale della macchina non sono supportati. Quando si tenta di eseguire un flusso contenente un nodo di elenco file, tutti i nomi file o directory contenenti questi caratteri causeranno un errore di esecuzione del flusso. Questo può accadere con nomi di directory o nomi file di lingua straniera, ad esempio un nome file giapponese su una locale francese.

Supporto dati locali. Se si è connessi a un IBM SPSS Modeler Text Analytics Server remoto e si ha un flusso con un nodo Elenco file, i dati devono risiedere sulla stessa macchina IBM SPSS Modeler Text Analytics Server o accertarsi che la macchina server abbia accesso alla cartella in cui i dati di origine nel nodo Elenco file vengono memorizzati.

Nota: Non è possibile utilizzare il nodo Elenco file per il calcolo del punteggio all'interno di una configurazione IBM SPSS Collaboration and Deployment Services - Scoring.

Nodo Elenco file: scheda Impostazioni

In questa scheda è possibile definire le directory, le estensioni dei file, e l'output desiderato da questo nodo.

Nota: l'estrazione testo non può elaborare Microsoft Office e Adobe PDF in file in piattaforme non Microsoft Windows. È tuttavia sempre possibile elaborare file XML, HTML o in formato testo.

Tutti i nomi di directory e i nomi file contenenti caratteri che non sono inclusi nella codifica locale della macchina non sono supportati. Quando si tenta di eseguire un flusso contenente un nodo di elenco file, tutti i nomi file o directory contenenti questi caratteri causeranno un errore di esecuzione del flusso. Questo può accadere con nomi di directory o nomi file di lingua straniera, ad esempio un nome file giapponese su una locale francese.

Directory. Specifica la cartella root contenente i documenti che si desidera elencare.

- **Includi sottocategorie.** Specifica che le sottodirectory devono inoltre essere sottoposte a scansione.

Tipi file da includere nell'elenco: è possibile selezionare o deselezionare i tipi di file e le estensioni che si desidera utilizzare. Deselezionando un'estensione file, i file con tale estensione sono ignorati. È possibile filtrare per le seguenti estensioni:

Tabella 1. Filtri di tipo file per estensione file.

• *.rtf, .doc, .docx, .docm*	• *.xls, .xlsx, .xlsm*	• *.ppt, .pptx, .pptm*	• *.txt, .text*
• *.htm, .html, .shtml*	• *.xml*	• *.pdf*	• *.$*

Nota: per ulteriori informazioni consultare la sezione “Nodo Elenco file” a pagina 11.

Se si dispone di file senza estensione, o con punto finale (ad esempio `File01` o `File01.`), utilizzare l'opzione **Nessuna estensione** per selezionarli.

Il campo di output rappresenta. Selezionare il formato del campo di output. Le scelte sono:

- **Testo reale.** Selezionare questa opzione se il campo conterrà testo esatto. L'utente quindi è abilitato a scegliere il valore **Codifica input** dal seguente elenco:
  - Automatico (Europeo)
  - Automatico (Giapponese)
  - UTF-8
  - UTF-16
  - ISO-8859-1
  - US ascii
  - CP850
  - Shift-JIS
- **Nomi percorso dei documenti.** Selezionare questa opzione se il campo di output conterrà uno o più nomi percorso per l'ubicazione in cui sono presenti i documenti.

Importante! Dalla versione 14, l'opzione 'Elenco directory' non è più disponibile e l'unico output sarà un elenco di file.

Nodo di elenco file: altre schede

La scheda Tipi è una scheda standard dei nodi IBM SPSS Modeler uguale alla scheda Annotazioni.

Uso del nodo elenco file nell'estrazione testo

Il nodo elenco file viene utilizzato quando i dati di testo risiedono in documenti esterni non strutturati in formati quali Microsoft Word, Microsoft Excele Microsoft PowerPoint, oltre che Adobe PDF, XML, HTML e altri. Inoltre, per inserire automaticamente il testo reale, è possibile anche utilizzare questo nodo per generare un elenco di documenti o cartelle come input per il processo di estrazione testo (come ad esempio un nodo successivo di estrazione testo o un nodo di analisi di collegamento del testo).

Se per generare un elenco di documenti invece di utilizzare il testo reale, si utilizza il nodo Elenco file, quando successivamente si utilizza il nodo di estrazione testo o di analisi di collegamento del testo, specificare che il **campo Testo** rappresenta **Nomi percorso dei documenti** per indicare che invece di contenere il testo reale da estrarre, il campo selezionato contiene percorsi per i documenti in cui il testo è presente.

Ad esempio, si supponga di collegare un nodo Elenco file a un nodo di estrazione testo per fornire un testo che risiede in documenti esterni:

1. **Nodo di elenco file (Scheda Impostazioni).** In primo luogo, è stato aggiunto questo nodo al flusso per specificare dove i documenti di testo sono memorizzati. È stata selezionata la directory contenente tutti i documenti su cui si desidera eseguire l'estrazione di testo.
2. **Nodo di estrazione testo (Scheda Campi).** Inoltre, è stato aggiunto e connesso un nodo Estrazione testo al nodo Elenco file. In questo nodo, è stato definito il formato di input, il modello di risorsa e il formato di output. È stato selezionato il nome campo ottenuto dal nodo Elenco file e selezionata l'opzione in cui il campo di testo rappresenta **i nomi percorso dei documenti** oltre ad altre impostazioni. Per ulteriori informazioni, consultare la sezione “Utilizzo del nodo di estrazione testo in un flusso” a pagina 31.

Per ulteriori informazioni sull'uso del nodo di estrazione testo, consultare “Nodo di modellazione di estrazione testo” a pagina 20.

Nodo di flusso Web

Il nodo Flusso Web può essere utilizzato per preparare i dati di testo da flussi Web per il processo di estrazione testo. Questo nodo accetta i flussi Web in due formati:

- Formato RSS. RSS è un formato semplice standardizzato basato su XML per il contenuto Web. L'URL per questo formato punta ad una pagina che presenta una serie di articoli collegati come fonti di notizie associate e blog. Poiché RSS è un formato standardizzato, ogni articolo collegato viene automaticamente identificato e trattato come un record separato nel flusso di dati risultante. Non sono necessari ulteriori input per essere in grado di identificare i dati di testo importanti e i record dal flusso a meno che non si desidera applicare una tecnica di filtro al testo.
- Formato HTML. È possibile definire una o più URL sulle pagine HTML nella scheda Input. Quindi, nella scheda Record, definire il tag di inizio record come identificare le tag che delimitano i contenuti di destinazione e assegnare tali tag ai campi di output a scelta (descrizione, titolo, data di modifica e così via). Consultare la sezione “Nodo di flusso Web: scheda Record” a pagina 15 per ulteriori informazioni.

Importante! Se si sta tentando di richiamare le informazioni sul Web attraverso un server proxy, è necessario abilitare il server proxy nel file `net.properties` per il client e server di IBM SPSS Modeler Text Analytics. Seguire le istruzioni dettagliate in questo file. Ciò si applica quando si accede al Web attraverso il nodo Web Feed o viene richiamata una licenza SaaS (SDL Software as a Service), poiché queste connessioni attraversano Java™. Questo file è presente in *C:\Program Files\IBM\SPSS\Modeler\17.1\jre\lib\net.properties* per impostazione predefinita.

L'output di questo nodo è una serie di campi utilizzati per descrivere i record. Il campo **Descrizione** è più comunemente utilizzato poiché contiene la maggior parte del contenuto del testo. Tuttavia, si può

essere anche interessati ad altri campi, come la descrizione breve di un record (campo **Desc breve**) o il titolo del record (campo **Titolo**). Ognuno dei campi di output può essere selezionato come input per un successivo nodo di estrazione testo.

Nota: Non è possibile utilizzare il nodo di flusso Web per il calcolo del punteggio all'interno di una configurazione IBM SPSS Collaboration and Deployment Services - Scoring.

È possibile rilevare questo nodo sulla scheda IBM SPSS Modeler Text Analytics della tavolozza dei nodi nella parte inferiore della finestra IBM SPSS Modeler. Per ulteriori informazioni, consultare la sezione "IBM SPSS Modeler Text Analytics Nodi" a pagina 8.

Nodo di flusso Web: scheda Input

La scheda Input viene utilizzata per specificare uno o più indirizzi Web, o URL, per acquisire i dati di testo. In un contesto di estrazione testo, è possibile specificare gli URL per i flussi che contengono dati di testo.

Importante! Quando si gestiscono dati non RSS, è preferibile utilizzare uno strumento di raschiatura Web, ad esempio WebQL®, per automatizzare la raccolta di contenuto e quindi riferire l'output a questo strumento utilizzando un nodo origine differente.

È possibile impostare i seguenti parametri:

Immetti o incolla URL. In questo campo, è possibile immettere o incollare uno o più URL. Se se ne sta immettendo più di uno, immettere solo uno per riga e utilizzare il tasto **Invio/Ritorno unitario** su righe separate. Immettere il percorso URL completo per il file. Questi URL possono essere in uno di due formati:

- *Formato RSS.* RSS è un formato semplice standardizzato basato su XML per il contenuto Web. L'URL per questo formato punta ad una pagina che presenta una serie di articoli collegati come fonti di notizie associate e blog. Poiché RSS è un formato standardizzato, ogni articolo collegato viene automaticamente identificato e trattato come un record separato nel flusso di dati risultante. Non sono necessari ulteriori input per essere in grado di identificare i dati di testo importanti e i record dal flusso a meno che non si desidera applicare una tecnica di filtro al testo.
- *Formato HTML.* È possibile definire una o più URL sulle pagine HTML nella scheda Input. Quindi, nella scheda Record, definire il tag di inizio record come identificare le tag che delimitano i contenuti di destinazione e assegnare tali tag ai campi di output a scelta (descrizione, titolo, data di modifica e così via). Quando si gestiscono dati non RSS, è preferibile utilizzare uno strumento di raschiatura Web, ad esempio WebQL®, per automatizzare la raccolta di contenuto e quindi riferire l'output a questo strumento utilizzando un nodo origine differente. Consultare la sezione "Nodo di flusso Web: scheda Record" a pagina 15 per ulteriori informazioni.

Numero di voci più recenti da leggere per URL. Questo campo specifica il numero massimo di record da leggere per ogni URL elencato nel campo a partire dal primo record trovato nel flusso. La quantità di testo influenza la velocità di elaborazione durante l'estrazione a valle in un nodo di estrazione testo o di analisi collegamento del testo.

Salvare e riutilizzare i flussi precedenti quando possibile. Con questa opzione, i flussi Web vengono sottoposti a scansione e i risultati elaborati vengono memorizzati nella cache. Quindi, su esecuzioni di flusso successive, se il contenuto di un flusso specificato non è stato modificato o se il flusso è inaccessibile (un'interruzione Internet, ad esempio), la versione memorizzata viene utilizzata per velocizzare il tempo di elaborazione. Qualsiasi nuovo contenuto rilevato in questi flussi viene inoltre memorizzato nella cache per la volta successiva in cui si desidera eseguire il nodo.

- **Etichetta.** Se si seleziona **Salva e riutilizza flussi Web precedenti quando possibile**, è necessario specificare un nome etichetta per i risultati. Questa etichetta viene utilizzata per descrivere il flusso memorizzato sul server. Se non viene specificata alcuna etichetta o l'etichetta non viene riconosciuta, il riuso non sarà possibile. È possibile gestire queste cache di flusso Web nella tabella di sessione di IBM

SPSS Text Analytics Administration Console . Fare riferimento alla guida per l'utente di IBM SPSS Text Analytics Administration Console per ulteriori informazioni.

Nodo di flusso Web: scheda Record

La scheda Record viene utilizzata per specificare il contenuto di testo dei flussi non RSS identificando dove inizia ogni nuovo record, oltre ad altre informazioni pertinenti relative a ogni record. Se si sa che un flusso non RSS (HTML) contiene il testo che si trova in più record, è necessario identificare la tag di inizio record oppure il testo che verrà trattato come un record. Mentre i flussi RSS sono standardizzati e non richiedono alcuna specifica tag in questa scheda è possibile ancora visualizzare l'anteprima del contenuto nella scheda Anteprima.

Importante! Quando si gestiscono dati non RSS, è preferibile utilizzare uno strumento di raschiatura Web, ad esempio WebQL®, per automatizzare la raccolta di contenuto e quindi riferire l'output a questo strumento utilizzando un nodo origine differente.

URL. Questo elenco a discesa contiene un elenco di URL inseriti nella scheda Input. Sono presenti i flussi formattati HTML e RSS. Se l'indirizzo URL è troppo lungo per l'elenco a discesa, verrà automaticamente tagliato a metà utilizzando una sospensione per sostituire il testo tagliato, ad esempio *http://www.ibm.com/example/start-of-address...rest-of-address/path.htm*.

- Con **flussi formattati HTML**, se il flusso contiene più di un record (o voce), è possibile definire quali tag HTML contengono i dati corrispondenti al campo visualizzato nella tabella. Ad esempio, è possibile definire la tag di inizio che indica che un nuovo record è stato avviato, una tag data di modifica o il nome dell'autore.
- Con **Flusso formattato RSS**, non viene richiesto di immettere tutte le tag poiché RSS è un formato standardizzato. Tuttavia, se si desidera, è possibile visualizzare i risultati di esempio nella scheda Anteprima. Tutti i flussi RSS riconosciuti sono preceduti da un'immagine del logo RSS.

Scheda Origine. In questa scheda, è possibile visualizzare il codice origine per qualsiasi flusso HTML. Tale codice non è modificabile. È possibile utilizzare il campo Trova per individuare le tag specifiche o informazioni su questa pagina che è possibile quindi copiare e incollare nella tabella riportata di seguito. Il campo di ricerca non è sensibile al maiuscolo / minuscolo e corrisponderà a stringhe parziali.

Scheda Anteprima. In questa scheda, è possibile visualizzare in anteprima come un record verrà letto dal nodo di flusso Web. Ciò è particolarmente utile per flussi HTML poiché è possibile modificare il modo in cui un record verrà letto definendo tag HTML nella tabella al di sotto della scheda Anteprima.

Tag di avvio record non RSS. Questa opzione è valida solo per flussi non RSS. Se il flusso contiene più testo che si desidera interrompere in più record, specificare la tag HTML che segnala l'inizio di un record (come un articolo o una voce di blog). Se non se ne definisce uno per un flusso non RSS, l'intera pagina viene trattata come un singolo record, l'intero contenuto è l'output nel campo **Descrizione** e la data di esecuzione del nodo viene utilizzata come **Data di modifica** e **Data di pubblicazione**.

Tabella Campo. Questa opzione è valida solo per flussi non RSS. In questa tabella, è possibile suddividere il contenuto di testo in campi di output specifici immettendo una tag di inizio per tutti i campi di output predefiniti. Immettere solo la tag di inizio. Tutte le corrispondenze vengono eseguite esaminando l'HTML e adattando il contenuto della tabella ai nomi e attributi di tag trovati nell'HTML. È possibile utilizzare i pulsanti nella parte inferiore per copiare le tag definite e riutilizzarle per altri flussi.

Tabella 2. Campi di output possibili per flussi non RSS (formati HTML)

Nome campo di output	Contenuto tag previsto
Titolo	La tag di delimitazione del titolo del record. (facoltativo).
Breve descrizione	La tag che delimita la descrizione breve o etichetta. (facoltativo).

Tabella 2. Campi di output possibili per flussi non RSS (formati HTML) (Continua)

Nome campo di output	Contenuto tag previsto
Descrizione	La tag di delimitazione del testo principale. Se lasciato vuoto, questo campo conterrà tutto il resto del contenuto nella tag <body> (se è presente un record singolo) o il contenuto trovato all'interno del record corrente (quando viene specificato un delimitatore di record).
Autore	La tag di delimitazione dell'autore del testo. (facoltativo).
Contributi	La tag di delimitazione dei nomi dei collaboratori al testo. (facoltativo).
Data di pubblicazione	La tag che delimita la data in cui il testo è stato pubblicato. Se lasciato vuoto, questo campo conterrà la data in cui il nodo legge i dati.
Data ultima modifica	La tag che delimita la data di quando il testo è stato modificato. Se lasciato vuoto, questo campo conterrà la data in cui il nodo legge i dati.

Quando si immette una tag nella tabella, il flusso viene eseguito utilizzando questa tag come tag minima di corrispondenza piuttosto che una corrispondenza esatta. Cioè, se è stato immesso per il campo Titolo, questo dovrebbe corrispondere a qualsiasi tag nel flusso, compreso quelle con gli attributi specificati (ad esempio,), tale che è uguale alla tag principale ( ) e a tutti i derivati che includono un attributo e che utilizzano tale contenuto per il campo di output del titolo. Se si immette una tag principale, viene incluso anche qualsiasi ulteriore attributo.

Tabella 3. Esempi di tag HTML utilizzato per identificare il testo per i campi di output

Se si immette:	Corrisponde a:	Corrisponde anche a:	Ma non corrisponde a:
			qualsiasi altra tag
<p class="auth">	<p class="auth">	<p color="black" class="auth" id="85643">	<p color="black">

Nodo di flusso Web: scheda Filtro contenuto

La scheda Filtro contenuto viene utilizzata per applicare una tecnica di filtro al contenuto di flusso RSS. Questa scheda non si applica ai flussi HTML. È possibile che si desideri filtrare se il flusso contiene molto testo nella forma di intestazioni, piè di pagina, menu, pubblicità e così via. È possibile utilizzare questa scheda per eliminare tag HTML indesiderate, JavaScript e parole o righe brevi dal contenuto.

Filtrare il contenuto. Se non si desidera applicare una tecnica di pulizia, selezionare **Nessuno**. Altrimenti, selezionare **Pulizia contenuto RSS**.

Opzioni di pulizia del contenuto RSS. Se si seleziona **Pulizia contenuto RSS** è possibile scegliere di eliminare le righe in base a determinati criteri. Una riga è delimitato da un tag HTML come <p> e <li> ma escludendo tag in linea come , <b> e <font>. Tenere presente che le tag
 vengono elaborate come interruzioni di riga.

- **Scarta righe brevi**. Questa opzione ignora le righe che non contengono il **numero minimo di parole** definite qui.
- **Scarta righe con parole brevi**. Questa opzione ignora le righe che contengono più della **lunghezza media minima di parole** definite qui.
- **Scarta righe con molte parole di un solo carattere**. Questa opzione ignora le righe che contengono più di una certa **proporzione di parole a singolo carattere**.
- **Scarta righe contenenti tag specifiche**. Questa opzione ignora il testo nelle righe che contengono una qualsiasi delle tag specificate nel campo.
- **Scarta righe contenenti testo specifico**. Questa opzione ignora le righe che contengono il testo specificato nel campo.

Uso del nodo Flusso Web nell'estrazione testo

Il nodo Flusso Web può essere utilizzato per preparare i dati di testo da flussi Web Internet per il processo di estrazione testo. Questo nodo accetta flussi Web in un formato HTML o RSS. Questi flussi servono come input nel processo di estrazione testo (un nodo successivo di estrazione testo o di analisi di collegamento del testo).

Se si utilizza il nodo Flusso Web, accertarsi di specificare che il campo Testo rappresenta **il testo attuale** nel nodo di estrazione testo o di analisi di collegamento del testo per indicare che questi campi sono collegati direttamente ad ogni articolo o voce di blog.

Importante! Se si sta tentando di richiamare le informazioni sul Web attraverso un server proxy, è necessario abilitare il server proxy nel file `net.properties` per il client e server di IBM SPSS Modeler Text Analytics. Seguire le istruzioni dettagliate in questo file. Ciò si applica quando si accede al Web attraverso il nodo Web Feed o viene richiamata una licenza SaaS (SDL Software as a Service), poiché queste connessioni attraversano Java. Questo file è presente in *C:\Program Files\IBM\SPSS\Modeler\17.1\jre\lib\net.properties* per impostazione predefinita.

Esempio: nodo Flusso Web (flusso RSS) con il nodo di modellazione di estrazione testo

Ad esempio, si supponga di connettere un nodo Flusso Web ad un nodo di estrazione testo per fornire dati di testo da un flusso RSS nel processo di estrazione testo.

1. **Nodo Flusso Web (scheda Input).** In primo luogo, questo nodo è stato aggiunto al flusso per specificare dove si trova il contenuto del flusso e per verificare la struttura del contenuto. Sulla prima scheda, è stato fornito l'URL per un flusso RSS. Poiché l'esempio è per un flusso RSS, la formattazione è già definita e non è necessario apportare alcuna modifica nella scheda Record. Un algoritmo di filtraggio contenuto facoltativo è disponibile per i flussi RSS, tuttavia in questo caso non è stata applicato.
2. **Nodo di estrazione testo (Scheda Campi).** Inoltre, è stato aggiunto e connesso un nodo di estrazione testo al nodo Flusso Web. Su questa scheda, è stato definito l'output di campo di testo in base al nodo di flusso Web. In questo caso, si è utilizzato il campo **Descrizione**. Per l'esempio è anche selezionato l'opzione Campo testo rappresenta **il testo effettivo**, oltre ad altre impostazioni.
3. **Nodo di estrazione testo (scheda Modello).** Quindi, nella scheda Modello, è stata scelta la modalità di creazione e le risorse. In questo esempio, abbiamo scelto di costruire un modello di concetto direttamente da questo nodo utilizzando il modello di risorsa predefinito.

Per ulteriori informazioni sull'uso del nodo di estrazione testo, consultare “Nodo di modellazione di estrazione testo” a pagina 20.

Capitolo 3. Estrazione per concetti e categorie

Il nodo di modellazione di estrazione testo è utilizzato per generare uno dei due nugget del modello di estrazione testo:

- I *nugget del modello di concetto* scoprono ed estraggono i concetti importanti dai propri dati di testo strutturati o non strutturati.
- I *nugget del modello di categoria* calcolano il punteggio e assegnano documenti e record alle categorie, che sono composta dai concetti estratti (e modelli).

I concetti e i modelli estratti come pure le categorie del proprio nugget del modello possono essere combinati con dati strutturati esistenti, per esempio dati demografici, e applicarli utilizzando la serie completa degli strumenti da IBM SPSS Modeler per produrre decisioni migliori e più mirate. Ad esempio, se i clienti spesso riportano questioni come il principale ostacolo per il completamento delle attività di gestione account online, si potrebbe voler incorporare i "problemi di accesso" nei propri modelli.

Inoltre, il nodo di modellazione di estrazione testo è completamente integrato all'interno di IBM SPSS Modeler in modo che è possibile distribuire flussi di estrazione testo tramite IBM SPSS Modeler Solution Publisher per il calcolo del punteggio in tempo reale dei dati non strutturati in applicazioni quali PredictiveCallCenter. La possibilità di distribuire questi flussi garantisce implementazioni efficaci di estrazione testo a circuito chiuso. Ad esempio, l'organizzazione può analizzare ora note da chiamanti in entrata o in uscita applicando i modelli predittivi per aumentare la precisione dei propri messaggi di marketing in tempo reale. È dimostrato che l'uso dei risultati del modello di estrazione testo nei flussi migliora la precisione dei dati di modello predittivi.

Nota: per eseguire IBM SPSS Modeler Text Analytics con IBM SPSS Modeler Solution Publisher, aggiungere la directory `<install_directory>/ext/bin/spss.TMWBServer` alla variabile d'ambiente `$LD_LIBRARY_PATH`.

In IBM SPSS Modeler Text Analytics si fa spesso riferimento a concetti e categorie estratti. È importante comprendere il significato di concetti e categorie poiché possono aiutare l'utente a prendere decisioni avendo più informazioni a disposizione durante il lavoro esplorativo e la creazione del modello.

Concetti e nugget del modello di concetto

Durante il processo di estrazione, i dati di testo vengono sottoposti a scansione e analizzati per identificare singole parole interessanti o pertinenti, come `elezione` o `pace` e frasi di parole quali `elezione presidenziale`, `elezione del presidente` o `trattati di pace`. Queste parole o frasi sono indicate collettivamente come *termini*. Utilizzando le risorse linguistiche, i termini rilevanti vengono estratti e termini simili vengono raggruppati sotto un termine principale, denominato **concetto**.

In questo modo, un concetto potrebbe rappresentare più termini sottostanti a seconda del proprio testo e della serie di risorse linguistiche che si stanno utilizzando. Ad esempio, si consideri una relazione di un impiegato da cui viene estratto il concetto `salario`. Quando si osservano i record associati con `salario`, si nota che `salario` non è sempre presente nel testo ma piuttosto alcuni record contengono qualcosa di simile, come i termini `paga`, `paghe` e `salari`. Questi termini vengono raggruppati in `salario` poiché il motore di estrazione li considera come simili o sinonimi in base alle regole di elaborazione o alle risorse linguistiche. In questo caso, qualsiasi documento o record che contiene uno qualsiasi di tali termini viene trattato come se contenesse la parola `salario`.

Se si desidera vedere quali condizioni vengono raggruppate sotto un concetto, è possibile esplorare il concetto all'interno di un workbench interattivo o esaminare quali sinonimi vengono visualizzati nel modello di concetto. Consultare la sezione “Termini sottostanti in modelli di concetto” a pagina 34 per ulteriori informazioni.

Un **nugget del modello di concetto** contiene una serie di concetti che possono essere utilizzati per identificare i record o i documenti che anche contengono il concetto (inclusi tutti i suoi sinonimi o i termini raggruppati). Un modello di concetto può essere utilizzato in due modi. Il primo è quello di esplorare e analizzare i concetti che sono stati rilevati nel testo sorgente di origine o per identificare rapidamente i documenti interessanti. Il secondo sarebbe di applicare questo modello a nuovi record o documenti di testo per identificare rapidamente gli stessi concetti chiave nei nuovi documenti/record, come il rilevamento in tempo reale dei concetti chiave nei dati di un call center.

Consultare la sezione “Nugget di estrazione testo: modello di concetto” a pagina 32 per ulteriori informazioni.

Categorie e nugget del modello di categoria

È possibile creare **categorie** che rappresentano, in sostanza, concetti di livello superiore o argomenti per acquisire le idee chiave, la conoscenza e le posizioni espresse nel testo. Le categorie sono costituite da una serie di descrittori, per esempio *concetti*, *tipi* e *regole*. Tutti insieme questi descrittori vengono utilizzati per identificare se un record o documento appartiene o meno a una determinata categoria. Un documento o record può essere analizzato per vedere se una parte del suo testo corrisponde a un descrittore. Se viene rilevata una corrispondenza, il documento/record viene assegnato a tale categoria. Questo processo viene definito **categorizzazione**.

Le categorie possono essere create automaticamente utilizzando una serie completa del prodotto di tecniche automatizzate o utilizzando manualmente informazioni aggiuntive che potrebbero avere rilevanza con i dati o una combinazione di entrambi. È anche possibile caricare una serie di categorie predefinite da un pacchetto di analisi del testo tramite la scheda Modello di questo nodo. La creazione manuale di categorie o il perfezionamento delle categorie può essere eseguita solo tramite il workbench interattivo. Consultare la sezione “Nodo di estrazione testo: scheda Modello” a pagina 24 per ulteriori informazioni.

Un **nugget del modello di categoria** contiene una serie di categorie insieme ai relativi descrittori. Il modello può essere utilizzato per categorizzare un insieme di documenti o record in base al testo in ogni documento/record. Ogni documento o record viene letto e poi assegnato a ogni categoria per la quale è stata trovata una corrispondenza di descrittore. In questo modo, un documento o record potrebbe essere assegnato a più di una categoria. È possibile utilizzare i nugget del modello di categoria per visualizzare le idee essenziali in risposte ai sondaggi o in una serie di voci di blog, ad esempio.

Consultare la sezione “Nugget di estrazione testo: Modello di categoria” a pagina 41 per ulteriori informazioni.

Nodo di modellazione di estrazione testo

Il nodo di estrazione testo utilizza tecniche linguistiche e di frequenza per estrarre concetti chiave dal testo e creare categorie con tali concetti e altri dati. Il nodo può essere utilizzato per esaminare i contenuti dei dati di testo o per produrre un nugget del modello di concetto o del modello di categoria. Quando si esegue questo nodo di modellazione, un motore di estrazione linguistico interno estrae e organizza i concetti, modelli e/o le categorie utilizzando metodi di elaborazione lingua naturale.

È possibile eseguire il nodo di estrazione testo e produrre automaticamente un nugget del modello di concetto o categoria utilizzando l'opzione **Genera direttamente**. In alternativa, è possibile utilizzare un approccio sperimentale, manuale utilizzando la modalità di **creazione interattiva** in cui non solo si possono estrarre concetti, creare categorie e perfezionare le proprie risorse linguistiche, ma anche eseguire

l'analisi di collegamento del testo ed esplorare i cluster. Consultare la sezione “Nodo di estrazione testo: scheda Modello” a pagina 24 per ulteriori informazioni.

È possibile rilevare questo nodo sulla scheda IBM SPSS Modeler Text Analytics della tavolozza dei nodi nella parte inferiore della finestra IBM SPSS Modeler. Per ulteriori informazioni, consultare la sezione “IBM SPSS Modeler Text Analytics Nodi” a pagina 8.

Requisiti. I nodi di modellazione di estrazione testo accettano i dati di testo da un nodo di avanzamento web, di elenco file o uno dei nodi di origine standard. Questo nodo viene installato con IBM SPSS Modeler Text Analytics e vi si può accedere dalla tavolozza di IBM SPSS Modeler Text Analytics.

Nota: Questo nodo sostituisce il nodo di estrazione testo per tutti gli utenti e il vecchio nodo di estrazione testo per gli utenti giapponesi, che sono stati forniti nelle versioni precedenti di Text Mining for Clementine. Se si dispone di flussi più vecchi che utilizzano questi nodi o nugget del modello, è necessario ricreare i flussi utilizzando il nuovo nodo di estrazione testo.

Nodo di estrazione testo: scheda Campi

La scheda Campi viene utilizzata per specificare le impostazioni dei campi per i dati da cui si estraggono concetti. Considerare l'utilizzo di un nodo campione di flusso a monte da questo nodo quando si lavora con dataset più estesi per accelerare i tempi di elaborazione. Consultare la sezione “Esempio di flusso a monte per risparmiare tempo” a pagina 31 per ulteriori informazioni.

È possibile impostare i seguenti parametri:

Campo Testo. Selezionare il campo che contiene il testo da estrarre, il nome percorso del documento o il nome percorso della directory per i documenti. Questo campo dipende dall'origine dati.

Il campo Testo rappresenta. Indica cosa contiene il campo di testo specificato nella precedente impostazione. Le scelte sono:

- **Testo reale.** Selezionare questa opzione se il campo contiene il testo esatto da cui devono essere estratti i concetti.
- **Nomi percorso dei documenti.** Selezionare questa opzione se il campo contiene uno o più nomi percorso per l'ubicazione dei documenti di testo.

Tipo documento. Questa opzione è disponibile solo se è stato indicato che il campo di testo rappresenta **nomi percorso di documenti**. Il tipo di documento specifica la struttura del testo. Selezionare uno dei seguenti tipi:

- **Testo.** Da utilizzare per la maggior parte dei documenti o origini di testo. Tutto l'insieme del testo viene analizzato per l'estrazione. A differenza delle altre opzioni, non vi sono ulteriori impostazioni per questa opzione.
- **Testo strutturato.** Utilizzato per moduli di bibliografia, brevetti e qualsiasi file contenente strutture regolari che possono essere identificate e analizzate. Questo tipo di documento viene utilizzato per ignorare tutto o parte del processo di estrazione. Esso consente di definire i separatori, assegnare i tipi ed imporre un valore di frequenza minima. Se viene selezionata questa opzione, fare clic sul pulsante **Impostazioni** ed immettere i separatori di testo nell'area **Formattazione testo strutturato** della finestra di dialogo Impostazioni documento. Consultare la sezione “Impostazioni di documento per la scheda Campi” a pagina 22 per ulteriori informazioni.
- **Testo XML.** Utilizzare per specificare i tag XML che contengono il testo da estrarre. Tutti gli altri tag vengono ignorati. Se si seleziona questa opzione, fare clic sul pulsante **Impostazioni** e specificare esplicitamente gli elementi XML che contengono il testo da leggere durante il processo di estrazione nell'area **Formattazione testo XML** della finestra di dialogo Impostazioni documento. Consultare la sezione “Impostazioni di documento per la scheda Campi” a pagina 22 per ulteriori informazioni.

Unità di testo. Questa opzione è disponibile solo se è stato specificato che il campo di testo rappresenta **Nomi percorso di documenti** e selezionato **Testo completo** come tipo di documento. Selezionare la modalità di estrazione tra i seguenti:

- **Modo Documento.** Utilizzare per i documenti brevi e semanticamente omogenei, come articoli da agenzie stampa.
- **Modo Paragrafo.** Utilizzare per le pagine Web e documenti non taggati. Il processo di estrazione divide semanticamente i documenti, approfittando delle caratteristiche quali tag interne e la sintassi. Se viene selezionata questa modalità, il calcolo del punteggio viene applicato paragrafo per paragrafo. Pertanto, ad esempio, la regola `mela & arancia` è true solo se `mela` e `arancia` si trovano nello stesso paragrafo.

  Nota: In base al criterio con cui viene estratto il testo dai documenti PDF, la **Modalità paragrafo** non funziona per questi documenti. Ciò accade perché l'estrazione elimina il marker di ritorno a capo.

Impostazioni di modo paragrafo. Questa opzione è disponibile solo se è stato specificato che il campo di testo rappresenta **nomi percorso di documenti** e se è stata impostata l'opzione unità di testo per **modo Paragrafo**. Specificare le soglie dei caratteri da utilizzare in qualsiasi estrazione. La dimensione effettiva è arrotondata per eccesso o per difetto al periodo più vicino. Per verificare che le associazioni di parole prodotte dal testo della raccolta di documenti sono rappresentative, evitare di specificare una dimensione di estrazione troppo piccola.

- **Minimo.** Specifica il numero minimo di caratteri da utilizzare in ogni estrazione.
- **Massimo.** Specifica il numero massimo di caratteri da utilizzare in ogni estrazione.

Codifica di input. Questa opzione è disponibile solo se è stato indicato che il campo di testo rappresenta **nomi percorso di documenti**. Specifica la codifica di testo predefinita. Per tutte le lingue, tranne giapponese, viene eseguita una conversione dalla codifica specificata o riconosciuta in `ISO-8859-1`. Quindi anche se si desidera specificare un'altra codifica, il motore di estrazione la converte in `ISO-8859-1` prima che venga elaborato. Qualsiasi carattere che non rientra nella definizione di codifica `ISO-8859-1` verrà convertito in spazio. Per il testo giapponese, è possibile scegliere una tra le diverse opzioni di codifica: `SHIFT_JIS`, `EUC_JP`, `UTF-8` o `ISO-2022-JP`.

Modo Partizione. Utilizzare la modalità di partizione per scegliere la suddivisione in partizioni in base alle impostazioni del nodo tipo o per selezionare un'altra partizione. La suddivisione separa i dati in campioni di addestramento e di test.

Impostazioni di documento per la scheda Campi

Formattazione del testo strutturato

Per saltare in parte o tutto il processo di estrazione in quanto si dispone di dati strutturati o si vogliono imporre regole su come gestire il testo, utilizzare l'opzione di tipo documento del **testo strutturato** e dichiarare i campi o tag che contengono il testo nella sezione **Formattazione testo strutturato** della finestra di dialogo Impostazioni documento. I termini estratti derivano solo dal testo contenuto all'interno dei campi dichiarati o tag (e tag derivati). Qualsiasi campo o tag non dichiarati verranno ignorati.

In determinati contesti, l'elaborazione linguistica non è richiesta e il motore di estrazione linguistica può essere sostituita da dichiarazioni esplicite. In un file di bibliografia in cui i campi di parola chiave sono separati da separatori come punto e virgola (;) o virgola (,), è sufficiente estrarre la stringa tra due separatori. Per questo motivo, è possibile sospendere il processo di estrazione e definire invece regole di trattamento speciale per dichiarare i separatori di termini, assegnare i tipi al testo estratto o imporre un conteggio della frequenza minima per l'estrazione.

Utilizzare le seguenti regole quando si dichiarano elementi di testo strutturato:

- Solo un campo, tag o elemento per riga può essere dichiarato. Non devono essere presenti nei dati.
- Le dichiarazioni distinguono tra caratteri maiuscoli e minuscoli.

- Se viene dichiarata una tag che contiene attributi, come `<title id="1234">` e si desidera includere tutte le variazioni o, in questo caso, tutti gli ID, aggiungere la tag senza l'attributo o le virgolette di chiusura (>), come `<title`
- Aggiungere due punti dopo il campo o il nome tag per indicare che si tratta di testo strutturato. Aggiungere questi due punti direttamente dopo il campo o tag, ma prima di qualsiasi separatore, tipi o valori di frequenza, come `author:` o `<place>:`.
- Per indicare che più termini sono contenuti nel campo o tag e che un separatore è utilizzato per indicare i singoli termini, dichiarare il separatore dopo i due punti, ad esempio `author:,` o `<section>:;`.
- Per assegnare un tipo al contenuto individuato nella tag, dichiarare il nome tipo dopo i due punti e un separatore, come `author:,Persona` o `<place>:;Ubicazione`. Dichiarare il tipo utilizzando i nomi come appaiono nell'Editor delle risorse.
- Per definire un conteggio di frequenza minimo per un campo o tag, dichiarare un numero alla fine della riga, come `author:,Persona1` o `<place>:;Ubicazione5`. `n` è il conteggio della frequenza definito, i termini trovati nel campo o nella tag devono ricorrere almeno `n` volte nell'intera serie di documenti o record da estrarre. Ciò richiede inoltre di definire un separatore.
- Se si dispone di una tag che contiene una virgola, è necessario anteporre ai due punti un carattere barra retroversa in modo che la dichiarazione non viene ignorata. Per esempio, se si dispone di un campo denominato `<topic:source>`, immettere come `<topic\:source>`.

Per illustrare la sintassi, si supponga di disporre dei seguenti campi bibliografici ricorrenti

```
author:Morel, Kawashima
abstract:Descrive come sono dichiarati i campi.
publication:Documentazione di Text Mining
datepub:Marzo 2010
```

Per questo esempio, se si vuole che il processo di estrazione si focalizzi su author e abstract ma ignori il resto del contenuto, si possono dichiarare solo i seguenti campi:

```
author:,Person1
abstract:
```

In questo esempio, la dichiarazione di campo `author:,Persona1` indica che l'elaborazione linguistica è stata sospesa sul contenuto del campo. Al contrario, afferma che il campo autore contiene più di un nome, che è separato dal successivo da un separatore di virgola e tali nomi devono essere assegnati al tipo di persona e che se il nome ricorre almeno una volta nell'intera serie di documenti o record, esso deve essere estratto. Poiché il campo `abstract:` è elencato senza altre dichiarazioni, sul campo verrà eseguita la scansione durante l'estrazione e verrà applicata l'elaborazione linguistica standard.

Formattazione testo XML

Se si desidera limitare il processo di estrazione a solo il testo all'interno della tag XML specifica, utilizzare l'opzione **testo XML** del tipo di documento e dichiarare le tag contenenti il testo nella sezione **Formattazione testo XML** della finestra di dialogo Impostazioni documento. I termini estratti sono derivati solo dal testo contenuto all'interno di tali tag o dai tag derivati.

Importante! Se si desidera ignorare il processo di estrazione e imporre norme sui separatori, assegnare i tipi al testo estratto o imporre un conteggio di frequenza di termini estratti, utilizzare l'opzione **di testo strutturato** descritta di seguito.

Utilizzare le seguenti regole quando si dichiarano tag per la formattazione del testo XML:

- È possibile dichiarare una sola tag XML per riga.
- Gli elementi della tag sono sensibili al maiuscolo/minuscolo.
- Se una tag contiene attributi, come `<title id="1234">` e si desidera includere tutte le variazioni o, in questo caso, tutti gli ID, aggiungere la tag senza l'attributo o le virgolette di chiusura (>), come `<title`

Per illustrare la sintassi, si supponga di disporre del seguente documento XML

```
<section>Codice della strada
     <title id="01234">Segnali</title>
     <p>I segnali stradali sono utili.</p>
</section>
<p>Apprendere le regole è importante.</p>
```

In questo esempio, si dichiarano le seguenti tag:

```
<section>
<title
```

In questo esempio, poiché è stata dichiarata la tag `<section>`, il testo in questa tag e la sue tag concatenate, `Segnali` e `I segnali stradali sono utili`, vengono sottoposti a scansione durante il processo di estrazione. Tuttavia, `Apprendere le regole è importante` viene ignorato poiché la tag `<p>` non è stata dichiarata esplicitamente né è stata concatenata all'interno di una tag dichiarata.

Nodo di estrazione testo: scheda Modello

La scheda Modello viene utilizzata per specificare il metodo di creazione e le impostazioni del modello generale per l'output del nodo.

È possibile impostare i seguenti parametri:

Nome modello È possibile generare il nome modello automaticamente in base al campo ID o obiettivo (o in base al tipo di modello nei casi in cui non è specificato tale campo) o specificare un nome personalizzato.

Utilizza dati partizionati. Se è definito un campo partizione, questa opzione garantisce che per la creazione del modello vengano utilizzati dati solo dalla partizione di addestramento.

Modo creazione. Specifica il modo in cui il nugget del modello verrà prodotto quando viene eseguito un flusso con questo nodo di estrazione testo. In alternativa, è possibile utilizzare un approccio sperimentale, manuale utilizzando la modalità di **creazione interattiva** in cui non solo si possono estrarre concetti, creare categorie e perfezionare le proprie risorse linguistiche, ma anche eseguire l'analisi di collegamento del testo ed esplorare i cluster.

- **Creazione interattiva.** Durante l'esecuzione di un flusso, questa opzione avvia un'interfaccia interattiva in cui è possibile estrarre concetti e modelli, esplorare e ottimizzare i risultati estratti, creare e affinare le categorie, ottimizzare le risorse linguistiche (modelli, sinonimi, i tipi, librerie, ecc.)e creare nugget del modello di categoria. Consultare la sezione “Creazione interattiva” a pagina 25 per ulteriori informazioni.
- **Genera direttamente.** Questa opzione indica che, quando viene eseguito il flusso, un modello deve essere creato automaticamente e aggiunto alla tavolozza Modelli. A differenza del workbench interattivo, non è necessaria alcuna ulteriore manipolazione al momento dell'esecuzione oltre alle impostazioni definite nel nodo. Se si seleziona questa opzione, le opzioni specifiche del modello vengono visualizzate e con esse è possibile definire il tipo di modello che si desidera generare. Consultare la sezione “Genera direttamente” a pagina 26 per ulteriori informazioni.

Copia risorse da. Quando si estrae testo, l'estrazione si basa non solo sulle impostazioni nella scheda Avanzate ma anche sulle risorse linguistiche. Tali risorse servono come base per come gestire ed elaborare il testo durante l'estrazione in modo da ottenere i concetti, i tipi e a volte i modelli. È possibile copiare le risorse in questo nodo da un modello di risorsa o da un pacchetto di analisi del testo (TAP). Selezionarne uno e quindi fare clic su **Carica** per definire il pacchetto o il modello da cui le risorse verranno copiate. Al momento in cui si desidera caricare, una copia delle risorse selezionate viene memorizzata nel nodo. Pertanto, se mai si desiderava utilizzare un modello aggiornato o TAP, è necessario ricaricare qui o in una

sessione workbench interattiva. Per comodità, la data e l'ora in cui le risorse sono state copiate e caricate vengono visualizzate nel nodo. Consultare la sezione “Copia risorse da modelli e TAP” a pagina 27 per ulteriori informazioni.

Lingua di testo. Identifica la lingua del testo estratto. Le risorse copiate nel nodo controllano le opzioni di lingua riportate. È possibile selezionare la lingua per la quale le risorse sono state regolate o scegliere l'opzione **TUTTI**. Si consiglia vivamente di specificare la lingua esatta per i dati di testo; tuttavia, se non si è sicuri, è possibile scegliere l'opzione **TUTTI**. **TUTTI** non è disponibile per il testo giapponese. Questa opzione **TUTTI** incrementa il tempo di esecuzione poiché il riconoscimento automatico della lingua viene utilizzato per eseguire la scansione di tutti i documenti e record in modo da identificare innanzitutto la lingua di testo. Con questa opzione, tutti i record o i documenti in una lingua supportata e con licenza vengono letti dal motore di estrazione utilizzando i dizionari interni appropriati per la lingua. Per ulteriori informazioni, consultare la sezione “Identificativo di lingua” a pagina 213. Rivolgersi al rappresentante delle vendite se si è interessati all'acquisto di una licenza per una lingua supportata alla quale attualmente non si ha accesso.

Creazione interattiva

Nella scheda Modello del nodo modellazione di estrazione testo, è possibile scegliere una modalità di generazione per i nugget del modello. Se si sceglie **creazione interattiva** viene aperta un'interfaccia interattiva quando si esegue il flusso. In questo workbench interattivo, è possibile:

- Estrarre ed esplorare i risultati di estrazione, compresi i concetti e rilevare le idee salienti dei propri dati di testo.
- Utilizzare una varietà di metodi per creare ed estendere le categorie dai concetti, i tipi, o modelli TLA e le regole in modo che sia possibile calcolare il punteggio dei propri documenti e record in queste categorie.
- Perfezionare le proprie risorse linguistiche (modelli di risorse, librerie, dizionari, sinonimi e altro) in modo da poter migliorare i risultati attraverso un processo iterativo in cui i concetti vengono estratti, analizzati e perfezionati.
- Eseguire l'analisi di collegamento del testo (TLA) e utilizzare i modelli TLA rilevati per costruire nugget del modello di categoria migliori. Il nodo Analisi di collegamento del testo non offre le stesse opzioni preliminari o capacità di modellazione.
- Generare i cluster per rilevare nuove relazioni ed esplorare le relazioni tra concetti, tipi, modelli e categorie nel riquadro Visualizzazione.
- Generare nugget del modello di categoria perfezionati per la tavolozza Modelli in IBM SPSS Modeler e utilizzarli in altri sistemi.

Nota: Non è possibile creare un modello interattivo se si sta creando un lavoro IBM SPSS Collaboration and Deployment Services.

Utilizzare una sessione di lavoro (categorie, TLA, risorse, ecc.) dall'ultimo aggiornamento del nodo. Quando si lavora in una sessione workbench interattiva, è possibile aggiornare il nodo con dati sessione (parametri di estrazione, risorse, definizioni di categoria, ecc.). L'opzione **Utilizza lavoro di sessione** consente di rilanciare il workbench interattivo utilizzando i dati di sessione salvate. Questa opzione è disabilitata la prima volta che si utilizza questo nodo, poiché i dati di sessione non sono stati salvati. Per conoscere in che modo aggiornare il nodo con i dati della sessione in modo da poter utilizzare questa opzione, consultare “Aggiornamento dei nodi di modellazione e salvataggio” a pagina 84.

Se si avvia una sessione *con* questa opzione, le impostazioni di estrazione, le categorie, le risorse e qualsiasi altro lavoro dell'ultima volta in cui è stato eseguito un aggiornamento del nodo da una sessione workbench interattiva, sono disponibili quando si avvia una sessione. Poiché i dati di sessione salvati vengono utilizzati con questa opzione, parti del contenuto, ad esempio le risorse copiate dal modello che segue e altre schede vengono disabilitate e ignorate. Ma se si avvia una sessione *senza* questa opzione, viene utilizzato solo il contenuto del nodo come ora è definito, il che significa che qualsiasi lavoro precedente eseguito nel workbench non sarà disponibile.

Nota: se si modifica il nodo di origine per il flusso dopo che i risultati di estrazione sono stati memorizzati con l'opzione **Utilizza lavoro di sessione...** , è necessario eseguire una nuova estrazione una volta che la sessione di workbench interattivo viene avviata, se si desidera ottenere risultati di estrazione aggiornati.

Ignorare l'estrazione e riutilizzare i dati memorizzati nella cache e i risultati. È possibile riutilizzare qualsiasi risultato di estrazione memorizzato nella cache e i dati nella sessione workbench interattiva. Questa opzione è particolarmente utile quando si desidera risparmiare tempo e riutilizzare i risultati dell'estrazione piuttosto che aspettare un'estrazione completamente nuova per essere eseguita quando la sessione viene avviata. Per utilizzare questa opzione, è necessario avere precedentemente aggiornato questo nodo dall'interno di una sessione workbench interattiva e scelto l'opzione **Mantieni i lavori di sessione e i dati cache con i risultati di estrazione per riutilizzare il testo**. Per conoscere in che modo aggiornare il nodo con i dati della sessione in modo da poter utilizzare questa opzione, consultare "Aggiornamento dei nodi di modellazione e salvataggio" a pagina 84.

Inizia sessione per. Seleziona l'opzione che indica la vista e l'azione da intraprendere per prime all'avvio della sessione workbench interattiva. Indipendentemente dalla vista in cui si avvia, è possibile passare a qualsiasi vista una volta nella sessione.

- **Uso dei risultati di estrazione per creare le categorie.** Questa opzione avvia il workbench interattivo nella vista Categorie e concetti e, se applicabile, esegue un'estrazione. In questa vista, è possibile creare categorie e generare un modello di categoria. È inoltre possibile passare ad un'altra vista. Consultare la sezione Capitolo 8, "Modalità Workbench interattivo", a pagina 73 per ulteriori informazioni.
- **Esame dei risultati TLA (text link analysis).** Questa opzione avvia e inizia con l'estrazione e identificazione delle relazioni tra concetti all'interno del testo, come ad esempio le opinioni o altri collegamenti nella vista Analisi di collegamento del testo . È necessario selezionare un pacchetto di modello o di testo che contiene le regole del modello TLA per utilizzare questa opzione e ottenere risultati. Se si sta lavorando con insiemi di dati più estesi, l'estrazione TLA possono richiedere del tempo. In questo caso, è possibile che si consideri di utilizzare il flusso a monte di un nodo campione. Consultare la sezione Capitolo 12, "Esplorazione di analisi di collegamento del testo", a pagina 153 per ulteriori informazioni.
- **Analisi dei cluster di parole associate.** Questa opzione viene avviata nella vista Cluster ed aggiorna tutti i risultati di estrazione superati. In questa vista, è possibile eseguire l'analisi dei cluster di parole associate, che produce una serie di cluster. Il cluster di parole associate è un processo che inizia con la valutazione della forza del valore del collegamento tra due concetti in base alla loro ricorrenza in un record o un determinato documento e termina con il raggruppamento di concetti strettamente collegati in cluster. Consultare la sezione Capitolo 8, "Modalità Workbench interattivo", a pagina 73 per ulteriori informazioni.

Genera direttamente

Nella scheda Modello del nodo modellazione di estrazione testo, è possibile scegliere una modalità di generazione per i nugget del modello. Se si sceglie **Genera direttamente**, è possibile impostare le opzioni nel nodo e quindi eseguire il proprio flusso. L'output è un nugget del modello di concetto, che è stato inserito direttamente nella tavolozza Modelli. A differenza del workbench interattivo, non è necessaria alcuna ulteriore manipolazione al momento dell'esecuzione oltre alle impostazioni di frequenza definite per questa opzione nel nodo.

Numero massimo di concetti da includere nel modello. Questa opzione, che si applica solo quando si crea un modello automaticamente (non interattivo), indica che si desidera creare un modello di concetto. Afferma inoltre che questo modello non deve contenere più del numero specificato di concetti.

- **Controllare i concetti in base a frequenza più alta. Numero più alto di concetti.** A partire dal concetto con la massima frequenza, questo è il numero di concetti che verranno controllati. Qui, frequenza si riferisce al numero di volte che un concetto (e tutti i suoi termini sottostanti) viene visualizzato nell'intera serie dei documenti/record. Questo numero può essere superiore al numero di record, poiché un concetto può essere visualizzato più volte in un record.

- **Deselezionare i concetti che ricorrono in troppi record. Percentuale di record.** Deselezionare i concetti con una percentuale di conteggio di record superiore al numero specificato. Questa opzione è utile per escludere i concetti che ricorrono frequentemente nel testo o in ogni record ma non hanno alcun significato nell'analisi.

Ottimizza per velocità di calcolo del punteggio. Questa opzione è selezionata automaticamente e assicura che il modello creato sia compatto ed esegua punteggi ad alta velocità. Se si deseleziona questa opzione viene creato un modello molto più ampio che calcola i punteggi più lentamente. Tuttavia, il modello garantisce che i punteggi visualizzati inizialmente nel modello di concetto generato sono gli stessi di quelli ottenuti durante il calcolo del punteggio sullo stesso testo con il nugget del modello.

Copia risorse da modelli e TAP

Quando si estrae testo, l'estrazione si basa non solo sulle impostazioni nella scheda Avanzate ma anche sulle risorse linguistiche. Tali risorse servono come base per come gestire ed elaborare il testo durante l'estrazione in modo da ottenere i concetti, i tipi e a volte i modelli. È possibile copiare le risorse in questo nodo da un *modello di risorsa*e se si sta nel nodo di estrazione testo, è possibile anche selezionare un *TAP (text analysis package)*.

Per impostazione predefinita, le risorse vengono copiate nel nodo dal modello di base per la lingua della licenza del prodotto quando si aggiunge il nodo all'area. Se si dispone di licenze per più lingue, la prima lingua selezionata viene utilizzata per determinare il modello da caricare automaticamente.

Al momento in cui si desidera caricare, una copia delle risorse selezionate viene memorizzata nel nodo. Solo il contenuto del modello o TAP viene copiato mentre il modello o TAP stesso non viene collegato al nodo. Ciò significa che se questo modello o TAP viene aggiornato, tali aggiornamenti non sono automaticamente disponibili nel nodo. In breve, le risorse caricate nel nodo vengono sempre utilizzate a meno che non venga ricaricata una copia di un modello o TAP o a meno che non si aggiorna un nodo di estrazione testo e si seleziona l'opzione **Usa lavoro di sessione**. Per ulteriori informazioni su **Usa lavoro di sessione**, andare avanti in questa sezione.

Quando si seleziona un modello o TAP, sceglierne uno con la stessa lingua dei dati di testo. È possibile utilizzare solo i modelli o TAP nelle lingue previste dalla licenza. Se si desidera eseguire l'analisi di collegamento del testo, è necessario selezionare un modello che contiene modelli TLA. Se un modello contiene modelli TLA, un'icona viene visualizzata nella colonna TLA della finestra di dialogo Carica modello di risorsa.

Nota: Non è possibile caricare TAP nel nodo Analisi di collegamento del testo.

Modelli di risorsa

Un modello di risorsa è un insieme predefinito di librerie e risorse linguistiche e non linguistiche avanzate che sono state definite per un particolare dominio o utilizzo. Nel nodo di modellazione di estrazione testo, una copia delle risorse da un modello di base è già stata caricata nel nodo quando si aggiunge il nodo al flusso, ma è possibile modificare i modelli o caricare un pacchetto di analisi testo selezionando **Modello di risorsa** o **TAP** e quindi facendo clic su **Carica**. Per i modelli, è possibile selezionare il modello nella finestra di dialogo Carica risorsa di modello.

Nota: se il modello che si desidera non viene visualizzato nell'elenco ma si dispone di una copia esportata sulla macchina, è possibile importarlo. È anche possibile esportare da questa finestra di dialogo per condividere con altri utenti. Consultare la sezione “Importazione ed esportazione di modelli” a pagina 177 per ulteriori informazioni.

TAP (Text Analysis Package)

Un TAP (text analysis package) è una serie di librerie predefinite e risorse linguistiche e non linguistiche avanzate raggruppate insieme con una o più serie di categorie predefinite. IBM SPSS Modeler Text

Analytics fornisce diversi TAP predefiniti per il testo in lingua inglese e anche per il testo in lingua giapponese, ognuno dei quali viene ottimizzato per un dominio specifico. Non è possibile modificare questi TAP ma è possibile utilizzarli per avviare la creazione del modello di categoria. È inoltre possibile creare i propri TAP nella sessione interattiva. Consultare la sezione “Caricamento dei pacchetti di analisi del testo (TAP)” a pagina 142 per ulteriori informazioni.

Nota: Non è possibile caricare TAP nel nodo Analisi di collegamento del testo.

Utilizzando l'opzione "Usa lavoro di sessione" (scheda Modello)

Se le risorse vengono copiate nel nodo nella scheda Modello, si potrebbero anche apportare modifiche alle risorse successivamente in una sessione interattiva e aggiornare il nodo modellazione estrazione testo con queste ultime modifiche. In questo caso, selezionare l'opzione **Usa lavoro di sessione** nella scheda Modello del nodo di modellazione estrazione testo.

Se si seleziona **Usa lavoro di sessione**, il pulsante **Carica** è disabilitato nel nodo per indicare che quelle risorse che provengono dal workbench interattivo verranno utilizzate al posto delle risorse che sono state caricate qui precedentemente.

Per apportare modifiche alle risorse una volta selezionata l'opzione **Usa lavoro di sessione**, è possibile modificare o passare a risorse direttamente all'interno della sessione workbench interattiva tramite la vista Editor risorse. Consultare la sezione “Aggiornamento delle risorse del nodo dopo il caricamento” a pagina 175 per ulteriori informazioni.

Nodo di estrazione testo: scheda Avanzate

La scheda Avanzate contiene determinati parametri avanzati che influenzano il modo in cui il testo viene estratto e gestito. I parametri contenuti in questa finestra di dialogo controllano il comportamento di base, nonché una qualche funzionalità avanzata, del processo di estrazione. Tuttavia, essi rappresentano solo una parte delle opzioni disponibili. Vi sono inoltre alcune risorse linguistiche e le opzioni che influenzano i risultati di estrazione, che sono controllate dal modello di risorsa è selezionato nella scheda Modello. Consultare la sezione “Nodo di estrazione testo: scheda Modello” a pagina 24 per ulteriori informazioni.

Nota: La scheda intera è disabilitata se è stata selezionata la modalità **creazione interattiva** utilizzando le informazioni salvate nel workbench interattivo nella scheda Modello, nel qual caso le impostazioni di estrazione vengono prese dall'ultima sessione del workbench salvato.

Per testo olandese, inglese, francese, tedesco, italiano, portoghese e spagnolo

È possibile impostare i seguenti parametri quando l'estrazione per lingue diverse dal giapponese come l'inglese, lo spagnolo, il francese, il tedesco e così via:

Nota: Andare avanti nella sezione per informazioni relative alle impostazioni avanzate per il testo giapponese.

Limita estrazione ai concetti con una frequenza globale di almeno [n]. Specifica il numero minimo di volte in cui una parola o una frase deve ricorrere nel testo per essere estratta. In questo modo, un valore di 5 limita l'estrazione a quelle parole o frasi che ricorrono almeno cinque volte nell'intera serie di record o documenti.

In alcuni casi, la modifica di questo limite può fare una grande differenza nei risultati di estrazione e, di conseguenza, nelle categorie. Se, ad esempio, si stanno gestendo i dati di un ristorante, il limite 1 è adeguato per questa opzione. In questo caso, è possibile trovare *pizza (1), pizza piccola (2), pizza spinaci (2)*e *pizza preferita (2)* nei risultati di estrazione. Tuttavia, se si stava per limitare l'estrazione ad una frequenza globale di 5 o superiore e si riesegue l'estrazione, non sarebbe più possibile ottenere tre di questi concetti. Invece si potrebbe ottenere *pizza (7)*, poiché *pizza* è la forma più semplice e anche perché questo termine

già esisteva come possibile candidato. E a seconda del resto del testo, si potrebbe effettivamente avere una frequenza di oltre sette, a seconda se vi sono ancora altre frasi con pizza nel testo. Inoltre, se *pizza spinaci* è stato già un descrittore di categoria, potrebbe essere necessario aggiungere *pizza* come descrittore invece di acquisire tutti i record. Per questo motivo, modificare questo limite con attenzione quando le categorie sono già state create.

Si tratta di un'estrazione-solo funzione; se il modello contiene i termini (cosa che succede di solito) e un termine per il modello si trova nel testo, il termine verrà indicizzato indipendentemente dalla sua frequenza.

Ad esempio, si supponga di utilizzare un modello di risorse di base che include "los angeles" nel tipo `<Ubicazione>` nella libreria principale; se il documento contiene Los Angeles solo una volta, Los Angeles sarà parte dell'elenco di concetti. Per evitare ciò sarà necessario impostare un filtro per visualizzare i concetti che ricorrono almeno lo stesso numero di volte indicato dal valore immesso nel campo **Limita estrazione ai concetti con una frequenza globale di almeno [n]**.

Correggi errori di punteggiatura. Questa opzione normalizza temporaneamente testo contenente errori di punteggiatura (ad esempio, l'uso improprio) durante l'estrazione per migliorare la possibilità di estrazione dei concetti. Questa opzione risulta estremamente utile quando il testo è breve e di scarsa qualità (come, ad esempio, le risposte del sondaggio aperto, e-mail e i dati CRM), o quando il testo contiene molte abbreviazioni.

Correzione di errori ortografici per un limite minimo di caratteri root [n]. Questa opzione applica una tecnica di raggruppamento che consente di raggruppare parole errate o parole simili in un unico concetto. L'algoritmo di raggruppamento toglie temporaneamente tutte le vocali (tranne la prima) e stacca consonanti doppie/triple da parole estratte e poi le confronta per vedere se sono uguali in modo che `modeli` e `modelli` verrebbero raggruppate insieme. Tuttavia, se ogni termine viene assegnato ad un tipo differente, escluso il tipo `<Sconosciuto>`, la tecnica di raggruppamento confuso non verrà applicata.

È possibile inoltre definire il numero minimo di caratteri *radice* richiesti prima di utilizzare il raggruppamento. Il numero di caratteri radice in un termine è calcolato sommando tutti i caratteri e sottraendo quelli che formano i suffissi di desinenza e, nel caso di parole composte, i determinativi e le preposizioni. Ad esempio, il termine `esercizi` potrebbe essere conteggiato come 9 caratteri radice nel formato "esercizio", poiché la lettera *i* alla fine della parola è un flesso (forma plurale). Allo stesso modo, `succo di mela` conta 11 caratteri radice (“succo di mela”) e `fabbrica di automobili` conta 20 caratteri radice. Questo metodo di conteggio viene utilizzato solo per verificare se il raggruppamento deve essere applicato ma non influenza il modo in cui le parole sono corrispondenti.

Nota: se si riscontra che determinate parole vengono successivamente raggruppate in modo non corretto, è possibile escludere da questa tecnica le coppie di parole esplicitamente dichiarate nella sezione **Raggruppamento confuso: eccezioni** nella scheda Avanzate. Per ulteriori informazioni, consultare la sezione “Raggruppamento confuso” a pagina 206.

Estrazione termini univoci. Questa opzione estrae parole singole (termini univoci) purché il termine non è già parte di una parola composta e se è un nome o una parte del discorso non riconosciuta.

Estrazione entità non linguistiche. Questa opzione estrae entità non linguistiche, ad esempio numeri di telefono, numeri di codice fiscale, orari, date, valute, cifre, percentuali, indirizzi e-mail e indirizzi HTTP. È possibile includere o escludere alcuni tipi di entità non linguistiche nella sezione **Entità non linguistiche: configurazione** della scheda Avanzate. Disabilitando qualsiasi entità non necessaria, il motore di estrazione non spreca tempo di elaborazione. Per ulteriori informazioni, consultare la sezione “Configurazione” a pagina 210.

Algoritmo maiuscolo. Questa opzione estrae i termini semplici e composti che non si trovano nei dizionari incorporati, purché la prima lettera del termine sia in maiuscolo. Questa opzione fornisce un modo efficace di estrarre i sostantivi più appropriati.

Raggruppare insieme nomi di persona parziali e completi quando possibile. Questa opzione raggruppa i nomi che vengono visualizzati insieme diversamente nel testo. Questa funzione è utile poiché i nomi vengono spesso definiti nel loro formato completo all'inizio del testo e poi solo da una versione abbreviata. Questa opzione tenta la corrispondenza con qualsiasi termine univoco con il tipo `<Sconosciuto>` per l'ultima parola di tutti i termini composti immessi come `<Persona>`. Ad esempio, se viene rilevato *rossi* e inizialmente immesso come `<Sconosciuto>`, il motore controlla l'estrazione per vedere se tutti i termini composti nel tipo `<Persona>` includono *rossi* come ultima parola, per esempio *giovanni rossi*. Questa opzione non si applica ai nomi di persona poiché la maggior parte non sono mai estratti come termini univoci.

Numero massimo di permutazioni di parole non funzionali. Questa opzione specifica il numero massimo di parole non funzionali che possono essere presenti quando si applica la tecnica di permutazione. Questa tecnica di permutazione raggruppa frasi simili che differiscono tra loro solo per parole non funzionali (ad esempio, di e il) contenute, indipendentemente dall'inflessione. Ad esempio, si imposta questo valore su massimo due parole e vengono estratte `funzionari` e `funzionari dell'azienda`. In questo caso, entrambi i termini estratti verrebbero raggruppati insieme nell'elenco dei concetti poiché entrambi i termini vengono considerati uguali quando `dell'` viene ignorato.

Nota: Per abilitare l'estrazione di risultati di Analisi di collegamento del testo è necessario avviare la sessione con l'opzione di **esplorazione dei risultati dell'analisi di collegamento del testo** e scegliere anche le risorse che contengono le definizioni TLA. È sempre possibile estrarre i risultati TLA successivamente durante una sessione workbench interattiva tramite la finestra di dialogo Impostazioni di estrazione. Consultare la sezione “Estrazione di dati” a pagina 88 per ulteriori informazioni.

Per il testo giapponese

La finestra di dialogo contiene diverse opzioni per il testo giapponese, dal momento che il processo di estrazione presenta alcune differenze. Al fine di lavorare con il testo giapponese, è necessario anche selezionare un modello o un pacchetto di analisi del testo (TAP) ottimizzati per la lingua giapponese nella scheda Modello di questo nodo. Consultare la sezione “Copia risorse da modelli e TAP” a pagina 27 per ulteriori informazioni.

Analisi secondaria. Quando viene lanciata un'estrazione, ha luogo l'estrazione di parole chiave di base utilizzando la serie predefinita di tipi. Tuttavia, quando si seleziona un analizzatore secondario, è possibile ottenere molti più concetti e più arricchiti poiché l'estrazione comprenderà particelle e verbi ausiliari come parte del concetto. Nel caso di analisi di opinione, viene incluso anche un numero elevato di tipi aggiuntivi. Inoltre, la scelta di un'analisi secondaria consente di generare anche dei risultati di analisi di collegamento del testo.

Nota: Quando un analizzatore secondario viene richiamato, il processo di estrazione richiede più tempo per completare il processo.

- **Analisi di dipendenza**. La scelta di questa opzione produce particelle estese per i concetti di estrazione dal tipo di base ed estrazione di parola chiave. È anche possibile ottenere risultati di modello più arricchiti dall'analisi TLA di dipendenza.
- **Analisi di opinione**. La scelta di questo tipo di analisi genera ulteriori concetti estratti e, dove applicabile, l'estrazione dei risultati del modello TLA. Oltre ai tipi di base, è possibile anche usufruire di più di 80 tipi di opinioni:. Tali tipi vengono utilizzati per rilevare i concetti e i modelli nel testo mediante l'espressione delle emozioni, opinioni e opinioni. Sono disponibili tre opzioni che regolano il focus per l'analisi di opinione: **Tutti i opinioni**, **Solo opinione rappresentativo** e **Solo conclusioni**.
- **Nessun analizzatore secondario**. Questa opzione disattiva tutti gli analizzatori secondari. Questa opzione viene nascosta se è stata selezionata l'opzione di **esplorazione dei risultati TLA (text link analysis)** nella scheda Modello poiché è richiesta un'analisi secondaria per ottenere risultati TLA. Se viene selezionata questa opzione ma si sceglie successivamente l'opzione di **esplorazione dei risultati TLA (text link analysis)**, si verifica un errore durante l'esecuzione del flusso.

Esempio di flusso a monte per risparmiare tempo

Quando si dispone di una grande quantità di dati, i tempi di elaborazione possono richiedere da minuti a ore, in particolare quando si utilizzano le sessioni workbench interattive. Maggiore è la dimensione dei dati, più tempo è necessario affinché abbiano luogo i processi di estrazione e di categorizzazione. Per lavorare in modo più efficiente, è possibile aggiungere uno dei flussi di nodi campione di IBM SPSS Modeler dal nodo di estrazione testo. Utilizzare questo nodo Campione per prendere un campione casuale utilizzando una sottoserie più piccola di documenti o record per fare i primi passi.

Un campione più piccolo è spesso perfettamente adeguato per decidere come modificare le proprie risorse e persino creare la maggior parte se non tutte le categorie. E dopo avere eseguito operazioni su dataset più piccoli e si è soddisfatti dei risultati, è possibile applicare la stessa tecnica per la creazione di categorie all'intera serie di dati. Quindi, è possibile ricercare documenti o record che non soddisfano le categorie create e apportare modifiche in base alle necessità.

Nota: il nodo Campione è un nodo IBM SPSS Modeler standard.

Utilizzo del nodo di estrazione testo in un flusso

Il nodo di modellazione di estrazione testo viene utilizzato per accedere ai dati ed estrarre concetti in un flusso. È possibile utilizzare qualsiasi nodo origine per accedere ai dati, come ad esempio un nodo Database, Var. Nodo file, nodo internet o nodo file fisso. Per il testo che risiede in documenti esterni, può essere utilizzato un nodo file dell'elenco.

Esempio 1: nodo elenco file e nodo estrazione testo per generare un nugget del modello di concetto direttamente

Il seguente esempio mostra come utilizzare il nodo elenco file insieme al nodo di modellazione di estrazione testo per generare il nugget del modello di concetto. Per ulteriori informazioni sull'uso del nodo di elenco file, consultare “Nodo Elenco file” a pagina 11.

1. **Nodo di elenco file (Scheda Impostazioni).** In primo luogo, è stato aggiunto questo nodo al flusso per specificare dove i documenti di testo sono memorizzati. È stata selezionata la directory contenente tutti i documenti su cui si desidera eseguire l'estrazione di testo.
2. **Nodo di estrazione testo (Scheda Campi).** Inoltre, è stato aggiunto e connesso un nodo Estrazione testo al nodo Elenco file. In questo nodo, è stato definito il formato di input, il modello di risorsa e il formato di output. È stato selezionato il nome campo ottenuto dal nodo Elenco file e selezionata l'opzione in cui il campo di testo rappresenta **i nomi percorso dei documenti** oltre ad altre impostazioni. Consultare la sezione “Utilizzo del nodo di estrazione testo in un flusso” per ulteriori informazioni.
3. **Nodo di estrazione testo (scheda Modello).** Quindi, nella scheda Modello, è stata selezionata la modalità di creazione per generare un nugget del modello di concetto direttamente da questo nodo. È possibile selezionare un modello di risorsa differente oppure conservare le risorse di base.

Esempio 2: nodi file di Excel e di estrazione testo per creare un modello di categoria in modo interattivo

Questo esempio mostra come il nodo di estrazione testo può avviare una sessione workbench interattiva. Per ulteriori informazioni sul workbench interattivo, consultare Capitolo 8, “Modalità Workbench interattivo”, a pagina 73.

1. **Nodo di origine Excel (scheda Dati).** In primo luogo, è stato aggiunto questo nodo al flusso per specificare dove il testo è archiviato.
2. **Nodo di estrazione testo (Scheda Campi).** Inoltre, è stato aggiunto e connesso un nodo di estrazione testo. In questa prima scheda, è stato definito il formato di input. È stato selezionato un nome campo dal nodo di origine e l'opzione. Il campo Testo rappresenta **il testo effettivo** poiché i dati provengono direttamente dal nodo origine Excel.

3. **Nodo di estrazione testo (scheda Modello).** Successivamente, nella scheda Modello, è stato scelto di creare un nugget del modello di categoria in modo interattivo e di utilizzare i risultati di estrazione per creare categorie automaticamente. In questo esempio, è stata caricata una copia di risorse e una serie di categorie da un pacchetto di analisi del testo.
4. **Sessione workbench interattiva**. È stato eseguito il flusso ed è stata aperta l'interfaccia workbench interattiva. Dopo è stata eseguita un'estrazione ed è iniziata l'esplorazione dei dati, migliorando così la categoria.

Nugget di estrazione testo: modello di concetto

Un nuggett del modello di concetto viene creato ogni volta che si esegue correttamente un nodo di modello di estrazione testo dove è stata selezionata l'opzione per **generare un modello direttamente** nella scheda Modello. Un nugget del modello di concetto di estrazione testo è utilizzato per il rilevamento in tempo reale dei concetti chiave in altri dati di testo, come i dati di riempimento da un call center.

Lo stesso nugget del modello di concetto comprende un elenco di concetti, che sono stati assegnati ai tipi. È possibile selezionare uno o tutti i concetti in questo modello per il calcolo del punteggio rispetto ad altri dati. Quando si esegue un flusso contenente un nugget del modello di estrazione testo, vengono aggiunti nuovi campi ai dati in base alla modalità di creazione selezionata nella scheda Modello del nodo di modellazione di estrazione testo prima della creazione del modello. Consultare la sezione "Modello di concetto: scheda Modello" per ulteriori informazioni.

Se il nugget del modello è stato generato utilizzando documenti tradotti, il calcolo del punteggio verrà eseguito nella lingua tradotta. Allo stesso modo, se il nugget del modello è stato generato utilizzando l'inglese come lingua, è possibile specificare una lingua di traduzione nel nugget del modello, poiché i documenti verranno poi tradotti in inglese.

Quando vengono generati, i nugget del modello di estrazione testo sono posizionati nella tavolozza del nugget del modello (ubicata nella scheda Modelli nella parte superiore destra della finestra IBM SPSS Modeler).

Visualizzazione di risultati

Per visualizzare le informazioni sui nugget del modello, fare clic con il tasto destro del mouse sul nodo nella tavolozza dei nugget del modello e scegliere **Sfoglia** dal menu di scelta rapida (oppure **Modifica** per i nodi in un flusso).

Aggiunta di modelli ai flussi

Per aggiungere il nugget del modello al flusso, fare clic sull'icona nella tavolozza dei nugget del modello e quindi fare clic sull'area del flusso in cui si desidera posizionare il nodo. Oppure fare clic con il tasto destro del mouse sull'icona e selezionare **Aggiungi a flusso** dal menu di scelta rapida. Collegare quindi il flusso al nodo. A questo punto è possibile passare i dati per generare previsioni.

Attenzione: se si desidera utilizzare un nugget di calcolo del punteggio per rigenerare un nodo di modellazione che contiene sia il modello di categoria che il modello utilizzato, si consiglia di creare un TAP e di utilizzarlo in una sessione interattiva in sostituzione del nodo di modellazione prima di generare il nugget di calcolo del punteggio.

Modello di concetto: scheda Modello

Nei modelli di concetto, la scheda Modello visualizza la serie di concetti che sono stati estratti. I concetti vengono presentati in un formato tabella con una riga per ciascun concetto. L'obiettivo in questa scheda è selezionare quale dei concetti verranno utilizzati per il calcolo del punteggio.

Nota: se è stato invece generato un nugget di modello di categoria, questa tabella conterrà informazioni differenti. Consultare la sezione "Nugget del modello di categoria: scheda Modello" a pagina 42 per ulteriori informazioni.

Tutti concetti vengono selezionati per il calcolo del punteggio per impostazione predefinita, come mostrato nelle caselle di spunta nella colonna più a sinistra. Una casella selezionata indica che il concetto verrà utilizzato per il calcolo del punteggio. Una casella deselezionata indica che il concetto verrà escluso dal calcolo del punteggio. È possibile selezionare più righe selezionandole e facendo clic su una delle caselle di spunta nella selezione.

Per saperne di più su ciascun concetto, è possibile esaminare le informazioni aggiuntive fornite in ognuna delle seguenti colonne:

Concetto. Questo è la parola o frase di riferimento che è stata estratta. In alcuni casi, questo concetto rappresenta il nome concetto e alcuni altri termini sottostanti associati a questo concetto. Per vedere quali termini sottostanti sono parte di un concetto, visualizzare il riquadro Termini sottostanti all'interno di questa scheda e selezionare il concetto per vedere i termini corrispondenti nella parte inferiore della finestra di dialogo. Consultare la sezione "Termini sottostanti in modelli di concetto" a pagina 34 per ulteriori informazioni.

Globale. Qui, globale (frequenza) fa riferimento al numero di volte che un concetto (e tutti i suoi termini sottostanti) viene visualizzato nell'intera serie dei documenti/record.

- **Grafico a barre.** La frequenza globale di questo concetto nei dati di testo presentata come un grafico a barre. La barra prende il colore del tipo a cui il concetto è assegnato per distinguere visivamente i tipi.
- **%.** La frequenza globale di questo concetto nei dati di testo presentata come una percentuale.
- **N.** Il numero reale di ricorrenze di questo concetto nei dati di testo.

Documenti. Qui, Documenti fa riferimento al conteggio dei documenti, ovvero è il numero di documenti o record in cui il concetto (e tutti i suoi termini sottostanti) appare.

- **Grafico a barre.** Il conteggio documenti per questo concetto presentato come un grafico a barre. La barra prende il colore del tipo a cui il concetto è assegnato per distinguere visivamente i tipi.
- **%.** Il conteggio documenti per questo concetto presentato come una percentuale.
- **N.** Il numero effettivo di documenti o record contenente questo concetto.

Tipo. Il tipo a cui il concetto è assegnato. Per ogni concetto, le colonne Globale e Documenti vengono visualizzate in un colore per indicare il tipo a cui è assegnato questo concetto. Un **tipo** è un raggruppamento semantico di concetti. Consultare la sezione "Dizionari di tipo" a pagina 189 per ulteriori informazioni.

Gestione dei concetti

Premendo il tastino destro del mouse su una tabella, è possibile visualizzare un menu di scelta rapida in cui sono possibili le seguenti azioni:

- **Seleziona tutti.** Tutte le righe nella tabella vengono selezionate.
- **Copia.** I concetti selezionati vengono copiati negli Appunti.
- **Copia con campi** I concetti selezionati vengono copiati negli appunti insieme all'intestazione di colonna.
- **Attiva selezionati.** Attiva tutte le caselle di spunta per le righe selezionate nella tabella, compresi i concetti per il calcolo del punteggio.
- **Non attivare selezionati.** Deseleziona tutte le caselle di spunta per le righe selezionate nella tabella.
- **Attiva tutti.** Attiva tutte le caselle di spunta nella tabella. Questo è il risultato in tutti i concetti in uso nell'output finale.

- **Disattiva tutti.** Deseleziona tutte le caselle di spunta nella tabella. La deselezione di un concetto indica che non verrà utilizzato nell'output finale.
- **Includi concetti.** Visualizza la finestra di dialogo del nodo Includi concetto. Consultare la sezione "Opzioni per l'inclusione dei concetti per il calcolo del punteggio" per ulteriori informazioni.

Opzioni per l'inclusione dei concetti per il calcolo del punteggio

Per selezionare o deselezionare rapidamente i concetti che verranno utilizzato per il calcolo del punteggio, fare clic sul pulsante della barra degli strumenti **Includi concetti.**.

Figura 1. Pulsante della barra degli strumenti Includi concetti

Facendo clic su questo pulsante della barra degli strumenti verrà visualizzata la finestra di dialogo Includi concetti che consente di selezionare i concetti basati su regole. Tutti i concetti che presentano un segno di spunta nella scheda Modello verranno inclusi per il calcolo del punteggio. Applicare una regola in questa finestra secondaria per modificare quali concetti che verranno utilizzati per il calcolo del punteggio.

Sono disponibili le opzioni seguenti:

Controllare i concetti in base a frequenza più alta. Numero più alto di concetti. A partire dal concetto con la massima frequenza globale, questo è il numero di concetti che verranno controllati. Qui frequenza si riferisce al numero di volte che un concetto (e tutti i suoi termini sottostanti) viene visualizzato nell'intera serie dei documenti/record. Questo numero può essere superiore al numero di record, poiché un concetto può essere visualizzato più volte in un record.

Controllare i concetti in base al conteggio di documenti. Conteggio minimo. Indica il conteggio di documenti minimo necessario affinché i concetti vengano controllati. Il conteggio di documenti fa riferimento al numero di documenti/record in cui appare il concetto (e tutti i suoi termini sottostanti).

Controllare i concetti assegnati al tipo. Selezionare un tipo dall'elenco a discesa per controllare tutti i concetti che sono assegnati a questo tipo. I concetti vengono assegnati ai tipi automaticamente durante il processo di estrazione. Un **tipo** è un raggruppamento semantico di concetti. I tipi includono elementi come i concetti di livello superiore, le parole positive e negative e i qualificativi, i qualificativi di contesto, i nomi, i luoghi, le organizzazioni e molto altro ancora. Consultare la sezione "Dizionari di tipo" a pagina 189 per ulteriori informazioni.

Deselezionare i concetti che ricorrono in troppi record. Percentuale di record. Deselezionare i concetti con una percentuale di conteggio di record superiore al numero specificato. Questa opzione è utile per escludere i concetti che ricorrono frequentemente nel testo o in ogni record ma non hanno alcun significato nell'analisi.

Deselezionare i concetti assegnati al tipo. Deseleziona i concetti corrispondenti al tipo selezionati dall'elenco a discesa.

Termini sottostanti in modelli di concetto

È possibile visualizzare i termini sottostanti definiti per i concetti selezionati nella tabella. Facendo clic sul pulsante di attivazione/disattivazione dei termini sottostanti sulla barra degli strumenti, è possibile visualizzare la tabella dei termini sottostanti in un pannello suddiviso, nella parte inferiore della finestra di dialogo.

Tali termini sottostanti comprendono i sinonimi definiti nelle risorse linguistiche (indipendentemente dal fatto che sono stati rilevati nel testo o meno) così come gli eventuali forme trovate nel testo

singolari/plurali, utilizzate per generare nugget di modello, termini permutati, termini da raggruppamenti casuali e così via.

Figura 2. Pulsante della barra degli strumenti Termini sottostanti

Nota: non è possibile modificare l'elenco dei termini sottostanti. Questo elenco viene generato tramite le sostituzioni, le definizioni di sinonimi (nel dizionario di sostituzione), raggruppamento confuso e altro - tutti i dati vengono definiti nelle risorse linguistiche. Per effettuare modifiche su come i termini vengono raggruppati sotto un concetto o su come vengono gestiti, è necessario apportare le modifiche direttamente nelle risorse (modificabili in Editor risorse nel workbench interattivo o in Editor di modelli e quindi ricaricare nel nodo) e rieseguire il flusso per ottenere un nuovo nugget del modello con i risultati aggiornati.

Premendo il tastino destro del mouse sulla cella contenente un termine o concetto sottostante, è possibile visualizzare un menu di scelta rapida nel quale è possibile:

- **Copiare.** La cella selezionata viene copiata negli appunti.
- **Copia con campi.** La cella selezionata viene copiata negli appunti insieme alle intestazioni di colonna.
- **Seleziona tutti.** Tutte le celle nella tabella vengono selezionate.

Modello di concetto: scheda Impostazioni

La scheda Impostazioni viene utilizzata per definire il valore del campo di testo per i dati di input nuovi, se necessario. Questo è anche il luogo in cui è possibile definire il modello di dati per l'output (in modalità di calcolo del punteggio).

Nota: Questa scheda viene visualizzata solo quando il nugget del modello è posizionato in un'area. Esso non viene mostrato quando si accede a questa finestra di dialogo direttamente nella tavolozza Modelli.

Modalità calcolo del punteggio: concetti come record

Con questo modo di calcolo del punteggio, viene creato un nuovo record per ogni coppia `concetto/documento`. Di solito, ci sono più record nell'output di quanti ce ne sono stati nell'input.

Oltre ai campi di input, i seguenti nuovi campi vengono aggiunti ai dati:

Tabella 4. Campi di output per "Concetti come record".

Campo	Descrizione
`Concetto`	Contiene il nome concetto estratto trovato nel campo dei dati di testo.
`Type`	Memorizza il tipo del concetto come nome tipo completo, come *Ubicazione* o *Persona*. Un **tipo** è un raggruppamento semantico di concetti. Consultare la sezione “Dizionari di tipo” a pagina 189 per ulteriori informazioni.
`Count`	Visualizza il numero di ricorrenze per quel concetto (e i suoi termini sottostanti) nel corpo del testo (record/documento).

Quando si seleziona questa opzione, tutte le altre opzioni eccetto **Correggi errori di punteggiatura** sono disabilitati.

Modalità calcolo del punteggio: concetti come campi

In modelli di concetto, per ogni record di input, un nuovo record viene creato per ogni concetto che si trova in un determinato documento. Di conseguenza, il numero di record di output è uguale a quello dei record di input. Tuttavia, ogni record (riga) contiene ora un nuovo campo (colonna) per ogni concetto che

è stato selezionato (tramite il segno di spunta) nella scheda Modello. Il valore per ogni campo concetto dipende da se si seleziona **Flag** o **Conteggi** come valore del campo in questa scheda.

Nota: Se si utilizzano dataset di grandi dimensioni, ad esempio, con un database DB2, l'utilizzo di **Concetti come campi** può rilevare problemi di elaborazione a causa della quantità di dati. In questo caso, invece, si consiglia l'utilizzo di **Concetti come record**.

Valori di campo. Scegliere se il nuovo campo per ciascun concetto conterrà un valore conteggio o un valore di flag.

- **Flag.** Questa opzione viene utilizzata per ottenere flag con due valori distinti nell'output, per esempio *Sì/No*, *True/False*, *T/F*, o *1* e *2*. I tipi di archiviazione vengono impostati automaticamente in modo da rispecchiare i valori scelti. Ad esempio, se si immettono valori numerici per i flag, questi verranno automaticamente gestiti come un valore intero. I tipi di archiviazione per i flag possono essere: stringa, numero intero, numero reale o data/ora. Immettere un valore flag per **True** e per **False**.
- **Conteggi.** Utilizzato per ottenere un conteggio di quante volte il concetto si è verificato in un dato record.

Estensione nome campo. Specificare un'estensione per il nome campo. I nomi campo vengono generati utilizzando il nome concetto più questa estensione.

- **Aggiungi come.** Specificare dove l'estensione deve essere aggiunta al nome campo. Scegliere **Prefisso** per aggiungere l'estensione all'inizio della stringa. Scegliere **Suffisso** per aggiungere l'estensione alla fine della stringa.

Correggi errori di punteggiatura. Questa opzione normalizza temporaneamente testo contenente errori di punteggiatura (ad esempio, l'uso improprio) durante l'estrazione per migliorare la possibilità di estrazione dei concetti. Questa opzione risulta estremamente utile quando il testo è breve e di scarsa qualità (come, ad esempio, le risposte del sondaggio aperto, e-mail e i dati CRM), o quando il testo contiene molte abbreviazioni.

Nota: L'opzione **Correggi errori di punteggiatura** non si applica quando si gestisce il testo giapponese.

Modello di concetto: scheda Campi

La scheda Campi viene utilizzata per definire il valore del campo di testo per i nuovi dati di input, se necessario.

Nota: questa scheda viene visualizzata solo quando il nugget del modello viene collocato nel flusso. Esso non esiste quando si accede a questo output direttamente nella tavolozza Modelli.

Campo Testo. Selezionare il campo che contiene il testo da estrarre, il nome percorso del documento o il nome percorso della directory per i documenti. Questo campo dipende dall'origine dati.

Il campo Testo rappresenta. Indica cosa contiene il campo di testo specificato nella precedente impostazione. Le scelte sono:

- **Testo reale.** Selezionare questa opzione se il campo contiene il testo esatto da cui devono essere estratti i concetti.
- **Nomi percorso dei documenti.** Selezionare questa opzione se il campo contiene uno o più nomi percorso per l'ubicazione dei documenti di testo.

Tipo documento. Questa opzione è disponibile solo se è stato indicato che il campo di testo rappresenta **nomi percorso di documenti**. Il tipo di documento specifica la struttura del testo. Selezionare uno dei seguenti tipi:

- **Testo.** Da utilizzare per la maggior parte dei documenti o origini di testo. Tutto l'insieme del testo viene analizzato per l'estrazione. A differenza delle altre opzioni, non vi sono ulteriori impostazioni per questa opzione.

- **Testo strutturato.** Utilizzato per moduli di bibliografia, brevetti e qualsiasi file contenente strutture regolari che possono essere identificate e analizzate. Questo tipo di documento viene utilizzato per ignorare tutto o parte del processo di estrazione. Esso consente di definire i separatori, assegnare i tipi ed imporre un valore di frequenza minima. Se viene selezionata questa opzione, fare clic sul pulsante **Impostazioni** ed immettere i separatori di testo nell'area **Formattazione testo strutturato** della finestra di dialogo Impostazioni documento. Consultare la sezione "Impostazioni di documento per la scheda Campi" a pagina 22 per ulteriori informazioni.
- **Testo XML.** Utilizzare per specificare i tag XML che contengono il testo da estrarre. Tutti gli altri tag vengono ignorati. Se si seleziona questa opzione, fare clic sul pulsante **Impostazioni** e specificare esplicitamente gli elementi XML che contengono il testo da leggere durante il processo di estrazione nell'area **Formattazione testo XML** della finestra di dialogo Impostazioni documento. Consultare la sezione "Impostazioni di documento per la scheda Campi" a pagina 22 per ulteriori informazioni.

Codifica di input. Questa opzione è disponibile solo se è stato indicato che il campo di testo rappresenta **nomi percorso di documenti**. Specifica la codifica di testo predefinita. Per tutte le lingue, tranne giapponese, viene eseguita una conversione dalla codifica specificata o riconosciuta in `ISO-8859-1`. Quindi anche se si desidera specificare un'altra codifica, il motore di estrazione la converte in `ISO-8859-1` prima che venga elaborato. Qualsiasi carattere che non rientra nella definizione di codifica `ISO-8859-1` verrà convertito in spazio. Per il testo giapponese, è possibile scegliere una tra le diverse opzioni di codifica: `SHIFT_JIS`, `EUC_JP`, `UTF-8` o `ISO-2022-JP`.

Lingua di testo. Identifica la lingua del testo che viene estratto; questa è la lingua principale rilevata durante l'estrazione. Rivolgersi al rappresentante delle vendite se si è interessati all'acquisto di una licenza per una lingua supportata alla quale attualmente non si ha accesso.

Modello di concetto: scheda Riepilogo

La scheda Riepilogo visualizza le informazioni relative al modello stesso (cartella *Analisi*), i campi utilizzati nel modello (cartella *Campi*), le impostazioni utilizzate durante la creazione del modello (cartella *Impostazioni di creazione*) e l'addestramento del modello (cartella *Riepilogo addestramento*).

Quando si sfoglia per la prima volta un nodo di modellazione, le cartelle nella scheda Riepilogo sono compresse. Per visualizzare i risultati utilizzare il controllo dell'espansore a sinistra della cartella se si desidera visualizzare solo i risultati oppure fare clic sul pulsante **Espandi tutto** per visualizzare tutti i risultati. Per nascondere i risultati dopo la visualizzazione, utilizzare il controllo dell'espansore se si desidera comprimere la cartella specifica che si desidera nascondere oppure fare clic sul pulsante **Comprimi tutto** per comprimere tutte le cartelle.

Uso dei nugget del modello di concetto in un flusso

Quando si utilizza un nodo di modellazione di estrazione testo, è possibile generare un nugget del modello di concetto o un nugget del modello di categoria (attraverso una sessione workbench interattiva). Il seguente esempio mostra come utilizzare un modello di concetto in un flusso semplice.

Esempio: nodo file delle statistiche con il nugget del modello di concetto

Il seguente esempio mostra come utilizzare il nugget del modello di concetto di estrazione testo.

Figura 3. Esempio di flusso: nodo file delle statistiche con il nugget del modello di concetto di estrazione testo

1. **Nodo file delle statistiche (scheda Dati).** In primo luogo, è stato aggiunto questo nodo al flusso per specificare dove i documenti di testo sono memorizzati.

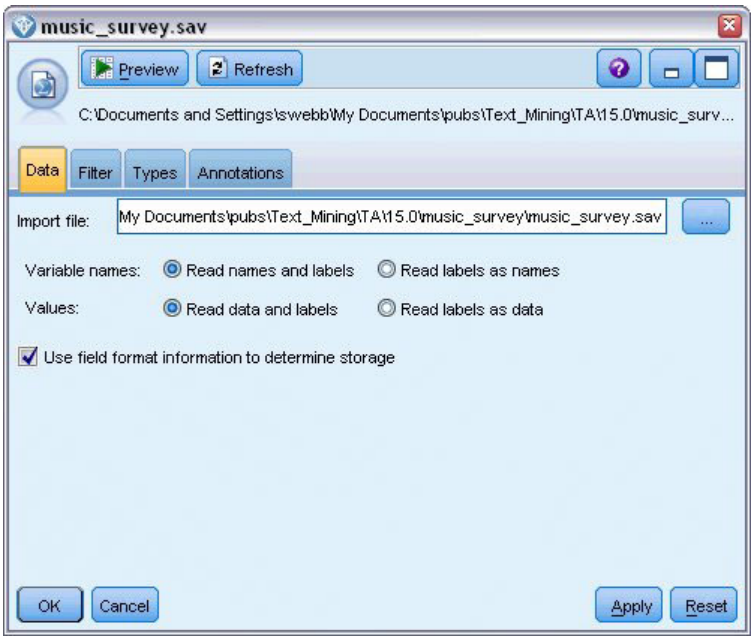


Figura 4. Finestra di dialogo del nodo file delle statistiche: scheda Dati

2. **Nugget del modello di categoria di estrazione testo (scheda Modello).** Inoltre, è stato aggiunto e connesso un nugget del modello di concetto al nodo file delle statistiche. Sono stati selezionati dei concetti da utilizzare per il calcolo del punteggio dei dati.

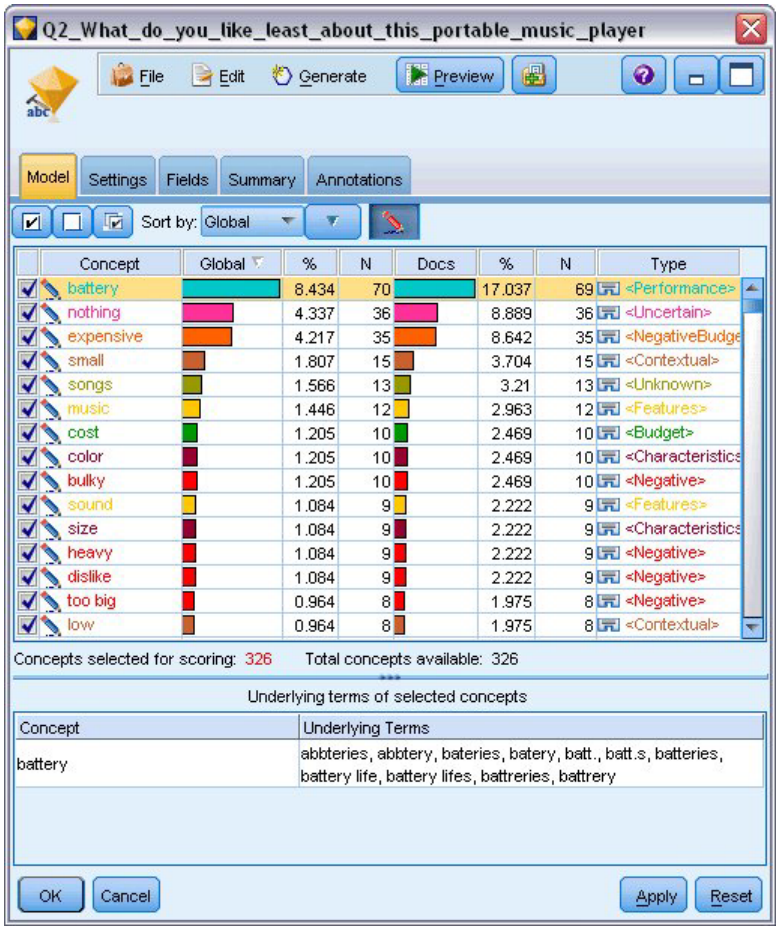


Figura 5. Finestra di dialogo del nugget del modello di estrazione testo: scheda Modello

3. **Nugget del modello di concetto di estrazione testo (scheda Impostazioni).** Successivamente, è stato definito il formato di output ed è stato selezionato *Concetti come campi*. Un nuovo campo verrà creato nell'output per ciascun concetto selezionato nella scheda Modello. Ciascun nome campo viene creato dal nome concetto e dal prefisso "*Concept_*"

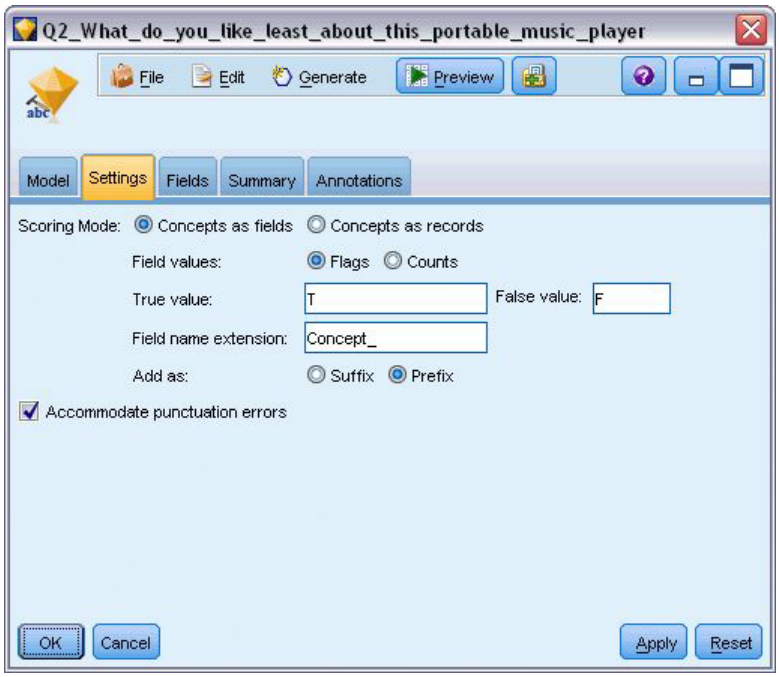


Figura 6. Finestra di dialogo del nugget del modello di concetto di estrazione testo: scheda Impostazioni

4. **Nugget del modello di concetto di estrazione testo (scheda Campi).** Quindi, è stato selezionato il campo di testo, **Q2_Cosa_ti_piace_di_meno_di_questo_stereo_portatile**, che è il nome campo proveniente dal nodo del file delle statistiche. Viene anche selezionata l'opzione **Campo testo indica: testo reale**.

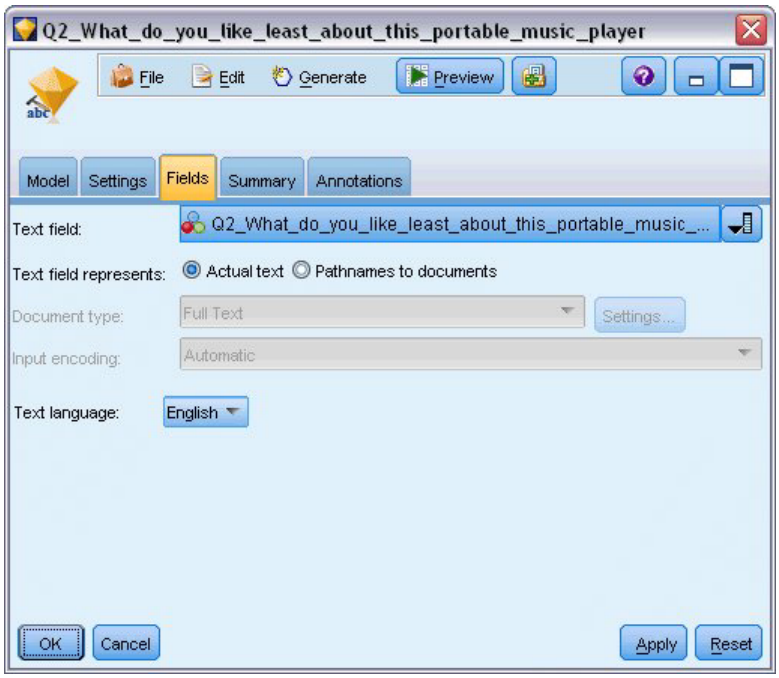
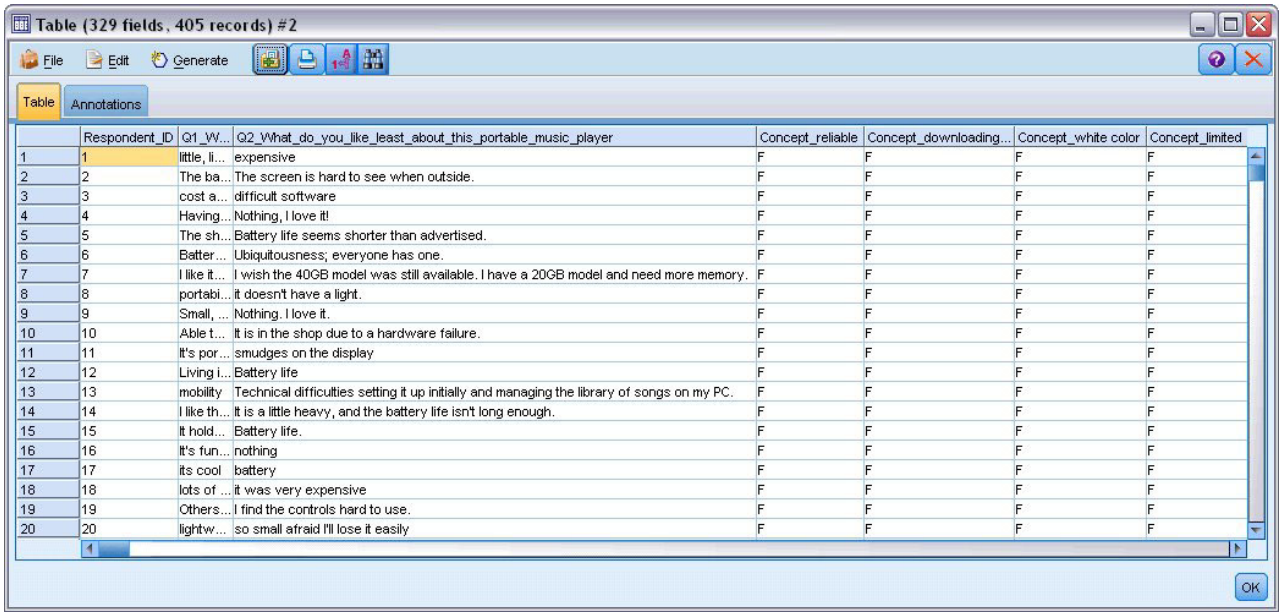


Figura 7. Finestra di dialogo del nugget del modello di concetto estrazione testo: scheda Campi

5. **Nodo tabella.** Inoltre, è stato collegato un nodo di tabella per vedere i risultati ed è stato eseguito il flusso. L'output della tabella viene visualizzata sullo schermo.



	Respondent_ID	Q1_W...	Q2_What_do_you_like_least_about_this_portable_music_player	Concept_reliable	Concept_downloading...	Concept_white color	Concept_limited
1	1	little, li...	expensive	F	F	F	F
2	2	The ba...	The screen is hard to see when outside.	F	F	F	F
3	3	cost a...	difficult software	F	F	F	F
4	4	Having...	Nothing, I love it!	F	F	F	F
5	5	The sh...	Battery life seems shorter than advertised.	F	F	F	F
6	6	Batter...	Ubiquitousness; everyone has one.	F	F	F	F
7	7	I like it...	I wish the 40GB model was still available. I have a 20GB model and need more memory.	F	F	F	F
8	8	portabi...	it doesn't have a light.	F	F	F	F
9	9	Small, ...	Nothing. I love it.	F	F	F	F
10	10	Able t...	It is in the shop due to a hardware failure.	F	F	F	F
11	11	It's por...	smudges on the display	F	F	F	F
12	12	Living i...	Battery life	F	F	F	F
13	13	mobility	Technical difficulties setting it up initially and managing the library of songs on my PC.	F	F	F	F
14	14	I like th...	It is a little heavy, and the battery life isn't long enough.	F	F	F	F
15	15	It hold...	Battery life.	F	F	F	F
16	16	It's fun...	nothing	F	F	F	F
17	17	its cool	battery	F	F	F	F
18	18	lots of ...	it was very expensive	F	F	F	F
19	19	Others...	I find the controls hard to use.	F	F	F	F
20	20	lightw...	so small afraid I'll lose it easily	F	F	F	F

Figura 8. L'output della tabella scorre per mostrare i flag di concetto

Nugget di estrazione testo: Modello di categoria

Il nugget del modello di categoria di estrazione testo viene creato quando si genera un modello di categoria dall'interno del workbench interattivo. Questo nugget di modellazione contiene una serie di categorie, la cui definizione è costituita da concetti, tipi, modelli TLA modelli e/o regole di categoria. Il nugget viene utilizzato per categorizzare le risposte, le voci blog, gli altri flussi Web e tutti gli altri dati di testo.

Se si avvia una sessione workbench interattiva nel nodo di modellazione, è possibile esaminare i risultati di estrazione, perfezionare le risorse, regolare le categorie prima di generare modelli di categoria. Quando si esegue un flusso contenente un nugget del modello di estrazione testo, vengono aggiunti nuovi campi per i dati in base alla modalità di creazione selezionata nella scheda Modello del nodo di modellazione di estrazione testo prima della creazione del modello. Consultare la sezione “Nugget del modello di categoria: scheda Modello” a pagina 42 per ulteriori informazioni.

Se il nugget del modello è stato generato utilizzando documenti tradotti, il calcolo del punteggio verrà eseguito nella lingua tradotta. Allo stesso modo, se il nugget del modello è stato generato utilizzando l'inglese come lingua, è possibile specificare una lingua di traduzione nel nugget del modello, poiché i documenti verranno poi tradotti in inglese.

Quando vengono generati, i nugget del modello di estrazione testo sono posizionati nella tavolozza del nugget del modello (ubicata nella scheda Modelli nella parte superiore destra della finestra IBM SPSS Modeler).

Visualizzazione di risultati

Per visualizzare le informazioni sui nugget del modello, fare clic con il tasto destro del mouse sul nodo nella tavolozza dei nugget del modello e scegliere **Sfoglia** dal menu di scelta rapida (oppure **Modifica** per i nodi in un flusso).

Aggiunta di modelli ai flussi

Per aggiungere il nugget del modello al flusso, fare clic sull'icona nella tavolozza dei nugget del modello e quindi fare clic sull'area del flusso in cui si desidera posizionare il nodo. Oppure fare clic con il tasto destro del mouse sull'icona e selezionare **Aggiungi a flusso** dal menu di scelta rapida. Collegare quindi il flusso al nodo. A questo punto è possibile passare i dati per generare previsioni.

Attenzione: se si desidera utilizzare un nugget di calcolo del punteggio per rigenerare un nodo di modellazione che contiene sia il modello di categoria che il modello utilizzato, si consiglia di creare un TAP e di utilizzarlo in una sessione interattiva in sostituzione del nodo di modellazione prima di generare il nugget di calcolo del punteggio.

Nugget del modello di categoria: scheda Modello

Per i modelli di categoria, la scheda Modello visualizza l'elenco di categorie nel modello di categoria a sinistra e i descrittori per una categoria selezionata sulla destra. Ogni categoria è costituita da un numero di descrittori. Per ciascuna categoria selezionata, i descrittori associati vengono visualizzati nella tabella. Questi descrittori possono comprendere concetti, regole di categoria, tipi e modelli TLA. Viene inoltre mostrato il tipo di ciascun descrittore, così come alcuni esempi di cosa ciascun descrittore rappresenta.

In questa scheda, l'obiettivo è quello di selezionare le categorie che si desidera utilizzare per il calcolo del punteggio. Per un modello di categoria, i documenti e i record vengono conteggiati nelle categorie. Se un documento o record contiene uno o più descrittori nel suo testo o qualsiasi termine sottostante, tale documento o record viene assegnato alla categoria alla quale il descrittore appartiene. Tali termini sottostanti comprendono i sinonimi definiti nelle risorse linguistiche (indipendentemente dal fatto che sono stati rilevati nel testo o meno) così come gli eventuali termini trovati nel testo singolari/plurali, utilizzati per generare nugget di modello, termini permutati, termini da raggruppamenti casuali e così via.

Nota: se è stato invece generato un nugget di modello di concetto, questa tabella conterrà risultati differenti. Consultare la sezione “Modello di concetto: scheda Modello” a pagina 32 per ulteriori informazioni.

Struttura ad albero di categoria

Per ulteriori informazioni relative a ciascuna categoria, selezionare tale categoria e rivedere le informazioni visualizzate per i descrittori in tale categoria. Per ciascun descrittore, è possibile visualizzare le seguenti informazioni:

- Nome **descrittore**. Questo campo contiene un'icona che presenta che tipo di descrittore è, così come il nome del descrittore.

Tabella 5. Icone di descrittore

	Concetti		Modelli TLA
	Tipi		Regole di categoria

- **Tipo.** Questo campo contiene il nome tipo per il descrittore. I tipi sono raccolte di concetti simili (raggruppamenti semantici), ad esempio nomi di organizzazione, prodotti o opinioni positive. Le regole non sono assegnate a tipi.
- **Dettagli.** Questo campo contiene un elenco di cosa è incluso in tale descrittore. In base al numero di corrispondenze, potrebbe non essere possibile visualizzare l'elenco completo per ogni descrittore a causa di limitazioni di dimensioni nella finestra di dialogo.

Selezione e copia di categorie

Tutte le categorie principali vengono selezionate per il calcolo del punteggio per impostazione predefinita, come mostrato nelle caselle di spunta nel riquadro sinistro. Una casella selezionata indica che

la categoria verrà utilizzata per il calcolo del punteggio. Una casella deselezionata indica che la categoria verrà esclusa dal punteggio. È possibile selezionare più righe selezionandole e facendo clic su una delle caselle di spunta nella selezione. Inoltre, se una categoria o una sottocategoria è selezionata ma una delle relative categorie secondarie non è selezionata, la casella di spunta mostra uno sfondo blu per indicare che esiste solo una parziale selezione nei secondari della categoria selezionata.

Premendo il tastino destro del mouse su una categoria nella struttura ad albero, è possibile visualizzare un menu di scelta rapida da cui sono possibili le seguenti azioni:

- **Attiva selezionati.** Attiva tutte le caselle di spunta per le righe selezionate nella tabella.
- **Non attivare selezionati.** Deseleziona tutte le caselle di spunta per le righe selezionate nella tabella.
- **Attiva tutti.** Attiva tutte le caselle di spunta nella tabella. Questo è il risultato in tutte le categorie in uso nell'output finale. È inoltre possibile utilizzare l'icona corrispondente di casella di spunta sulla barra degli strumenti.
- **Disattiva tutti.** Deseleziona tutte le caselle di spunta nella tabella. La deselezione di una categoria significa che non verrà utilizzata nell'output finale. È possibile anche utilizzare la corrispondente icona di casella di spunta vuota sulla barra degli strumenti.

Premendo il tastino destro del mouse su una cella nella tabella di descrittori, è possibile visualizzare un menu di scelta rapida da cui sono possibili le seguenti azioni:

- **Copia.** I concetti selezionati vengono copiati negli Appunti.
- **Copia con campi.** Il descrittore selezionato viene copiato negli appunti insieme alle intestazioni di colonna.
- **Seleziona tutti.** Tutte le righe nella tabella vengono selezionate.

Nugget del modello di categoria: scheda Impostazioni

La scheda Impostazioni viene utilizzata per definire il valore del campo di testo per i dati di input nuovi, se necessario. Questo è anche il luogo in cui è possibile definire il modello di dati per l'output (in modalità di calcolo del punteggio).

Nota: Questa scheda viene visualizzata nella finestra di dialogo Nodo solo quando il nugget del modello è posizionato in un'area o in un flusso. Esso non esiste quando si accede a questo nugget direttamente nella tavolozza Modelli.

Modalità calcolo del punteggio: categorie come campi

Con questa opzione, il numero di record di output è uguale a quello dei record di input. Tuttavia, ogni record contiene ora un nuovo campo per ogni categoria che è stata selezionata (tramite il segno di spunta) nella scheda Modello. Per ogni campo, immettere un valore di flag per **True** e per **False**, per esempio *Sì/No*, *True/False*, *T/F*, o *1* e *2*. I tipi di archiviazione vengono impostati automaticamente in modo da rispecchiare i valori scelti. Ad esempio, se si immettono valori numerici per i flag, questi verranno automaticamente gestiti come un valore intero. I tipi di archiviazione per i flag possono essere: stringa, numero intero, numero reale o data/ora.

Nota: Se si stanno utilizzando dataset di grandi dimensioni, ad esempio, con un database DB2, l'utilizzo di **Categorie come campi** può rilevare problemi di elaborazione a causa della quantità di dati. In questo caso, invece, si consiglia l'utilizzo di **Categorie come record**.

Estensione nome campo. È possibile scegliere di specificare un prefisso/suffisso di estensione per il nome campo o è possibile scegliere di utilizzare i codici della categoria. I nomi di campo vengono generati utilizzando il nome categoria e questa estensione.

- **Aggiungi come.** Specificare dove l'estensione deve essere aggiunta al nome campo. Scegliere **Prefisso** per aggiungere l'estensione all'inizio della stringa. Scegliere **Suffisso** per aggiungere l'estensione alla fine della stringa.

Se non è selezionata una sottocategoria. Questa opzione consente di specificare come verranno gestiti i descrittori appartenenti alle sottocategorie che non sono state selezionate per il calcolo del punteggio. Sono possibili due opzioni.

- L'opzione **Escludi i descrittori completamente dal calcolo del punteggio** farà sì che i descrittori di sottocategorie che non hanno segni di spunta (non selezionati) vengono ignorati e non utilizzati durante il calcolo del punteggio.
- Con l'opzione **Aggrega descrittori con quelli della categoria principale**, i descrittori di sottocategorie che non hanno segni di spunta (non selezionati) vengono utilizzati come descrittori per la categoria principale (la categoria al di sopra di questa sottocategoria). Se diversi livelli di sottocategorie non sono selezionati, i descrittori verranno riepilogati sotto la prima categoria principale disponibile.

Correggi errori di punteggiatura. Questa opzione normalizza temporaneamente testo contenente errori di punteggiatura (ad esempio, l'uso improprio) durante l'estrazione per migliorare la possibilità di estrazione dei concetti. Questa opzione risulta estremamente utile quando il testo è breve e di scarsa qualità (come, ad esempio, le risposte del sondaggio aperto, e-mail e i dati CRM), o quando il testo contiene molte abbreviazioni.

Nota: L'opzione **Correggi errori di punteggiatura** non si applica quando si gestisce testo giapponese.

Modalità calcolo del punteggio: categorie come record

Con questa opzione, viene creato un nuovo record per ogni coppia `categoria`, `documento`. Di solito, ci sono più record nell'output di quanti ce ne sono stati nell'input. Oltre ai campi di input, i nuovi campi vengono inoltre aggiunti ai dati a seconda del tipo di modello che è.

Tabella 6. Campi di output per "categorie come record".

Ricava di output	Descrizione
`Categoria`	Contiene il nome della categoria alla quale il documento di testo è stato assegnato. Se la categoria è sottocategoria di un'altra, il percorso completo per il nome della categoria è controllato dal valore scelto in questa finestra di dialogo.

Valori di categorie gerarchiche. Questa opzione controlla come i nomi delle sottocategorie vengono visualizzati nell'output.

- **Percorso completo di categoria.** Questa opzione emette il nome della categoria e il percorso completo di categorie principali, se applicabile, utilizzando barre per separare i nomi categoria dai nomi sottocategoria.
- **Percorso abbreviato di categoria.** Questa opzione emette solo il nome della categoria, ma utilizza le ellissi per visualizzare il numero di categorie principali per la categoria in questione.
- **Categoria di livello inferiore.** Questa opzione emette solo il nome della categoria senza il percorso completo o senza mostrare le categorie principali.

Se non è selezionata una sottocategoria. Questa opzione consente di specificare come verranno gestiti i descrittori appartenenti alle sottocategorie che non sono state selezionate per il calcolo del punteggio. Sono possibili due opzioni.

- L'opzione **Escludi i descrittori completamente dal calcolo del punteggio** farà sì che i descrittori di sottocategorie che non hanno segni di spunta (non selezionati) vengono ignorati e non utilizzati durante il calcolo del punteggio.
- Con l'opzione **Aggrega descrittori con quelli della categoria principale**, i descrittori di sottocategorie che non hanno segni di spunta (non selezionati) vengono utilizzati come descrittori per la categoria principale (la categoria al di sopra di questa sottocategoria). Se diversi livelli di sottocategorie non sono selezionati, i descrittori verranno riepilogati sotto la prima categoria principale disponibile.

Correggi errori di punteggiatura. Questa opzione normalizza temporaneamente testo contenente errori di punteggiatura (ad esempio, l'uso improprio) durante l'estrazione per migliorare la possibilità di estrazione dei concetti. Questa opzione risulta estremamente utile quando il testo è breve e di scarsa qualità (come, ad esempio, le risposte del sondaggio aperto, e-mail e i dati CRM), o quando il testo contiene molte abbreviazioni.

Nota: L'opzione **Correggi errori di punteggiatura** non si applica quando si gestisce testo giapponese.

Nugget del modello di categoria: altre schede

La scheda Campi e scheda Impostazioni per il nugget del modello di categoria sono uguali a quelli per il nugget del modello di concetto.

- Scheda Campi. Per ulteriori informazioni, consultare la sezione "Modello di concetto: scheda Campi" a pagina 36.
- Scheda Riepilogo. Per ulteriori informazioni, consultare la sezione "Modello di concetto: scheda Riepilogo" a pagina 37.

Uso dei nugget del modello di categoria in un flusso

Il nugget del modello di categoria di estrazione testo è generato da una sessione workbench interattiva. È possibile utilizzare questo nugget del modello in un flusso.

Esempio: nodo file delle statistiche con il nugget del modello di categoria

Il seguente esempio mostra come utilizzare il nugget del modello di estrazione testo.

Figura 9. Esempio di flusso: nodo file delle statistiche con il nugget del modello di categoria di estrazione testo

1. **Nodo file delle statistiche (scheda Dati).** In primo luogo, è stato aggiunto questo nodo al flusso per specificare dove i documenti di testo sono memorizzati.

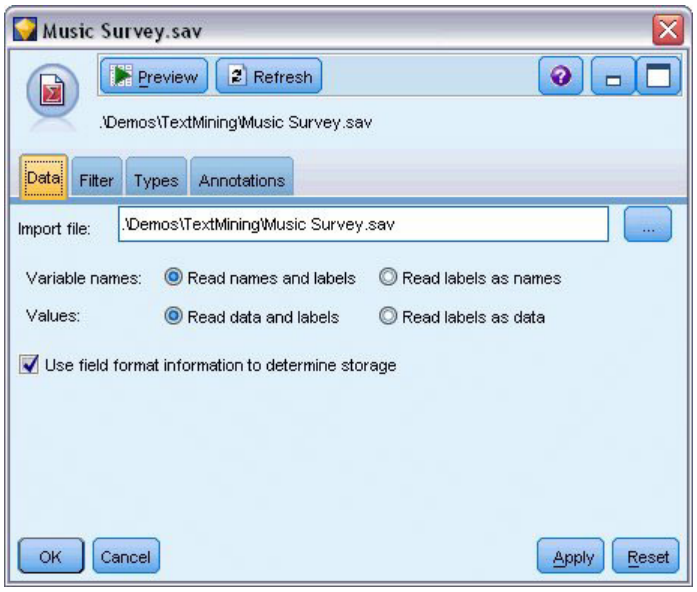


Figura 10. Finestra di dialogo del nodo file delle statistiche: scheda Dati

2. **Nugget del modello di categoria di estrazione testo (scheda Modello).** Inoltre, è stato aggiunto e connesso un nugget del modello di categoria al nodo file delle statistiche. Sono state selezionate delle categorie da utilizzare per il calcolo del punteggio dei dati.

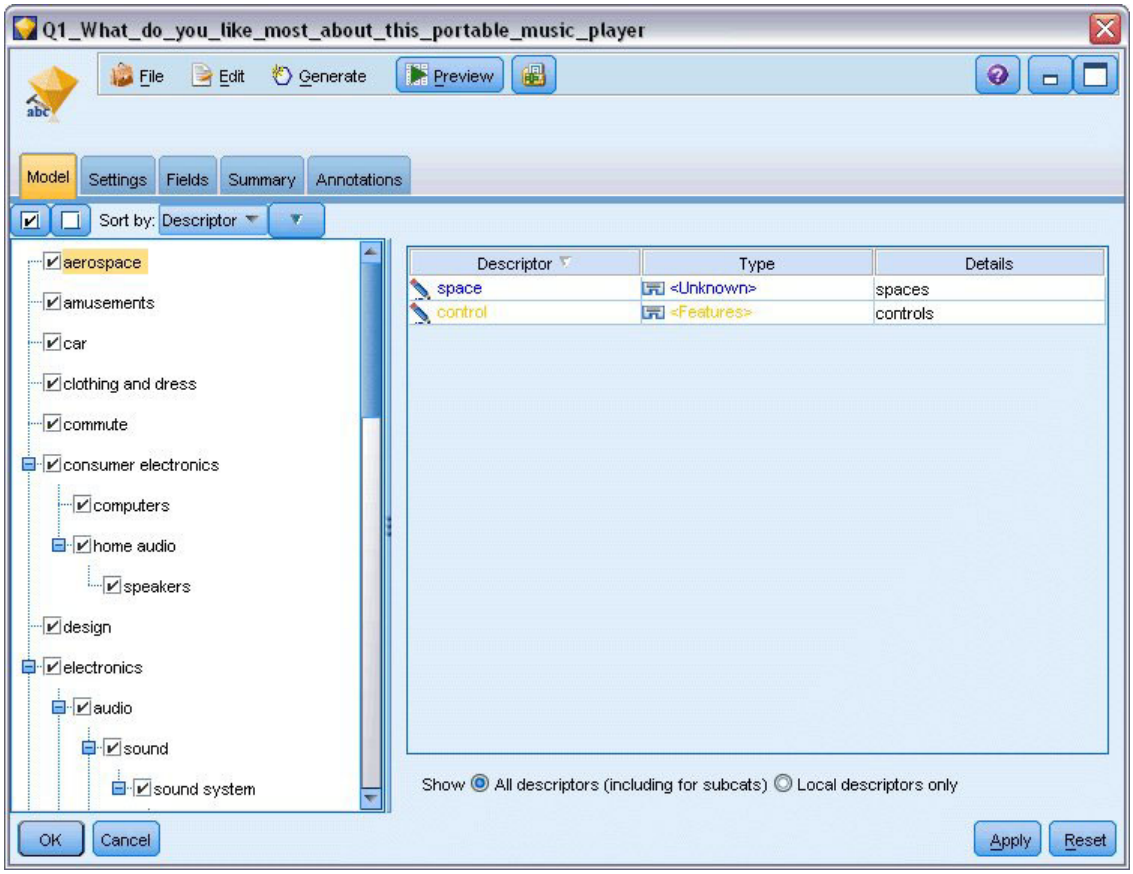


Figura 11. finestra di dialogo del nugget del modello di estrazione testo: scheda Modello

3. **Nugget del modello di estrazione testo (scheda Impostazioni).** Successivamente, è stato definito il formato di output **Categorie come campi**.

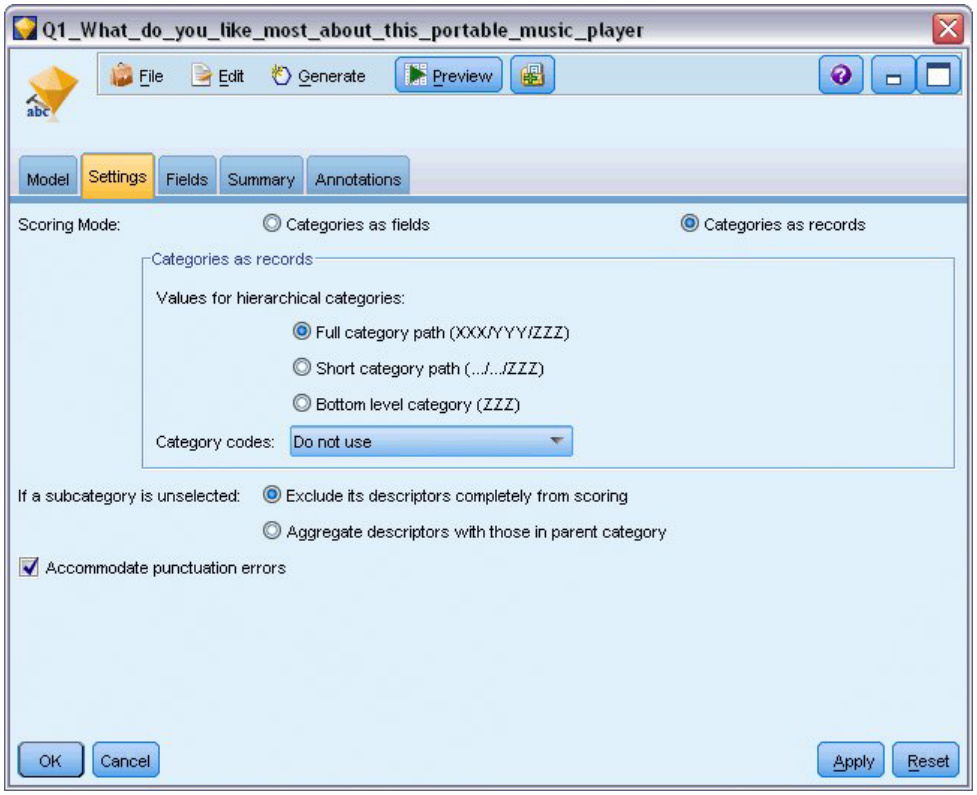


Figura 12. Finestra di dialogo del nugget del modello di categoria: scheda Impostazioni

4. **Nugget del modello di categoria di estrazione testo (scheda Campi).** Successivamente è stata selezionata la variabile del campo di testo, che è il nome campo che scaturisce del nodo file delle statistiche e il campo Testo dell'opzione che rappresenta il **testo reale**, oltre ad altre impostazioni.

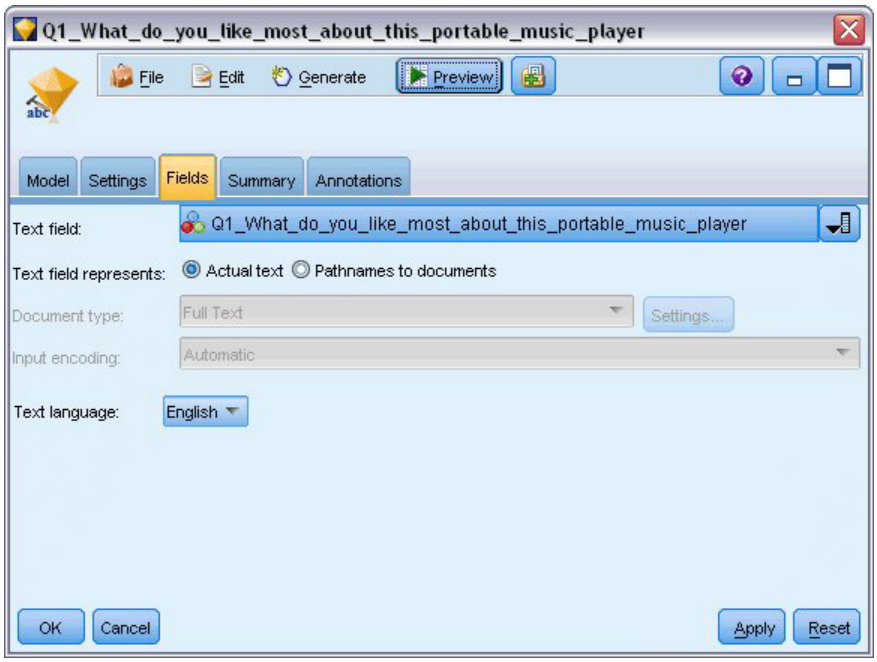
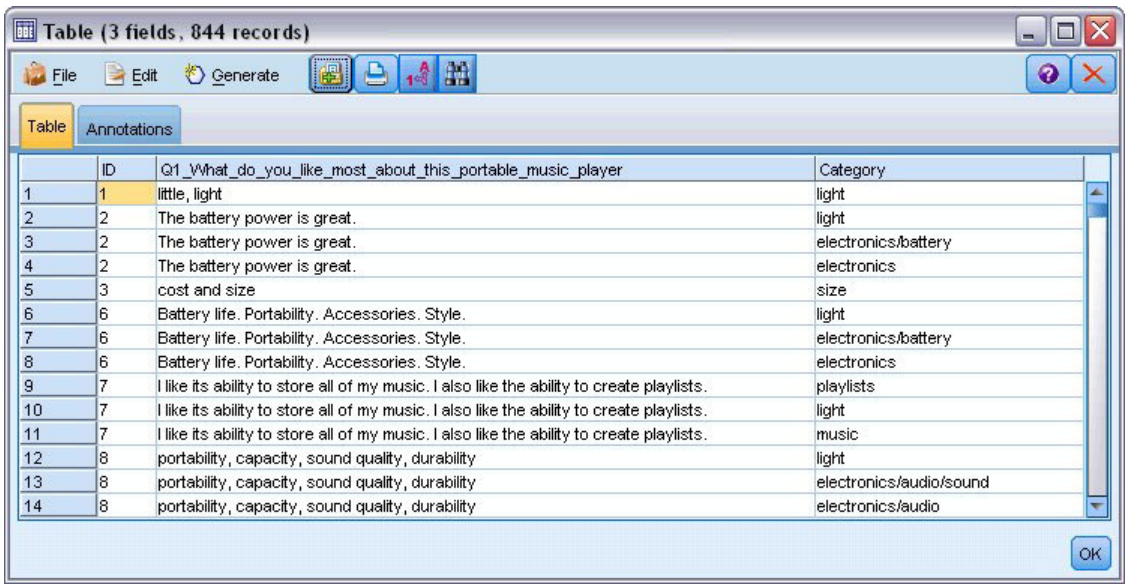


Figura 13. Finestra di dialogo del nugget del modello di estrazione testo: scheda Campi

5. **Nodo tabella.** Inoltre, è stato collegato un nodo di tabella per vedere i risultati ed è stato eseguito il flusso.



	ID	Q1_What_do_you_like_most_about_this_portable_music_player	Category
1	1	little, light	light
2	2	The battery power is great.	light
3	2	The battery power is great.	electronics/battery
4	2	The battery power is great.	electronics
5	3	cost and size	size
6	6	Battery life. Portability. Accessories. Style.	light
7	6	Battery life. Portability. Accessories. Style.	electronics/battery
8	6	Battery life. Portability. Accessories. Style.	electronics
9	7	I like its ability to store all of my music. I also like the ability to create playlists.	playlists
10	7	I like its ability to store all of my music. I also like the ability to create playlists.	light
11	7	I like its ability to store all of my music. I also like the ability to create playlists.	music
12	8	portability, capacity, sound quality, durability	light
13	8	portability, capacity, sound quality, durability	electronics/audio/sound
14	8	portability, capacity, sound quality, durability	electronics/audio

Figura 14. Output di tabella

Capitolo 4. Estrazione per link di testo

Nodo Analisi di collegamento del testo

Il nodo di analisi collegamenti del testo (TLA) aggiunge una tecnologia di corrispondenza modello per l'estrazione di concetti di estrazione testo per identificare le relazioni tra i concetti nei dati di testo in base a modelli conosciuti. Queste relazioni possono descrivere in che modo un cliente si sente su un prodotto, quali aziende stanno lavorando insieme o anche le relazioni tra geni o agenti farmaceutici.

Ad esempio l'estrazione del nome prodotto del proprio concorrente potrebbe non risultare interessante. Utilizzando questo nodo è possibile inoltre illustrare cosa i cittadini sentono su questo prodotto, se tali opinioni esistono nei dati. Le relazioni e le associazioni vengono identificate ed estratte da modelli corrispondenti noti ai propri dati di testo.

È possibile utilizzare le regole del modello TLA all'interno di alcuni modelli di risorsa forniti con IBM SPSS Modeler Text Analytics o creare/modificare le proprie. Le regole del modello sono costituite da macro, elenchi di parole e differenze di parole per formare una interrogazione booleana o una regola, che vengono confrontate con il testo di input. Ogni volta che una regola di modello TLA corrisponde al testo, questo testo può essere esatto come risultato TLA e ristrutturato come dati di output. Consultare la sezione Capitolo 19, “Informazioni sulle regole di collegamento del testo”, a pagina 215 per ulteriori informazioni.

Il nodo analisi collegamenti del testo offre un modo più diretto per identificare ed estrarre i risultati di modelli TLA dal testo e quindi aggiungere i risultati al dataset nel flusso. Ma il nodo di analisi collegamenti del testo non è l'unico modo in cui è possibile eseguire l'analisi di collegamento del testo. Ma è anche possibile eseguire TLA utilizzando una sessione workbench interattiva nel nodo di modellazione di estrazione testo.

Nel workbench interattivo, è possibile esplorare i risultati del modello TLA e utilizzarli come descrittori di categoria e/o per ulteriori informazioni sui risultati mediante analizzatori e grafici. Consultare la sezione Capitolo 12, “Esplorazione di analisi di collegamento del testo”, a pagina 153 per ulteriori informazioni. In realtà, l'uso del nodo di estrazione testo per estrarre i risultati TLA è un ottimo modo per esplorare e perfezionare i modelli ai dati per un utilizzo successivo direttamente nel nodo TLA.

L'output può essere rappresentato fino a 6 slot, o parti. I modelli giapponesi sono solo output come uno o due slot. Consultare la sezione “Output del nodo TLA” a pagina 53 per ulteriori informazioni.

È possibile rilevare questo nodo sulla scheda IBM SPSS Modeler Text Analytics della tavolozza dei nodi nella parte inferiore della finestra IBM SPSS Modeler. Consultare la sezione “IBM SPSS Modeler Text Analytics Nodi” a pagina 8 per ulteriori informazioni.

Requisiti. Il nodo Analisi collegamenti del testo accetta i dati di testo letto in un campo utilizzando uno dei nodi di origine standard (nodo Database, nodo File semplice, ecc.) oppure letti in un elenco di percorsi a documenti esterni in un campo generato da un nodo di elenco file o un nodo di flusso web.

Efficacia. Il nodo di analisi collegamenti del testo va oltre il concetto di base di estrazione per fornire informazioni sulle relazioni *tra* concetti, così come opinioni o qualificativi correlati che possono essere rivelati nei dati.

Nodo di analisi di collegamenti del testo: scheda Campi

La scheda Campi viene utilizzata per specificare le impostazioni dei campi per i dati da cui si estraggono concetti. È possibile impostare i seguenti parametri:

ID campo. Selezionare il campo contenente l'identificativo per i record di testo. Gli identificativi devono essere numeri interi. Il campo ID funge da indice per i singoli record di testo. Utilizzare un ID campo se il campo di testo rappresenta il testo che deve essere estratto. Non utilizzare un campo ID se il campo di testo rappresenta **Nomi percorso di documenti**.

Campo Testo. Selezionare il campo che contiene il testo da estrarre, il nome percorso del documento o il nome percorso della directory per i documenti. Questo campo dipende dall'origine dati.

Il campo Testo rappresenta. Indica cosa contiene il campo di testo specificato nella precedente impostazione. Le scelte sono:

- **Testo reale.** Selezionare questa opzione se il campo contiene il testo esatto da cui devono essere estratti i concetti.
- **Nomi percorso dei documenti.** Selezionare questa opzione se il campo contiene uno o più nomi percorso per l'ubicazione dei documenti di testo.

Tipo documento. Questa opzione è disponibile solo se è stato indicato che il campo di testo rappresenta **nomi percorso di documenti**. Il tipo di documento specifica la struttura del testo. Selezionare uno dei seguenti tipi:

- **Testo.** Da utilizzare per la maggior parte dei documenti o origini di testo. Tutto l'insieme del testo viene analizzato per l'estrazione. A differenza delle altre opzioni, non vi sono ulteriori impostazioni per questa opzione.
- **Testo strutturato.** Utilizzato per moduli di bibliografia, brevetti e qualsiasi file contenente strutture regolari che possono essere identificate e analizzate. Questo tipo di documento viene utilizzato per ignorare tutto o parte del processo di estrazione. Esso consente di definire i separatori, assegnare i tipi ed imporre un valore di frequenza minima. Se viene selezionata questa opzione, fare clic sul pulsante **Impostazioni** ed immettere i separatori di testo nell'area **Formattazione testo strutturato** della finestra di dialogo Impostazioni documento. Consultare la sezione “Impostazioni di documento per la scheda Campi” a pagina 22 per ulteriori informazioni.
- **Testo XML.** Utilizzare per specificare i tag XML che contengono il testo da estrarre. Tutti gli altri tag vengono ignorati. Se si seleziona questa opzione, fare clic sul pulsante **Impostazioni** e specificare esplicitamente gli elementi XML che contengono il testo da leggere durante il processo di estrazione nell'area **Formattazione testo XML** della finestra di dialogo Impostazioni documento. Consultare la sezione “Impostazioni di documento per la scheda Campi” a pagina 22 per ulteriori informazioni.

Unità di testo. Questa opzione è disponibile solo se è stato specificato che il campo di testo rappresenta **Nomi percorso di documenti** e selezionato **Testo completo** come tipo di documento. Selezionare la modalità di estrazione tra i seguenti:

- **Modo Documento.** Utilizzare per i documenti brevi e semanticamente omogenei, come articoli da agenzie stampa.
- **Modo Paragrafo.** Utilizzare per le pagine Web e documenti non taggati. Il processo di estrazione divide semanticamente i documenti, approfittando delle caratteristiche quali tag interne e la sintassi. Se viene selezionata questa modalità, il calcolo del punteggio viene applicato paragrafo per paragrafo. Pertanto, ad esempio, la regola `mela & arancia` è true solo se `mela` e `arancia` si trovano nello stesso paragrafo.

  Nota: In base al criterio con cui viene estratto il testo dai documenti PDF, la **Modalità paragrafo** non funziona per questi documenti. Ciò accade perché l'estrazione elimina il marker di ritorno a capo.

Impostazioni di modo paragrafo. Questa opzione è disponibile solo se è stato specificato che il campo di testo rappresenta **nomi percorso di documenti** e se è stata impostata l'opzione unità di testo per **modo Paragrafo**. Specificare le soglie dei caratteri da utilizzare in qualsiasi estrazione. La dimensione effettiva è arrotondata per eccesso o per difetto al periodo più vicino. Per verificare che le associazioni di parole prodotte dal testo della raccolta di documenti sono rappresentative, evitare di specificare una dimensione di estrazione troppo piccola.

- **Minimo.** Specifica il numero minimo di caratteri da utilizzare in ogni estrazione.
- **Massimo.** Specifica il numero massimo di caratteri da utilizzare in ogni estrazione.

Codifica di input. Questa opzione è disponibile solo se è stato indicato che il campo di testo rappresenta **nomi percorso di documenti**. Specifica la codifica di testo predefinita. Per tutte le lingue, tranne giapponese, viene eseguita una conversione dalla codifica specificata o riconosciuta in `ISO-8859-1`. Quindi anche se si desidera specificare un'altra codifica, il motore di estrazione la converte in `ISO-8859-1` prima che venga elaborato. Qualsiasi carattere che non rientra nella definizione di codifica `ISO-8859-1` verrà convertito in spazio. Per il testo giapponese, è possibile scegliere una tra le diverse opzioni di codifica: `SHIFT_JIS`, `EUC_JP`, `UTF-8` o `ISO-2022-JP`.

Copia risorse da. Quando si estrae testo, l'estrazione si basa non solo sulle impostazioni nella scheda Avanzate ma anche sulle risorse linguistiche. Tali risorse servono come base per come gestire ed elaborare il testo durante l'estrazione in modo da ottenere i concetti, i tipi e a volte i modelli TLA. È possibile copiare le risorse in questo nodo da un modello di risorsa.

Un modello di risorsa è un insieme predefinito di librerie e risorse linguistiche e non linguistiche avanzate che sono state definite per un particolare dominio o utilizzo. Tali risorse servono come base per le modalità di gestione e di elaborazione dei dati durante l'estrazione. Fare clic su **Carica** e selezionare il modello da cui copiare le risorse.

I modelli vengono caricati quando si seleziona e non quando viene eseguito il flusso. Al momento in cui si desidera caricare, una copia delle risorse selezionate viene memorizzata nel nodo. Pertanto, se mai si desiderava utilizzare un modello aggiornato, è necessario ricaricarlo qui. Consultare la sezione “Copia risorse da modelli e TAP” a pagina 27 per ulteriori informazioni.

Lingua di testo. Identifica la lingua del testo estratto. Le risorse copiate nel nodo controllano le opzioni di lingua riportate. È possibile selezionare la lingua per la quale le risorse sono state regolate o scegliere l'opzione **TUTTI**. Si consiglia vivamente di specificare la lingua esatta per i dati di testo; tuttavia, se non si è sicuri, è possibile scegliere l'opzione **TUTTI**. **TUTTI** non è disponibile per il testo giapponese. Questa opzione **TUTTI** incrementa il tempo di esecuzione poiché il riconoscimento automatico della lingua viene utilizzato per eseguire la scansione di tutti i documenti e record in modo da identificare innanzitutto la lingua di testo. Con questa opzione, tutti i record o i documenti in una lingua supportata e con licenza vengono letti dal motore di estrazione utilizzando i dizionari interni appropriati per la lingua. Per ulteriori informazioni, consultare la sezione “Identificativo di lingua” a pagina 213. Rivolgersi al rappresentante delle vendite se si è interessati all'acquisto di una licenza per una lingua supportata alla quale attualmente non si ha accesso.

Nodo di analisi di collegamenti del testo : scheda Modello

La scheda Modello contiene una singola opzione che riguarda la velocità e la precisione del processo di estrazione.

Ottimizza per velocità di calcolo del punteggio. Questa opzione è selezionata automaticamente e assicura che il modello creato sia compatto ed esegua punteggi ad alta velocità. Se si deseleziona questa opzione viene creato un modello che garantisce i punteggi più lentamente ma assicura la congruenza completa del tipo di concetto e cioè garantisce che un determinato concetto non venga mai assegnato a più di un tipo.

Nodo di analisi di collegamenti del testo: scheda Avanzate

In questo nodo, l'estrazione dei risultati del modello di analisi collegamento del testo (TLA) viene abilitata automaticamente. La scheda Avanzate contiene determinati parametri aggiuntivi che influenzano il modo in cui il testo viene estratto e gestito. I parametri contenuti in questa finestra di dialogo controllano il comportamento di base, nonché una qualche funzionalità avanzata, del processo di estrazione. Vi sono inoltre alcune risorse e opzioni linguistiche che influenzano anche i risultati di estrazione, che sono controllate dal modello di risorsa selezionato.

Per testo olandese, inglese, francese, tedesco, italiano, portoghese e spagnolo

Correggi errori di punteggiatura. Questa opzione normalizza temporaneamente testo contenente errori di punteggiatura (ad esempio, l'uso improprio) durante l'estrazione per migliorare la possibilità di estrazione dei concetti. Questa opzione risulta estremamente utile quando il testo è breve e di scarsa qualità (come, ad esempio, le risposte del sondaggio aperto, e-mail e i dati CRM), o quando il testo contiene molte abbreviazioni.

Correzione di errori ortografici per un limite minimo di caratteri root [n]. Questa opzione applica una tecnica di raggruppamento che consente di raggruppare parole errate o parole simili in un unico concetto. L'algoritmo di raggruppamento toglie temporaneamente tutte le vocali (tranne la prima) e stacca consonanti doppie/triple da parole estratte e poi le confronta per vedere se sono uguali in modo che `modeli` e `modelli` verrebbero raggruppate insieme. Tuttavia, se ogni termine viene assegnato ad un tipo differente, escluso il tipo `<Sconosciuto>`, la tecnica di raggruppamento confuso non verrà applicata.

È possibile inoltre definire il numero minimo di caratteri *radice* richiesti prima di utilizzare il raggruppamento. Il numero di caratteri radice in un termine è calcolato sommando tutti i caratteri e sottraendo quelli che formano i suffissi di desinenza e, nel caso di parole composte, i determinativi e le preposizioni. Ad esempio, il termine `esercizi` potrebbe essere conteggiato come 9 caratteri radice nel formato "esercizio", poiché la lettera *i* alla fine della parola è un flesso (forma plurale). Allo stesso modo, `succo di mela` conta 11 caratteri radice (“succo di mela”) e `fabbrica di automobili` conta 20 caratteri radice. Questo metodo di conteggio viene utilizzato solo per verificare se il raggruppamento deve essere applicato ma non influenza il modo in cui le parole sono corrispondenti.

Nota: se si riscontra che determinate parole vengono successivamente raggruppate in modo non corretto, è possibile escludere da questa tecnica le coppie di parole esplicitamente dichiarate nella sezione **Raggruppamento confuso: eccezioni** nella scheda Avanzate. Per ulteriori informazioni, consultare la sezione “Raggruppamento confuso” a pagina 206.

Estrazione termini univoci. Questa opzione estrae parole singole (termini univoci) purché il termine non è già parte di una parola composta e se è un nome o una parte del discorso non riconosciuta.

Estrazione entità non linguistiche. Questa opzione estrae entità non linguistiche, ad esempio numeri di telefono, numeri di codice fiscale, orari, date, valute, cifre, percentuali, indirizzi e-mail e indirizzi HTTP. È possibile includere o escludere alcuni tipi di entità non linguistiche nella sezione **Entità non linguistiche: configurazione** della scheda Avanzate. Disabilitando qualsiasi entità non necessaria, il motore di estrazione non spreca tempo di elaborazione. Per ulteriori informazioni, consultare la sezione “Configurazione” a pagina 210.

Algoritmo maiuscolo. Questa opzione estrae i termini semplici e composti che non si trovano nei dizionari incorporati, purché la prima lettera del termine sia in maiuscolo. Questa opzione fornisce un modo efficace di estrarre i sostantivi più appropriati.

Raggruppare insieme nomi di persona parziali e completi quando possibile. Questa opzione raggruppa i nomi che vengono visualizzati insieme diversamente nel testo. Questa funzione è utile poiché i nomi vengono spesso definiti nel loro formato completo all'inizio del testo e poi solo da una versione abbreviata. Questa opzione tenta la corrispondenza con qualsiasi termine univoco con il tipo `<Sconosciuto>` per l'ultima parola di tutti i termini composti immessi come `<Persona>`. Ad esempio, se viene rilevato *rossi* e inizialmente immesso come `<Sconosciuto>`, il motore controlla l'estrazione per vedere se tutti i termini composti nel tipo `<Persona>` includono *rossi* come ultima parola, per esempio *giovanni rossi*. Questa opzione non si applica ai nomi di persona poiché la maggior parte non sono mai estratti come termini univoci.

Numero massimo di permutazioni di parole non funzionali. Questa opzione specifica il numero massimo di parole non funzionali che possono essere presenti quando si applica la tecnica di permutazione. Questa tecnica di permutazione raggruppa frasi simili che differiscono tra loro solo per

parole non funzionali (ad esempio, di e il) contenute, indipendentemente dall'inflessione. Ad esempio, si imposta questo valore su massimo due parole e vengono estratte `funzionari` e `funzionari dell'azienda`. In questo caso, entrambi i termini estratti verrebbero raggruppati insieme nell'elenco dei concetti poiché entrambi i termini vengono considerati uguali quando `dell'` viene ignorato.

Per il testo giapponese

Con il testo giapponese è possibile scegliere l'analizzatore secondario da applicare.

Analisi secondaria. Quando viene lanciata un'estrazione, ha luogo l'estrazione di parole chiave di base utilizzando la serie predefinita di tipi. Tuttavia, quando si seleziona un analizzatore secondario, è possibile ottenere molti più concetti e più arricchiti poiché l'estrazione comprenderà particelle e verbi ausiliari come parte del concetto. Nel caso di analisi di opinione, viene incluso anche un numero elevato di tipi aggiuntivi. Inoltre, la scelta di un'analisi secondaria consente di generare anche dei risultati di analisi di collegamento del testo.

Nota: Quando un analizzatore secondario viene richiamato, il processo di estrazione richiede più tempo per completare il processo.

- **Analisi di dipendenza**. La scelta di questa opzione produce particelle estese per i concetti di estrazione dal tipo di base ed estrazione di parola chiave. È anche possibile ottenere risultati di modello più arricchiti dall'analisi TLA di dipendenza.
- **Analisi di opinione**. La scelta di questo tipo di analisi genera ulteriori concetti estratti e, dove applicabile, l'estrazione dei risultati del modello TLA. Oltre ai tipi di base, è possibile anche usufruire di più di 80 tipi di opinioni:. Tali tipi vengono utilizzati per rilevare i concetti e i modelli nel testo mediante l'espressione delle emozioni, opinioni e opinioni. Sono disponibili tre opzioni che regolano il focus per l'analisi di opinione: **Tutti i opinioni**, **Solo opinione rappresentativo** e **Solo conclusioni**.

Output del nodo TLA

Dopo l'esecuzione del nodo di analisi collegamenti del testo, i dati vengono ristrutturati. È importante comprendere il modo in cui l'estrazione testo ristruttura i dati. Se si desidera una struttura differente per il data mining, è possibile utilizzare i nodi sulla tavolozza delle operazioni di campo. Ad esempio, se si stava lavorando con dati in cui ogni riga rappresenta un record di testo, viene creata una riga per ciascun modello scoperto nei dati di testo di origine. Per ogni riga nell'output, vi sono 15 campi:

- Sei campi (**Concetto#**, come **Concetto1**, **Concetto2**, ..., e **Concetto6**) rappresentano qualsiasi concetto trovato nella corrispondenza del modello.
- Sei campi (**Tipo#**, come **Tipo1**, **Tipo2**, ..., e **Tipo6**) rappresentano il tipo per ogni concetto.
- **Nome regola** rappresenta il nome della regola di collegamento di testo utilizzato per corrispondere al testo e produrre l'output.
- Un campo che utilizza il nome dell'ID campo specificato nel nodo e che rappresenta l'ID record o documento come era nei dati di input
- **Testo corrispondente** rappresenta la porzione di dati di testo nel record o documento originale che è stata associata al modello TLA.

Nota: le regole del modello di analisi di collegamento del testo per testo giapponese producono solo uno o due risultati di modello slot.

Nota: qualsiasi flusso preesistente che contiene un nodo di analisi collegamenti del testo da un rilascio precedente alla Versione 5,0 potrebbe non essere completamente eseguibile fino a che non si aggiornano i nodi. Alcuni miglioramenti nelle versioni successive di IBM SPSS Modeler richiedono che i nodi più vecchi siano sostituiti con versioni più recenti, che garantiscono una distribuzione più facile oltre a essere più potenti.

È anche possibile eseguire una traduzione automatica di alcune lingue. Questa funzione consente di estrarre documenti in una lingua che si potrebbe non parlare o leggere. Se si desidera utilizzare la funzione di conversione, è necessario disporre dell'accesso a SaaS (SDL Software as a Service). Consultare la sezione “Impostazioni di traduzione” a pagina 58 per ulteriori informazioni.

Cache dei risultati TLA

Se viene eseguita la memorizzazione in cache, i risultati di analisi di collegamento del testo si trovano nel flusso. Per evitare di ripetere l'estrazione di analisi di collegamento del testo ogni volta che viene eseguito il flusso, selezionare il nodo Analisi di collegamento del testo e dai menu scegliere, **Modifica > Nodo > Cache > Abilita**. La volta successiva in cui viene eseguito il flusso, l'output viene memorizzato nel nodo. L'icona del nodo visualizza un piccolo grafico "documento" che cambia da bianco a verde quando la memoria cache è piena. La cache viene mantenuta per tutta la durata della sessione. Per preservare la cache per un altro giorno (dopo che il flusso viene chiuso e riaperto), selezionare il nodo e dai menu scegliere, **Modifica > Nodo > Cache > Salva cache**. La volta successiva che si apre il flusso, è possibile ricaricare la cache salvata piuttosto che eseguire di nuovo la conversione.

In alternativa, è possibile salvare o abilitare una cache del nodo facendo clic con il tastino destro del mouse sul nodo e scegliendo **Cache** dal menu di contesto.

Uso del nodo di analisi di collegamento del testo in un flusso

Il nodo di analisi di collegamento del testo viene utilizzato per accedere ai dati ed estrarre concetti in un flusso. È possibile utilizzare qualsiasi nodo di origine per accedere ai dati.

Esempio: nodo file delle statistiche con il nodo di analisi collegamento del testo

Il seguente esempio mostra come utilizzare il nodo di analisi collegamento del testo.

Figura 15. Esempio: nodo file delle statistiche con il nodo di analisi collegamento del testo

1. **Nodo file delle statistiche (scheda Dati).** In primo luogo, è stato aggiunto questo nodo al flusso per specificare dove il testo è archiviato.

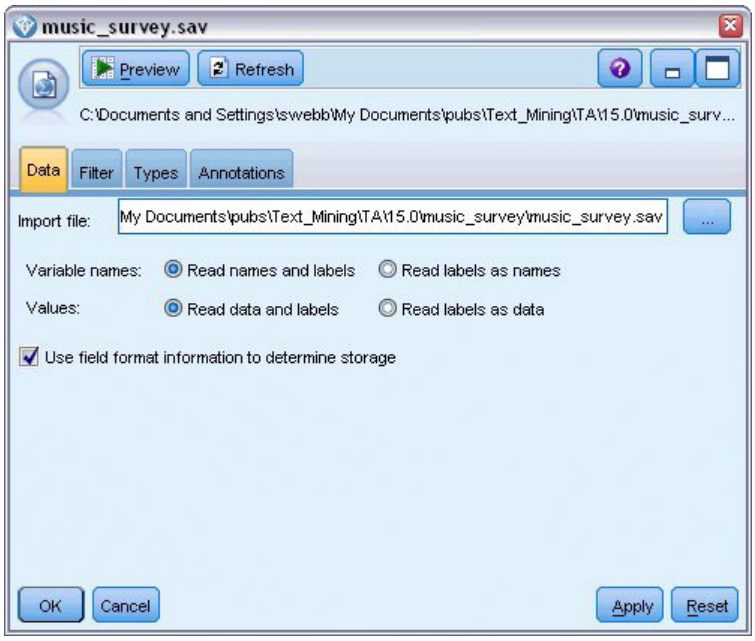


Figura 16. Finestra di dialogo del nodo file delle statistiche: scheda Dati

2. **Nodo di analisi collegamento del testo (Scheda Campi).** Questo nodo è stato collegato al flusso per estrarre concetti per la modellazione o la visualizzazione verso il basso. Occorre specificare l'ID campo e il nome del campo di testo contenente i dati, oltre ad altre impostazioni.

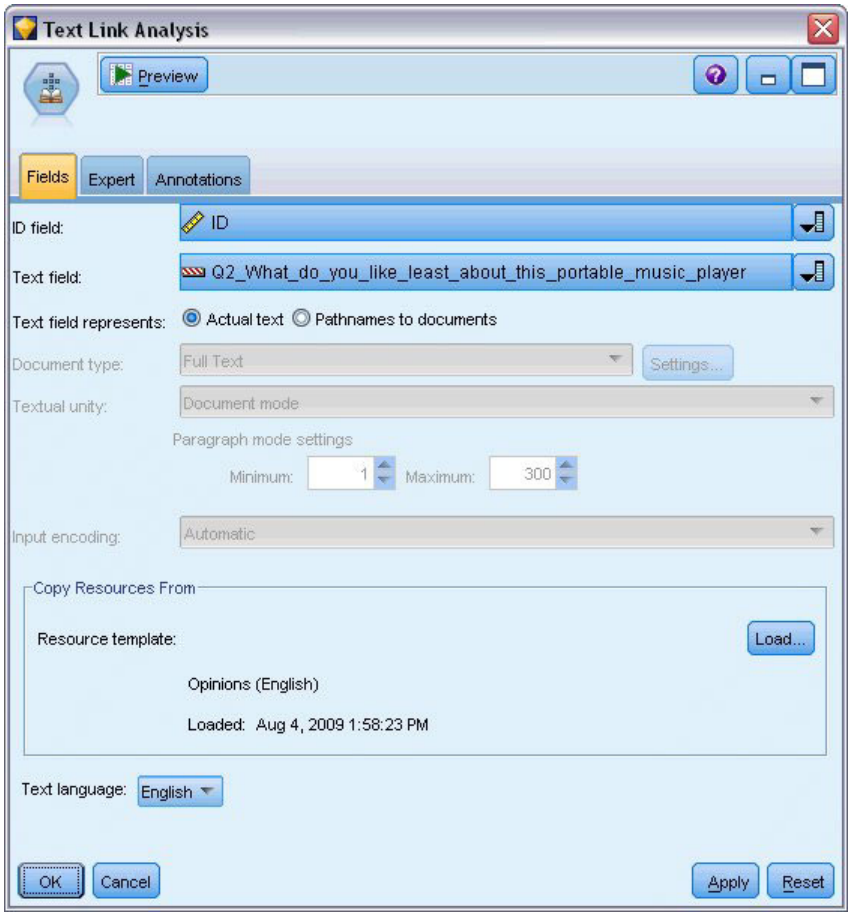
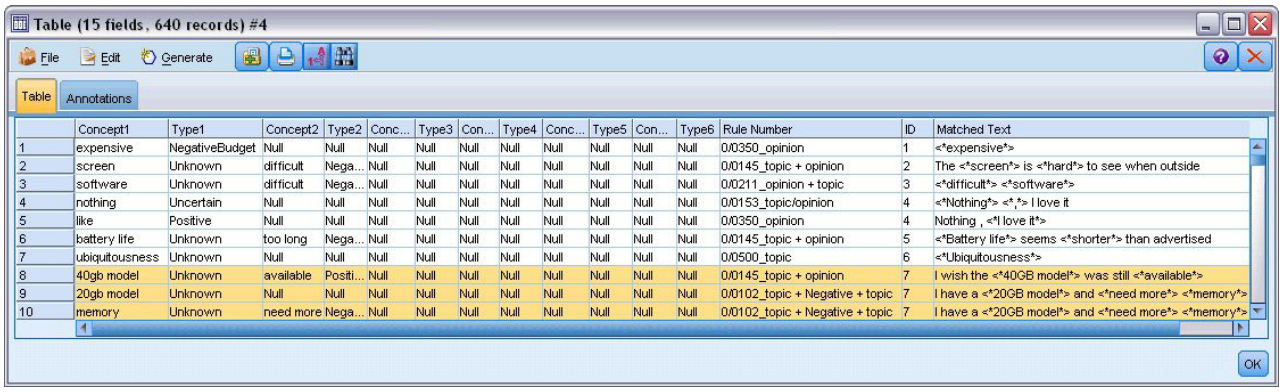


Figura 17. finestra di dialogo del nodo di analisi collegamento del testo: scheda Campi

3. **Nodo tabella.** Inoltre, è stato collegato un nodo di tabella per vedere i concetti estratti dai propri documenti di testo. Nell'output di tabella mostrato, è possibile visualizzare i risultati del modello TLA trovato nei dati dopo che questo flusso è stato eseguito con un nodo di analisi collegamento di testo. Alcuni risultati mostrano solo un concetto/tipo in corrispondenza. In altri, i risultati sono più complessi e contengono diversi tipi e concetti. Inoltre, come risultato dell'esecuzione dei dati attraverso il nodo di analisi collegamento del testo e l'estrazione di concetti, diversi aspetti dei dati vengono modificati. I dati originali nell'esempio contenevano 8 campi e 405 record. Dopo l'esecuzione del nodo di analisi collegamento del testo, ci sono ora 15 campi e 640 record. Vi è ora una riga per ogni risultato di modello TLA trovato. Ad esempio, `ID 7` è diventato tre righe dall'originale poiché sono stati estratti tre risultati di modello TLA. È possibile utilizzare un nodo Unione se si desidera unire questi dati di output ai propri dati originali.



	Concept1	Type1	Concept2	Type2	Conc...	Type3	Con...	Type4	Conc...	Type5	Con...	Type6	Rule Number	ID	Matched Text
1	expensive	NegativeBudget	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0350_opinion	1	<*expensive*>
2	screen	Unknown	difficult	Nega...	Null	Null	Null	Null	Null	Null	Null	Null	0/0145_topic + opinion	2	The <*screen*> is <*hard*> to see when outside
3	software	Unknown	difficult	Nega...	Null	Null	Null	Null	Null	Null	Null	Null	0/0211_opinion + topic	3	<*difficult*> <*software*>
4	nothing	Uncertain	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0153_topic/opinion	4	<*Nothing*> <*,*> I love it
5	like	Positive	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0350_opinion	4	Nothing , <*I love it*>
6	battery life	Unknown	too long	Nega...	Null	Null	Null	Null	Null	Null	Null	Null	0/0145_topic + opinion	5	<*Battery life*> seems <*shorter*> than advertised
7	ubiquitousness	Unknown	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0500_topic	6	<*Ubiquitousness*>
8	40gb model	Unknown	available	Positi...	Null	Null	Null	Null	Null	Null	Null	Null	0/0145_topic + opinion	7	I wish the <*40GB model*> was still <*available*>
9	20gb model	Unknown	Null	Null	Null	Null	Null	Null	Null	Null	Null	Null	0/0102_topic + Negative + topic	7	I have a <*20GB model*> and <*need more*> <*memory*>
10	memory	Unknown	need more	Nega...	Null	Null	Null	Null	Null	Null	Null	Null	0/0102_topic + Negative + topic	7	I have a <*20GB model*> and <*need more*> <*memory*>

Figura 18. Nodo di output Tabella

Capitolo 5. Conversione testo per estrazione

Nodo di traduzione

Il nodo di traduzione può essere utilizzato per tradurre il testo dalle lingue supportate, come l'arabo, il cinese e persiano, in inglese lingue per analisi che utilizzano IBM SPSS Modeler Text Analytics. Questo rende possibile l'estrazione di documenti in lingue a doppio byte che non sarebbero altrimenti supportate e permette agli analisti di estrarre concetti dai documenti in lingua straniera anche se non conoscono la lingua in questione. Tenere presente che è necessario essere in grado di connettersi a SaaS (SDL Software as a Service) per poter utilizzare il nodo di traduzione.

Quando si estrae testo in una di queste lingue, aggiungere semplicemente un nodo di traduzione prima del nodo di modellazione di estrazione testo nel proprio flusso. È possibile anche abilitare la cache nel nodo di traduzione per evitare di ripetere la traduzione ogni volta che il flusso viene eseguito.

È possibile rilevare questo nodo sulla scheda IBM SPSS Modeler Text Analytics della tavolozza dei nodi nella parte inferiore della finestra IBM SPSS Modeler. Per ulteriori informazioni, consultare la sezione “IBM SPSS Modeler Text Analytics Nodi” a pagina 8.

Memorizzazione della traduzione nella cache Se si memorizza in cache la traduzione, il testo tradotto viene memorizzato nel flusso invece che in file esterni. Per evitare di ripetere la traduzione ogni volta che viene eseguito il flusso, selezionare il nodo di traduzione e dai menu scegliere **Modifica > Nodo > Cache > Abilita**. La volta successiva in cui viene eseguito il flusso, l'output della traduzione viene memorizzato nel nodo. L'icona del nodo visualizza un piccolo grafico "documento" che cambia da bianco a verde quando la memoria cache è piena. La cache viene mantenuta per tutta la durata della sessione. Per preservare la cache per un altro giorno (dopo che il flusso viene chiuso e riaperto), selezionare il nodo e dai menu scegliere, **Modifica > Nodo > Cache > Salva cache**. La volta successiva che si apre il flusso, è possibile ricaricare la cache salvata piuttosto che eseguire di nuovo la conversione.

In alternativa, è possibile salvare o abilitare una cache del nodo facendo clic con il tastino destro del mouse sul nodo e scegliendo **Cache** dal menu di contesto.

Importante! Se si sta tentando di richiamare le informazioni sul Web attraverso un server proxy, è necessario abilitare il server proxy nel file `net.properties` per il client e server di IBM SPSS Modeler Text Analytics. Seguire le istruzioni dettagliate in questo file. Ciò si applica quando si accede al Web attraverso il nodo Web Feed o viene richiamata una licenza SaaS (SDL Software as a Service), poiché queste connessioni attraversano Java. Questo file è presente in *C:\Program Files\IBM\SPSS\Modeler\17.1\jre\lib\net.properties* per impostazione predefinita.

Nota: Non è possibile utilizzare il nodo di traduzione per il calcolo del punteggio all'interno di una configurazione IBM SPSS Collaboration and Deployment Services - Scoring.

Nodo traduzione: scheda Traduzione

Campo testo Selezionare il campo che contiene il testo da estrarre, il nome percorso del documento o il nome percorso della directory per i documenti. Questo campo dipende dall'origine dati. È possibile specificare qualsiasi campo stringa, anche quelli con `Direction=None` o `Type=Typeless`.

Il campo di testo rappresenta. Indica cosa contiene il campo di testo specificato nella precedente impostazione. Le scelte sono:

- **Testo effettivo** Selezionare questa opzione se il campo contiene il testo esatto da cui devono essere estratti i concetti.
- **Percorsi dei documenti** Selezionare questa opzione se il campo contiene uno o più percorsi in cui i documenti esterni, che contengono il testo per l'estrazione, risiedono. Ad esempio, se un nodo di elenco file viene utilizzato per leggere in un elenco di documenti, questa opzione deve essere selezionata. Consultare la sezione “Nodo Elenco file” a pagina 11 per ulteriori informazioni.

Codifica di input Selezionare la codifica del testo di origine. È possibile iniziare selezionando l'opzione **Automatico** ma se si nota che alcuni file non vengono elaborati correttamente, si consiglia di selezionare la codifica effettiva dall'elenco. L'opzione Automatico può identificare in maniera errata la codifica quando si tratta testo breve come ad esempio i record brevi del database. L'emissione testo da questo nodo è codificata come `UTF-8`.

Impostazioni Specifica le impostazioni di traduzione per il flusso..

- **Connessione di coppia di lingua**. Selezionare la coppia di lingua che si desidera utilizzare; coppie di lingua disponibili vengono automaticamente visualizzate in questo elenco dopo aver impostato il collegamento al servizio SDL nella finestra di dialogo **Impostazioni per la traduzione**. Consultare la sezione “Impostazioni di traduzione” per ulteriori informazioni.
- **Punto di contatto**. Se sono stati precedentemente creati dei *punti di contatto SDL (TouchPoint)*, selezionarne uno da utilizzare insieme alla traduzione.
- **Salva e riutilizza il testo tradotto in precedenza quando possibile** Specifica che i risultati della traduzione devono essere salvati e se lo stesso numero di record/i documenti sono presenti la volta successiva in cui viene eseguito il flusso, il contenuto viene considerato lo stesso e i risultati della traduzione vengono riutilizzati per risparmiare il tempo di elaborazione. Se questa opzione è selezionata al momento dell'esecuzione e il numero di record non corrisponde a ciò che è stato salvato l'ultima volta, il testo viene completamente tradotto e quindi salvato con il nome etichetta per la successiva esecuzione. Questa opzione è disponibile solo se è stata selezionata una lingua di traduzione SDL.

  Nota: Se il testo è memorizzato nel flusso, è anche possibile abilitare la cache in un nodo di traduzione. In questo caso, non solo i risultati della traduzione vengono riutilizzati ma anche tutto ciò che è a monte del flusso viene ignorato ogni volta che la cache è disponibile.
- **Etichetta** Se si seleziona **Salva e riutilizza testo precedentemente tradotto quando possibile**, è necessario specificare un nome etichetta per i risultati. Questa etichetta viene utilizzata per identificare il testo precedentemente tradotto. Se non viene specificata alcuna etichetta, un'avvertenza viene aggiunta alle proprietà del flusso quando si esegue il flusso e il riutilizzo non è possibile.

Impostazioni di traduzione

In questa casella di dialogo, è possibile definire e gestire la connessione di traduzione SaaS (SDL Software as a Service) che è possibile riutilizzare in qualsiasi momento della traduzione. Una volta definita una connessione, è possibile selezionare rapidamente una connessione di coppia di lingue al momento della traduzione senza dover immettere nuovamente tutte le impostazioni di connessione.

Una connessione coppia di lingue identifica le lingue di origine e di traduzione così come i dettagli della connessione URL per il server. Ad esempio, *Cinese - Inglese* significa che il testo origine è in cinese e la traduzione risultante sarà in inglese. È necessario definire manualmente ogni connessione a cui si accede tramite i servizi SDL in linea.

Importante! Se si sta tentando di richiamare le informazioni sul Web attraverso un server proxy, è necessario abilitare il server proxy nel file `net.properties` per il client e server di IBM SPSS Modeler Text Analytics. Seguire le istruzioni dettagliate in questo file. Ciò si applica quando si accede al Web attraverso il nodo Web Feed o viene richiamata una licenza SaaS (SDL Software as a Service), poiché queste connessioni attraversano Java. Questo file è presente in *C:\Program Files\IBM\SPSS\Modeler\17.1\jre\lib\net.properties* per impostazione predefinita.

URL di connessione Immettere l'URL per SDL Software come connessione di servizio.

Chiave API Immettere la chiave fornita tramite SDL.

ID account Immettere l'ID univoco fornito da SDL.

ID utente Immettere l'ID univico ID fornito da SDL.

Test Fare clic su **Verifica** per verificare che la connessione sia configurata correttamente e per visualizzare la coppia o le coppie di lingua rilevate su tale connessione.

Uso del nodo di traduzione

Per estrarre i concetti dalle lingue di traduzione, come arabo, cinese o persiano, aggiungere prima un nodo di traduzione ad ogni nodo di estrazione testo del flusso.

Se il testo che deve essere tradotto è contenuto in uno o più file esterni, è possibile utilizzare un nodo di elenco file dell'elenco per leggere in un elenco di nomi. In questo caso, il nodo di traduzione verrebbe aggiunto tra il nodo di elenco file e tutti i nodi di estrazione testo successivi e l'output sarà l'ubicazione in cui si trova il testo tradotto.

Capitolo 6. Visione di testo di origine esterna

Nodo Visualizzatore file

Quando si estrae una raccolta di documenti, è possibile specificare i nomi percorso completi dei file direttamente nei propri nodi di modellazione estrazione testo e di traduzione. Tuttavia, quando di emette su un nodo Tabella, verrà visualizzato solo il nome percorso completo di un documento piuttosto che il testo al suo interno. Il nodo Visualizzatore file può essere utilizzato come un analogo del nodo Tabella e consente di accedere al testo effettivo all'interno di ciascuno dei documenti senza doverli unire tutti insieme in un singolo file.

Il nodo Visualizzatore file può aiutare a comprendere meglio i risultati dall'estrazione del testo fornendo l'accesso al testo di origine o non tradotto, da cui sono stati estratti i concetti, altrimenti inaccessibile nel flusso. Questo nodo viene aggiunto al flusso, dopo un nodo Elenco file, per ottenere un elenco di collegamenti a tutti i file.

Il risultato di questo nodo è una finestra che mostra tutti gli elementi del documento che sono stati letti e utilizzati per estrarre concetti. Da questa finestra, è possibile fare clic su un'icona della barra degli strumenti per avviare il report in un browser esterno che elenca i nomi documento come collegamenti ipertestuali. È possibile fare clic su un collegamento per aprire il documento corrispondente nella raccolta. Consultare la sezione “Uso del nodo Visualizzatore file” per ulteriori informazioni.

È possibile rilevare questo nodo sulla scheda IBM SPSS Modeler Text Analytics della tavolozza dei nodi nella parte inferiore della finestra IBM SPSS Modeler. Per ulteriori informazioni, consultare la sezione “IBM SPSS Modeler Text Analytics Nodi” a pagina 8.

Nota: quando si lavora in modalità client-server e i nodi Visualizzatore file sono parte del flusso, le raccolte di documenti devono essere memorizzate in una directory del server Web sul server. Poiché il nodo di output di estrazione testo produce un elenco di documenti memorizzati nella directory del server Web, le impostazioni della sicurezza del server Web gestiscono le autorizzazioni per questi documenti.

Impostazioni del nodo Visualizzatore file

È possibile specificare le seguenti impostazioni per il nodo Visualizzatore file.

Campo Documento. Selezionare il campo dai dati che contengono il nome ed il percorso completi dei documenti da visualizzare.

Titolo della pagina HTML generata. Crea un titolo da visualizzare nella parte superiore della pagina che contiene l'elenco di documenti.

Uso del nodo Visualizzatore file

Il seguente esempio mostra come utilizzare il nodo Visualizzatore file.

Esempio: nodo Elenco file e un nodo Visualizzatore file

Figura 19. Flusso che illustra l'uso di un nodo Visualizzatore file

1. **Nodo di elenco file (Scheda Impostazioni).** In primo luogo, è stato aggiunto questo nodo per specificare dove i documenti di testo sono memorizzati.

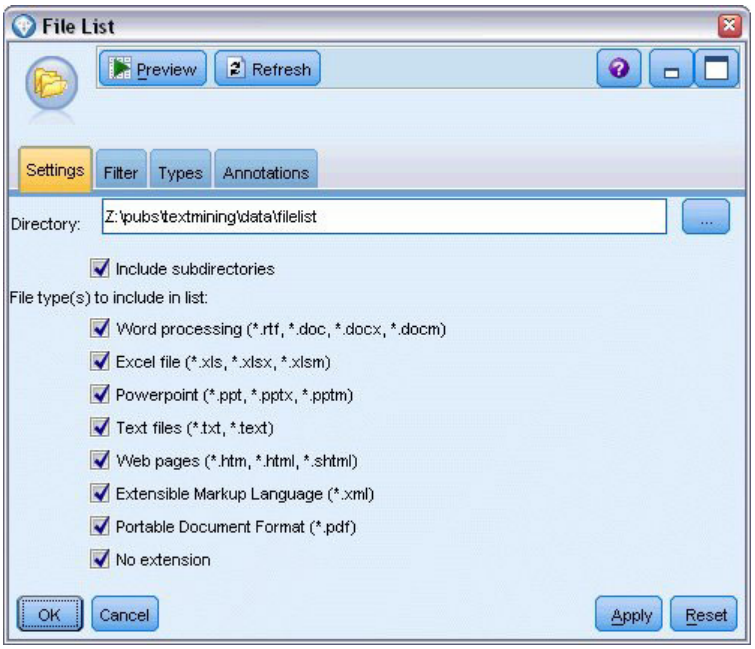


Figura 20. Finestra di dialogo del nodo Elenco file: scheda Impostazioni

2. **Nodo Visualizzatore file (scheda Impostazioni).** Quindi, è stato allegato il nodo Visualizzatore file per produrre un elenco HTML di documenti.

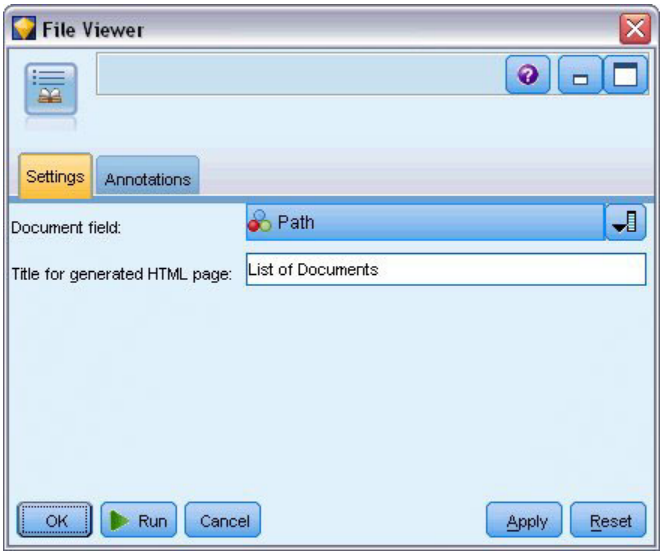


Figura 21. Finestra di dialogo del nodo Visualizzatore file: scheda Impostazioni

3. **Finestra di dialogo dell'output di Visualizzatore file.** Successivamente viene eseguito il flusso che emette l'elenco di documenti in una nuova finestra.

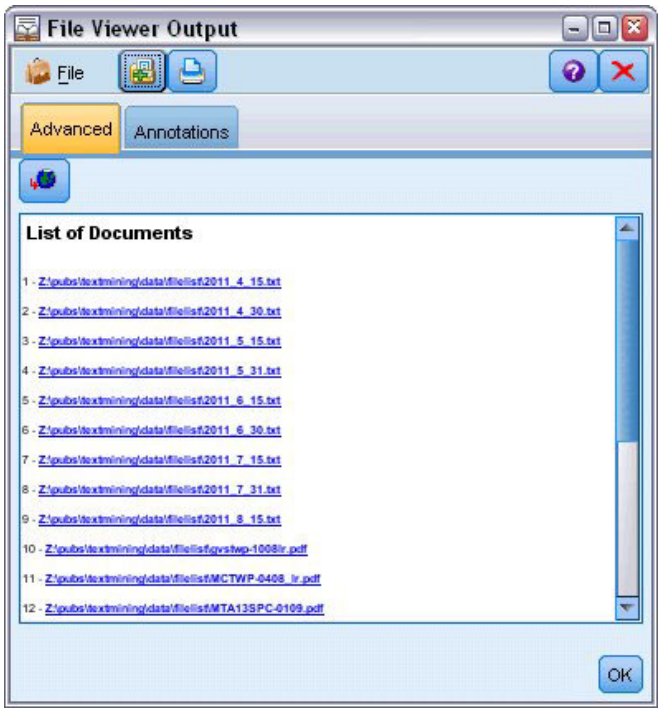


Figura 22. Output di Visualizzatore file

4. Per visualizzare i documenti, si fa clic sul pulsante della barra degli strumenti che mostra un mondo con una freccia rossa. Viene aperto un elenco di collegamenti ipertestuali del documento nel browser.

Capitolo 7. Proprietà dei nodi per lo script

IBM SPSS Modeler dispone di un linguaggio di script che consente di eseguire flussi dalla riga comandi. Qui, è possibile ottenere ulteriori informazioni sulle proprietà del nodo che sono specifiche per ciascuno dei nodi forniti con IBM SPSS Modeler Text Analytics. Per ulteriori informazioni sulla serie standard di nodi forniti con IBM SPSS Modeler, fare riferimento al manuale Scripting and Automation Guide.

Nodo di elenco file: filelistnode

È possibile utilizzare le proprietà nella seguente tabella per gli script. Il nodo è denominato `filelistnode`.

Tabella 7. Proprietà di script del nodo di elenco file

Proprietà script	Tipo di dati
`path`	*stringa*
`recurse`	*flag*
`word_processing`	*flag*
`excel_file`	*flag*
`powerpoint_file`	*flag*
`text_file`	*flag*
`web_page`	*flag*
`xml_file`	*flag*
`pdf_file`	*flag*
`no_extension`	*flag*

Nota: il parametro 'Create list' non è più disponibile e tutti gli script che contiene questa opzione verranno automaticamente convertiti in un output 'Files'.

Nodo di flusso Web: webfeednode

È possibile utilizzare le proprietà nella seguente tabella per gli script. Il nodo è denominato `webfeednode`.

Tabella 8. Proprietà di script del nodo di flusso Web

Proprietà script	Tipo di dati	Descrizione proprietà
`url`	*string1 string2 ...stringn*	Ogni URL viene specificato nella struttura di elenco. Elenco di URL separati da "\n"
`recent_entries`	*flag*	
`limit_entries`	*integer*	Numero di voci più recenti da leggere per URL.
`use_previous`	*flag*	Per salvare e riutilizzare la cache del flusso Web.
`use_previous_label`	*stringa*	Nome per la cache Web salvata.
`start_record`	*stringa*	Tag di avvio non RSS.
`url` *n* `.title`	*stringa*	Per ciascun URL nell'elenco, è necessario definirne uno anche qui. La prima sarà `url1.title`, dove il numero corrisponde alla sua posizione nell'elenco URL. Questo è il tag iniziale che contiene il titolo del contenuto.
`url` *n* `.short_description`	*stringa*	Come per `url` *n* `.title`.
`url` *n* `.description`	*stringa*	Come per `url` *n* `.title`.

Tabella 8. Proprietà di script del nodo di flusso Web (Continua)

Proprietà script	Tipo di dati	Descrizione proprietà
`url` *n* `.authors`	*stringa*	Come per `url` *n* `.title`.
`url` *n* `.contributors`	*stringa*	Come per `url` *n* `.title`.
`url` *n* `.published_date`	*stringa*	Come per `url` *n* `.title`.
`url` *n* `.modified_date`	*stringa*	Come per `url` *n* `.title`.
`html_alg`	`Nessuno` `HTMLCleaner`	Metodo di filtraggio del contenuto.
`discard_lines`	*flag*	Scarta righe brevi. Utilizzato con `min_words`
`min_words`	*integer*	Numero minimo di parole.
`discard_words`	*flag*	Scarta righe brevi. Utilizzato con `min_avg_len`
`min_avg_len`	*integer*	
`discard_scw`	*flag*	Scarta righe con molte parole di un solo carattere. Utilizzato con `max_scw`
`max_scw`	*integer*	Proporzione massima 0-100 percentuale di parole a un solo carattere in una riga
`discard_tags`	*flag*	Scarta righe contenenti determinate tag.
`tag`	*stringa*	I caratteri speciali devono essere preceduti da un carattere barra retroversa \.
`discard_spec_words`	*flag*	Scarta righe contenenti stringhe specifiche
`words`	*stringa*	I caratteri speciali devono essere preceduti da un carattere barra retroversa \.

Nodo di estrazione testo: TextMiningWorkbench

È possibile utilizzare i seguenti parametri per definire o aggiornare un nodo mediante gli script. Il nodo stesso è denominato `TextMiningWorkbench`.

Importante! Non è possibile specificare un modello di risorsa diverso mediante gli script. Se si pensa che è necessario un modello, selezionarlo nella finestra di dialogo del nodo.

Tabella 9. Proprietà script del nodo di modellazione di estrazione testo

Proprietà script	Tipo di dati	Descrizione proprietà
`testo`	*campo*	
`metodo`	`ReadText` `ReadPath`	
`docType`	*integer*	Valori possibili (0,1,2) in cui 0 = `Testo completo`, 1 = `Testo strutturato` e 2 = `XML`

`encoding`	`Automatico`
	`"UTF-8"`
	`"UTF-16"`
	`"ISO-8859-1"`
	`"US-ASCII"`
	`"CP850"`
	`"EUC-JP"`
	`"SHIFT-JIS"`
	`"ISO2022-JP"`
		Si noti che i valori con caratteri speciali, quali `"UTF-8"`, devono essere racchiusi tra virgolette, in modo da evitare confusioni con operatori matematici.

`unity`	*integer*	Valori possibili (0,1) in cui 0 = `Paragrafo` e 1 = `Documento`

Tabella 9. Proprietà script del nodo di modellazione di estrazione testo (Continua)

Proprietà script	Tipo di dati	Descrizione proprietà
para_min	*integer*	
para_max	*integer*	
mtag	*stringa*	Contiene tutte le impostazioni mtag (dalla finestra di dialogo Impostazioni per i file XML)
mclef	*stringa*	Contiene tutte le impostazioni mclef (dalla finestra di dialogo Impostazioni per i file di testo strutturato)
partition	*campo*	
custom_field	*flag*	Indica se verrà specificato un campo partizione.
use_model_name	*flag*	
model_name	*stringa*	
use_partitioned_data	*flag*	Se è definito un campo partizione, per la creazione del modello verranno utilizzati solo i dati di addestramento.
model_output_type	Interattivo Modello	Risultati interattivi in un modello di categoria. Risultati di modello in un modello di concetto.
use_interactive_info	*flag*	Per la creazione interattiva solo in una sessione workbench.
reuse_extraction_results	*flag*	Per la creazione interattiva solo in una sessione workbench.
interactive_view	Categorie TLA Cluster	Per la creazione interattiva solo in una sessione workbench.
extract_top	*integer*	Questo parametro viene utilizzato quando model_type = Concept
use_check_top	*flag*	
check_top	*integer*	
use_uncheck_top	*flag*	
uncheck_top	*integer*	
lingua	de en es fr it ja nl pt	
frequency_limit	*integer*	Obsoleto in 14.0.
concept_count_limit	*integer*	Limita l'estrazione ai concetti con frequenza globale di almeno questo valore. Non disponibile per testo giapponese
fix_punctuation	*flag*	Non disponibile per testo giapponese
fix_spelling	*flag*	Non disponibile per testo giapponese
spelling_limit	*integer*	Non disponibile per testo giapponese
extract_uniterm	*flag*	Non disponibile per testo giapponese

Tabella 9. Proprietà script del nodo di modellazione di estrazione testo (Continua)

Proprietà script	Tipo di dati	Descrizione proprietà
extract_nonlinguistic	*flag*	Non disponibile per testo giapponese
upper_case	*flag*	Non disponibile per testo giapponese
group_names	*flag*	Non disponibile per testo giapponese
permutation	*integer*	Numero massimo di permutazioni di parole (il valore predefinito è 3). Non disponibile per testo giapponese.
conclusioni jp_algorithmset solo rappresentativo di Tutte le opinioni	0 1 2	Solo per estrazione testo giapponese. 0 = Estrazione secondaria di opinione 1 = Estrazione di dipendenza 2 = Nessuna serie di analizzatore secondaria.
jp_algorithm_sense_mode	0 1 2	Solo per estrazione testo giapponese. 0 = Solo conclusioni 2 = Solo rappresentativo 3 = Tutte le opinioni.

Nugget del modello di estrazione testo: TMWBModelApplier

È possibile utilizzare le proprietà nella seguente tabella per gli script. Il nugget è denominato TMWBModelApplier.

Tabella 10. Proprietà dei nugget del modello di estrazione testo

Proprietà script	Tipo di dati	Descrizione proprietà
scoring_mode	Campi Record	
field_values	Flag Counts	Questa opzione non è disponibile nel nugget del modello di categoria. Per Flag, impostato su TRUE o FALSE
true_value	*stringa*	Con Flag, definire il valore per true.
false_value	*stringa*	Con Flag, definire il valore per false.
extension_concept	*stringa*	Specificare un'estensione per il nome campo. I nomi campo vengono generati utilizzando il nome concetto più questa estensione. Specificare dove inserire questa estensione utilizzando il valore add_as.
extension_category	*stringa*	Estensione nome campo. È possibile scegliere di specificare un prefisso/suffisso di estensione per il nome campo o è possibile scegliere di utilizzare i codici della categoria. I nomi campo vengono generati utilizzando il nome categoria e questa estensione. Specificare dove inserire questa estensione utilizzando il valore add_as.
add_as	Suffisso Prefisso	
fix_punctuation	*flag*	

Tabella 10. Proprietà dei nugget del modello di estrazione testo (Continua)

Proprietà script	Tipo di dati	Descrizione proprietà
excluded_subcategories_descriptors	RollUpToParent Ignora	Solo per i modelli di categoria. Se una sottocategoria è deselezionata. Questa opzione consente di specificare come verranno gestiti i descrittori appartenenti alle sottocategorie che non sono state selezionate per il calcolo del punteggio. Sono possibili due opzioni. • Ignora. L'opzione Escludi i descrittori completamente dal calcolo del punteggio farà sì che i descrittori di sottocategorie che non hanno segni di spunta (non selezionati) vengono ignorati e non utilizzati durante il calcolo del punteggio. • RollUpToParent. Con l'opzione Aggrega descrittori con quelli della categoria principale, i descrittori di sottocategorie che non hanno segni di spunta (non selezionati) vengono utilizzati come descrittori per la categoria principale (la categoria al di sopra di questa sottocategoria). Se diversi livelli di sottocategorie non sono selezionati, i descrittori verranno riepilogati sotto la prima categoria principale disponibile.
check_model	*flag*	Cancellato nella versione 14.
testo	*campo*	
metodo	ReadText ReadPath	
docType	*integer*	Valori possibili (0,1,2) in cui 0 = Testo completo, 1 = Testo strutturato e 2 = XML
encoding	Automatico "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	Si noti che i valori con caratteri speciali, quali "UTF-8", devono essere racchiusi tra virgolette, in modo da evitare confusioni con operatori matematici.
lingua	de en es fr it ja nl pt	

Nodo di analisi di collegamento del testo: textlinkanalysis

È possibile utilizzare i parametri nella seguente tabella per definire o aggiornare un nodo mediante gli script. Il nodo è denominato textlinkanalysis.

Importante! Non è possibile specificare un modello di risorsa diverso mediante gli script. Per selezionare un modello, è necessario farlo dall'interno della finestra di dialogo del nodo.

Tabella 11. Proprietà degli script del nodo TLA (Text Link Analysis)

Proprietà script	Tipo di dati	Descrizione proprietà
id_field	*campo*	
testo	*campo*	
metodo	ReadText ReadPath	
docType	*integer*	Valori possibili (0,1,2) in cui 0 = Testo completo, 1 = Testo strutturato e 2 = XML
encoding	Automatico "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	Si noti che i valori con caratteri speciali, quali "UTF-8", devono essere racchiusi tra virgolette, in modo da evitare confusioni con operatori matematici.
unity	*integer*	Valori possibili (0,1) in cui 0 = Paragrafo e 1 = Documento
para_min	*integer*	
para_max	*integer*	
mtag	*stringa*	Contiene tutte le impostazioni mtag (dalla finestra di dialogo Impostazioni per i file XML)
mclef	*stringa*	Contiene tutte le impostazioni mclef (dalla finestra di dialogo Impostazioni per i file di testo strutturato)
lingua	de en es fr it ja nl pt	
concept_count_limit	*integer*	Limita l'estrazione ai concetti con frequenza globale di almeno questo valore. Non disponibile per testo giapponese
fix_punctuation	*flag*	Non disponibile per testo giapponese
fix_spelling	*flag*	Non disponibile per testo giapponese
spelling_limit	*integer*	Non disponibile per testo giapponese
extract_uniterm	*flag*	Non disponibile per testo giapponese
extract_nonlinguistic	*flag*	Non disponibile per testo giapponese
upper_case	*flag*	Non disponibile per testo giapponese
group_names	*flag*	Non disponibile per testo giapponese
permutation	*integer*	Numero massimo di permutazioni di parole (il valore predefinito è 3). Non disponibile per testo giapponese.

Tabella 11. Proprietà degli script del nodo TLA (Text Link Analysis) (Continua)

Proprietà script	Tipo di dati	Descrizione proprietà
conclusioni jp_algorithmset solo rappresentativo di Tutte le opinioni	0 1 2	Solo per estrazione testo giapponese. 0 = Estrazione secondaria di opinione 1 = Estrazione di dipendenza 2 = Nessuna serie di analizzatore secondaria.
jp_algorithm_sense_mode	0 1 2	Solo per estrazione testo giapponese. 0 = Solo conclusioni 2 = Solo rappresentativo 3 = Tutte le opinioni.

Nodo traduzione: translatenode

È possibile utilizzare le proprietà nella seguente tabella per gli script. Il nodo stesso è denominato translatenode.

Tabella 12. Proprietà del nodo di traduzione

Proprietà script	Tipo di dati	Descrizione proprietà
testo	*campo*	
metodo	ReadText ReadPath	
encoding	Automatico "Big5", "Big5-HKSCS" , "UTF-8", "UTF-16", "US-ASCII", "Latin1", "CP850", "CP874", "CP1250", "CP1251", "CP1252", "CP1253", "CP1254","CP1255", "CP1256","CP1257", "CP1258","GB18030", "GB2312","GBK", "eucJP","JIS7", "SHIFT_JIS", "eucKR","TSCII", "ucs2","KOI8-R", "KOI8-U","ISO8859-1", "ISO8859-2","ISO8859-3", "ISO8859-4","ISO8859-5", "ISO8859-6","ISO8859-7", "ISO8859-8", "ISO8859-8-i", "ISO8859-9","ISO8859-10", "ISO8859-13","ISO8859-14", "ISO8859-15","IBM 850", "IBM 866","Apple Roman", "TIS-620"	Si noti che i valori con caratteri speciali, quali "UTF-8", devono essere racchiusi tra virgolette, in modo da evitare confusioni con operatori matematici.
lw_server_type	LOC WAN HTTP	
lw_hostname	*stringa*	
lw_port	*integer*	
url	*stringa*	url of the translation server
apiKey	*stringa*	
user_id	*stringa*	

Tabella 12. Proprietà del nodo di traduzione (Continua)

Proprietà script	Tipo di dati	Descrizione proprietà
lpid	*integer*	Non utilizzato se è impostato *language_from* o *language_from_id*.
translate_from	Arabo, Cinese, Cinese tradizionale, Ceco, Danese,Olandese, Inglese,Francese, Tedesco, Greco, Hindi, Ungherese, Italiano, Giapponese, Coreano, Persiano, Polacco, Portoghese, Rumeno, Russo, Spagnolo, Somalo, Svedese	
translate_from_id	ara, chi, cht, cze, dan, dut, eng, fra, ger, gre, hin, hun, ita, jpn, kor, per, pol, por, rum, rus, som, spa, swe	
translate_to	Inglese	
translate_to_id	eng	
translation_accuracy	*integer*	Specifica il livello di precisione desiderato per il processo di traduzione - scegliere un valore compreso tra 1 e 3
use_previous_translation	*flag*	Specifica che i risultati della traduzione esistono già da una precedente esecuzione e possono essere riutilizzati.
translation_label	*stringa*	Immettere un'etichetta per identificare i risultati della traduzione per il riutilizzo

Capitolo 8. Modalità Workbench interattivo

Da un nodo di modellazione di estrazione testo, è possibile scegliere di avviare una sessione workbench interattiva durante l'esecuzione del flusso. In questo workbench, è possibile estrarre i concetti dai dati di testo, creare categorie ed esplorare i modelli di analisi di collegamento del testo e i cluster e generare modelli di categoria. In questo capitolo, si discute l'interfaccia workbench da una prospettiva di alto livello insieme ai principali elementi da gestire tra cui:

- **Risultati di estrazione.** Dopo che viene eseguita un'estrazione, queste sono le parole chiave e frasi identificate ed estratte dai dati di testo, note anche come *concetti*. Questi concetti sono raggruppati in *tipi*. Utilizzando questi concetti e tipi, è possibile esplorare i dati, oltre a creare le proprie categorie. Questi vengono gestiti nella vista **Categorie e concetti**.
- **Categorie.** Utilizzando i descrittori (ad esempio risultati di estrazione, modelli e regole) come una definizione, è possibile creare manualmente o automaticamente una serie di categorie a cui documenti e record sono assegnati in base al fatto che contengano o meno una parte della definizione della categoria. Questi vengono gestiti nella vista **Categorie e concetti**.
- **Cluster.** I *Cluster* sono un raggruppamento di concetti tra i quali sono stati scoperti collegamenti che indicano una relazione tra loro. I concetti sono raggruppati utilizzando un algoritmo complesso che utilizza, tra gli altri fattori, quanto spesso due concetti vengono visualizzati insieme rispetto a quanto spesso vengono visualizzati separatamente. Questi vengono gestiti nella vista **Cluster**. È anche possibile aggiungere i concetti che costituiscono un cluster alle categorie.
- **Modelli di analisi di collegamento del testo.** Se si hanno regole di modelli di analisi di collegamento del testo (TLA) nelle proprie risorse linguistiche o si sta utilizzando un modello di risorse che già dispone di alcune regole TLA, è possibile estrarre i modelli dai dati di testo. Tali modelli possono aiutare a scoprire relazioni interessanti tra i concetti nei propri dati. È anche possibile utilizzare tali modelli come descrittori nelle proprie categorie. Questi vengono gestiti nella vista **TLA (Text Link Analysis)**. Per il testo giapponese, è necessario selezionare un analizzatore secondario e attivare l'estrazione TLA.
- **Risorse linguistiche.** Il processo di estrazione si basa su una serie di parametri e definizioni linguistiche per gestire il modo in cui il testo viene estratto e manipolato. Questi vengono gestiti in forma di modelli e librerie nella vista **Editor di risorse**.

Vista Categorie e concetti

L'interfaccia dell'applicazione è composta da diverse viste. La vista Categorie e concetti è la finestra in cui è possibile creare e analizzare le categorie come esplorare e regolare i risultati di estrazione. **Categorie** si riferisce a un gruppo di idee e modelli strettamente correlati a ai documenti e record assegnati tramite un processo di calcolo del punteggio. Mentre i **concetti** si riferiscono al livello più basilare dei risultati di estrazione disponibili da utilizzare come blocchi di creazione, denominati descrittori, per le categorie.

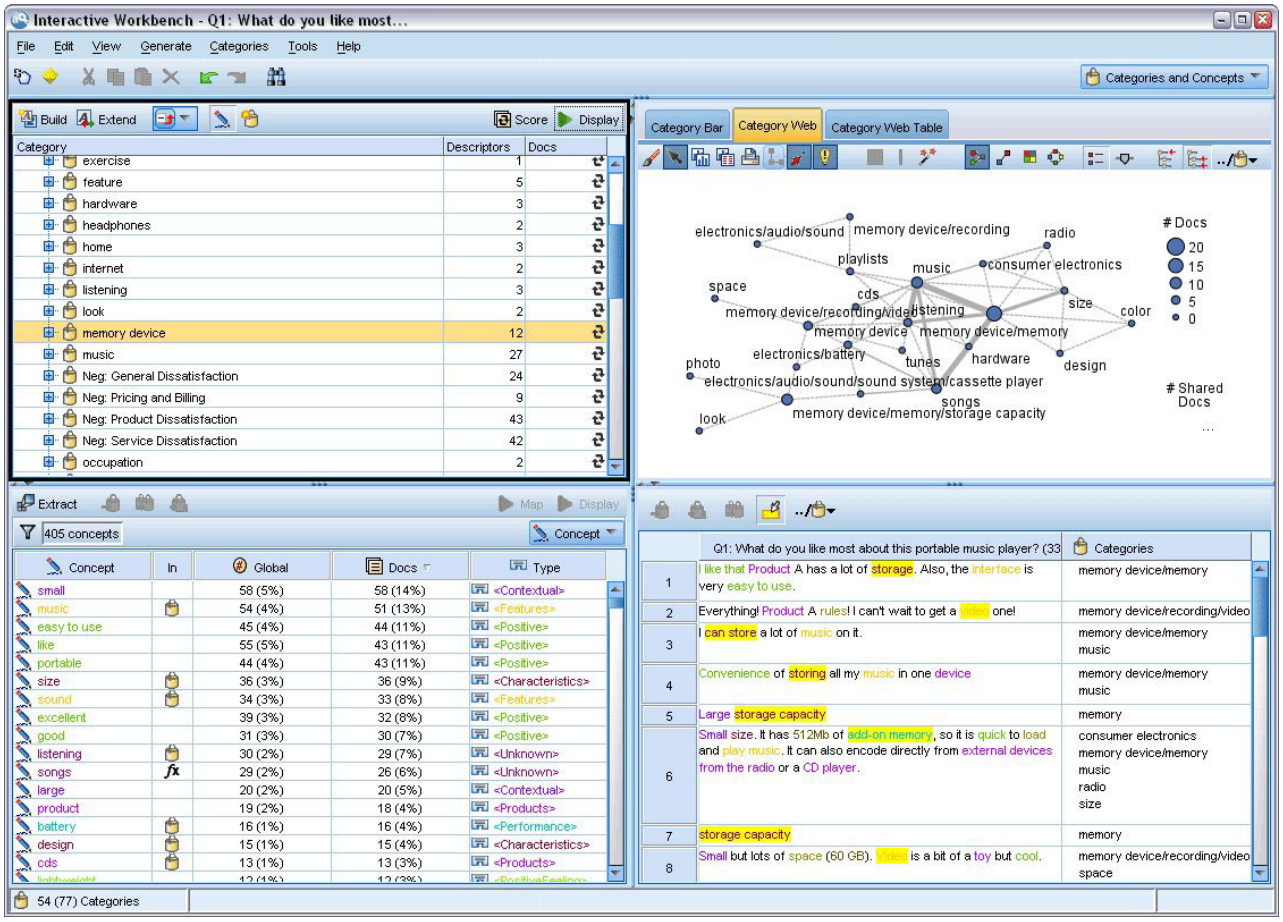


Figura 23. Vista Categorie e concetti

La vista Categorie e concetti è organizzata in quattro riquadri, ognuno dei quali può essere nascosto o mostrato selezionandone il nome dal menu Visualizza. Consultare la sezione Capitolo 10, “Categorizzazione dei dati di testo”, a pagina 101 per ulteriori informazioni.

riquadro Categorie

Situata nell'angolo in alto a sinistra, questa area presenta una tabella in cui è possibile gestire qualsiasi categoria creata. Dopo aver estratto i concetti e i tipi di dati di testo, è possibile iniziare la creazione di categorie utilizzando tecniche come le reti semantiche e di inclusione o creandole manualmente. Se si fa doppio clic con il mouse su un nome di categoria, viene aperta la finestra di dialogo Definizioni di categoria che visualizza tutti i descrittori che costituiscono la sua definizione, come ad esempio i concetti, i tipi, i modelli e le regole di categoria. Consultare la sezione Capitolo 10, “Categorizzazione dei dati di testo”, a pagina 101 per ulteriori informazioni. Non tutte le tecniche automatiche sono disponibili per tutte le lingue.

Quando si seleziona una riga nel riquadro, è possibile visualizzare le informazioni sui documenti/record corrispondenti o i descrittori nei riquadri Dati e Visualizzazione.

Riquadro Risultati di estrazione

Situato nell'angolo in basso a sinistra, questa area presenta i risultati di estrazione. Quando si esegue un'estrazione, il motore di estrazione legge attraverso i dati di testo, individua i concetti pertinenti e assegna un tipo ad ognuno. I **concetti** sono parole o frasi estratte dai dati di testo. I **tipi** sono raggruppamenti semantici dei concetti memorizzati sotto forma di dizionari di tipo. Una volta completata

l'estrazione, i concetti e i tipi sono visualizzati con codifica colori nel riquadro Risultati di estrazione. Consultare la sezione “Risultati estrazione: concetti e tipi” a pagina 87 per ulteriori informazioni.

È possibile visualizzare la serie di termini sottostante per un concetto, passando il puntatore del mouse sul nome concetto. In questo modo viene visualizzato un suggerimento che mostra il nome concetto e diverse righe di termini che sono raggruppate in tale concetto. Tali termini sottostanti comprendono i sinonimi definiti nelle risorse linguistiche (indipendentemente dal fatto che sono stati rilevati nel testo o meno) così come gli eventuali termini estratti singolari/plurali, termini permutati, termini da raggruppamenti casuali e così via. È possibile copiare tali termini o consultare la serie completa di termini sottostanti facendo clic con il tasto destro sul nome del concetto e scegliendo l'opzione di menu di scelta rapida.

L'estrazione testo è un processo iterativo in cui i risultati di estrazione vengono esaminati in base al contesto dei dati di testo, ottimizzato per produrre risultati nuovi e poi rivalutato. I risultati dell'estrazione possono essere ottimizzati modificando le risorse linguistiche. Questa ottimizzazione può essere eseguite in parte direttamente dal riquadro Risultati di estrazione o Dati ma anche direttamente nella vista Editor di risorsa. Consultare la sezione “Vista Editor delle risorse” a pagina 80 per ulteriori informazioni.

riquadro Visualizzazione

Situato nell'angolo superiore destro, questa area presenta più prospettive sulle accomunanze nella categorizzazione del documento/record. Ogni grafico e diagramma presenta informazioni simili ma in modo differente o con un livello di dettaglio differente. Questi diagrammi e grafici in questo riquadro possono essere utilizzati per analizzare i risultati di categorizzazione e come aiuto per regolare le categorie o la notifica. Ad esempio, in un grafico è possibile scoprire le categorie che sono troppo simili (ad esempio, condividono più del 75% dei relativi record) o troppo distinte. Il contenuto in un grafico o diagramma corrisponde ai criteri di selezione negli altri riquadri. Consultare la sezione “Grafici e diagrammi di categoria” a pagina 159 per ulteriori informazioni.

Riquadro Dati

Il riquadro Dati si trova nell'angolo in basso a destra. Questo riquadro mostra una tabella contenente documenti o record che corrispondono ad una selezione di un'altra area della vista. In base alla selezione, solo il testo corrispondente viene visualizzato nel riquadro Dati. Una volta effettuata la selezione, fare clic sul pulsante **Visualizza** per compilare il riquadro dei dati con il testo corrispondente.

Se è stata effettuata una selezione in un altro riquadro, i relativi documenti o record mostrano i concetti evidenziati con colore per aiutare l'utente a identificarli facilmente nel testo. È inoltre possibile passare il mouse su elementi di colore per visualizzare un suggerimento che mostra il nome del concetto da cui è stato estratto e il tipo a cui è stato assegnato. Per ulteriori informazioni, consultare la sezione “Riquadro Dati” a pagina 110.

Cerca e Trova nella vista Categorie e concetti

In alcuni casi, potrebbe essere necessario individuare rapidamente le informazioni in una sezione particolare. Utilizzando la barra degli strumenti Trova, è possibile immettere la stringa che si desidera ricercare e definire altri criteri di ricerca come la sensibilità al maiuscolo/minuscolo o direzione della ricerca. Quindi, è possibile scegliere il riquadro in cui si desidera effettuare la ricerca.

Per utilizzare la funzione Trova

1. Nella vista Categorie e concetti, scegliere **Modifica > Trova** dal menu. La barra degli strumenti Trova viene visualizzata sui riquadri Categorie e Visualizzazione.

2. Immettere la stringa di parole che si desidera ricercare nella casella di testo. È possibile utilizzare il pulsante della barra degli strumenti per controllare la sensibilità al maiuscolo/minuscolo, la corrispondenza parziale e la direzione della ricerca.
3. Nella barra degli strumenti, fare clic sul nome del riquadro in cui si desidera effettuare la ricerca. Se viene individuata una corrispondenza, il testo viene evidenziato nella finestra.
4. Per ricercare la corrispondenza successiva, fare clic sul nome del riquadro di nuovo.

Vista Cluster

Nella vista Cluster, è possibile creare ed esplorare i risultati del cluster trovati nei dati di testo. I **cluster** sono raggruppamenti di concetti generati da algoritmi di cluster in base a quanto spesso si verificano concetti e quanto spesso vengono visualizzati insieme. L'obiettivo dei cluster è di raggruppare concetti che ricorrono insieme in base a come il testo che contengono corrisponda ai descrittori (concetti, regole, modelli) per ogni categoria.

Più spesso i concetti all'interno di un cluster ricorrono insieme con quelli con la minore frequenza che ricorrono con altri concetti, tanto più il cluster è abile a identificare le relazioni dei concetti. Due concetti ricorrono insieme quando entrambi vengono visualizzati (o uno dei relativi termini o sinonimi) nello stesso documento o record. Consultare la sezione Capitolo 11, “Analisi dei cluster”, a pagina 147 per ulteriori informazioni.

È possibile creare cluster ed esplorarli in una serie di diagrammi e grafici che potrebbe aiutare l'utente a scoprire le relazioni tra concetti che impiega troppo tempo per trovare. Mentre non è possibile aggiungere l'intero cluster alle categorie, è possibile aggiungere i concetti in un cluster ad una categoria mediante la finestra Definizioni cluster. Consultare la sezione “Definizioni di cluster” a pagina 151 per ulteriori informazioni.

È possibile apportare delle modifiche alle impostazioni per il raggruppamento in cluster per influenzare i risultati. Consultare la sezione “Creazione di cluster” a pagina 148 per ulteriori informazioni.

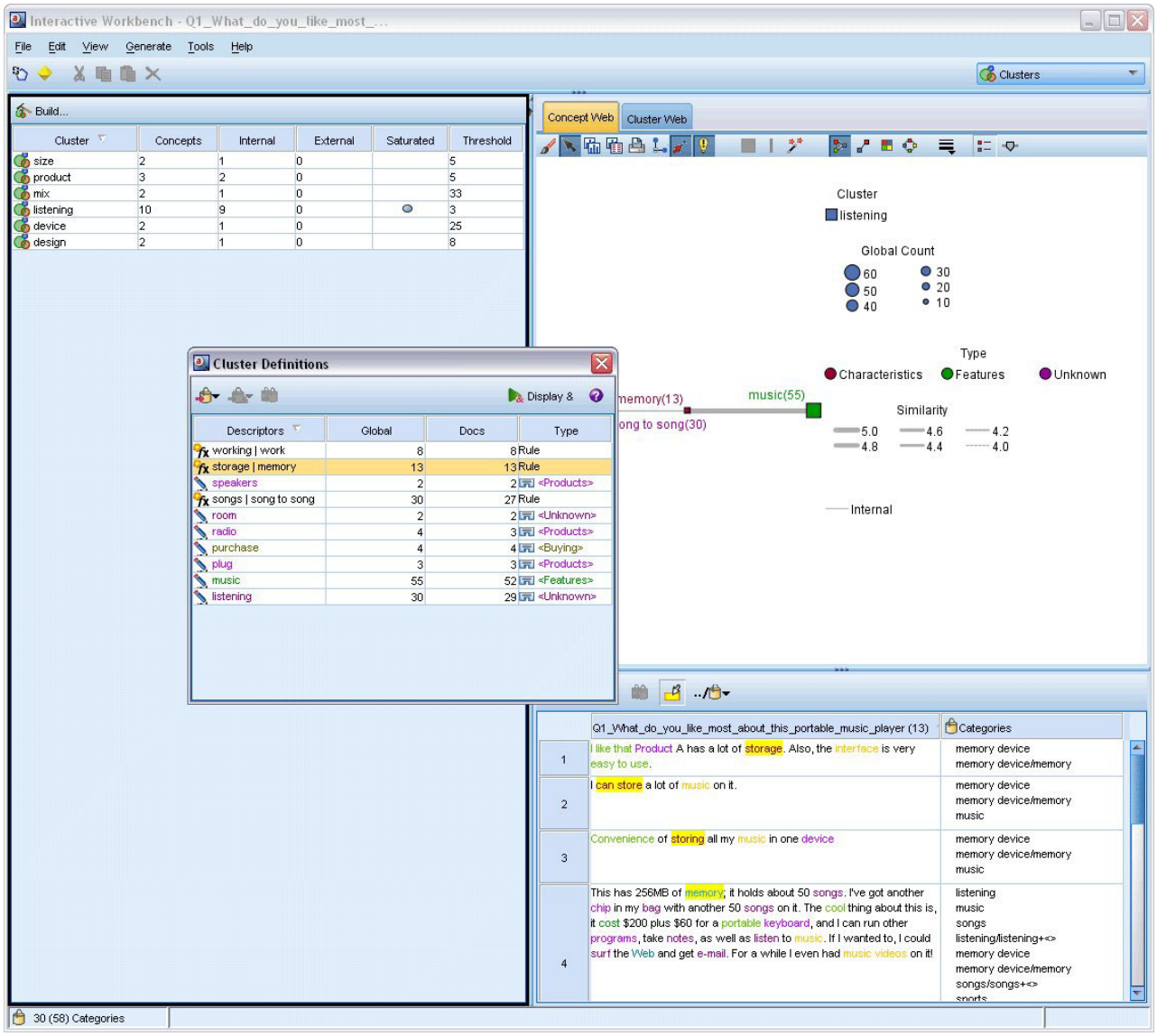


Figura 24. Vista Cluster

La vista Cluster è organizzata in tre riquadri, ognuno dei quali può essere nascosto o mostrato selezionandone il nome dal menu Visualizza. Di solito, solo il riquadro Cluster e Visualizzazione sono visibili.

Riquadro Cluster

Situato sul lato sinistro, questo riquadro visualizza i cluster rilevati nei dati di testo. È possibile creare risultati di raggruppamento in cluster facendo clic sul pulsante **Crea**. I cluster sono costituiti da un algoritmo di raggruppamento, che tenta di identificare i concetti che si verificano spesso insieme.

Ogni volta che una nuova estrazione avviene, i risultati del cluster vengono cancellati ed è possibile ricostruire il cluster per ottenere gli ultimi risultati. Quando si crea il cluster, è possibile modificare alcune impostazioni, quali il numero massimo di cluster da creare, il numero massimo di concetti che può contenere o il numero massimo di collegamenti con concetti esterni può avere. Consultare la sezione “Esplorazione dei cluster” a pagina 151 per ulteriori informazioni.

Riquadro Visualizzazione

Situato nell'angolo superiore destro, questo riquadro offre due prospettive sui cluster: un grafico Web di concetto e un grafico Web di cluster. Se non è già visibile, è possibile accedere a questo riquadro dal menu Visualizza (**Visualizza > Visualizzazione**). In base a quanto è selezionato nel riquadro del cluster, è possibile visualizzare le interazioni di corrispondenza tra cluster o all'interno di uno. I risultati sono presentati in più formati:

- **Web di concetto.** Questo grafico Web presenta tutti i concetti all'interno del cluster selezionato oltre a concetti collegati al di fuori del cluster.
- **Web di cluster.** Grafico Web che mostra il collegamento dal cluster selezionato per altri cluster, così come eventuali collegamenti tra altri cluster.

Note: per visualizzare il grafico Web di cluster, è necessario avere già creato i cluster con i collegamenti esterni. I collegamenti esterni sono collegamenti tra coppie concetto in cluster separati (un concetto all'interno di un cluster e un concetto esterno in un altro cluster). Consultare la sezione “Grafici del cluster” a pagina 161 per ulteriori informazioni.

Riquadro Dati

Il riquadro Dati si trova nell'angolo in basso a destra ed è nascosto per impostazione predefinita. Non è possibile visualizzare risultati del riquadro Dati dal riquadro Cluster poiché questi cluster abbracciano più documenti/record, rendendo i dati dei risultati non interessanti. Tuttavia, è possibile visualizzare i dati corrispondenti ad una selezione all'interno della finestra di dialogo Definizioni cluster. In base alla selezione, solo il testo corrispondente viene visualizzato nel riquadro Dati. Una volta effettuata una selezione, fare clic sul pulsante **Visualizza &** per compilare il riquadro Dati con i documenti o record che contengono tutti i concetti.

I documenti o record corrispondenti mostrano i concetti evidenziati con un colore per aiutare l'utente a identificarli facilmente nel testo. È inoltre possibile passare il mouse su elementi di colore per visualizzare il concetto da cui è stato estratto e il tipo a cui era assegnato. Il riquadro Dati può contenere più colonne ma la colonna del campo di testo viene sempre visualizzata. Esso reca il nome del campo di testo che è stato utilizzato durante l'estrazione o un nome documento se i dati di testo sono situati in molti file differenti. Sono disponibili altre colonne. Consultare la sezione “Riquadro Dati” a pagina 110 per ulteriori informazioni.

La vista Analisi di collegamento del testo

Nella vista TLA (Text Link Analysis) è possibile creare e analizzare i modelli di analisi di collegamento del testo trovati nei dati di testo. TLA (Text link analysis) è una tecnologia di corrispondenza modello che consente di definire le regole TLA e confrontarle con i concetti estratti e le relazioni trovati nel testo.

I modelli sono molto utili quando si tenta di rilevare le relazioni su un particolare argomento. Alcuni esempi includono l'estrazione di opinioni sui prodotti provenienti da dati di indagine, le relazioni genomiche da documenti di ricerca medica o le relazioni tra le persone o i luoghi dai dati di intelligence.

Una volta estratti alcuni modelli TLA, è possibile esplorarli nel riquadro Dati o Visualizzazione e persino aggiungerli alle categorie nella vista Categorie e concetti. Sono presenti alcune regole di modello TLA definite nel modello di risorsa o nelle librerie che si sta utilizzando per estrarre i risultati TLA. Consultare la sezione Capitolo 19, “Informazioni sulle regole di collegamento del testo”, a pagina 215 per ulteriori informazioni.

Se si sceglie di estrarre i risultati del modello TLA, i risultati vengono visualizzati in questa vista. Se non si è scelto di farlo, è necessario utilizzare il pulsante **Estrai** e scegliere l'opzione per abilitare l'estrazione di modelli .

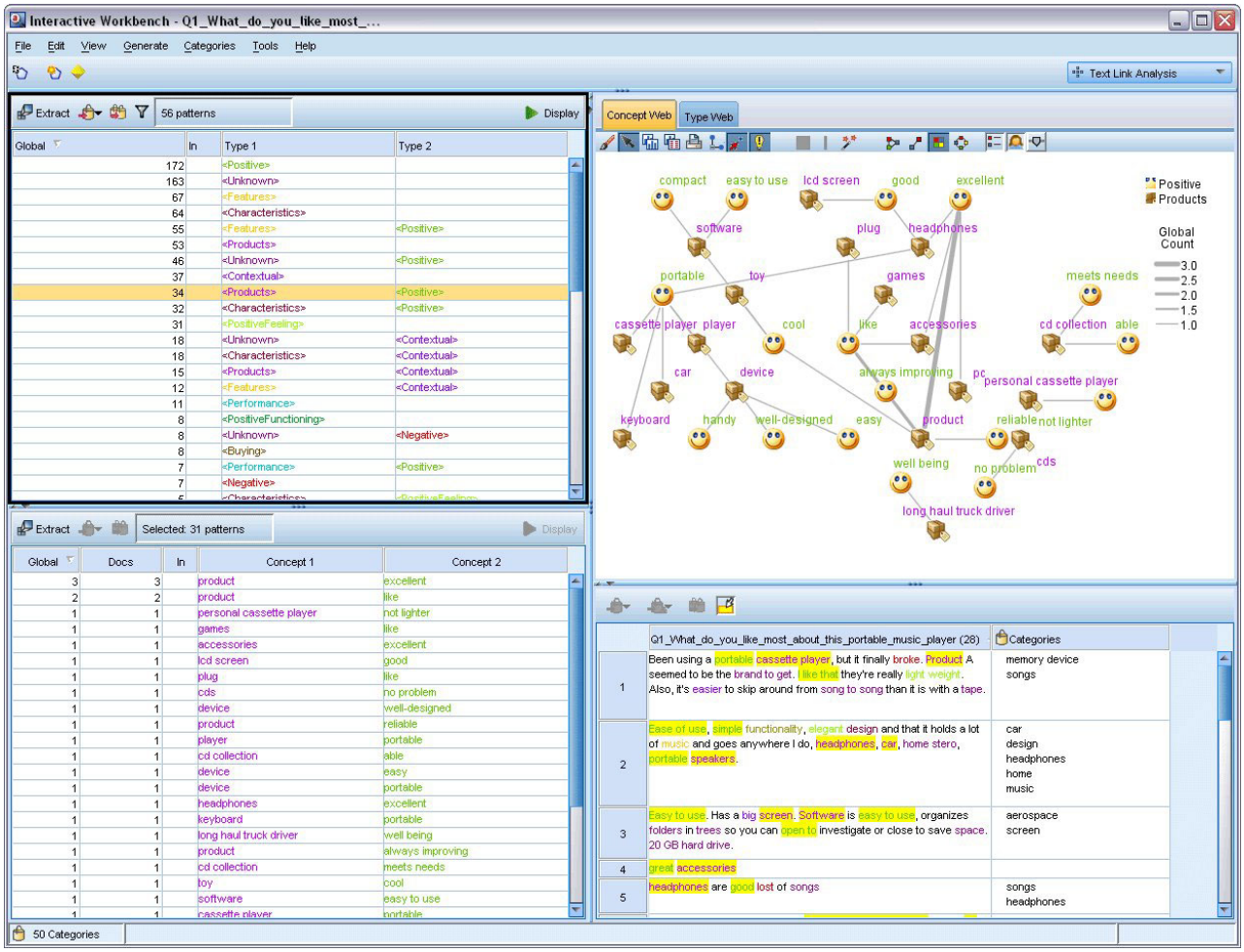


Figura 25. Visualizzazione Analisi di collegamento del testo

La vista TLA (Text Link Analysis) è organizzata in quattro riquadri, ognuno dei quali può essere nascosto o mostrato selezionandone il nome dal menu Visualizza. Consultare la sezione Capitolo 12, “Esplorazione di analisi di collegamento del testo”, a pagina 153 per ulteriori informazioni.

Riquadri dei modelli di tipo e concetto

Ubicato sul lato sinistro, i riquadri dei modelli di tipo e concetto sono due pannelli interconnessi in cui è possibile esaminare e selezionare i risultati del modello TLA. I modelli sono costituiti da una serie di un massimo di sei tipi o sei concetti. Si noti che per il testo giapponese, i modelli sono serie di solo fino a uno o due tipi o concetti. La regola del modello TLA come definito nella risorse linguistiche determina la complessità dei risultati del modello. Consultare la sezione Capitolo 19, “Informazioni sulle regole di collegamento del testo”, a pagina 215 per ulteriori informazioni.

I risultati del modello vengono prima raggruppati al livello di tipo e quindi suddivisi in modelli di concetto. Per questo motivo, vi sono due pannelli di risultati diversi: Modelli di tipo (in alto a sinistra) e Modelli di concetto (in basso a sinistra).

- **Modelli di tipo.** Questo pannello visualizza i modelli estratti costituiti da uno o più tipi correlati corrispondenti ad una regola di modello TLA. I modelli di tipo vengono visualizzati come `<Organizzazione>` + `<Ubicazione>` + `<Positivo>`, che potrebbe fornire un feedback positivo su un'organizzazione in una posizione specifica.

- **Modelli di concetto.** Questo pannello presenta i risultati dei modelli estratti al livello di concetto per tutti i modelli di tipo attualmente selezionati nel riquadro Modelli di tipo di cui sopra. I modelli di concetto seguono una struttura come `hotel + parigi + bellissimo`.

Così come con i risultati di estrazione nella vista Categorie e concetti, è possibile anche qui visualizzare i risultati. Se si desidera visualizzare tutti i miglioramenti che si desidera apportare ai tipi e concetti che costituiscono questi modelli, è possibile farlo nel riquadro Risultati di estrazione nella vista Categorie e concetti o direttamente nell'Editor di risorsa ed estrarre di nuovo i modelli.

Riquadro Visualizzazione

Situato nell'angolo in alto a destra della vista TLA, questo riquadro visualizza un grafico Web dei modelli selezionati come tipo di tipo o concetto. Se non è già visibile, è possibile accedere a questo riquadro dal menu Visualizza (**Visualizza** > **Visualizzazione**). In base a quanto è selezionato negli altri riquadri, è possibile visualizzare le interazioni di corrispondenza documenti/record e i modelli.

I risultati sono presentati in più formati:

- **Grafico di concetto.** Questo grafico presenta tutti i concetti nel modello selezionato. La larghezza della riga e le dimensioni del nodo (se le icone tipo non vengono visualizzate) in un grafico concetto mostrano il numero di ricorrenze globali nella tabella selezionata.
- **Grafico di tipo.** Questo grafico presenta tutti i tipi nel modello selezionato. La larghezza della riga e le dimensioni del nodo (se le icone tipo non vengono visualizzate) in un grafico mostrano il numero di ricorrenze globali nella tabella selezionata. I nodi sono rappresentati da un colore di tipo da un'icona.

Consultare la sezione “Grafici di analisi di collegamento del testo” a pagina 162 per ulteriori informazioni.

Riquadro Dati

Il riquadro Dati si trova nell'angolo in basso a destra. Questo riquadro mostra una tabella contenente documenti o record che corrispondono ad una selezione di un'altra area della vista. In base alla selezione, solo il testo corrispondente viene visualizzato nel riquadro Dati. Una volta effettuata la selezione, fare clic sul pulsante **Visualizza** per compilare il riquadro dei dati con il testo corrispondente.

Se è stata effettuata una selezione in un altro riquadro, i relativi documenti o record mostrano i concetti evidenziati con colore per aiutare l'utente a identificarli facilmente nel testo. È inoltre possibile passare il mouse su elementi di colore per visualizzare un suggerimento che mostra il nome del concetto da cui è stato estratto e il tipo a cui è stato assegnato. Per ulteriori informazioni, consultare la sezione “Riquadro Dati” a pagina 110.

Vista Editor delle risorse

IBM SPSS Modeler Text Analytics cattura rapidamente e in modo accurato i concetti chiave dai dati di testo utilizzando un motore di estrazione molto efficiente. Questo motore si basa essenzialmente su risorse linguistiche per indicare come analizzare e interpretare grandi quantità di dati non strutturati, di testo.

Nella vista Editor risorse è possibile visualizzare e regolare le risorse linguistiche utilizzate per estrarre concetti, raggrupparli in tipi, rilevare modelli nei dati di testo e molto altro ancora. IBM SPSS Modeler Text Analytics fornisce diversi modelli di risorsa preconfigurati. Inoltre, in alcune lingue, è possibile anche utilizzare le risorse in un pacchetto di analisi del testo. Per ulteriori informazioni, consultare la sezione “Uso dei pacchetti di analisi del testo (TAP)” a pagina 141.

Poiché queste risorse potrebbero non essere sempre perfettamente adattate al contesto dei dati, è possibile creare, modificare e gestire le proprie risorse per un determinato contesto o di dominio nel Editor risorse. Per ulteriori informazioni, consultare la sezione Capitolo 16, “Gestione delle librerie”, a pagina 179.

Per semplificare il processo di ottimizzazione delle risorse linguistiche, è possibile eseguire attività comuni di dizionario dalla vista Categorie e concetti tramite il menu di scelta rapida dei riquadri Risultati di estrazione e Dati. Per ulteriori informazioni, consultare la sezione “Perfezionamento dei risultati di estrazione” a pagina 95.

Nota: l'interfaccia per le risorse ottimizzate per il testo giapponese differisce leggermente.

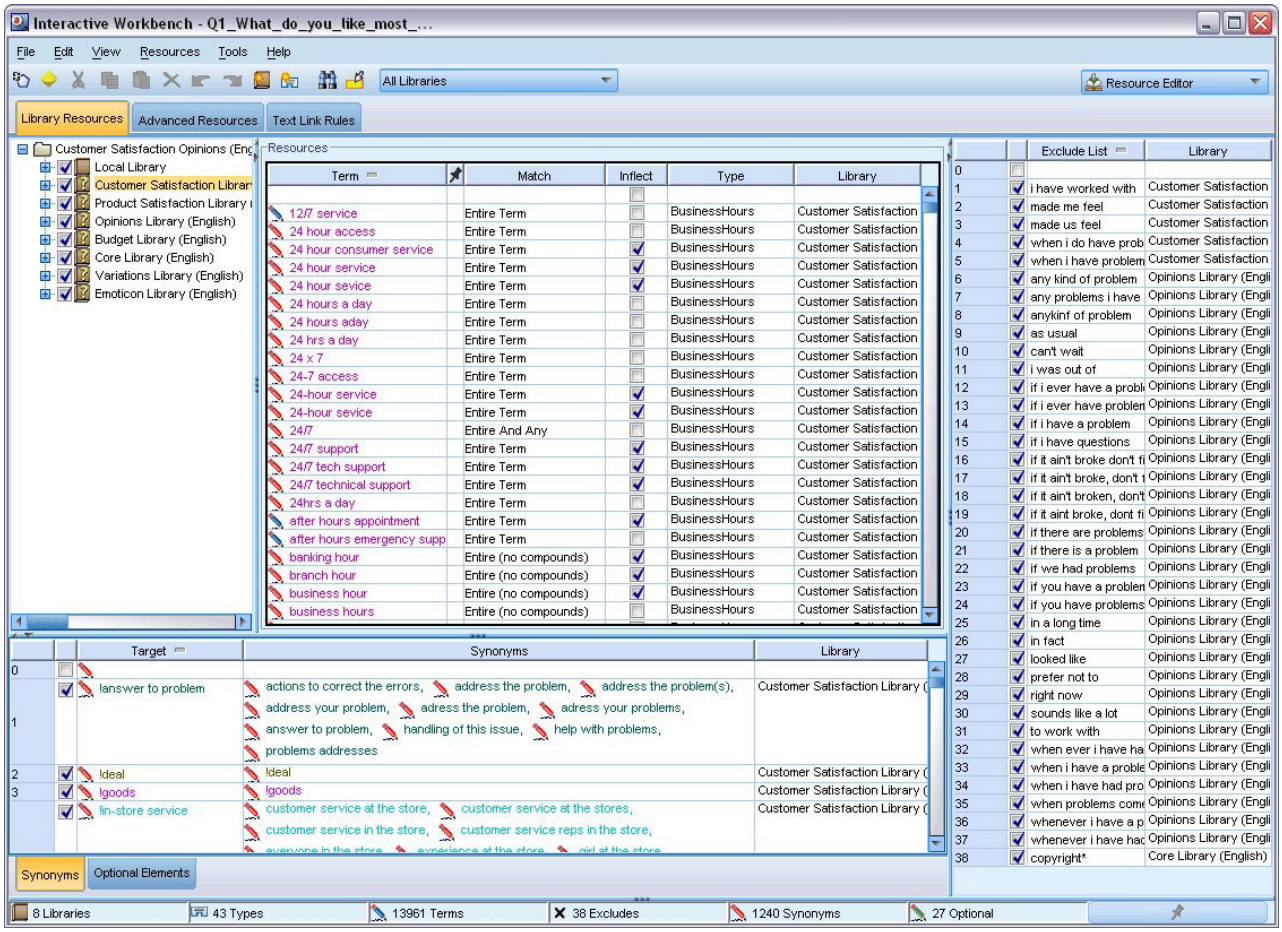


Figura 26. Vista Editor delle risorse

Le operazioni che è possibile eseguire nella vista Editor risorse si articolano intorno alla gestione e ottimizzazione delle risorse linguistiche. Queste risorse sono memorizzate sotto forma di modelli e librerie. La vista Editor risorse è organizzata in quattro parti: riquadro ad albero Libreria, riquadro Dizionario di tipo, riquadro Dizionario di sostituzione e riquadro Dizionario di esclusione.

Nota: per ulteriori informazioni consultare la sezione “L'interfaccia dell'Editor” a pagina 170.

Impostazione delle opzioni

Le opzioni generali per IBM SPSS Modeler Text Analytics possono essere impostate nella casella di dialogo Opzioni. Questa casella di dialogo contiene le seguenti schede:

- **Sessione.** Questa scheda contiene le opzioni generali e i delimitatori.
- **Visualizzazione.** Questa scheda contiene le opzioni per le colorazioni utilizzate nell'interfaccia.
- **Suoni.** Questa scheda contiene le opzioni generali per i segnali sonori.

Per modificare le opzioni

1. Dai menu scegliere **Strumenti > Opzioni**. Si apre la finestra di dialogo Opzioni.
2. Selezionare la scheda contenente le informazioni che si desidera modificare.
3. Modificare una qualsiasi delle opzioni.
4. Fare clic su **OK** per salvare le modifiche.

Opzioni: scheda Sessione

In questa scheda, è possibile definire alcune delle impostazioni di base.

Riquadro Dati e visualizzazione del grafico di categoria. Queste opzioni influenzano il modo in cui i dati vengono visualizzati nel pannello Dati e nel pannello di visualizzazione nella vista categorie e concetti.

- **Limite di visualizzazione per il pannello Dati e il Web di categoria.** Questa opzione imposta il numero massimo di documenti da visualizzare o utilizzare per compilare i riquadri di dati o i grafici e diagrammi nella vista Categorie e concetti.
- **Mostra categorie per i documenti/record in fase di visualizzazione.** Se selezionata, i documenti o record vengono conteggiati ogni volta che si fa clic su Visualizza in modo che tutte le categorie a cui appartengono possano essere visualizzate nella colonna Categorie, nel riquadro Dati così come nei grafici della categoria. In alcuni casi, soprattutto per grandi insiemi, si potrebbe voler disattivare questa opzione in modo che i dati ed i grafici vengano visualizzati molto più velocemente.

Aggiungi a Categoria dal riquadro Dati. Queste opzioni influenzano ciò che viene aggiunto quando i documenti e i record vengono aggiunti dal pannello Dati.

- **Nella vista Categorie e concetti, copia.** Un documento o record aggiunto dal pannello Dati in questa vista sarà copiato su **Concetti** o **Concetti e modelli**.
- **Nella vista Analisi di collegamento del testo, copia.** Un documento o record aggiunto dal pannello Dati in questa vista sarà copiato su **Concetti** o **Concetti e modelli**.

Delimitatore dell'editor delle risorse. Selezionare il carattere da utilizzare come delimitatore quando si immettono gli elementi, come concetti, sinonimi e elementi facoltativi nella vista Editor delle risorse.

Opzioni: scheda Visualizza

In questa scheda, è possibile modificare le opzioni che influenzano l'aspetto generale e aspetto dell'applicazione e i colori utilizzati per distinguere elementi.

Nota: per cambiare l'aspetto del prodotto su un aspetto classico o uno da un rilascio precedente, aprire la finestra di dialogo Opzioni utente nel menu Strumenti nella finestra principale di IBM SPSS Modeler .

Colori personalizzati. Modifica i colori per gli elementi visualizzati. Per ognuno degli elementi nella tabella, è possibile modificare il colore. Per specificare un colore personalizzato, fare clic sull'area colore a destra dell'elemento che si desidera modificare e selezionare un colore dall'elenco a discesa dei colori.

- **Testo non estratto.** I dati di testo non estratti sono ancora visibili nel riquadro Dati.
- **Evidenzia in sfondo.** Colore di sfondo di selezione testo quando si selezionano gli elementi nei riquadri o testo nel riquadro Dati.

- **Sfondo per estrazione necessaria.** Colore di sfondo dei risultati di estrazione, modelli e pannelli Cluster che indica che modifiche sono state apportate alle librerie e un'estrazione è necessaria.
- **Sfondo per sommario categoria.** Colore di sfondo di categoria che viene visualizzata dopo un'operazione.
- **Tipo predefinito.** Colore predefinito per i tipi e i concetti che compaiono nel riquadro Dati e nel riquadro Risultati di estrazione. Questo colore verrà applicato a qualsiasi tipo personalizzato che si desidera creare nell'Editor di risorsa. È possibile sovrascrivere questo colore predefinito per il tipo di dizionari personalizzati modificando le proprietà per questi dizionari di tipo in Editor risorse. Consultare la sezione “Creazione di tipi” a pagina 191 per ulteriori informazioni.
- **Tabella 1 a strisce.** Primo di due colori utilizzati in un modo alternativo nella tabella nella finestra di dialogo Modifica concetti forzati per distinguere ciascuna serie di righe.
- **Tabella 2 a strisce.** Secondo di due colori utilizzati in un modo alternativo nella tabella nella finestra di dialogo Modifica concetti forzati per distinguere ciascuna serie di righe.

Nota: se si fa clic su **Reimposta su valori predefiniti**, tutte le opzioni in questa finestra di dialogo sono reimpostate ai valori che avevano quando è stato prima installato questo prodotto.

Opzioni: scheda Audio

In questa scheda, è possibile modificare le opzioni che interessano l'audio. In Eventi audio, è possibile specificare un suono da utilizzare per notificare all'utente il verificarsi di un evento. Sono disponibili un certo numero di suoni. Utilizzare il pulsante con i puntini (**...**) per ricercare e selezionare un suono. I file *.wav* utilizzati per creare i suoni per IBM SPSS Modeler Text Analytics sono memorizzati nella sottodirectory *media* della directory di installazione. Se non si desidera eseguire suoni, selezionare **Disattiva suoni**. L'impostazione predefinita prevede che i suoni siano disattivati.

Nota: se si fa clic su **Reimposta su valori predefiniti**, tutte le opzioni in questa finestra di dialogo sono reimpostate ai valori che avevano quando è stato prima installato questo prodotto.

Microsoft Internet Explorer Impostazioni per la guida

Impostazioni di Microsoft Internet Explorer

La maggior parte delle funzioni della Guida in linea di questa applicazione utilizza una tecnologia basata su Microsoft Internet Explorer. Alcune versioni di Internet Explorer (compresa quella fornita con Microsoft Windows XP, Service Pack 2) bloccano per default ciò che viene considerato "contenuto attivo" nelle finestre di Internet Explorer sul computer locale. Questa impostazione potrebbe far sì che parte del contenuto della Guida in linea risulti bloccato. Per visualizzare tutto il contenuto della Guida in linea, modificare le impostazioni di default di Internet Explorer.

1. Dai menu di Internet Explorer, selezionare:

   Strumenti > **Opzioni Internet...**
2. Selezionare la scheda **Avanzate**.
3. Scorrere l'elenco verso il basso fino alla sezione **Protezione**.
4. Selezionare **Consenti l'esecuzione di contenuto attivo in file nel computer**.

Generazione dei nugget del modello e nodi di modellazione

Quando ci si trova in una sessione interattiva, è possibile utilizzare il lavoro svolto per generare:

- **Un nodo di modellazione di estrazione testo.** Un nodo di modellazione generato da una sessione workbench interattiva è un nodo di estrazione testo le cui impostazioni e opzioni riflettono quelle memorizzate nella sessione interattiva aperta. Ciò può essere utile quando non si dispone più del nodo di estrazione testo originale o quando si desidera crearne una nuova versione. Consultare la sezione Capitolo 3, “Estrazione per concetti e categorie”, a pagina 19 per ulteriori informazioni.

- **Nugget del modello di categoria.** Un nugget del modello generato da una sessione workbench interattiva è un nugget del modello di categoria. È necessaria almeno una categoria nella vista Categorie e concetti per generare un nugget del modello di categoria. Consultare la sezione "Nugget di estrazione testo: Modello di categoria" a pagina 41 per ulteriori informazioni.

Per generare un nodo di modellazione di estrazione testo

1. Dal menu scegliere **Genera > Genera nodo di modellazione**. Viene aggiunto un nodo di estrazione testo all'area di lavoro utilizzando tutte le impostazioni correnti nella sessione del workbench. Il nodo è denominato dopo il campo di testo.

Per generare un nugget del modello di categoria

1. Dal menu scegliere **Genera > Genera modello**. Un nugget del modello viene generato direttamente sulla tavolozza del modello con il nome predefinito.

Aggiornamento dei nodi di modellazione e salvataggio

Mentre si lavora in una sessione interattiva, si raccomanda di aggiornare il nodo di modellazione di tanto in tanto per salvare le modifiche. È inoltre necessario aggiornare il nodo di modellazione quando si è terminato di lavorare nella sessione del workbench interattivo e si desidera salvare il lavoro. Quando si aggiorna il nodo di modellazione, il contenuto della sessione del workbench viene salvato sul nodo di estrazione testo che ha originato la sessione workbench interattiva. Ciò non fa chiudere la finestra di output.

Importante! Questo aggiornamento non salverà il flusso. Per salvare il flusso, effettuare questa operazione nella finestra principale di IBM SPSS Modeler dopo aver aggiornato il nodo di modellazione.

Per aggiornare un nodo di modellazione

1. Dal menu scegliere **File > Aggiorna nodo di modellazione**. Il nodo di modellazione viene aggiornato con le impostazioni di creazione e di estrazione, insieme ad eventuali opzioni e categorie di cui si dispone.

Chiusura e fine delle sessioni

Quando si è terminato di lavorare nella sessione, è possibile lasciare la sessione in tre modi diversi:

- **Salva.** Questa opzione consente di salvare il proprio lavoro nel nodo di origine per sessioni future, così come pubblicare le librerie da riutilizzare in altre sessioni. Consultare la sezione "Condivisione librerie" a pagina 184 per ulteriori informazioni. Dopo aver salvato la finestra della sessione viene chiusa e la sessione viene cancellata dal gestore di output nella finestra IBM SPSS Modeler .
- **Esci.** Questa opzione consente di annullare qualsiasi lavoro non salvato, chiudere la finestra della sessione ed eliminare la sessione dal gestore di output nella finestra IBM SPSS Modeler . Per liberare memoria, si consiglia di salvare qualsiasi lavoro importante ed uscire dalla sessione.
- **Chiudi.** Questa opzione non salva o eliminare qualsiasi lavoro. Questa opzione chiude la finestra della sessione, ma la sessione continuerà ad essere eseguita. È possibile aprire la finestra della sessione di nuovo selezionando questa sessione nel gestore di output nella finestra IBM SPSS Modeler .

Per chiudere una sessione workbench

1. Dal menu scegliere **File > Chiudi**.

Accesso facilitato mediante tastiera

L'interfaccia workbench interattiva offre le scelte rapide da tastiera per rendere più accessibile la funzionalità del prodotto. Di base, è possibile premere Alt più il tasto pertinente per attivare i menu delle finestre (per esempio, premendo Alt+F si accede al menu File) oppure premere il tasto Tab per spostarsi fra i controlli delle finestre di dialogo. In questa sezione verranno descritte le scelte rapide da tastiera per la navigazione alternativa. Vi sono altre scelte rapide da tastiera diverse per l'interfaccia di IBM SPSS Modeler.

Tabella 13. Tasti di scelta rapida generici

Tasto di scelta rapida	Funzione
Ctrl+1	Visualizza la prima scheda in un pannello con schede.
Ctrl+2	Visualizza la seconda scheda in un pannello con schede.
Ctrl+A	Selezionare tutti gli elementi per il riquadro che ha lo stato attivo.
Ctrl+C	Copia il testo selezionato negli Appunti.
Ctrl+E	Avvia l'estrazione nelle vista Categorie e concetti e Analisi di collegamento del testo.
Ctrl+F	Visualizza la barra degli strumenti Trova in Editor risorse/Editor di modelli, se non è già visibile e la inserisce in stato attivo.
Ctrl+I	Nella vista Categorie e concetti, avvia la finestra di dialogo Definizioni di categoria per la categoria selezionata. Nella vista Cluster, avvia la finestra di dialogo Definizioni di cluster per il cluster selezionato.
Ctrl+R	Apre la finestra di dialogo Aggiungi Termini in Editor risorse/Editor di modelli.
Ctrl+T	Apre la finestra di dialogo Proprietà tipo per creare un nuovo tipo in Editor risorse/Editor di modelli.
Ctrl+V	Incolla il contenuto degli Appunti.
Ctrl+X	Taglia gli elementi selezionati da Editor risorse/Editor di modelli.
Ctrl+Y	Ripete l'ultima azione nella vista.
Ctrl+Z	Annulla l'ultima azione nella vista.
F1	Visualizza la guida o, in una finestra di dialogo, visualizza la guida di contesto per un elemento.
F2	Attiva / disattiva la modalità di modifica nelle celle della tabella.
F6	Sposta lo stato attivo tra i principali pannelli nella vista attiva.
F8	Sposta lo stato attivo nel riquadro per il ridimensionamento delle barre di suddivisione.
F10	Espande il menu File principale.
freccia su, freccia giù	Ridimensiona il pannello in verticale quando la barra di suddivisione è selezionata.
freccia sinistra, freccia destra	Ridimensiona il pannello in orizzontale quando la barra di suddivisione è selezionata.
Home, Fine	Ridimensiona i pannelli sulla dimensione minima o massima quando la barra di suddivisione è selezionata.
Tabulatore	Sposta in avanti attraverso gli elementi nella finestra, riquadro o finestra di dialogo.
Maiusc+F10	Visualizza il menu di scelta rapida per un elemento.
Maiusc+Tab	Sposta indietro attraverso gli elementi nella finestra o finestra di dialogo.
Maiusc+freccia	Seleziona i caratteri nel campo di modifica in modalità di modifica (F2).
Ctrl+Tab	Sposta lo stato attivo in avanti alla successiva area principale nella finestra.
Maiusc+Ctrl+Tab	Sposta lo stato attivo indietro alla precedente area principale nella finestra.

Tasti di scelta rapida per finestre di dialogo

Alcuni tasti di scelta rapida e tasti per il lettore di schermo sono utili quando si lavora con finestre di dialogo. All'immissione di una finestra di dialogo, potrebbe essere necessario premere il tasto Tab per spostare lo stato sul primo controllo e avviare il lettore di schermo. Nella tabella che segue è riportato un elenco completo di tasti di scelta rapida per tastiera e per lettore schermo.

Tabella 14. Tasti di scelta rapida per finestra di dialogo

Tasto di scelta rapida	**Funzione**
Tabulatore	Sposta in avanti attraverso gli elementi nella finestra o finestra di dialogo.
Ctrl+Tab	Passaggio da una casella di testo all'elemento successivo.
Maiusc+Tab	Sposta indietro attraverso gli elementi nella finestra o finestra di dialogo.
Maius+Ctrl+Tab	Passaggio da una casella di testo all'elemento precedente.
barra spaziatrice	Seleziona il controllo o il pulsante attivo.
Esc	Annulla le modifiche e chiude la finestra di dialogo.
Invio	Convalida le modifiche e chiude la finestra di dialogo (equivalente al pulsante OK). Se si sta in una casella di testo, è necessario prima premere Ctrl+Tab per uscire dalla casella di testo.

Capitolo 9. Estrazione di concetti e tipi

Ogni volta che si esegue un flusso che avvia il workbench interattivo, viene eseguita automaticamente un'estrazione sui dati di testo nel flusso. Il risultato finale di questa estrazione è una serie di concetti, tipi e, nel caso in cui esistono i modelli TLA nelle risorse linguistiche, modelli. È possibile visualizzare e gestire i concetti e i tipi nel pannello Risultati di estrazione. Consultare la sezione “Come funziona il processo di estrazione” a pagina 5 per ulteriori informazioni.

Se si desidera ottimizzare i risultati di estrazione, è possibile modificare le risorse linguistiche ed estrarre di nuovo. Consultare la sezione “Perfezionamento dei risultati di estrazione” a pagina 95 per ulteriori informazioni. Il processo di estrazione si basa sulle risorse e su tutti i parametri nella finestra di dialogo Estrai per indicare come estrarre e organizzare i risultati. È possibile utilizzare i risultati di estrazione per definire la maggior parte, se non tutte le definizioni di categoria.

Risultati estrazione: concetti e tipi

Durante il processo di estrazione, tutti i dati di testo vengono sottoposti a scansione e i concetti sono identificati, estratti ed assegnati a tipi. Una volta completata l'estrazione, i risultati vengono visualizzati nel riquadro Risultati di estrazione che si trova nell'angolo in basso a sinistra della vista Categorie e concetti. La prima volta che si avvia la sessione, il modello di risorsa linguistica selezionata nel nodo viene utilizzata per estrarre e organizzare tali concetti e tipi.

I concetti, i tipi e modelli TLA che vengono estratti sono collettivamente indicati come **risultati di estrazione** e possono servire come descrittori o blocchi di creazione, per le categorie. È anche possibile utilizzare concetti, tipi e modelli nelle proprie regole di categoria. Inoltre, per costruire le categorie, le tecniche automatiche utilizzano concetti e tipi.

L' di estrazione testo è un processo iterativo in cui i risultati di estrazione vengono esaminati in base al contesto dei dati di testo, ottimizzato per produrre risultati nuovi e poi rivalutato. Dopo l'estrazione, è necessario esaminare i risultati e apportare le modifiche che si ritengono necessarie modificando le risorse linguistiche. È possibile ottimizzare le risorse in parte direttamente dal riquadro Risultati di estrazione, dal riquadro dei dati, dalla finestra di dialogo Definizioni delle categorie o Definizioni del cluster. Per ulteriori informazioni, consultare la sezione “Perfezionamento dei risultati di estrazione” a pagina 95. Questa operazione è possibile anche direttamente nella vista Editor risorse. Per ulteriori informazioni, consultare la sezione “Vista Editor delle risorse” a pagina 80.

Dopo l'ottimizzazione, è possibile poi riestrarre per visualizzare i nuovi risultati. Ottimizzando i risultati di estrazione dall'inizio, è possibile assicurarsi che ogni volta che viene rieseguita l'estrazione, si otterranno risultati identici nelle definizioni di categoria, perfettamente adattati al contesto dei dati. In questo modo, documenti/record verranno assegnati alle proprie definizioni di categoria in modo più accurato e ripetibile.

Concetti

Durante il processo di estrazione, i dati di testo vengono sottoposti a scansione e analizzati per identificare le singole parole interessanti o pertinenti (come ad esempio `elezione` o `pace`) e frasi di parole (ad esempio, `elezioni presidenziali`, `elezione del presidente` o `trattative di pace`) nel testo. Queste parole o frasi sono indicate collettivamente come *termini*. Utilizzando le risorse linguistiche, i termini rilevanti vengono estratti e quindi termini simili vengono raggruppati sotto un termine principale, denominato **concetto**.

È possibile visualizzare la serie di termini sottostante per un concetto, passando il puntatore del mouse sul nome concetto. In questo modo viene visualizzato un suggerimento che mostra il nome concetto e

diverse righe di termini che sono raggruppate in tale concetto. Tali termini sottostanti comprendono i sinonimi definiti nelle risorse linguistiche (indipendentemente dal fatto che sono stati rilevati nel testo o meno) così come gli eventuali termini estratti singolari/plurali, termini permutati, termini da raggruppamenti casuali e così via. È possibile copiare tali termini o consultare la serie completa di termini sottostanti facendo clic con il tasto destro sul nome del concetto e scegliendo l'opzione di menu di scelta rapida.

Per impostazione predefinita, i concetti sono visualizzati in minuscolo e ordinati in ordine decrescente in base al conteggio del documento (Doc. colonna) . Quando vengono estratti i concetti, vengono assegnati ad un tipo per contribuire a raggruppare concetti simili. Essi sono codificati a colori in base a questo tipo. I colori vengono definiti nelle proprietà di tipo all'interno di Editor risorse. Per ulteriori informazioni, consultare la sezione "Dizionari di tipo" a pagina 189.

Ogni volta che un concetto, tipo o modello viene utilizzato in una definizione di categoria, viene visualizzata un'icona **nella** colonna .

Tipi

I **tipi** sono raggruppamenti semantici di concetti. Quando vengono estratti i concetti, vengono assegnati ad un tipo per contribuire a raggruppare concetti simili. Con IBM SPSS Modeler Text Analytics sono forniti diversi tipi incorporati, come `<Ubicazione>`, `<Organizzazione>`, `<Persona>`, `<Positivo>`, `<Negativo>` e così via. Ad esempio il tipo `<Ubicazione>` raggruppa le parole chiave e i posti. Questo tipo potrebbe essere assegnato a concetti come `chicago`, `paris` e `tokyo`. Per la maggior parte delle lingue, concetti che non sono presenti nei dizionari ma vengono estratti dal testo vengono automaticamente inserite come `<Sconosciuto>` Per ulteriori informazioni consultare la sezione "Tipi incorporati" a pagina 190.

Quando si seleziona la vista Tipo, i tipi estratti vengono visualizzati per impostazione predefinita in ordine decrescente per frequenza globale. Notare che i tipi sono codificati con colore per aiutare a distinguerli. I colori sono parte delle proprietà di tipo. Per ulteriori informazioni, consultare la sezione "Creazione di tipi" a pagina 191. È anche possibile creare dei tipi propri.

Modelli

I modelli possono essere estratti anche dai dati di testo. Tuttavia, è necessario predisporre una libreria che contenga le regole di modello TLA (Text Link Analysis) in Editor risorse. È inoltre necessario scegliere di estrarre questi modelli nell'impostazione nodo di IBM SPSS Modeler Text Analytics o nella finestra di dialogo Estrai utilizzando l'opzione **Abilita estrazione di modello di analisi di collegamento del testo** . Per ulteriori informazioni, consultare la sezione Capitolo 12, "Esplorazione di analisi di collegamento del testo", a pagina 153.

Estrazione di dati

Ogni volta che un'estrazione è necessaria, il riquadro Risultati di estrazione diventa di colore giallo e viene visualizzato il messaggio **Premi pulsante Estrai per estrarre concetti** al di sotto della barra degli strumenti in questo riquadro.

Potrebbe essere necessario estrarre se non si dispone ancora di risultati di estrazione, sono state apportate delle modifiche alle risorse linguistiche e bisogna aggiornare i risultati di estrazione o si è riaperto una sessione in cui non sono stati salvati i risultati di estrazione (**Strumenti** > **Opzioni**).

Nota: Se si modifica il nodo di origine per il flusso dopo che i risultati di estrazione sono stati memorizzati con l'opzione **Utilizza lavoro di sessione...** , è necessario eseguire una nuova estrazione una volta che la sessione di workbench interattivo viene avviata, se si desidera ottenere risultati di estrazione aggiornati.

Quando si esegue un'estrazione, viene visualizzato un indicatore di avanzamento per fornire un feedback sullo stato dell'estrazione. Durante questo periodo, il motore di estrazione legge attraverso tutti i dati di testo e individua i termini e i modelli, li estrae e li assegna ad un tipo. Quindi, il motore raggruppa i termini sinonimi sotto un termine principale, chiamato concetto. Quando il processo è completo, i concetti risultanti, tipi e modelli vengono visualizzati nel riquadro Risultati di estrazione.

Il processo di estrazione ha come risultato una serie di concetti e di tipi, come i modelli TLA, se abilitati. È possibile visualizzare e gestire questi concetti e tipi nel riquadro Risultati di estrazione nella vista Categorie e concetti. Se sono stati estratti i modelli TLA, è possibile visualizzarli nella vista Analisi di collegamento del testo .

Nota: si tratta di una relazione tra la dimensione del proprio dataset e il tempo impiegato per completare il processo di estrazione. È possibile sempre considerare l'inserimento di un nodo Campione upstream o ottimizzare la configurazione della macchina.

Per il data mining

1. Dai menu scegliere **Strumenti** > **Estrai**. In alternativa, fare clic sul pulsante della barra degli strumenti **Estrai**.
2. Se si sceglie di visualizzare sempre la finestra di dialogo Impostazioni di estrazione, essa viene visualizzata per apportare tutte le modifiche. Consultare questa sezione per ulteriori descrittori di ciascuna delle impostazioni.
3. Fare clic su **Estrai** per avviare il processo di estrazione. Una volta avviata l'estrazione, viene visualizzata la finestra di dialogo di avanzamento. Dopo l'estrazione, i risultati vengono visualizzati nel riquadro Risultati di estrazione. Per impostazione predefinita, i concetti sono visualizzati in minuscolo e ordinati in ordine decrescente in base al conteggio del documento (Doc. colonna) .

È possibile visualizzare i risultati utilizzando le opzioni della barra degli strumenti per ordinare i risultati in modo diverso, per filtrare i risultati o per passare a una vista differente (concetti o tipi). È inoltre possibile restringere i propri risultati di estrazione utilizzando le risorse linguistiche. Per ulteriori informazioni, consultare la sezione “Perfezionamento dei risultati di estrazione” a pagina 95.

Per testo olandese, inglese, francese, tedesco, italiano, portoghese e spagnolo

La finestra di dialogo Impostazioni di estrazione contiene alcune opzioni di estrazione di base.

Attiva estrazione del modello di analisi di collegamento del testo . Specifica che si desidera estrarre i modelli TLA dai dati di testo. Si suppone inoltre che in una delle librerie nell'Editor di risorsa siano presenti le regole del modello TLA. Questa opzione può aumentare in modo significativo il tempo di estrazione. Consultare la sezione Capitolo 12, “Esplorazione di analisi di collegamento del testo”, a pagina 153 per ulteriori informazioni.

Correggi errori di punteggiatura. Questa opzione normalizza temporaneamente testo contenente errori di punteggiatura (ad esempio, l'uso improprio) durante l'estrazione per migliorare la possibilità di estrazione dei concetti. Questa opzione risulta estremamente utile quando il testo è breve e di scarsa qualità (come, ad esempio, le risposte del sondaggio aperto, e-mail e i dati CRM), o quando il testo contiene molte abbreviazioni.

Correzione di errori ortografici per un limite minimo di caratteri root [n]. Questa opzione applica una tecnica di raggruppamento che consente di raggruppare parole errate o parole simili in un unico concetto. L'algoritmo di raggruppamento toglie temporaneamente tutte le vocali (tranne la prima) e stacca consonanti doppie/triple da parole estratte e poi le confronta per vedere se sono uguali in modo che `modeli` e `modelli` verrebbero raggruppate insieme. Tuttavia, se ogni termine viene assegnato ad un tipo differente, escluso il tipo `<Sconosciuto>`, la tecnica di raggruppamento confuso non verrà applicata.

È possibile inoltre definire il numero minimo di caratteri *radice* richiesti prima di utilizzare il raggruppamento. Il numero di caratteri radice in un termine è calcolato sommando tutti i caratteri e sottraendo quelli che formano i suffissi di desinenza e, nel caso di parole composte, i determinativi e le preposizioni. Ad esempio, il termine `esercizi` potrebbe essere conteggiato come 9 caratteri radice nel formato "esercizio", poiché la lettera *i* alla fine della parola è un flesso (forma plurale). Allo stesso modo, `succo di mela` conta 11 caratteri radice ("succo di mela") e `fabbrica di automobili` conta 20 caratteri radice. Questo metodo di conteggio viene utilizzato solo per verificare se il raggruppamento deve essere applicato ma non influenza il modo in cui le parole sono corrispondenti.

Nota: se si riscontra che determinate parole vengono successivamente raggruppate in modo non corretto, è possibile escludere da questa tecnica le coppie di parole esplicitamente dichiarate nella sezione **Raggruppamento confuso: eccezioni** nella scheda Avanzate. Per ulteriori informazioni, consultare la sezione "Raggruppamento confuso" a pagina 206.

Estrazione termini univoci. Questa opzione estrae parole singole (termini univoci) purché il termine non è già parte di una parola composta e se è un nome o una parte del discorso non riconosciuta.

Estrazione entità non linguistiche. Questa opzione estrae entità non linguistiche, ad esempio numeri di telefono, numeri di codice fiscale, orari, date, valute, cifre, percentuali, indirizzi e-mail e indirizzi HTTP. È possibile includere o escludere alcuni tipi di entità non linguistiche nella sezione **Entità non linguistiche: configurazione** della scheda Avanzate. Disabilitando qualsiasi entità non necessaria, il motore di estrazione non spreca tempo di elaborazione. Per ulteriori informazioni, consultare la sezione "Configurazione" a pagina 210.

Algoritmo maiuscolo. Questa opzione estrae i termini semplici e composti che non si trovano nei dizionari incorporati, purché la prima lettera del termine sia in maiuscolo. Questa opzione fornisce un modo efficace di estrarre i sostantivi più appropriati.

Raggruppare insieme nomi di persona parziali e completi quando possibile. Questa opzione raggruppa i nomi che vengono visualizzati insieme diversamente nel testo. Questa funzione è utile poiché i nomi vengono spesso definiti nel loro formato completo all'inizio del testo e poi solo da una versione abbreviata. Questa opzione tenta la corrispondenza con qualsiasi termine univoco con il tipo `<Sconosciuto>` per l'ultima parola di tutti i termini composti immessi come `<Persona>`. Ad esempio, se viene rilevato *rossi* e inizialmente immesso come `<Sconosciuto>`, il motore controlla l'estrazione per vedere se tutti i termini composti nel tipo `<Persona>` includono *rossi* come ultima parola, per esempio *giovanni rossi*. Questa opzione non si applica ai nomi di persona poiché la maggior parte non sono mai estratti come termini univoci.

Numero massimo di permutazioni di parole non funzionali. Questa opzione specifica il numero massimo di parole non funzionali che possono essere presenti quando si applica la tecnica di permutazione. Questa tecnica di permutazione raggruppa frasi simili che differiscono tra loro solo per parole non funzionali (ad esempio, di e il) contenute, indipendentemente dall'inflessione. Ad esempio, si imposta questo valore su massimo due parole e vengono estratte `funzionari` e `funzionari dell'azienda`. In questo caso, entrambi i termini estratti verrebbero raggruppati insieme nell'elenco dei concetti poiché entrambi i termini vengono considerati uguali quando `dell'` viene ignorato.

Opzione di indice per la mappa dei concetti Specifica che si desidera generare l'indice della mappa in fase di estrazione in modo che le associazioni di concetto possano essere elaborate più rapidamente in seguito. Per modificare le impostazioni dell'indice, fare clic su **Impostazioni**. Consultare la sezione "Creazione di indici di mappa di concetti" a pagina 95 per ulteriori informazioni.

Mostra sempre questa finestra di dialogo prima di avviare un'estrazione. Specifica se si desidera visualizzare la finestra di dialogo Impostazioni di estrazione ogni volta che si estrae, se non si desidera di non visualizzarla mai, a meno che non si desidera andare al menu Strumenti o se si desidera che venga richiesto ogni volta che si esegue un'estrazione se si desidera modificare eventuali impostazioni di estrazione.

Per il testo giapponese

La finestra di dialogo Impostazioni di estrazione contiene alcune opzioni di estrazione di base per la lingua di testo giapponese. Per impostazione predefinita, le impostazioni selezionate nella finestra di dialogo sono le stesse di quelle selezionate nella scheda Livello avanzato del nodo di modellazione dell'estrazione testo. Per poter gestire il testo in giapponese, è necessario utilizzare il testo come input, oltre a scegliere un modello in lingua giapponese o un pacchetto di analisi del testo nella scheda Modello del nodo di estrazione testo. Consultare la sezione “Copia risorse da modelli e TAP” a pagina 27 per ulteriori informazioni.

Analisi secondaria. Quando viene lanciata un'estrazione, ha luogo l'estrazione di parole chiave di base utilizzando la serie predefinita di tipi. Tuttavia, quando si seleziona un analizzatore secondario, è possibile ottenere molti più concetti e più arricchiti poiché l'estrazione comprenderà particelle e verbi ausiliari come parte del concetto. Nel caso di analisi di opinione, viene incluso anche un numero elevato di tipi aggiuntivi. Inoltre, la scelta di un'analisi secondaria consente di generare anche dei risultati di analisi di collegamento del testo.

Nota: Quando un analizzatore secondario viene richiamato, il processo di estrazione richiede più tempo per completare il processo.

- **Analisi di dipendenza**. La scelta di questa opzione produce particelle estese per i concetti di estrazione dal tipo di base ed estrazione di parola chiave. È anche possibile ottenere risultati di modello più arricchiti dall'analisi TLA di dipendenza.
- **Analisi di opinione**. La scelta di questo tipo di analisi genera ulteriori concetti estratti e, dove applicabile, l'estrazione dei risultati del modello TLA. Oltre ai tipi di base, è possibile anche usufruire di più di 80 tipi di opinioni:. Tali tipi vengono utilizzati per rilevare i concetti e i modelli nel testo mediante l'espressione delle emozioni, opinioni e opinioni. Sono disponibili tre opzioni che regolano il focus per l'analisi di opinione: **Tutti i opinioni**, **Solo opinione rappresentativo** e **Solo conclusioni**.
- **Nessun analizzatore secondario**. Questa opzione disattiva tutti gli analizzatori secondari. Questa opzione non può essere selezionata se è stata selezionata l'opzione **Abilita analisi di collegamento del testo** poiché è sempre richiesto un analizzatore secondario per ottenere risultati di TLA.

Attiva estrazione del modello di analisi di collegamento del testo . Specifica che si desidera estrarre i modelli TLA dai dati di testo. Si suppone inoltre che in una delle librerie nell'Editor di risorsa siano presenti le regole del modello TLA. Questa opzione può aumentare in modo significativo il tempo di estrazione. Inoltre, deve essere sempre selezionato un'analizzatore secondario per estrarre i risultati del modello TLA. Consultare la sezione Capitolo 12, “Esplorazione di analisi di collegamento del testo”, a pagina 153 per ulteriori informazioni.

Filtro dei risultati di estrazione

Quando si lavora con insiemi di dati di grandi dimensioni, il processo di estrazione potrebbe produrre milioni di risultati. Per molti utenti, tale quantità può rendere più difficile esaminare i risultati effettivi. Pertanto, per ingrandire quelli che sono più interessanti, è possibile filtrare tali risultati tramite la finestra di dialogo Filtro disponibile nel riquadro Risultati di estrazione.

Tenere presente che tutte le impostazioni di questa finestra di dialogo Filtro vengono utilizzate insieme per filtrare i risultati di estrazione che sono disponibili per le categorie.

Filtro per frequenza. È possibile filtrare in modo da visualizzare solo i risultati con un certo valore di frequenza globale o del documento.

- La **frequenza globale** è il numero totale di volte in cui un concetto viene visualizzato nell'intera serie di documenti o record e viene visualizzata nella colonna **Globale**.
- La **frequenza del documento** è il numero totale di documenti o record in cui un concetto viene visualizzato e viene mostrata nella colonna **Documenti**.

Ad esempio, se il concetto `nato` è apparso 800 volte in 500 record, questo concetto avrebbe una frequenza globale di 800 e una frequenza di documento di 500.

E per tipo. È possibile filtrare per visualizzare solo i risultati appartenenti a determinati tipi. È possibile scegliere tutti i tipi o solo tipi specifici.

E per testo corrispondente. È anche possibile filtrare per visualizzare solo i risultati che corrispondono alla regola definita qui. Immettere la serie di caratteri corrispondenti nel campo **Testo di corrispondenza** e quindi selezionare la condizione in cui applicare la corrispondenza.

Tabella 15. Condizioni del testo di corrispondenza

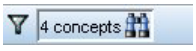
Condizione	Descrizione
Contiene	Il testo è esatto se la stringa ricorre in qualsiasi punto. (Scelta predefinita)
Inizia con	Il testo è esatto solo se il concetto o il tipo inizia con il testo specificato.
Termina con	Il testo è esatto solo se il concetto o il tipo termina con il testo specificato.
Corrispondenza esatta	L'intera stringa deve corrispondere al nome tipo o nome concetto.

E per punteggio. È anche possibile filtrare per visualizzare solo i concetti più in alto in base alla frequenza globale (**Globale**) o frequenza documento (**Docs**) in ordine crescente o decrescente.

Risultati visualizzati nel riquadro Risultati di estrazione

Di seguito sono riportati alcuni esempi di come i risultati potrebbero essere visualizzati, in inglese, nella barra degli strumenti del riquadro Risultati di estrazione in base ai filtri.

Tabella 16. Esempi di feedback di filtro

Feedback di filtro	Descrizione
	La barra degli strumenti mostra il numero di risultati. Poiché non c'è filtro di corrispondenza di testo e il valore massimo non è stato raggiunto, non vengono visualizzate icone aggiuntive.
	La barra degli strumenti mostra che i risultati sono stati limitati al massimo specificato nel filtro, che in questo caso era 300. Se è presente un'icona viola significa che è stato raggiunto il numero massimo di concetti. Passare con il mouse sull'icona per ulteriori informazioni.
	La barra degli strumenti mostra che i risultati sono stati limitati utilizzando un filtro di testo. Viene visualizzata un'icona con lente di ingrandimento.

Per filtrare i risultati

1. Dai menu scegliere **Strumenti > Filtro**. Si apre la finestra di dialogo Filtro.
2. Selezionare e perfezionare i filtri che si desidera utilizzare.
3. Fare clic su **OK** per applicare i filtri e vedere i nuovi risultati nel riquadro dei risultati di estrazione.

Esplorazione di mappe di concetto

È possibile creare una mappa per esplorare come i concetti sono correlati. Selezionando un concetto unico e facendo clic su **Mappa**, una finestra di mappa concetto viene aperta in modo da poter esplorare la serie di concetti che sono correlati al concetto selezionato. È possibile filtrare i concetti che sono visualizzati modificando le impostazioni come ad esempio quali tipi includere, quale tipo di relazione ricercare e così via.

Importante! Prima di poter creare una mappa, deve essere generato un indice. L'operazione può richiedere qualche minuto. Tuttavia, dopo aver generato l'indice, non è necessario generarlo nuovamente fino a quando non si estrae di nuovo. Se si vuole che l'indice venga generato automaticamente ad ogni estrazione, selezionare quella opzione nelle impostazioni di estrazione. Consultare la sezione “Estrazione di dati” a pagina 88 per ulteriori informazioni.

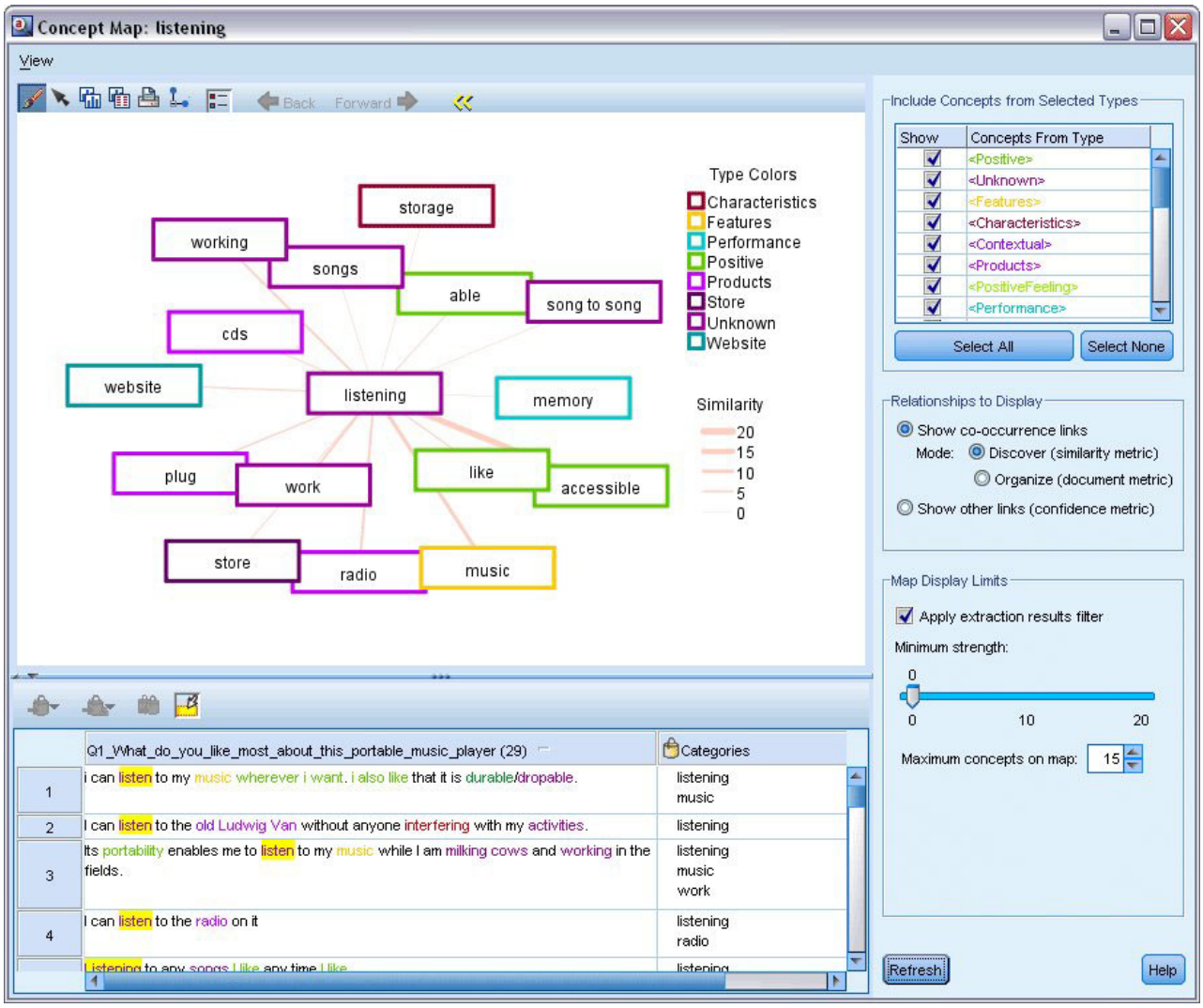


Figura 27. Una mappa di concetto per il concetto selezionato

Per visualizzare una mappa di concetto

1. Nel riquadro Risultati di estrazione, selezionare un concetto singolo.
2. Nella barra degli strumenti di questo pannello, fare clic sul pulsante **Mappa**. Se l'indice della mappa è stato già generato la mappa di concetto viene aperta in una finestra di dialogo distinta. Se l'indice di mappa non è stato generato o è scaduto, l'indice deve essere ricreato. Questo processo può richiedere qualche minuto.
3. Fare clic intorno alla mappa da esplorare. Se si fa doppio clic su un concetto collegato, la mappa verrà ridisegnata e mostrerà i concetti legati per il concetto su cui si è fatto doppio clic.
4. La barra degli strumenti superiore offre alcuni strumenti di mappa di base, ad esempio il passaggio a una mappa precedente, il filtro di collegamenti in base ai rapporti di forza, e inoltre l'apertura della finestra di dialogo del filtro per controllare i tipi di concetti che compaiono come i tipi di relazioni da

rappresentare. La seconda riga di barra degli strumenti contiene strumenti di modifica del grafico. Consultare la sezione "Uso delle barre degli strumenti e tavolozze dei grafici" a pagina 163 per ulteriori informazioni.

5. Se non si è soddisfatti con i tipi di collegamenti trovati, esaminare le impostazioni per questa mappa mostrati sul lato destro della mappa.

Impostazioni mappa: Includi concetti da tipi selezionati

Solo quei concetti appartenenti al tipi selezionati nella tabella vengono visualizzati nella mappa. Per nascondere i concetti di un certo tipo, deselezionare tale tipo nella tabella.

Impostazioni mappa: relazioni da visualizzare

Mostra collegamenti di ricorrenza. Se si desidera mostrare i collegamenti di ricorrenza, selezionare la modalità. La modalità influenza il modo in cui è stata calcolata la forza di collegamento.

- *Scopri (metrica di similitudine)*. Con questa metrica, la forza del collegamento viene calcolata utilizzando un calcolo più complesso che tiene conto di quanto spesso due concetti vengono visualizzati distanti oltre a quanto spesso vengono visualizzati insieme. Un valore elevato indica che una coppia di concetti tende ad apparire più spesso insieme che da soli. Con la formula seguente, tutti i valori del punto fluttuante vengono convertiti in numeri interi.

$$\text{similarity coefficient} = \frac{(C_{IJ})^2}{(C_I \times C_J)}$$

Figura 28. Formula del coefficiente di similitudine

  In questa formula C_I è il numero di documenti o record in cui ricorre il concetto `I`.

  C_J è il numero di documenti o record in cui ricorre il concetto `J`.

  C_{IJ} è il numero di documenti o record in cui la coppia concetto `I` e `J` ricorre nella serie di documenti.

- *Organizza (metrica del documento)*. La forza dei collegamenti con questa metrica è stabilita dal conteggio non ordinato di ricorrenza. In generale, più frequenti sono i due concetti più è probabile che ricorrano a volte insieme. Un valore elevato indica che una coppia di concetti vengono spesso visualizzati insieme.

Mostra altri collegamenti (metrica di affidabilità). È possibile scegliere altri collegamenti da visualizzare; questi possono essere semantici, di derivazione (morfologica) o inclusione (della sintassi) e sono correlati a quanti passi rimossi il concetto è dal concetto al quale è collegato. Questi consentono di ottimizzare le risorse, in particolare la sinonimia o l'ambiguità. Per brevi descrizioni di ciascuna di queste tecniche di raggruppamento, consultare "Impostazioni linguistiche avanzate" a pagina 114

Nota: tenere presente che se queste non sono state selezionate quando l'indice è stato creato o se non sono state trovate le relazioni, nessuna verrà visualizzato. Consultare la sezione "Creazione di indici di mappa di concetti" a pagina 95 per ulteriori informazioni.

Impostazioni mappa: limiti di visualizzazione di mappa

Applica filtro dei risultati di estrazione. Se non si desidera utilizzare tutti i concetti, è possibile utilizzare il filtro nel riquadro dei risultati di estrazione per limitare ciò che viene visualizzato. Quindi, selezionare questa opzione e IBM SPSS Modeler Text Analytics ricercherà concetti correlati utilizzando questa serie filtrata. Consultare la sezione "Filtro dei risultati di estrazione" a pagina 91 per ulteriori informazioni.

Forza minima. Impostare il livello minimo di collegamento. Tutti i concetti correlati con una forza di relazione inferiore a questo limite verranno nascosti dalla mappa.

Numero massimo di concetti sulla mappa. Specificare il numero massimo di relazioni da mostrare sulla mappa.

Creazione di indici di mappa di concetti

Prima di poter creare una mappa, deve essere generato un indice delle relazioni dei concetti. Quando si crea una mappa di concetti, IBM SPSS Modeler Text Analytics fa riferimento a questo indice. È possibile scegliere quali relazioni indicizzare selezionando le tecniche in questa finestra di dialogo.

Tecniche di raggruppamento. Scegliere uno o più tecniche. Per descrizioni brevi di ciascuna di queste tecniche, consultare "Informazioni sulle tecniche linguistiche" a pagina 116 Non tutte le tecniche sono disponibili per tutte le lingue di testo.

Previeni l'abbinamento di concetti specifici. Selezionare questa casella di spunta per arrestare il processo di raggruppamento o di accoppiamento di due concetti insieme nell'output. Per creare o gestire coppie di concetti, fare clic su **Gestisci Coppie**. Per ulteriori informazioni, consultare la sezione "Gestione delle coppie di eccezione di collegamento" a pagina 116.

La creazione dell'indice potrebbe richiedere alcuni minuti. Tuttavia, dopo avere generato l'indice, non è necessario generarlo di nuovo fino a quando non viene estratto nuovamente a meno che non si desideri modificare le impostazioni in modo da includere ulteriori relazioni. Se si desidera generare un indice ogni volta che si esegue un'estrazione, è possibile selezionare tale opzione nelle impostazioni di estrazione. Consultare la sezione "Estrazione di dati" a pagina 88 per ulteriori informazioni.

Perfezionamento dei risultati di estrazione

L'estrazione è un processo iterativo in cui è possibile estrarre, esaminare i risultati, apportarvi modifiche e quindi estrarre di nuovo per aggiornare i risultati. Poiché la precisione e continuità sono essenziali per l'estrazione testo e categorizzazione, ottimizzando l'estrazione dei risultati da estrarre si garantisce che ogni volta che si estrae si otterrà esattamente lo stesso risultato nelle definizioni di categoria. In questo modo, record e documenti verranno assegnati alle proprie definizioni di categoria in modo più accurato e ripetibile.

I risultati dell'estrazione servono come blocchi di creazione per le categorie. Quando si generano categorie utilizzando questi risultati di estrazione, i record o documenti vengono automaticamente assegnati a categorie se essi contengono testo che corrisponde ad uno o più descrittori di categoria. Sebbene sia possibile avviare la categorizzazione prima di apportare eventuali miglioramenti alle risorse linguistiche, è utile esaminare i risultati dell'estrazione almeno una volta prima di iniziare.

Appena è possibile esaminare i risultati, è possibile trovare gli elementi che si desidera che il motore di estrazione gestisca in modo diverso. Considerare i seguenti esempi:

- **Sinonimi non riconosciuti.** Si supponga di individuare diversi concetti che si considerano essere sinonimi, ad esempio `simpatico`, `intelligente`, `brillante`, e `aperto` e tutti sono visualizzati come singoli concetti nei risultati di estrazione. È possibile creare una definizione dei sinonimi in cui `intelligente`, `brillante` e `aperto` sono tutti raggruppati sotto il concetto di destinazione `simpatico`. In questo modo, vengono raggruppati tutti questi insieme a `simpatico` e anche il conteggio della frequenza globale potrebbe essere superiore. Consultare la sezione "Aggiunta di sinonimi" a pagina 96 per ulteriori informazioni.
- **Concetti tra più tipi.** Si supponga che i concetti nei propri risultati di estrazione vengono visualizzati in un tipo e si desidera che siano assegnati a un altro. In un altro esempio, si supponga che si desidera trovare 15 concetti di verdura nei risultati di estrazione e si desidera che tutti siano aggiunti a un nuovo tipo denominato `<Verdura>`. Per la maggior parte delle lingue, concetti che non sono presenti nei dizionari ma vengono estratti dal testo vengono automaticamente inserite come `<Sconosciuto>` È possibile aggiungere concetti ai tipi. Consultare la sezione "Aggiunta di concetti ai tipi" a pagina 97 per ulteriori informazioni.

- **Concetti non significativi.** Si supponga che si desidera trovare un concetto che è stato estratto e che ha un conteggio di frequenza che è molto elevato, viene, cioè, trovato in molti record o documenti. Tuttavia, è possibile considerare questo concetto irrilevante per l'analisi. È possibile escluderlo dall'estrazione. Consultare la sezione "Esclusione di concetti dall'estrazione" a pagina 99 per ulteriori informazioni.
- **Corrispondenze non corrette.** Si supponga che nella revisione dei record o documenti che contengono un certo concetto, si scopre che due parole sono state raggruppate insieme non correttamente, ad esempio `facoltà` e `facilità`. Questa corrispondenza può essere dovuta ad un algoritmo interno, indicato come raggruppamento confuso, che ignora temporaneamente consonanti e vocali doppie o triple per raggruppare errori ortografici comuni. È possibile aggiungere tali parole a un elenco di coppie di parole che non devono essere raggruppate. Per ulteriori informazioni, consultare la sezione "Raggruppamento confuso" a pagina 206. Il raggruppamento confuso non è disponibile per il testo giapponese.
- **Concetti non estratti.** Si supponga che si prevede di trovare alcuni concetti estratti e notare che alcune parole o frasi non sono state estratte quando si desidera rivedere il testo del record o documento. Spesso, queste parole sono verbi o aggettivi a cui non si è interessati. Tuttavia talvolta non si desidera utilizzare una parola o frase che non sia stata estratta come parte di una definizione di una categoria. Per estrarre il concetto, è possibile forzare un termine in un dizionario di tipo. Consultare la sezione "Forzatura di parole nell'estrazione" a pagina 99 per ulteriori informazioni.

Molte di queste modifiche possono essere eseguite direttamente dal riquadro dei risultati di estrazione , pannello Dati, finestra di dialogo Definizioni di categoria, o finestra di dialogo Definizioni cluster selezionando uno o più elementi e facendo clic il tasto destro del mouse per accedere ai menu di scelta rapida.

Dopo aver apportato le modifiche, il colore di sfondo del pannello cambia per mostrare che è necessario rieseguire l'estrazione per visualizzare le modifiche. Consultare la sezione "Estrazione di dati" a pagina 88 per ulteriori informazioni. Se si sta lavorando con dataset di dimensioni maggiori, potrebbe essere più efficace riestrarre dopo aver apportato diverse modifiche piuttosto che dopo ogni modifica.

Nota: è possibile visualizzare l'intera serie di risorse linguistiche modificabili utilizzate per produrre i risultati di estrazione nella vista Editor risorse (Visualizza > Editor risorse). Queste risorse vengono visualizzate sotto forma di librerie e dizionari in questa vista. È possibile personalizzare i concetti e i tipi direttamente all'interno delle librerie e dei dizionari. Consultare la sezione Capitolo 16, "Gestione delle librerie", a pagina 179 per ulteriori informazioni.

Aggiunta di sinonimi

I *sinonimi* associano due o più parole che hanno lo stesso significato. I sinonimi sono spesso anche utilizzati per raggruppare i termini con le loro abbreviazioni o per raggruppare parole comunemente errate con l'ortografia corretta. Se si utilizzano sinonimi, la frequenza per il concetto di destinazione è maggiore e ciò rende molto più facile rilevare le informazioni simili presentate in vari modi nei dati di testo.

I modelli di risorsa linguistica e le librerie fornite con il prodotto contengono molti sinonimi predefiniti. Tuttavia, se si scoprono sinonimi riconosciuti, è possibile definirli in modo che vengano riconosciute la volta successiva in cui si esegue l'estrazione.

Il primo passo consiste nel decidere quale sarà la destinazione o la direzione del concetto. Il *il concetto di destinazione* è la parola o la frase in cui si desidera raggruppare tutti i termini sinonimi nei risultati finali. Durante l'estrazione, i sinonimi sono raggruppati sotto questo concetto di destinazione. Il secondo passo è quello di identificare tutti i sinonimi per questo concetto. Il concetto di destinazione viene sostituito per tutti i sinonimi nell'estrazione finale. Per essere un sinonimo un termine deve essere estratto. Tuttavia, il concetto di destinazione non deve necessariamente essere estratto perché si verifichi la sostituzione. Ad esempio, se si desidera che `intelligente` sia sostituito da `simpatico`, `intelligente` è il sinonimo e `simpatico` è il concetto di destinazione.

Se si crea una nuova definizione di sinonimo, un nuovo concetto di destinazione viene aggiunto al dizionario. È quindi necessario aggiungere sinonimi di questo concetto di destinazione. Quando si creano o si modificano i sinonimi, tali modifiche vengono registrate nei dizionari dei sinonimi in Editor risorse. Se si desidera visualizzare l'interno contenuto di dizionari di sinonimi o eseguire un numero considerevole di modifiche, è preferibile lavorare direttamente in Editor risorse. Consultare la sezione “Dizionari dei sinonimi/di sostituzione” a pagina 196 per ulteriori informazioni.

Tutti i nuovi sinonimi verranno automaticamente memorizzati nella prima libreria elencata nella struttura ad albero delle librerie in Editor risorse: per impostazione predefinita, questa è la *Libreria locale*.

Nota: Se si desidera cercare una definizione di sinonimo e non si riesce a trovare tramite il menu di scelta rapida o direttamente in Editor risorse, è possibile che sia risultata una corrispondenza da una tecnica di raggruppamento confuso interna. Consultare la sezione “Raggruppamento confuso” a pagina 206 per ulteriori informazioni.

Per creare un nuovo sinonimo

1. In uno dei riquadri tra Risultati estrazione , pannello Dati, finestra di dialogo Definizioni di categoria o finestra di dialogo Definizioni cluster, selezionare i concetti per i quali si desidera creare un nuovo sinonimo.
2. Dai menu, scegliere **Modifica** > **Aggiungi a sinonimo** > **Nuovo**. Viene visualizzata la finestra di dialogo Crea sinonimo.
3. Immettere un concetto di destinazione nella casella di testo Destinazione. Questo è il concetto sotto il quale tutti i sinonimi verranno raggruppati.
4. Se si desidera aggiungere ulteriori sinonimi, immetterli nella casella di elenco Sinonimi. Utilizzare il separatore globale per separare ogni termine sinonimo. Per ulteriori informazioni, consultare la sezione “Opzioni: scheda Sessione” a pagina 82.
5. Se si lavora con testo in giapponese, specificare un tipo per tali sinonimi selezionando il nome tipo nel campo **Sinonimi da tipo**. Tuttavia la destinazione prende il tipo assegnato durante l'estrazione. Tuttavia, se la destinazione non è stata estratta come concetto, il tipo elencato in questa colonna viene assegnato alla destinazione nei risultati di estrazione.
6. Fare clic su **OK** per applicare le modifiche. La finestra di dialogo viene chiusa e il riquadro Risultati di estrazione cambia il colore di sfondo, indicando che è necessario estrarre di nuovo per visualizzare le modifiche. Se vengono fatte molte modifiche, eseguirle prima di una nuova estrazione.

Per aggiungere ad un sinonimo

1. In uno dei riquadri tra Risultati estrazione , pannello Dati, finestra di dialogo Definizioni di categoria o finestra di dialogo Definizioni cluster, selezionare i concetti che si desidera aggiungere a una definizione di sinonimo esistente.
2. Dai menu, scegliere **Modifica** > **Aggiungi a sinonimo**. Il menu presenta una serie di sinonimi con in cima all'elenco quella creata più di recente. Selezionare il nome del sinonimo a cui si desidera aggiungere i concetti selezionati. Se si desidera visualizzare il nome che si sta cercando, selezionarlo e i concetti selezionati vengono aggiunti a quella definizione di sinonimo. Se non viene visualizzato, selezionare **Altro** per visualizzare la finestra di dialogo Tutti i sinonimi.
3. Nella finestra di dialogo Tutti i sinonimi, è possibile impostare l'elenco di ordinamento naturale (ordine di creazione) in ordine crescente o decrescente. Selezionare il nome del sinonimo a cui si desidera aggiungere i concetti selezionati e fare clic su **OK**. La finestra di dialogo viene chiusa e i concetti vengono aggiunti alla definizione di sinonimo.

Aggiunta di concetti ai tipi

Ogni volta che viene eseguita un'estrazione, i concetti estratti sono assegnati ai tipi in uno sforzo per raggruppare i termini che hanno qualcosa in comune. IBM SPSS Modeler Text Analytics viene distribuito con molti tipi incorporati. Consultare la sezione “Tipi incorporati” a pagina 190 per ulteriori informazioni.

Per la maggior parte delle lingue, concetti che non sono presenti nei dizionari ma vengono estratti dal testo vengono automaticamente inserite come <Sconosciuto>

Quando si esaminano i risultati, è possibile individuare alcuni concetti che compaiono in un tipo che si desidera assegnare a un altro o è possibile che un gruppo di parole davvero appartiene da solo ad un nuovo tipo. In questi casi , si potrebbe decidere di assegnare di nuovo i concetti ad un altro tipo o creare insieme un nuovo tipo. Non è possibile creare nuovi tipi per testo giapponese.

Ad esempio, si supponga che si sta lavorando con dati di indagine relativi alle automobili e si è interessati a categorizzare concentrandosi su aree differenti dei veicoli. È possibile creare un tipo denominato <Cruscotto> per raggruppare tutti i concetti relativi agli indicatori e manopole che si trovano nel cruscotto dei veicoli. È quindi possibile assegnare concetti come indicatore carburante, riscaldamento, radio, e contachilometri a questo nuovo tipo.

In un altro esempio, si supponga che si sta lavorando con dati di indagine in materia di università e college e con l'estrazione di tipo Federico II (l'università) come tipo <Persona> piuttosto che come tipo <Organizzazione>. In questo caso, è possibile aggiungere questo concetto al tipo <Organizzazione> .

Quando si crea un tipo o si aggiungono concetti ad un elenco di termini di tipo, tali variazioni sono registrate nei dizionari di tipo all'interno delle proprie librerie di risorse linguistiche in Editor risorse. Se si desidera visualizzare il contenuto di queste librerie o eseguire un numero considerevole di modifiche, si potrebbe preferire di lavorare direttamente nel Editor risorse. Consultare la sezione “Aggiunta di termini” a pagina 192 per ulteriori informazioni.

Per aggiungere un concetto a un tipo

1. In uno dei riquadri tra Risultati estrazione , pannello Dati, finestra di dialogo Definizioni di categoria o finestra di dialogo Definizioni cluster, selezionare i concetti che si desidera aggiungere a un tipo esistente.
2. Fare clic con il pulsante destro del mouse per visualizzare il menu di scelta rapida.
3. Dai menu, scegliere **Modifica** > **Aggiungi a tipo**. Il menu presenta una serie di tipi con in cima all'elenco quella creata più recentemente. Selezionare il nome tipo a cui si desidera aggiungere i concetti selezionati. Se si desidera visualizzare il nome che si sta cercando, selezionarlo e i concetti selezionati vengono aggiunti a quel tipo. Se non viene visualizzato, selezionare **Altro** per visualizzare la finestra di dialogo Tutti i tipi.
4. Nella finestra di dialogo Tutti i tipi, è possibile ordinare l'elenco di ordinamento naturale (ordine di creazione) o in ordine crescente o decrescente. Selezionare il nome del tipo a cui si desidera aggiungere i concetti selezionati e fare clic su **OK**. La finestra di dialogo si chiude ed essi vengono aggiunti come termini al tipo.

Nota: con testo giapponese, esistono alcune istanze in cui la modifica di un tipo di un termine non modifica il tipo a cui verrà assegnato alla fine nell'elenco di estrazione finale. Ciò è dovuto a dizionari interni che hanno la precedenza durante l'estrazione per alcuni termini fondamentali.

Per creare un nuovo tipo

1. In uno dei riquadri tra Risultati estrazione , pannello Dati, finestra di dialogo Definizioni di categoria o finestra di dialogo Definizioni cluster, selezionare i concetti per i quali si desidera creare un nuovo tipo.
2. Dai menu, scegliere **Modifica** > **Aggiungi a tipo** > **Nuovo**. Viene visualizzata la finestra di dialogo Proprietà del tipo.
3. Immettere un nuovo nome per questo tipo nella casella di testo Nome e apportare qualsiasi modifica agli altri campi. Consultare la sezione “Creazione di tipi” a pagina 191 per ulteriori informazioni.
4. Fare clic su **OK** per applicare le modifiche. La finestra di dialogo viene chiusa e il riquadro Risultati di estrazione cambia il colore di sfondo, indicando che è necessario estrarre di nuovo per visualizzare le modifiche. Se vengono fatte molte modifiche, eseguirle prima di una nuova estrazione.

Esclusione di concetti dall'estrazione

Quando si esaminano i risultati, è possibile trovare occasionalmente concetti che non si desidera siano estratti o utilizzati da qualsiasi tecnica di creazione automatica delle categorie. In alcuni casi, questi concetti hanno un punteggio di frequenza molto elevato e sono completamente insignificanti per la propria analisi. In questi casi, è possibile contrassegnare un concetto da escludere dall'estrazione finale. Di solito, i concetti da aggiungere a questo elenco sono parole o frasi utilizzate per inserire nel testo per la continuità ma che non aggiungono nulla di importante e possono intasare i risultati di estrazione. Aggiungendo i concetti di dizionario di esclusione, è possibile assicurare che non vengano mai estratti.

Escludendo i concetti, tutte le variazioni del concetto escluso scompaiono dai risultati di estrazione la volta successiva in cui si esegue l'estrazione. Se questo concetto già appare come un descrittore in una categoria, rimarrà nella categoria con un numero zero dopo la nuova estrazione.

Quando si esclude, tali modifiche vengono registrate in un dizionario di esclusione in Editor risorse. Se si desidera visualizzare tutte le definizioni di esclusione e modificarle direttamente, si potrebbe preferire di lavorare direttamente in Editor risorse. Consultare la sezione “Dizionari di esclusione” a pagina 200 per ulteriori informazioni.

Nota: Con testo giapponese, esistono alcune istanze in cui escludere un termine o un tipo non risulta in un'esclusione. Ciò è dovuto a dizionari interni che hanno la precedenza durante l'estrazione per alcuni termini fondamentali per il giapponese.

Per escludere concetti

1. In uno dei riquadri tra Risultati estrazione , pannello Dati, finestra di dialogo Definizioni di categoria o finestra di dialogo Definizioni cluster, selezionare i concetti che si desidera escludere dall'estrazione.
2. Fare clic con il pulsante destro del mouse per visualizzare il menu di scelta rapida.
3. Selezionare **Escludi da estrazione**. Il concetto viene aggiunto al dizionario di esclusione in Editor risorse e il colore di sfondo del riquadro Risultati di estrazione cambia, indicando che è necessario estrarre di nuovo per visualizzare le modifiche. Se vengono fatte molte modifiche, eseguirle prima di una nuova estrazione.

Nota: Le parole che si desidera escludere verranno automaticamente memorizzate nella prima libreria elencata nella struttura ad albero delle librerie in Editor risorse: per impostazione predefinita, questa è la *Libreria locale*.

Forzatura di parole nell'estrazione

Quando si esaminano i dati di testo nel riquadro dei dati dopo l'estrazione, è possibile scoprire che alcune parole o frasi non sono stati estratte. Spesso, queste parole sono verbi o aggettivi a cui non si è interessati. Tuttavia talvolta non si desidera utilizzare una parola o frase che non sia stata estratta come parte di una definizione di una categoria.

Se si desidera che queste parole e frasi siano estratte, è possibile forzare un termine in una libreria di tipo. Consultare la sezione “Forzatura di termini” a pagina 195 per ulteriori informazioni.

Importante! Il contrassegno di un termine in un dizionario come forzato non è fedele. Con questo si intende che anche se è stato aggiunto esplicitamente un termine ad un dizionario, ci sono momenti in cui potrebbe non essere presente nel riquadro Risultati di estrazione dopo aver rieseguito l'estrazione o essere visualizzato ma non esattamente come era stato dichiarato. Sebbene questa ricorrenza sia rara, può verificarsi quando una parola o una frase sia stata già estratta come parte di una frase più lunga. Per evitare ciò, applicare l'opzione di corrispondenza **intero (non composti)** a questo termine nel dizionario di tipo. Consultare la sezione “Aggiunta di termini” a pagina 192 per ulteriori informazioni.

Capitolo 10. Categorizzazione dei dati di testo

Nella vista Categorie e concetti, è possibile creare **categorie** che rappresentano, in sostanza, concetti di livello superiore o gli argomenti, che individuano le idee chiave, le conoscenze e gli atteggiamenti espressi nel testo.

A partire dalla release 14 di IBM SPSS Modeler Text Analytics, le categorie possono anche avere una struttura gerarchica, ossia possono contenere sottocategorie e quelle sottocategorie possono avere anche categorie secondarie proprie e così via. È possibile importare strutture di categoria predefinite, in precedenza definite frame di codice, con categorie gerarchiche oltre a creare queste categorie gerarchiche all'interno del prodotto.

In effetti, le categorie gerarchiche consentono di creare una struttura ad albero con una o più categorie secondarie per raggruppare le voci come concetto differente o aree di argomento più accurate. Un esempio semplice può essere correlato alle attività ricreative; rispondendo a una domanda come *Quali attività si desidera eseguire se si avesse più tempo a disposizione?* è possibile avere categorie principali come *sport*, *arte e artigianato*, *pesca*e così via; in un livello al di sotto di *sport*, è possibile avere sottocategorie come *giochi di palla*, *d'acqua* e così via.

Le categorie sono costituite da una serie di descrittori, per esempio *concetti*, *tipi*, *modelli* e *regole di categoria*. Tutti insieme questi descrittori vengono utilizzati per identificare se un documento o record appartiene o meno a una determinata categoria. Il testo all'interno di un documento o record può essere analizzato per vedere se esiste testo che corrisponde a un descrittore. Se viene rilevata una corrispondenza, il documento/record viene assegnato a tale categoria. Questo processo viene definito **categorizzazione**.

È possibile gestire, creare ed esplorare visivamente le proprie categorie utilizzando i dati presenti nei quattro riquadri della vista Categorie e concetti, ognuno dei quali può essere nascosto o mostrato selezionandone il nome dal menu Visualizza.

- **Riquadro Categorie.** Creare e gestire le categorie in questo riquadro. Consultare la sezione "Riquadro Categorie" a pagina 102 per ulteriori informazioni.
- **Riquadro Risultati di estrazione.** Esplorare e gestire i concetti e i tipi estratti in questo riquadro. Per ulteriori informazioni, consultare la sezione "Risultati estrazione: concetti e tipi" a pagina 87.
- **Riquadro Visualizzazione.** Esplorare visivamente le categorie e il modo in cui esse interagiscono in questo riquadro. Per ulteriori informazioni, consultare la sezione "Grafici e diagrammi di categoria" a pagina 159.
- **Riquadro Dati.** Esplorare e rivedere in questo riquadro il testo contenuto all'interno di documenti e record che corrispondono alle selezioni. Per ulteriori informazioni, consultare la sezione "Riquadro Dati" a pagina 110.

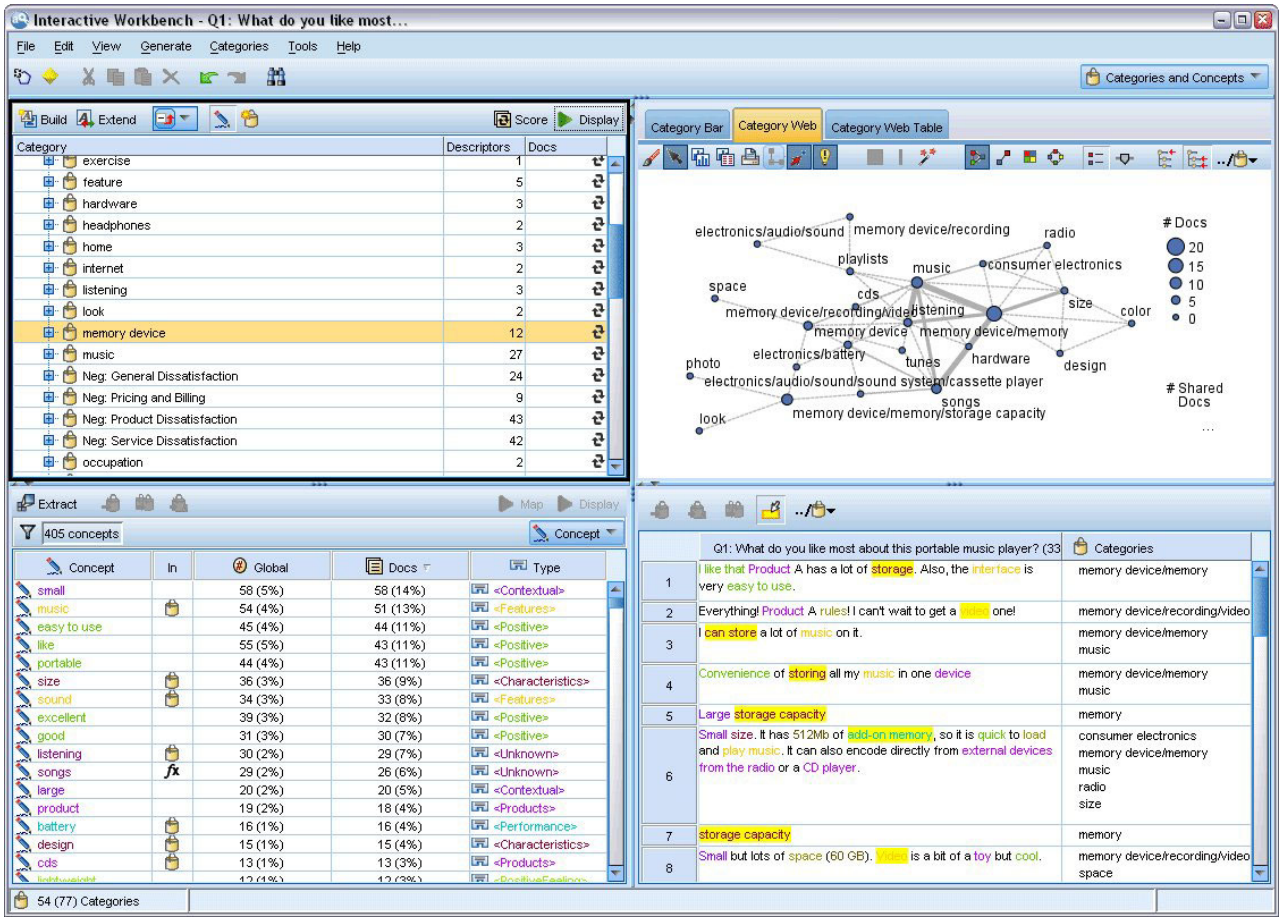


Figura 29. Vista Categorie e concetti

Mentre è possibile iniziare con una serie di categorie da un TAP (text analysis package) o importarle da un file di categoria predefinito, può rivelarsi necessario doverne creare di proprie. Le categorie possono essere create automaticamente utilizzando la serie completa del prodotto di tecniche automatiche, che utilizzano i risultati dell'estrazione (i concetti, i tipi e i modelli) per generare le categorie e i relativi descrittori. Le categorie possono anche essere create manualmente utilizzando informazioni aggiuntive che potrebbero riguardare i dati. Tuttavia, è possibile creare solo le categorie o regolarle manualmente attraverso il workbench interattivo. Per ulteriori informazioni, consultare la sezione “Nodo di estrazione testo: scheda Modello” a pagina 24. È possibile creare definizioni di categoria manualmente trascinando e rilasciando i risultati di estrazione nelle categorie. È possibile arricchire queste categorie o qualsiasi categoria vuota aggiungendo regole di categoria a una categoria, utilizzando le proprie categorie predefinite o una combinazione.

Ciascuna delle tecniche e dei metodi si adatta adeguatamente ad alcuni tipi di dati e situazioni, ma spesso è utile combinare le tecniche nella stessa analisi e acquisire la gamma completa di documenti o record. E nel corso della categorizzazione, è possibile individuare altre modifiche da apportare alle risorse linguistiche.

Riquadro Categorie

Il riquadro Categorie è l'area in cui è possibile creare e gestire le proprie categorie. Questo riquadro si trova nell'angolo in alto a sinistra della vista Categorie e concetti. Dopo aver estratto i concetti e i tipi dai dati di testo, è possibile iniziare a creare categorie automaticamente mediante tecniche quali l'inclusione concetto, la ricorrenza e così via o manualmente. Consultare la sezione “Creazione di categorie” a pagina 112 per ulteriori informazioni.

Ogni volta che una categoria viene creata o aggiornata i documenti o record possono essere valutati facendo clic sul pulsante **Punteggio** verificare se un testo qualsiasi corrisponde ad un descrittore in una data categoria. Se viene rilevata una corrispondenza, il documento/record viene assegnato a tale categoria. Il risultato finale è che la maggior parte, se non tutti, i documenti o i record vengono assegnati a categorie basate sui descrittori nelle categorie.

Tabella ad albero Categoria

La tabella con struttura ad albero di questo riquadro mostra l'insieme di categorie, sottocategorie e descrittori. La struttura ad albero presenta anche diverse colonne che riportano le informazioni per ciascun elemento della struttura. È possibile visualizzare le seguenti colonne:

- **Codice.** Elenca il valore di codice per ogni categoria. Questa colonna è nascosta per impostazione predefinita. È possibile visualizzare questa colonna mediante il menu: **Visualizza > riquadro Categorie**.
- **Categoria.** Contiene la struttura ad albero della categoria che mostra il nome della categoria e delle sottocategorie. Inoltre, se l'icona barra degli strumenti descrittori è selezionata, verrà visualizzata anche la serie di descrittori.
- **Descrittori.** Fornisce il numero di descrittori che costituiscono la sua definizione. Questo conteggio non include il numero di descrittori nelle sottocategorie. Nessun conteggio viene fornito quando un nome descrittore viene mostrato nella colonna **Categorie**. È possibile visualizzare o nascondere gli stessi descrittori nella struttura ad albero mediante i menu: **Visualizza > riquadro Categorie > Tutti i descrittori**.
- **Documenti.** Dopo il calcolo del punteggio, questa colonna fornisce il numero di documenti o record che sono raggruppati in una categoria con tutte le sue sottocategorie. Quindi se 5 record corrispondono alla categoria principale in base ai propri descrittori e 7 diversi record corrispondono a una sottocategoria in base ai propri descrittori, il conteggio totale dei documenti per la categoria principale è la somma dei due-- e in questo caso sarebbe 12. Tuttavia, se il record stesso corrisponde alla categoria principale e alla sua sottocategoria, il conteggio sarebbe 11.

Quando non esistono categorie, la tabella contiene ancora due righe. La riga superiore, denominata **Tutti i documenti**, è il numero totale di documenti o record. Una seconda riga, denominata **Non categorizzato**, mostra il numero di documenti/record che devono ancora essere categorizzati.

Per ogni categoria nel riquadro, una piccola icona con un carrello giallo precede il nome categoria. Se si fa doppio clic su una categoria oppure si sceglie **Visualizza > Definizioni di categoria** nel menu, la finestra di dialogo Definizioni di categoria si apre e presenta tutti gli elementi, denominati **descrittori**, che costituiscono la sua definizione, come ad esempio i concetti, i tipi, i modelli e le regole di categoria. Consultare la sezione “Informazioni sulle categorie” a pagina 109 per ulteriori informazioni. Per impostazione predefinita, la tabella ad albero della categoria non mostra i descrittori nelle categorie. Se si desidera visualizzare i descrittori direttamente nella struttura ad albero anziché nella finestra di dialogo Definizioni di categoria, fare clic sul pulsante con l'icona matita nella barra degli strumenti. Quando questo pulsante è selezionato, è possibile espandere la struttura ad albero per visualizzare i descrittori.

Calcolo del punteggio delle categorie

La colonna **Documento** nella tabella con struttura ad albero della categoria visualizza il numero di documenti o record che vengono classificati in tale categoria specifica. Se i numeri sono superati o non sono calcolati, in quella colonna viene visualizzata un'icona. È possibile fare clic su **Punteggio** sulla barra degli strumenti del riquadro per ricalcolare il numero di documenti. Tenere presente che il processo di calcolo del punteggio può richiedere del tempo quando si lavora con dataset estesi.

Selezione delle categorie nella struttura ad albero

Quando si seleziona nella struttura ad albero, è possibile selezionare solo le categorie gemelle, vale a dire, se si selezionano categorie di livello superiore, non è possibile selezionare anche una sottocategoria. In

alternativa, se si selezionano 2 sottocategorie di una certa categoria, non è possibile selezionare simultaneamente una sottocategoria di un'altra categoria. La selezione di una categoria non contigua determinerà la perdita della selezione precedente.

Esposizione nei riquadri Dati e Visualizzazione

Quando si seleziona una riga nella tabella, è possibile fare clic sul pulsante **Visualizza** per aggiornare i riquadri Visualizzazione e Dati con le informazioni corrispondenti alla propria selezione. Se un riquadro non è visibile, facendo clic su **Visualizza** il riquadro viene visualizzato.

Perfezionamento delle categorie

La categorizzazione potrebbe non portare immediatamente a risultati perfetti per i propri dati e si potrebbero altresì produrre categorie che si desidera eliminare o combinare con altre. È inoltre possibile trovare, attraverso un riesame dei risultati di estrazione, che vi sono alcune categorie che non sono state create e che potrebbero rivelarsi utili. In tal caso, è possibile apportare modifiche manuali ai risultati per meglio adattarli al proprio particolare contesto. Per ulteriori informazioni, consultare “Modifica e perfezionamento delle categorie” a pagina 143.

Metodi e strategie per la Creazione di categorie

Se non è stato ancora estratto nulla o i risultati di estrazione sono non aggiornati, l'uso di una delle tecniche di creazione ed estensione delle categorie richiederà automaticamente all'utente un'estrazione. Dopo aver applicato una tecnica, i concetti e i tipi che sono stati raggruppati in una categoria sono ancora disponibili per la creazione della categoria con altre tecniche. Ciò significa che è possibile visualizzare un concetto in più categorie a meno che non si scelga di non riutilizzarle.

Al fine di aiutare l'utente a creare le migliori categorie, consultare le seguenti sezioni:

- **Metodi per la creazione di categorie**
- **Strategie per la creazione di categorie**
- **Suggerimenti per la creazione di categorie**

Metodi per la creazione di categorie

Poiché ogni dataset è univoco, il numero di metodi di creazione categoria e l'ordine in cui si desidera applicarli possono cambiare nel tempo. Inoltre, poiché gli obiettivi di estrazione di testo possono essere diversi da una serie di dati a quella successiva, potrebbe essere necessario sperimentare metodi differenti per vedere quale produce i migliori risultati per i dati di testo forniti. Nessuna delle tecniche automatiche è in grado di classificare perfettamente i propri dati; di conseguenza si consiglia di rilevare ed applicare una o più tecniche automatiche che funzionano bene con i propri dati.

Oltre a utilizzare pacchetti di analisi del testo (TAP, **.tap*) con serie di categorie precostituite, è anche possibile categorizzare le risposte utilizzando qualsiasi combinazione dei seguenti metodi:

- **Tecniche di creazione automatica.** Per creare categorie automaticamente, sono disponibili diverse opzioni di categoria basate sulla linguistica e sulla frequenza. Consultare la sezione “Creazione di categorie” a pagina 112 per ulteriori informazioni.
- **Tecniche di estensione automatica.** Sono disponibili molte tecniche linguistiche per estendere le categorie esistenti aggiungendo e aumentando i descrittori in modo da catturare più record. Consultare la sezione “Estensione delle categorie” a pagina 122 per ulteriori informazioni.
- **Tecniche manuali.** Esistono diversi metodi manuali, come trascinamento e rilascio. Consultare la sezione “Creazione manuale di categorie” a pagina 125 per ulteriori informazioni.

Strategie per la creazione di categorie

Il seguente elenco di strategie non è in alcun modo esaustivo ma può fornire alcune idee su come affrontare la creazione delle proprie categorie.

- Quando si definire il nodo di estrazione testo, selezionare un insieme di categorie da un pacchetto di analisi del testo (TAP) in modo che sia possibile iniziare l'analisi con alcune categorie precostituite. Queste categorie possono classificare il testo sin dall'inizio. Tuttavia, se si desidera aggiungere più categorie, è possibile modificare le impostazioni di creazione categorie (**Categorie > Impostazioni di creazione**). Aprire la finestra di dialogo **Impostazioni avanzate: linguistica**, selezionare l'opzione di input di categoria **Risultati di estrazione non utilizzati** e creare le altre categorie.
- Quando si definisce il nodo, selezionare un insieme di categorie da un TAPnella vista Categorie e concetti nel Workbench interattivo. Quindi trascinare e rilasciare i concetti o i modelli inutilizzati nelle categorie che si considerano appropriate. Estendere poi le categorie esistenti appena modificate (**Categorie > Estendi categorie**) per ottenere ulteriori descrittori correlati agli attuali descrittori di categoria.
- Creare categorie automaticamente utilizzando le impostazioni linguistiche avanzate (**Categorie > Genera categorie**). Perfezionare poi le categorie manualmente eliminando descrittori, eliminando categorie o unendo categorie simili fino a che non si è soddisfatti delle categorie risultanti. Inoltre, se sono state originariamente create delle categorie **senza** utilizzare l'opzione **Generalizza con caratteri jolly dove possibile**, è possibile anche tentare di semplificare le categorie automaticamente utilizzando l'estensione delle categorie con l'opzione **Generalizza**.
- Importare un file di categoria predefinito con nomi categoria molto descrittivi e/o con annotazioni. Inoltre, se è stato originariamente importato **senza** scegliendo l'opzione per importare o generare descrittori dai nomi categoria, è possibile utilizzare la finestra di dialogo Estendi categorie e scegliere l'opzione **Estendi categorie vuote con i descrittori generati dal nome categoria** . Quindi, estendere tali categorie una seconda volta utilizzando stavolta le tecniche di raggruppamento.
- Creare manualmente una prima serie di categorie mettendo in ordine concetti o modelli concetto in base alla frequenza e poi trascinando e rilasciando le più interessanti sul riquadro Categorie. Una volta stabilita questa serie iniziale di categorie, utilizzare la funzione Estendi (**Categorie > Estendi categorie**) per espandere e perfezionare tutte le categorie selezionate in modo da includere altri descrittori correlati e corrispondere quindi a più record.

Dopo aver applicato queste tecniche, si consiglia di consultare la categorie risultanti e utilizzare tecniche manuali per effettuare piccoli adattamenti, rimuovere eventuali classificazioni errate o aggiungere record o parole che potrebbero essere stati omessi. Inoltre, poiché utilizzando tecniche diverse si possono produrre categorie ridondante, è anche possibile unire o eliminare categorie quando necessario. Per ulteriori informazioni, consultare “Modifica e perfezionamento delle categorie” a pagina 143.

Suggerimenti per la creazione di categorie

Al fine di aiutare a creare migliori categorie, è possibile esaminare alcuni suggerimenti che consentono di prendere decisioni sulla propria impostazione.

Suggerimenti sul rapporto categoria-documento

Le categorie in cui documenti e record sono assegnati spesso non sono mutualmente esclusivi nell'analisi di testo qualitativa per almeno due ragioni:

- In primo luogo, una regola pratica generale recita che quanto più è lungo il documento o record di testo, più vengono espresse opinioni e idee distinte. Quindi, le possibilità che un documento o record possa essere assegnato a più categorie aumentano di molto.
- In secondo luogo, spesso vi sono vari modi di raggruppare e interpretare i documenti o record di testo che non sono logicamente separati. Nel caso di un'indagine con un'interrogazione aperta sulle convinzioni politiche degli intervistati, si potrebbero creare categorie, quali *Liberale* e *Conservatore* o

Repubblicano e *Democratico*, così come categorie più specifiche, come ad esempio *Socialdemocratico*, *Ultraconservatore*, e così via. Queste categorie non devono necessariamente essere reciprocamente esclusive ed esaurienti.

Suggerimenti per Numero di categorie da creare

La creazione di categoria dovrebbe scorrere direttamente dai dati - come si vede qualcosa di interessante, con rispetto dei propri dati, è possibile creare una categoria per rappresentare tali informazioni. In generale, non vi è limite superiore consigliato sul numero di categorie da creare. Tuttavia è possibile indubbiamente creare troppe categorie da poter gestire. Si applicano due principi:

- **Frequenza di categoria.** Perché una categoria risulti utile, deve contenere un numero minimo di documenti o record. Uno o due documenti possono includere qualcosa di intrigante, ma se si tratta di uno o due su 1.000 documenti , le informazioni che essi contengono potrebbero non essere frequenti abbastanza per essere, nella pratica, utili.
- **Complessità.** Più categorie vengono create, più informazioni è necessario rivedere e riepilogare dopo il completamento dell'analisi. Tuttavia, troppe categorie, mentre si aggiunge complessità, potrebbero non aggiungere dettagli utili.

Purtroppo non esistono regole per determinare quante categorie sono troppe o per determinare il numero minimo di record per categoria. È necessario prendere tali decisioni in base alle esigenze della propria particolare situazione.

È possibile, tuttavia, fornire consigli su dove iniziare. Anche se il numero di categorie non dovrebbe essere eccessivo, nelle fasi iniziali dell'analisi è meglio avere troppe piuttosto che troppo poche categorie. E' più facile raggruppare le categorie che sono relativamente simili che dividere i casi in nuove categorie, quindi una strategia di lavoro da più a meno categorie è generalmente la migliore prassi. Data la natura iterativa dell'estrazione di testo e la facilità con cui può essere eseguita con questo programma software, creare più categorie all'inizio è accettabile.

Scelta dei migliori descrittori

Le seguenti informazioni contengono alcune linee guida per scegliere o creare i migliori descrittori (concetti, tipi, modelli TLA e regole di categoria) per le categorie. I descrittori sono i blocchi di creazione delle categorie. Quando parte o tutto il testo in un documento o record corrisponde a un descrittore, il documento o record è associato alla categoria.

A meno che un descrittore non contiene o non corrisponde a un concetto o modello estratti, non verrà associato ad alcun documento o record. Pertanto utilizzare concetti, tipi, modelli e regole di categoria come descritto nei seguenti paragrafi.

Poiché i concetti rappresentano non solo se stessi ma anche una serie di termini sottostanti che può variare da forme plurali a singolari, sinonimi, variazioni ortografiche, solo il concetto stesso dovrebbe essere utilizzato come descrittore o come parte di un descrittore. Per ulteriori informazioni sui termini sottostanti di qualsiasi concetto, fare clic sul nome del concetto nel riquadro di estrazione dei risultati della vista Categorie e Concetti. Quando si passa sul nome del concetto, viene visualizzato un suggerimento che riporta tutti i termini sottostanti rilevati durante l'ultima estrazione. Non tutti i concetti sono termini sottostanti. Ad esempio, se `auto` e `veicolo` sono stati sinonimi ma `auto` è stato estratto come concetto con `veicolo` come un termine sottostante, quindi si desidera solo utilizzare `auto` in un descrittore poiché il documento o il record corrisponderà automaticamente con `veicolo`.

Concetti e tipi come descrittori

Utilizzare un concetto come descrittore quando si desidera trovare tutti i documenti o record che contengono tale concetto (o uno qualsiasi dei suoi termini sottostanti). In questo caso, l'utilizzo di una regola di categoria più complessa non è necessario poiché il nome di concetto esatto è sufficiente. Tenere

presente che quando si utilizzano le risorse che estraggono opinioni, a volte i concetti possono cambiare durante l'estrazione di modelli TLA per catturare il senso più vero della frase (fare riferimento all'esempio nella sezione successiva su TLA).

Ad esempio, una risposta che indica la frutta preferita di un'indagine *Mela e ananas sono i migliori*" potrebbe risultare in un'estrazione di `mela` e `ananas`. Aggiungendo il concetto `mela` come descrittore per la categoria, tutte le risposte contenenti il concetto `mela` (o uno qualsiasi dei suoi termini sottostanti) corrispondono a tale categoria.

Tuttavia, se si è interessati a sapere semplicemente quali risposte menzionano *mele* in qualsiasi modo, è possibile scrivere una regola di categoria come `* mela *` e si acquisiranno le risposte che contengono concetti come `mela`, `succo di mela` o `torta di mele francese`.

È inoltre possibile catturare tutti i documenti o record che contengono concetti che sono stati immessi nello stesso modo utilizzando un tipo come un descrittore direttamente come `<Frutta>`. Notare che non è possibile utilizzare * con i tipi.

Per ulteriori informazioni, consultare “Risultati estrazione: concetti e tipi” a pagina 87.

Modelli TLA (Text Link Analysis) come descrittori

Utilizzare un risultato di modello TLA come un descrittore quando si desidera catturare idee più sottili, con più sfumature. Quando il testo viene analizzato durante l'estrazione TLA, il testo elabora una frase, o proposizione per volta piuttosto che cercare nell'intero testo (il documento o record). Considerando tutte le parti di una frase insieme, TLA può identificare, per esempio, opinioni, relazioni tra due elementi o una negazione e comprendere il senso più reale. È possibile utilizzare modelli di concetto o di tipo come descrittori. Per ulteriori informazioni, consultare “Modelli di tipo e concetto” a pagina 155.

Ad esempio, se avessimo il testo "*la stanza non era molto pulita*", possono essere estratti i seguenti concetti: `stanza` e `pulita`. Tuttavia, se l'estrazione TLA è stata abilitata nell'impostazione di estrazione, TLA potrebbe rilevare che `pulita` è stato utilizzato in maniera negativa e in realtà corrisponde a `non pulita`, che è un sinonimo del concetto `sporca`. Qui è possibile vedere che l'uso del concetto `pulita` come proprio descrittore può corrispondere a questo testo, ma può anche acquisire altri documenti o o record che parlano di pulizia. Pertanto, potrebbe essere preferibile utilizzare il modello di concetto TLA con `sporco` come concetto di output perché corrisponderebbe a questo testo e potrebbe essere un descrittore più adeguato.

Regole business di categoria come descrittori

Le regole di categoria sono istruzioni che vengono automaticamente classificate come documenti o record in una categoria basata su un'espressione logica utilizzando concetti estratti, tipi e modelli e operatori booleani. Ad esempio, è possibile scrivere un'espressione che significa *includi tutti i record che contengono i concetti estratti* `ambasciata` *ma non* `argentina` *in questa categoria*.

È possibile scrivere e utilizzare le regole di categoria come descrittori nelle proprie categorie per esprimere idee diverse utilizzando gli operatori booleani `&`, `|` e `!()`. Per informazioni dettagliate sulla sintassi di tali regole e su come scriverle e modificarle, consultare “Uso delle regole di categoria” a pagina 126.

- Utilizzare una regola di categoria con l'operatore booleano `&` (AND) per aiutare l'utente a trovare documenti o record in cui ricorrono 2 o più concetti. I 2 o più concetti connessi dagli operatori `&` non devono verificarsi nella stessa frase o periodo, ma possono ricorrere ovunque nello stesso documento o record perché vengano considerati una corrispondenza per la categoria. Ad esempio, se si crea la regola di categoria `vitto & a buon mercato` come descrittore, potrebbe corrispondere ad un record contenente il testo, "*il vitto era molto caro, ma le camere erano a buon mercato*" nonostante il fatto che `vitto` non era il nome per `a buon mercato` poiché il testo conteneva sia `vitto` che `a buon mercato`.

- Utilizzare una regola di categoria con l'operatore booleano `!()` (NOT) come descrittore ad aiutare l'utente a trovare documenti o record in cui alcuni elementi si verificano ma altri no. Questo può contribuire a evitare il raggruppamento delle informazioni che possono sembrare correlate in base alle parole ma non al contesto. Ad esempio, se si crea la regola di categoria `<Azienda> & !(ibm)` come descrittore, potrebbe corrispondere al seguente testo *SPSS Inc. era un'azienda fondata nel 1967* e non corrispondere al seguente testo *l'azienda di software è stata acquisita da IBM.*.
- Utilizzare una regola di categoria con l'operatore booleano `|` (OR) come descrittore per aiutare a trovare documenti o record contenenti uno dei diversi concetti o tipi. Ad esempio, se si crea la regola di categoria `(personale|staff|squadra|colleghi) & cattivi` come descrittore, sarebbe in corrispondenza con qualsiasi documento o record in cui nessuno di quei nomi vengono trovati con il concetto `cattivi`.
- Utilizzare i tipi di regole di categoria per renderli più generici e possibilmente più distribuibili. Ad esempio, se si sta lavorando su dati di un hotel, si potrebbe essere molto interessati ad apprendere il giudizio dei clienti sull'hotel. I termini correlati possono includere parole come centralinista, cameriere, cameriera, addetto al ricevimento, accoglienza e così via. Si potrebbe, in questo caso, creare un nuovo tipo denominato `<HotelStaff>` e aggiungere tutti i termini precedenti a quel tipo. Mentre è possibile creare una regola di categoria per ogni tipo di personale come `[* cameriera * & bella]`, `[* accoglienza * & buona]`, `[* centralinista * & disponibile]`, è possibile creare un'unica regola di categoria più generica utilizzando il tipo `<HotelStaff>` per catturare tutte le risposte che esprimono pareri favorevoli sul personale dell'hotel nel formato `[<HotelStaff> & <Positivo>]`.

Nota: è possibile utilizzare sia `+` che `&` nelle regole di categoria quando i modelli TLA vengono compresi in queste regole. Consultare la sezione "Uso dei modelli TLA nelle regole di categoria" a pagina 128 per ulteriori informazioni.

Esempio di come i concetti, TLA o le regole di categoria come i descrittori corrispondono in modo diverso

L'esempio che segue dimostra come un concetto come un descrittore, una regola di categoria come descrittore o utilizzando un modello TLA come descrittore influenzano il modo di categorizzazione di documenti o record. Si supponga che si dispone dei seguenti 5 record.

- A: *"personale meraviglioso del ristorante, cibo eccellente e stanze confortevoli e pulite."*
- B: *"il personale del ristorante è stato pessimo, ma le stanze erano pulite."*
- C: *"Stanze pulite, confortevoli."*
- D: *"La mia stanza non era tanto pulita."*
- E: *"Pulita."*

Poiché i record includono la parola *pulita* e si desidera catturare queste informazioni, è possibile creare uno dei descrittori mostrato nella seguente tabella. In base alla sostanza che si sta provando a catturare, è possibile capire come l'uso di un tipo di descrittore su un altro può produrre risultati diversi.

Tabella 17. Come i record di esempio sono corrispondenti di descrittori.

Descrittore	ERRORE! DATI DI SEGMENTO CORROTTI, SEGDATA=A	ERRORE! DATI DI SEGMENTO CORROTTI, SEGDATA=B	ERRORE! DATI DI SEGMENTO CORROTTI, SEGDATA=C	ERRORE! DATI DI SEGMENTO CORROTTI, SEGDATA=D	ERRORE! DATI DI SEGMENTO CORROTTI, SEGDATA=E	Spiegazione
`pulito`	*corrisponde*	*corrisponde*	*corrisponde*	*corrisponde*	*corrisponde*	Il descrittore è un concetto estratto. Ogni record contiene il concetto `pulito`, anche il record D poiché senza TLA; non viene riconosciuto automaticamente che "*non pulito*" significa `sporco` in base alle regole TLA.

Tabella 17. Come i record di esempio sono corrispondenti di descrittori (Continua).

Descrittore	ERRORE! DATI DI SEGMENTO CORROTTI, SEGDATA=A	ERRORE! DATI DI SEGMENTO CORROTTI, SEGDATA=B	ERRORE! DATI DI SEGMENTO CORROTTI, SEGDATA=C	ERRORE! DATI DI SEGMENTO CORROTTI, SEGDATA=D	ERRORE! DATI DI SEGMENTO CORROTTI, SEGDATA=E	Spiegazione
`pulito + .`	-	-	-	-	*corrisponde*	Il descrittore è un modello TLA che rappresenta da solo `pulito`. Corrisponde solo al record in cui è stato estratto con `pulito` senza concetto associato durante l'estrazione TLA.
`[pulito]`	*corrisponde*	*corrisponde*	*corrisponde*	-	*corrisponde*	Il descrittore è una regola di categoria che cerca una regola TLA che contiene da sola `pulito` o con altro. Corrisponde a tutti i record in cui un output TLA che contiene `pulito` è stato trovato indipendentemente dal fatto che `pulito` sia stato o meno collegato ad un altro concetto come `stanza` e in qualsiasi posizione nel blocco.

Informazioni sulle categorie

Per **categoria** si intende un gruppo di concetti, opinioni, o atteggiamenti strettamente correlati. Per essere utile, una categoria deve essere facilmente descritta da una frase breve o da un'etichetta che catturi il suo significato fondamentale.

Ad esempio, se si sta analizzando le risposte di indagine dai consumatori su un nuovo sapone per bucato, è possibile creare una categoria con etichetta *odore* che contenga tutte le risposte che descrivono l'odore del prodotto. Tuttavia, tale categoria non distingue tra chi trova l'odore gradevole e coloro che lo trovano sgradevole. Poiché IBM SPSS Modeler Text Analytics è in grado di estrarre opinioni quando si utilizzano le risorse adeguate, è possibile creare altre due categorie per identificare coloro che hanno *gradito l'odore* e coloro che non hanno *gradito l'odore*.

È possibile creare e gestire le categorie nel riquadro Categorie nel riquadro Categorie nel riquadro in alto a sinistra della vista Categorie e concetti . Ogni categoria viene definita da uno o più descrittori. **I descrittori** sono i concetti, i tipi e modelli, così come le regole di categoria che sono stati utilizzati per definire una categoria.

Se si desidera vedere i descrittori che costituiscono una determinata categoria, è possibile fare clic sull'icona matita nella barra degli strumenti del riquadro Categorie e quindi espandere la struttura ad albero per visualizzare i descrittori. In alternativa, selezionare la categoria e aprire la finestra di dialogo Definizioni di categoria (**Visualizza > Definizioni di categoria**).

Quando si creano categorie automaticamente utilizzando tecniche di creazione categoria come, ad esempio, l'inclusione di concetto, le tecniche utilizzeranno i concetti e i tipi come descrittori al fine di creare le proprie categorie. Se si estraggono modelli TLA, è possibile possibile aggiungere modelli o parti di tali modelli come descrittori di categoria. Consultare la sezione Capitolo 12, "Esplorazione di analisi di collegamento del testo", a pagina 153 per ulteriori informazioni. E se si desidera creare i cluster, è possibile aggiungere i concetti in un cluster per le categorie nuove o esistenti. Infine, è possibile creare manualmente regole di categoria da utilizzare come descrittori nelle proprie categorie. Consultare la sezione "Uso delle regole di categoria" a pagina 126 per ulteriori informazioni.

Proprietà della categoria

Oltre ai descrittori, anche le categorie hanno proprietà che è possibile modificare per ridenominare le categorie, aggiungere un'etichetta oppure aggiungere un'annotazione.

Sono presenti le seguenti proprietà:

- **Nome.** Questo nome viene visualizzato automaticamente nella struttura ad albero. Quando una categoria viene creata utilizzando una tecnica automatica, viene attribuito un nome automaticamente.
- **Etichetta.** Utilizzare le etichette è utile per la creazione di descrizioni di categoria più significative per l'uso in altri prodotti o in altre tabelle o grafici. Se si sceglie di visualizzare l'etichetta, l'etichetta viene utilizzata nell'interfaccia per identificare la categoria.
- **Codice.** Il numero del codice corrisponde al valore del codice per questa categoria. .
- **Annotazione.** È possibile aggiungere una breve descrizione per ciascuna categoria del campo. Quando una categoria viene generata dalla finestra di creazione categorie, una nota viene aggiunta a questa annotazione automaticamente. È anche possibile aggiungere testo di esempio ad un'annotazione direttamente dal pannello Dati selezionando il testo e scegliendo **Categorie > Aggiungi ad annotazione** dai menu.

Riquadro Dati

Quando si creano le categorie, ci potrebbero essere delle volte in cui si potrebbe voler rivedere alcuni dei dati di testo con cui si sta lavorando. Ad esempio, se si crea una categoria in cui i documenti categorizzati sono 640, è possibile che si desideri esaminare alcuni o tutti quei documenti per vedere cosa è scritto realmente sul testo. È possibile rivedere i record o i documenti nel riquadro Dati che si trova in basso a destra. Se non è visibile per impostazione predefinita, scegliere **Visualizza > Riquadri > Dati** dal menu.

Il pannello presenta una riga di dati per ogni documento o record corrispondente alla selezione nel riquadro Categorie, nel riquadro Risultati di estrazione o nella finestra di dialogo Definizioni categoria fino a un certo limite di visualizzazione. Per impostazione predefinita, il numero di documenti o record visualizzati nel riquadro Dati è limitata per consentire all'utente di visualizzare i dati più rapidamente. Tuttavia, è possibile regolare questo valore nella finestra Opzioni. Consultare la sezione “Opzioni: scheda Sessione” a pagina 82 per ulteriori informazioni.

Visualizzazione e aggiornamento del riquadro Dati

Il pannello Dati non viene aggiornato automaticamente in quanto con grandi insiemi di dati l'aggiornamento automatico potrebbe richiedere del tempo. Pertanto, ogni volta che si effettua una selezione in un altro riquadro in questa vista o nella finestra di dialogo Definizioni di categoria, fare clic su **Visualizza** per aggiornare il contenuto del pannello Dati.

Documenti di testo o record

Se i dati di testo è sotto forma di record e il testo è relativamente breve in lunghezza, il campo di testo nel pannello Dati visualizza i dati di testo nella loro interezza. Tuttavia, quando si lavora con record e grandi insiemi, la colonna campo di testo visualizza un breve del testo e apre un riquadro Anteprima testo a destra per visualizzare la maggior parte o tutto il testo del record selezionato nella tabella. Se i dati di testo sono sotto forma di documenti singoli, il riquadro Dati visualizza il nome file del documento. Quando si seleziona un documento, il riquadro Anteprima testo si apre con il testo del documento selezionato.

Colori ed evidenziazione

Ogni volta che si visualizzano i dati, i concetti e i descrittori che si trovano in quei documenti o record vengono evidenziati con un colore per aiutare l'utente a identificarli facilmente nel testo. Il codice colore

corrisponde ai tipi a cui i concetti appartengono. È inoltre possibile passare il mouse su elementi di colore per visualizzare il concetto da cui è stato estratto e il tipo a cui era assegnato. Tutto il testo non estratto appare in nero. Di solito, queste parole non estratte sono spesso connettori (*e* o *con*), pronomi (*me* o *essi*), e verbi (*è*, *ha* o *dai*).

Colonne del riquadro Dati

Mentre la colonna del campo di testo è sempre visibile, è possibile anche visualizzare altre colonne. Per visualizzare altre colonne, scegliere **Visualizza > riquadro Dati** dal menu e quindi selezionare la colonna che si desidera visualizzare nel riquadro Dati. È possibile visualizzare le seguenti colonne:

- **"Nome campo testo" (#)/Documenti.** Aggiunge una colonna per i dati di testo da cui i concetti e il tipo sono stati estratti. Se i dati sono nei documenti, la colonna viene denominata Documenti e solo il nome file documento o percorso completo è visibile. Per vedere il testo per tali documenti è necessario cercare nel riquadro Anteprima testo. Il numero di righe nel riquadro Dati viene visualizzato tra parentesi dopo il nome della colonna. Potrebbero esserci momenti in cui non tutti i documenti o record vengono visualizzati a causa di un limite nella finestra di dialogo Opzioni utilizzato per aumentare la velocità di caricamento. Se si raggiunge il massimo, il numero sarà seguito da **- Max**. Consultare la sezione “Opzioni: scheda Sessione” a pagina 82 per ulteriori informazioni.
- **Categorie.** Elenca ognuna delle categorie a cui il record appartiene. Quando questa colonna viene visualizzata, l'aggiornamento del riquadro dei dati potrebbe richiedere un po' più di tempo per poter visualizzare le informazioni più aggiornate.
- **Classifica di attinenza.** Fornisce una classificazione per ogni record in una categoria singola. Questa classificazione di attinenza indica quanto bene il record si inserisce nella categoria confrontata rispetto agli altri record di tale categoria. Selezionare una categoria nel riquadro Categorie (riquadro in alto a sinistra) per visualizzare la classificazione. Consultare la sezione “Attinenza tra categorie” per ulteriori informazioni.
- **Conteggio di categoria.** Elenca il numero di categorie a cui il record appartiene.

Attinenza tra categorie

Per facilitare la creazione di categorie migliori, è possibile ricercare attinenza tra documenti o record in ogni categoria e tra tutte le categorie a cui un documento o record appartiene.

Attinenza di una categoria a un record

Ogni volta che un documento o record viene visualizzato nel pannello Dati, tutte le categorie a cui appartiene sono elencate nella colonna Categorie. Quando un documento o record appartiene a più categorie, le categorie in questa colonna vengono visualizzate in ordine a partire dalla corrispondenza più rilevante. La categoria riportata per prima è ritenuta quella che meglio corrisponde a questo documento o record. Per ulteriori informazioni, consultare la sezione “Riquadro Dati” a pagina 110.

Attinenza di un record ad una categoria

Quando si seleziona una categoria, è possibile esaminare l'attinenza di ciascuno dei suoi record nella colonna Classifica di attinenza nel pannello Dati. Questa classificazione di attinenza indica quanto bene il documento o record si inserisce nella categoria selezionata rispetto agli altri record di tale categoria. Per visualizzare la classificazione del record per una singola categoria, selezionare questa categoria nel riquadro Categorie (riquadro superiore sinistro) e la classifica di documento o record viene visualizzata nella colonna. Questa colonna non è visibile per impostazione predefinita ma è possibile scegliere di visualizzarla. Per ulteriori informazioni, consultare la sezione “Riquadro Dati” a pagina 110.

Più basso è il numero per la classificazione del record, migliore è l'adattamento oppure più attinente è questo record alla categoria selezionata, tanto che 1 è la più adeguata. Se più di un record ha la stessa attinenza, ciascuno appare con la stessa classificazione seguita da un segno uguale (=) per indicare che

hanno attinenza uguale. Ad esempio, si potrebbe avere la seguente classifica 1=, 1=, 3, 4, e così via, che significa che vi sono due record che sono ugualmente considerati migliori corrispondenze per questa categoria.

Suggerimento: è possibile aggiungere il testo del record più attinente all'annotazione della categoria per fornire una migliore descrizione della categoria. Aggiungere il testo direttamente dal riquadro Dati selezionando il testo e scegliendo **Categorie > Aggiungi ad annotazione** dai menu.

Creazione di categorie

Mentre si dispone di categorie da un pacchetto di analisi del testo, è possibile anche creare categorie automaticamente utilizzando diverse tecniche linguistiche e di frequenza. Tramite la finestra di dialogo Impostazioni di creazione categorie, è possibile applicare tecniche linguistiche e di frequenza automatizzate per la produzione di categorie da entrambi i concetti o modelli di concetto.

In generale, le categorie possono essere composte da diversi tipi di descrittori (tipi, concetti, modelli TLA, regole di categoria). Quando si creano categorie utilizzando le tecniche di creazione automatica, le categorie risultanti vengono denominate su un modello di concetto o un concetto (a seconda dell'input selezionato), ciascuno contenente una serie di descrittori. Questi descrittori possono essere nel formato di regole di categoria o concetti ed includono tutti i concetti correlati rilevati dalle tecniche.

Dopo la creazione di categorie, è possibile imparare molto sulle categorie attraverso la loro revisione nel riquadro Categorie o l'esplorazione attraverso i grafici e i diagrammi. È possibile poi utilizzare tecniche manuali per effettuare piccoli adattamenti, rimuovere eventuali classificazioni errate o aggiungere record o parole che potrebbero essere stati omessi. Dopo aver applicato una tecnica, i concetti, i tipi e i modelli che sono stati raggruppati in una categoria sono ancora disponibili per altre tecniche. Inoltre, poiché utilizzando tecniche diverse si possono produrre categorie ridondanti o non appropriate, è anche possibile unire o eliminare categorie. Per ulteriori informazioni, consultare “Modifica e perfezionamento delle categorie” a pagina 143.

Importante! In rilasci precedenti, le regole di ricorrenza e sinonimo sono racchiuse tra parentesi quadre. In questo rilascio, le parentesi quadre indicano ora un risultato di modello di analisi di collegamento testo. Invece, le regole di ricorrenza e dei sinonimi verranno racchiuse da parentesi come `(sistemi di altoparlanti|altoparlanti)`.

Per creare le categorie

1. Dal menu, scegliere **Categorie > Crea categorie**. Se non si è scelto di non visualizzare richieste, viene visualizzata una casella di messaggio.
2. Scegliere se si desidera creare ora o modificare prima le impostazioni.

- Fare clic su **Crea ora** per iniziare la creazione di categorie utilizzando le impostazioni correnti. Le impostazioni selezionate per impostazione predefinita sono spesso sufficienti per iniziare il processo di categorizzazione. Il processo di creazione categoria inizia e viene visualizzata una finestra di dialogo di avanzamento.
- Fare clic su **Modifica** per rivedere e modificare le impostazioni di creazione.

Nota: il numero massimo di categorie che può essere visualizzato è 10.000. Viene visualizzato un avviso se questo numero viene raggiunto o superato. In tal caso è necessario modificare le opzioni di creazione o estensione delle categorie per ridurre il numero di categorie create.

Input

Le categorie sono generate dai descrittori derivati dai modelli di tipo o dai tipi. Nella tabella è possibile selezionare i singoli tipi o modelli da includere nel processo di creazione della categoria.

Modelli di tipo. Se si selezionano modelli di tipo, le categorie vengono costruite da modelli e non dai soli tipi o concetti. In questo modo, vengono categorizzati tutti i record o documenti contenenti un modello di concetto appartenente al modello di tipo selezionato. Pertanto, se si seleziona il tipo di modello `<Bilancio>` e `<Positivo>` nella tabella, potrebbero risultare categorie come `costo & <Positivo>` o `tassi & eccellenti`.

Quando si utilizzano i modelli di tipo come input per la creazione automatica della categoria, ci sono momenti in cui le tecniche identificano più modalità per formare la struttura della categoria. Tecnicamente, non vi è un unico modo giusto per produrre le categorie; tuttavia è possibile trovare una struttura più adatta alle proprie analisi rispetto a un'altra. Per facilitare, in questo caso, la personalizzazione dell'output, è possibile designare un tipo come fulcro preferito. Tutte le categorie di livello superiore prodotte scaturiranno da un concetto del tipo selezionato in questo punto (e nessun altro tipo). Ciascuna sottocategoria contiene un modello di collegamento di testo da questo tipo. Scegliere questo tipo nel campo **Struttura categorie in base al tipo di modello:** e la tabella verrà aggiornata in modo da visualizzare solo i modelli applicabili contenenti il tipo selezionato. In genere viene preselezionato `<Sconosciuto>`. In tal modo vengono selezionati tutti i modelli contenenti il tipo `<Sconosciuto>` (per testo non giapponese). . La tabella visualizza i tipi in ordine decrescente a partire da quello con il maggior numero di record o documenti (**Doc.** conteggio).

Tipi. Se si selezionano i tipi, le categorie verranno create dai concetti appartenenti ai tipi selezionati. Pertanto, se si seleziona il tipo `<Bilancio>` nella tabella, potrebbero essere prodotte categorie come `costo` o `prezzo` in quanto `costo` e `prezzo` sono concetti assegnati al tipo `<Bilancio>`.

Per impostazione predefinita, sono selezionati solo i tipi che acquisiscono la maggior parte dei record o documenti . Questa preselezione consente di concentrarsi sui tipi più interessanti ed evitare la creazione di categorie non importanti. La tabella visualizza i tipi in ordine decrescente a partire da quello con il maggior numero di record o documenti (**Doc.** conteggio). I tipi dalla libreria `Opinioni` vengono deselezionati per impostazione predefinita nella tabella dei tipi.

L'input scelto interessa le categorie ottenute. Quando si sceglie di utilizzare i tipi come input, è possibile visualizzare i concetti chiaramente correlati più facilmente. Ad esempio, se si desidera creare categorie utilizzando i tipi come input, è possibile ottenere una categoria `Frutta` con concetti come `mela`, `pera`, `agrumi`, `arancia` e così via. Se si sceglie invece come input i modelli di tipo e si seleziona il modello `<Sconosciuto> + <Positivo>` si potrebbe ottenere, ad esempio, una categoria `frutta + <Positivo>` con uno o due tipi di frutta come `frutta + gustoso` e `mela + buona`. Questo secondo risultato mostra solo 2 modelli di concetto perché le altre ricorrenze di frutta non sono necessariamente qualificate in modo positivo. E mentre ciò potrebbe essere sufficiente per i dati di testo correnti, negli studi longitudinali in cui è possibile utilizzare serie di documenti diverse, è possibile voler aggiungere manualmente altri descrittori come `agrume + positivo` o tipi di utilizzo. Utilizzando solo tipi come input, è più facile trovare tutta la frutta possibile.

Tecniche

Poiché ogni dataset è univoco, il numero di metodi e l'ordine in cui si applicano possono cambiare nel tempo. Inoltre, poiché gli obiettivi di estrazione di testo possono essere diversi da una serie di dati a quella successiva, potrebbe essere necessario sperimentare tecniche differenti per vedere quale produce i migliori risultati per i dati di testo forniti.

Non è necessario essere un esperto di queste impostazioni per poterle utilizzare. Per impostazione predefinita, le impostazioni più comuni e medie sono già selezionate. Pertanto, è possibile ignorare le finestre di impostazione avanzate e andare direttamente a costruire le proprie categorie. Allo stesso modo, se si apportano modifiche in questo punto, non è necessario tornare alla finestra di dialogo delle impostazioni ogni volta poiché le ultime impostazioni vengono sempre conservate.

Selezionare la frequenza o le tecniche linguistiche e fare clic sul pulsante Impostazioni avanzate per visualizzare le impostazioni per le tecniche selezionate. Nessuna delle tecniche automatiche è in grado di

classificare perfettamente i propri dati; di conseguenza si consiglia di rilevare ed applicare una o più tecniche automatiche che funzionano bene con i propri dati. Non è possibile creare utilizzando tecniche linguistiche e di frequenza simultaneamente.

- **Tecniche linguistiche avanzate**. Per ulteriori informazioni, consultare "Impostazioni linguistiche avanzate".
- **Tecniche di frequenza avanzate**. Per ulteriori informazioni, consultare "Impostazioni avanzate di frequenza" a pagina 121.

Impostazioni linguistiche avanzate

Quando si creano categorie, è possibile selezionare tra un certo numero di tecniche linguistiche di creazione categorie avanzate compreso la *derivazione principale di concetto* (non disponibile per giapponese), l'*inclusione di concetto*, le *reti semantiche* (solo testo inglese) e le *regole di ricorrenza*. Queste tecniche possono essere utilizzate singolarmente o in combinazione con altre per creare le categorie.

Ricordare che ogni dataset è univoco e, quindi, il numero di metodi e l'ordine in cui si applicano possono cambiare nel tempo. Inoltre, poiché gli obiettivi di estrazione di testo possono essere diversi da una serie di dati a quella successiva, potrebbe essere necessario sperimentare tecniche differenti per vedere quale produce i migliori risultati per i dati di testo forniti. Nessuna delle tecniche automatiche è in grado di classificare perfettamente i propri dati; di conseguenza si consiglia di rilevare ed applicare una o più tecniche automatiche che funzionano bene con i propri dati.

I seguenti campi e aree sono disponibili all'interno delle Impostazioni avanzate: finestra di dialogo Linguistiche:

Input e output

Input di categoria. Selezionare da cosa verranno create le categorie:

- **Risultati di estrazione inutilizzati**. Questa opzione consente di creare categorie dai risultati di estrazione che non vengono utilizzati in altre categorie esistenti. Ciò riduce la tendenza dei record di corrispondere più categorie e limita il numero di categorie prodotte.
- **Tutti i risultati di estrazione**. Questa opzione consente di creare categorie da tutti i risultati di estrazione. Questo è particolarmente utile quando non esistono già alcune o poche categorie.

Output di categoria. Selezionare la struttura generale per le categorie che verranno create:

- **Gerarchico con sottocategorie**. Questa opzione consente la creazione di categorie secondarie e sottodirectory secondarie. È possibile impostare la profondità delle proprie categorie scegliendo il numero massimo di livelli (campo **Numero massimo di livelli creati**) che possono essere creati. Se si sceglie 3, le categorie potrebbero contenere sottocategorie e anche quelle secondarie possono avere sottocategorie.
- **Categorie semplici (solo livello singolo)**. Questa opzione consente solo un livello di categorie da creare, ossia non verranno generate sottocategorie.

Tecniche di raggruppamento

Ciascuna delle tecniche disponibili si adatta adeguatamente ad alcuni tipi di dati e situazioni, ma spesso è utile combinare le tecniche nella stessa analisi e acquisire la gamma completa di documenti o record. Ciò significa che è possibile visualizzare un concetto in più categorie o trovare categorie ridondanti.

Derivazione principale di concetto. Questa tecnica crea categorie prendendo un concetto e cercando altri concetti che sono in relazione con esso analizzando se uno dei componenti del concetto sono morfologicamente correlati. Questa tecnica è molto utile per identificare i concetti di parola composta sinonimo, poiché i concetti in ciascuna categoria generata sono sinonimi o strettamente correlati nel significato. Essa lavora con dati di lunghezza variabile e genera un numero inferiore di categorie compatte. Ad esempio, la nozione di `opportunità di avanzare` potrebbe essere raggruppata con i concetti

di opportunità per l'avanzamento e opportunità di avanzamento. Consultare la sezione "Derivazione principale di concetto" a pagina 117 per ulteriori informazioni. Questa opzione non è disponibile per il testo giapponese.

Rete semantica. Questa tecnica inizia individuando i possibili sensi di ciascun concetto dal suo ampio indice di relazioni di parole e poi crea le categorie raggruppando concetti correlati. Questa tecnica è migliore quando i concetti sono noti alla rete semantica e non sono troppo ambigui. La tecnica è meno utile quando il testo contiene terminologia specialistica o gergo sconosciuto alla rete. In un esempio, il concetto mela della nonna potrebbe essere raggruppato con mela gala e mela poiché sono discendenti del primo concetto. In un altro esempio, il concetto animale potrebbe essere raggruppato con gatto e canguro poiché sono iponimi di animale. Questa tecnica è disponibile solo per il testo inglese in questo rilascio del prodotto. Consultare la sezione "Reti semantiche" a pagina 119 per ulteriori informazioni.

Inclusione di concetto. Questa tecnica crea categorie raggruppando concetti di più termini (parole composte) basata su se i concetti contengono parole che sono sottoserie o superserie di una parola nell'altra. Ad esempio, il concetto sedile viene raggruppato con sedile di sicurezza, cinture di sicurezza e fibbia di cintura di sicurezza. Consultare la sezione "Inclusione concetti" a pagina 118 per ulteriori informazioni.

Ricorrenza. Questa tecnica crea categorie da ricorrenze trovate nel testo. L'idea è che quando i concetti vengono spesso trovati insieme in documenti e record, tale ricorrenza riflette una relazione sottostante che ha probabilmente un valore nelle definizioni di categoria. Quando le parole ricorrono in modo significativo, viene creata una regola di ricorrenza e può essere utilizzata come descrittore di categoria per una sottocategoria nuova. Ad esempio, se molti record contengono le parole prezzo e disponibilità, questi concetti possono essere raggruppati in una regola di ricorrenza, (prezzo & disponibile) e assegnati, ad esempio, ad una sottocategoria della categoria prezzo. Consultare la sezione "Regole di ricorrenza" a pagina 120 per ulteriori informazioni.

Numero minimo di documenti Per determinare quanto interessanti possono essere le ricorrenze, definire il numero minimo di documenti o record che devono contenere una data ricorrenza per essere utilizzati come descrittori in una categoria.

Distanza massima di ricerca. Selezionare quanto si desidera che le tecniche effettuino la ricerca prima di produrre le categorie. Più basso è il valore, meno risultati vengono prodotti; e comunque questi risultati saranno meno clamorosi e avranno maggiori probabilità di essere significativamente associati o collegati tra loro. Maggiore è il valore, più risultati si ottengono; tuttavia questi risultati possono essere meno affidabili o pertinenti. Sebbene questa opzione venga applicata globalmente a tutte le tecniche, il suo effetto è maggiore sulle ricorrenze e reti semantiche.

Previeni l'abbinamento di concetti specifici. Selezionare questa casella di spunta per arrestare il processo di raggruppamento o di accoppiamento di due concetti insieme nell'output. Per creare o gestire coppie di concetto, fare clic su **Gestisci Coppie...**. Consultare la sezione "Gestione delle coppie di eccezione di collegamento" a pagina 116 per ulteriori informazioni.

Generalizzazione con caratteri jolly dove possibile. Selezionare questa opzione per consentire al prodotto di generare regole generiche in categorie utilizzando il carattere jolly asterisco. Ad esempio, invece di produrre più descrittori come [torta di mele + .] e [succo di mela + .], utilizzando caratteri jolly potrebbe produrre [mela * + .]. Se è necessario generalizzare con caratteri jolly, sarà spesso necessario ottenere esattamente lo stesso numero di record o documenti come è stato fatto in precedenza. Tuttavia, questa opzione ha il vantaggio di ridurre il numero e semplificare i descrittori di categoria. Inoltre, questa opzione aumenta la possibilità di classificare più record o documenti mediante queste categorie su nuovi dati di testo (ad esempio, in studi longitudinale/onda).

Altre opzioni per la creazione di categorie

Oltre alla selezione del raggruppamento delle tecniche da applicare, è possibile modificare diverse altre opzioni di creazione nel modo seguente:

Numero massimo di categorie create di livello superiore. Utilizzare questa opzione per limitare il numero di categorie che possono essere generate quando si fa clic successivamente sul pulsante Crea categorie. In alcuni casi, si potrebbero ottenere risultati migliori se si imposta questo valore elevato e poi si eliminano le categorie meno interessanti.

Numero minimo di descrittori e/o sottocategorie per categoria. Utilizzare questa opzione per definire il numero minimo di descrittori e sottocategorie che una categoria deve contenere per essere creata. Questa opzione consente di limitare la creazione di categorie che non catturano un numero significativo di record o documenti.

Consenti descrittori in più di una categoria. Se selezionata, questa opzione consente ai descrittori di essere utilizzati in più di una delle categorie che verranno create successivamente. Questa opzione è selezionata in genere dal momento in cui gli elementi comunemente o "naturalmente" rientrano in due o più categorie e consentendo loro di farlo di solito porta a categorie di qualità superiore. Se non si seleziona questa opzione, si riduce la sovrapposizione dei record in più categorie e a seconda del tipo di dati si dispone, ciò potrebbe essere auspicabile. Tuttavia, con la maggior parte dei tipi di dati, limitando i descrittori ad una singola categoria solitamente comporta una perdita di qualità o di copertura di categoria. Ad esempio, si supponga di avere il concetto di `fabbrica di sedili per auto`. Con questa opzione, questo concetto potrebbe apparire in una categoria basata sul testo `sedile per auto` e in un'altra basata su `fabbrica`. Ma se questa opzione non è selezionata, anche se si possono ancora ottenere entrambe le categorie, il concetto di `fabbrica di sedili per auto` verrà visualizzato solo come descrittore nella categoria a cui meglio corrisponde, in base a diversi fattori incluso il numero di record in cui `sedili per auto` e `fabbrica` ricorrono.

Risoluzione nomi di categoria duplicati con. Selezionare la modalità di gestione di eventuali nuove categorie o sottocategorie i cui nomi sono gli stessi di categorie esistenti. È possibile unire le nuove (e i relativi descrittori) con le categorie esistenti che hanno lo stesso nome. In alternativa, è possibile scegliere di ignorare la creazione di qualsiasi categoria se nelle categorie esistenti viene trovato un nome duplicato.

Gestione delle coppie di eccezione di collegamento

Durante la creazione , il raggruppamento in cluster della categoria e l'associazione di concetti, gli algoritmi interni raggruppano le parole in base ad associazioni riconosciute. Per evitare l'accoppiamento o il collegamento di due concetti, è possibile attivare questa funzione in **Impostazioni avanzate di Crea categorie**, **Crea cluster** e **Impostazioni di indice di associazione concetti** e fare clic sul pulsante **Gestisci Coppie**.

Nella finestra di dialogo risultante **Gestisci eccezioni di collegamento**, è possibile aggiungere, modificare o eliminare coppie di concetti. Immettere una coppia per riga. L'inserimento di coppie impedirà il verificarsi dell'accoppiamento quando si creano o si estendono le categorie, il raggruppamento in cluster e l'associazione dei concetti. Immettere le parole esattamente come si desidera, ad esempio la versione con accento delle parole non è uguale alla versione senza accento.

Ad esempio, se si desidera assicurare che `hot dog` e `dog` non sono raggruppati, è possibile aggiungere la coppia come una riga separata nella tabella.

Informazioni sulle tecniche linguistiche

Quando si creano o si estendono categorie, è possibile selezionare tra un certo numero di tecniche linguistiche di creazione categorie avanzate compreso la *derivazione principale di concetto* (non disponibile per giapponese), l'*inclusione di concetto*, le *reti semantiche* (solo testo inglese) e le *regole di ricorrenza*. Queste tecniche possono essere utilizzate singolarmente o in combinazione con altre per creare le categorie.

Non è necessario essere un esperto di queste impostazioni per poterle utilizzare. Per impostazione predefinita, le impostazioni più comuni e medie sono già selezionate. Pertanto, è possibile ignorare le

finestre di impostazione avanzate e andare direttamente a costruire o estendere le proprie categorie. Allo stesso modo, se si apportano modifiche in questo punto, non è necessario tornare alla finestra di dialogo delle impostazioni ogni volta poiché le ultime impostazioni vengono sempre memorizzate.

Ricordare tuttavia che ogni dataset è univoco e, quindi, il numero di metodi e l'ordine in cui si applicano possono cambiare nel tempo. Inoltre, poiché gli obiettivi di estrazione di testo possono essere diversi da una serie di dati a quella successiva, potrebbe essere necessario sperimentare tecniche differenti per vedere quale produce i migliori risultati per i dati di testo forniti. Nessuna delle tecniche automatiche è in grado di classificare perfettamente i propri dati; di conseguenza si consiglia di rilevare ed applicare una o più tecniche automatiche che funzionano bene con i propri dati.

Le principali tecniche linguistiche automatizzate per la creazione di categoria sono:

- **Derivazione principale di concetto.** Questa tecnica crea categorie prendendo un concetto e cercando altri concetti che sono in relazione con esso analizzando se uno dei componenti del concetto sono morfologicamente correlati. Consultare la sezione "Derivazione principale di concetto" per ulteriori informazioni. Questa opzione non è disponibile per il testo giapponese.
- **Inclusione di concetto.** Questa tecnica crea categorie prendendo un concetto e cercando altri concetti che lo includono. Consultare la sezione "Inclusione concetti" a pagina 118 per ulteriori informazioni.
- **Rete semantica.** Questa tecnica inizia individuando i possibili sensi di ciascun concetto dal suo ampio indice di relazioni di parole e poi crea le categorie raggruppando concetti correlati. Consultare la sezione "Reti semantiche" a pagina 119 per ulteriori informazioni. Questa opzione è disponibile solo per testo inglese.
- **Ricorrenza.** Questa tecnica crea regole di ricorrenza che è possibile utilizzare per creare una nuova categoria, estendere una categoria o come input per un'altra tecnica di categoria. Consultare la sezione "Regole di ricorrenza" a pagina 120 per ulteriori informazioni.

Derivazione principale di concetto

Nota: questa tecnica non è disponibile per il testo giapponese.

La tecnica di derivazione principale di concetto crea categorie prendendo un concetto e cercando altri concetti che sono in relazione con esso analizzando se uno dei componenti del concetto sono morfologicamente correlati. Un componente è una parola. La tecnica tenta di raggruppare concetti guardando alla fine (i suffissi) di ogni componente in un concetto e trovando altri concetti che possono essere derivati da essi. L'idea è che quando le parole derivano l'una dall'altra, hanno maggiore probabilità di condividere o di essere vicini di significato. Per identificare il fine, vengono utilizzate regole interne specifiche della lingua. Ad esempio, la nozione di `opportunità di avanzare` potrebbe essere raggruppata con i concetti di `opportunità per l'avanzamento` e `opportunità di avanzamento`.

È possibile utilizzare la derivazione principale di concetto su qualsiasi tipo di testo. Di per sé, produce piuttosto poche categorie e ciascuna categoria tende a contenere pochi concetti. I concetti in ciascuna categoria sono sinonimi o correlati nelle situazioni. Può essere utile utilizzare questo algoritmo anche se si creano le categorie manualmente; i sinonimi trovati possono essere sinonimi di quei concetti a cui si è particolarmente interessati.

Nota: è possibile impedire ai concetti di essere raggruppati insieme specificandoli esplicitamente. Consultare la sezione "Gestione delle coppie di eccezione di collegamento" a pagina 116 per ulteriori informazioni.

Suddivisione in componenti e deinflessione

Quando vengono applicate le tecniche di derivazione principale di concetto o di inclusione concetto, i termini vengono prima suddivisi in componenti (parole) e quindi i componenti vengono reintegrati. Quando viene applicata una tecnica, i concetti e i termini associati vengono caricati e suddivisi in componenti basati su separatori, come gli spazi, trattini e apostrofi. Ad esempio, il termine `amministratore di sistema` viene suddiviso in componenti come ad esempio {`amministratore`, `sistema`}.

Tuttavia, alcune parti del termine originale potrebbero non essere utilizzate e sono considerate parole di arresto. In italiano, alcuni di questi componenti non importanti potrebbero includere un, e, come, da, per, da, in, di, su, o, il, a e con.

Ad esempio, il termine esame dei dati ha una serie di componenti {dati, esame} e di e il sono considerati non importanti. Inoltre, l'ordine del componente non è in una serie di componenti. In questo modo, i seguenti tre termini potrebbero essere equivalenti: sciroppo per la tosse per bambini, sciroppo per bambini per la tosse e sciroppo per la tosse dei bambini poiché essi contengono tutti la stessa serie di componenti {bambino, tosse, sciroppo}. Ogni volta che una coppia di termini vengono identificati come equivalenti, i concetti corrispondenti vengono uniti per formare un nuovo concetto che faccia riferimento a tutti i termini.

Inoltre, poiché i componenti di un termine possono flettere, vengono applicate regole specifiche della lingua per identificare termini equivalenti indipendentemente dalla variazione di inflessione, come i moduli plurali. In questo modo, i termini livello di supporto e livelli di supporto possono essere identificati come equivalenti poiché il modulo singolare riunito sarebbe livello.

Come funziona la derivazione principale di concetto

Dopo che i termini sono stati suddivisi in componenti e poi riuniti (vedere la sezione precedente), l'algoritmo di derivazione principale di concetto analizza i fine componente o suffissi, per trovare la radice del componente e poi raggruppa i concetti con altri concetti che hanno radici uguali o simili. I fine componente sono identificati mediante una serie di regole di derivazione linguistiche specifiche per la lingua del testo. Ad esempio, è presente una regola di derivazione per il testo in lingua inglese che afferma che un componente di concetto che termina con il suffisso ical può essere ricavato da un concetto avente la stessa radice e che termina con il suffisso ic. Utilizzando questa regola (e la de-inflessione), l'algoritmo sarà in grado di raggruppare i concetti di epidemiologic study e epidemiological studies.

Poiché i termini sono già suddivisi in componenti e i componenti non importanti (ad esempio, in e di) sono stati individuati, l'algoritmo di derivazione deve essere inoltre in grado di raggruppare il concetto di studies in epidemiology con epidemiological studies.

La serie di regole di derivazione del componente è stato scelta in modo che la maggior parte dei concetti raggruppati da questo algoritmo sono sinonimi: i concetti di epidemiologic studies, epidemiological studies, studies in epidemiology sono tutti termini equivalenti. Per aumentare la completezza, vi sono alcune regole di derivazione che consentono all'algoritmo di raggruppare concetti che sono correlati per situazioni. Ad esempio, l'algoritmo può raggruppare concetti del gruppo come empire builder e empire building.

Inclusione concetti

La tecnica di inclusione concetto crea categorie prendendo un concetto e, utilizzando gli algoritmi di serie lessicali, identifica i concetti inclusi in altri concetti. L'idea è che quando le parole in un concetto sono un sottoinsieme di un altro concetto, si riflette un rapporto semantico sottostante. L'inclusione è una tecnica efficace che può essere utilizzata con qualsiasi tipo di testo.

Questa tecnica funziona bene in combinazione con le reti semantica ma può essere utilizzata separatamente. L'inclusione concetto può anche dare risultati migliori quando i documenti o record contengono molta terminologia o gergo del dominio. Questo vale soprattutto se si dispone di ottimizzare i dizionari, in modo che i termini speciali vengono estratti e raggruppati in modo appropriato (con sinonimi).

Come funziona l'inclusione di concetto

Prima che venga applicato l'algoritmo di inclusione concetto, i termini sono suddivisi in componenti e senza inflessione. Consultare la sezione “Derivazione principale di concetto” a pagina 117 per ulteriori

informazioni. Successivamente, l'algoritmo di inclusione di concetto analizza le serie di componenti. Per ogni serie di componenti, l'algoritmo cerca un'altra serie di componenti che è una sottoserie della prima serie di componenti.

Ad esempio, se si ha il concetto `colazione continentale`, che ha il componente di `{colazione, continentale}` ed è presente il concetto `colazione`, che ha la serie di componenti `{colazione}`, l'algoritmo potrebbe concludere che `colazione continentale` è un tipo di `colazione` e raggrupparli insieme.

In un esempio più esteso, nella lingua inglese, se si ha il concetto di `seat` nel riquadro Risultati di estrazione e si applica questo algoritmo, anche i concetti quali `safety seat`, `leather seat`, `seat belt`, `seat belt buckle`, `infant seat carrier` e `car seat laws` verranno raggruppati in quella categoria.

Poiché i termini sono già suddivisi in componenti e i componenti non importanti (ad esempio, `in` e `di`) sono stati identificati, l'algoritmo di inclusione concetto riconosce che il concetto `corso di spagnolo avanzato` include il concetto di `corso di spagnolo`.

Nota: è possibile impedire ai concetti di essere raggruppati insieme specificandoli esplicitamente. Consultare la sezione “Gestione delle coppie di eccezione di collegamento” a pagina 116 per ulteriori informazioni.

Reti semantiche

In questo release, la tecnica delle reti semantiche è disponibile solo per il testo in lingua inglese.

Questa tecnica crea le categorie utilizzando relazioni di reti di parole integrate. Per questo motivo, questa tecnica può produrre risultati molto buoni quando i termini sono concreti e non sono troppo ambigui. Tuttavia, si consiglia di non aspettarsi la tecnica per trovare molti collegamenti tra concetti molto tecnici/specialistici. Quando si affrontano tali concetti, possono essere utili il concetto di inclusione e le tecniche di derivazione principale di concetto.

Come funziona la rete semantica

L'idea della tecnica di rete semantica è di alzare i livello significativo delle relazioni conosciute tra parole in modo da creare categorie di sinonimi o iponimi. Per **iponimo** si intende quando un concetto è una sorta di secondo concetto in cui esiste una relazione gerarchica, nota anche come relazione ISA. Ad esempio, se `animale` è un concetto, `gatto` e `canguro` sono iponimi di `animale` poiché sono tipi di animali.

Oltre a relazioni di sinonimi e iponimi, la tecnica di rete semantica esamina inoltre una parte e per intero i collegamenti tra qualsiasi concetto dal tipo `<Ubicazione>`. Ad esempio, la tecnica raggrupperà i concetti di `normandia`, `provenzae francia` in un'unica categoria perché Normandia e Provenza sono parti della Francia.

Le reti semantiche iniziano con l'identificare i possibili sensi di ciascun concetto nella rete semantica. Quando i concetti sono identificati come sinonimi o iponimi, essi vengono raggruppati in una singola categoria. Ad esempio, la tecnica potrebbe creare una categoria singola contenente questi tre concetti: `mangiare la mela`, `mela per dessert` e `barilla` poiché la rete semantica contiene le informazioni che: 1) `mela per dessert` è un sinonimo di `mangiare la mela` e 2) `barilla` è una sorta di `mangiare la mela` (ovvero è iponimo di `mangiare la mela`).

Presi singolarmente, molti concetti, in particolare i termini univoci, sono ambigui. Ad esempio, il concetto di `buffet` può indicare una sorta di pasto o un pezzo di arredamento. Se la serie di concetti include `pasto`, `mobili` e `buffet`, l'algoritmo è costretto a scegliere tra il raggruppamento di `buffet` con `pasto` o con `mobili`. Si noti che in alcuni casi le scelte effettuate dall'algoritmo potrebbero non essere appropriate nel contesto di un particolare insieme di record o documenti.

La tecnica di rete semantica può essere più efficace dell'inclusione concetto con certi tipi di dati. Mentre sia la rete semantica che l'inclusione concetto riconoscono che `torta di mele` è una specie di `torta`, solo la rete semantica riconosce che anche `crostata` è una specie di `torta`.

Le reti semantiche lavorano insieme ad altre tecniche. Ad esempio, si supponga di avere selezionato ambedue le tecniche di rete semantica e di inclusione e che la rete semantica ha raggruppato il concetto `insegnante` con il concetto di `tutor` (poiché il tutor è un tipo di insegnante). L'algoritmo di inclusione può raggruppare il concetto `tutor laureato` con `tutor` e, come risultato, i due algoritmi collaborano a produrre una categoria di output che contiene tutti e tre i concetti: `tutor`, `tutor laureato` e `insegnante`.

Opzioni per la rete semantica

Sono possibili diverse impostazioni aggiuntive che potrebbero essere di interesse con questa tecnica.

- Modificare la **distanza massima di ricerca**. Selezionare quanto si desidera che le tecniche effettuino la ricerca prima di produrre le categorie. Più basso è il valore, meno risultati vengono prodotti; e comunque questi risultati saranno meno clamorosi e avranno maggiori probabilità di essere significativamente associati o collegati tra loro. Maggiore è il valore, più risultati si ottengono; tuttavia questi risultati possono essere meno affidabili o pertinenti.

Ad esempio, a seconda della distanza, l'algoritmo ricerca da `Sfogliatella` fino a `cornetto` (suo principale), quindi `brioscina` (derivato) e verso l'alto con `pane`.

Riducendo la distanza di ricerca, questa tecnica produce categorie più piccole che potrebbero essere più semplici da gestire se si ritiene che le categorie prodotte sono troppo grandi o raggruppano troppe cose insieme.

Importante! Si raccomanda inoltre di non applicare l'opzione **Correggi errori ortografici per un limite minimo di caratteri principali** (definita nella scheda Livello avanzato del nodo onella finestra di dialogo Estrai) per il raggruppamento indistinto quando si utilizza questa tecnica, perché alcuni falsi raggruppamenti possono avere un impatto in larga misura negativo sui risultati.

Regole di ricorrenza

Le regole di ricorrenza consentono di rilevare e raggruppare concetti che sono strettamente correlati all'interno della serie di documenti o record. L'idea è che quando i concetti vengono spesso trovati insieme in documenti e record, tale ricorrenza si riflette una relazione sottostante che ha probabilmente un valore nelle definizioni di categoria. Questa tecnica crea regole di ricorrenza che è possibile utilizzare per creare una nuova categoria, estendere una categoria o come input per un'altra tecnica di categoria. I due concetti sono strettamente connessi se frequentemente vengono visualizzati insieme in una serie di record e raramente separatamente in tutti gli altri record. Questa tecnica può produrre buoni risultati con insiemi di dati più estesi con almeno varie centinaia di documenti o record.

Ad esempio, se molti record contengono le parole `prezzo` e `disponibilità`, questi concetti possono essere raggruppati in una regola di ricorrenza, `(prezzo & disponibile)`. In un altro esempio, se i concetti di `burro`, `marmellata`, `sandwich` e vengono visualizzate più spesso insieme a parte, essi dovrebbero essere raggruppati in un concetto di regola di ricorrenza `(burro & marmellata & sandwich)`.

Importante! In rilasci precedenti, le regole di ricorrenza e sinonimo sono racchiuse tra parentesi quadre. In questo rilascio, le parentesi quadre indicano ora un risultato di modello di analisi di collegamento testo. Invece, le regole di ricorrenza e dei sinonimi verranno racchiuse da parentesi come `(sistemi di altoparlanti|altoparlanti)`.

Come funzionano le regole di ricorrenza

Questa tecnica analizza i documenti o record per due o più concetti che tendono a comparire insieme. Due o più concetti ricorrono molto di frequente se essi compaiono insieme in una serie di documenti o record e se raramente vengono visualizzati separatamente in altri documenti o record.

Quando vengono trovati concetti ricorrenti, si costituisce una regola di categoria. Tali regole sono costituite da due o più concetti connessi che utilizzano l'operatore booleano &. Tali regole sono istruzioni logiche che classificano automaticamente un documento o record in una categoria se la serie di concetti nella regola concorrono tutti in questo documento o record.

Le opzioni per le regole di ricorrenza

Se si sta utilizzando la tecnica della regola di ricorrenza, è possibile regolare diverse impostazioni che influenzano le regole risultanti:

- Modificare la **distanza massima di ricerca**. Selezionare in che misura si desidera che la tecnica ricerchi le ricorrenze. Poiché è possibile aumentare la distanza di ricerca, il valore di similitudine minima richiesta per ogni ricorrenza viene abbassato; come risultato, possono essere prodotte molte regole di ricorrenza, ma quelle che hanno un basso valore di somiglianza sono spesso di poca importanza. Man mano che si riduce la distanza di ricerca, il valore minimo di similarità richiesta aumenta; come risultato, vengono prodotte poche regole di ricorrenza che, però, tenderanno ad essere più significative (più forti).
- **Numero minimo di documenti.** Il numero minimo di record o documenti che devono contenere una determinata coppia di concetti per essere considerati come una ricorrenza; più basso è il valore per questa opzione, più facilmente emergono le ricorrenze. L'aumento del valore ha come risultato minori, ma più importanti, ricorrenze. Ad esempio, si supponga che i concetti "mela" e "pera" vengono trovati insieme in 2 record (e che nessuno dei due concetti si verifica in nessun altro record). Con **Numero minimo di documenti** impostato su 2 (valore predefinito), la tecnica di ricorrenza crea una regola di categoria (mela e pera). Se il valore è portato a 3, la regola non sarà più creata.

Nota: con dataset di piccole dimensioni (< 1000 risposte) è possibile che non vengano trovate ricorrenze con le impostazioni predefinite. In tal caso, tentare di aumentare il valore della distanza di ricerca.

Nota: è possibile impedire ai concetti di essere raggruppati insieme specificandoli esplicitamente. Consultare la sezione “Gestione delle coppie di eccezione di collegamento” a pagina 116 per ulteriori informazioni.

Impostazioni avanzate di frequenza

È possibile creare categorie in base alla tecnica di frequenza diretta e meccanica. Con questa tecnica, è possibile creare una sola categoria per ogni elemento (tipo, concetto o modello) che è stato trovato al di sopra di un determinato conteggio di record o documento. Inoltre, è possibile creare una singola categoria per tutti gli elementi che ricorrono meno di frequente. Per il conteggio, si fa riferimento al numero di record o documenti contenenti il concetto (e tutti i suoi sinonimi), tipo o modello estratti in questione, in opposizione al numero totale di ricorrenze nell'intero testo.

Raggruppando di frequente elementi ricorrenti è possibile produrre risultati interessanti, perché potrebbe indicare una risposta comune o importante. La tecnica è molto utile sui risultati di estrazione non utilizzati dopo che sono state applicate altre tecniche. Un'altra applicazione deve eseguire questa tecnica immediatamente dopo l'estrazione quando non esistono altre categorie, modificare i risultati per eliminare categorie non importanti e quindi estendere tali categorie in modo che corrispondano ad ancora più record o documenti. Consultare la sezione “Estensione delle categorie” a pagina 122 per ulteriori informazioni.

Invece di utilizzare questa tecnica, è possibile ordinare i concetti o i modelli di concetto per numero discendente di record o documenti nel riquadro Risultati di estrazione e quindi trascinare e rilasciare i primi nel riquadro Categorie per creare le categorie corrispondenti.

I seguenti campi sono disponibili all'interno delle Impostazioni avanzate: finestra di dialogo Frequenze:

Genera descrittori di categoria. Selezionare il tipo di input per i descrittori. Consultare la sezione “Creazione di categorie” a pagina 112 per ulteriori informazioni.

- **Livelli di concetti**. La selezione di questa opzione comporta che verranno utilizzate frequenze di concetti o di modelli di concetti. I concetti verranno utilizzati se sono state selezionati i tipi come input per la creazione di categorie e vengono utilizzati i modelli di concetto se sono stati selezionati modelli di tipo. In generale, applicando questa tecnica al livello di concetto si produrranno risultati più specifici, poiché i concetti e i modelli di concetto rappresentano un livello inferiore di misurazione.
- **Livelli di tipi**. La selezione di questa opzione comporta che verranno utilizzate frequenze di tipo o di modelli di tipo. I tipi verranno utilizzati se sono state selezionati i tipi come input per la creazione di categorie e vengono utilizzati i modelli di tipo se sono stati selezionati modelli di tipo. L'applicazione di questa tecnica al livello di tipo consente di ottenere una vista rapida relativa al tipo di informazioni presenti fornite.

Minimo doc. conteggio per gli elementi che hanno una propria categoria. Questa opzione consente di creare categorie da elementi che ricorrono di frequente. Questa opzione restringe l'output a solo quelle categorie che contengono un descrittore che ricorre in almeno un X numero di record o documenti, dove X è il valore da immettere per questa opzione.

Raggruppare tutte gli elementi rimanenti in una categoria richiamata. Questa opzione consente di raggruppare tutti i concetti o tipi che non ricorrono frequentemente in un'unica categoria 'pigliatutto' con il nome di propria scelta. Per impostazione predefinita, questa categoria si chiama *Altro*.

Input di categoria. Selezionare il gruppo a cui applicare le tecniche:
- **Risultati di estrazione inutilizzati**. Questa opzione consente di creare categorie dai risultati di estrazione che non vengono utilizzati in altre categorie esistenti. Ciò riduce la tendenza dei record di corrispondere più categorie e limita il numero di categorie prodotte.
- **Tutti i risultati di estrazione**. Questa opzione consente di creare categorie da tutti i risultati di estrazione. Questo è particolarmente utile quando non esistono già alcune o poche categorie.

Risoluzione nomi di categoria duplicati. Selezionare la modalità di gestione di eventuali nuove categorie o sottocategorie i cui nomi sono gli stessi di categorie esistenti. È possibile unire le nuove (e i relativi descrittori) con le categorie esistenti con lo stesso nome. In alternativa, è possibile scegliere di ignorare la creazione di qualsiasi categoria se nelle categorie esistenti viene trovato un nome duplicato.

Estensione delle categorie

L'estensione è un processo mediante il quale vengono aggiunti o migliorati automaticamente descrittori per 'far crescere' categorie esistenti. L'obiettivo è quello di elaborare una categoria migliore che catturi i record correlati o documenti che non erano originariamente assegnati a tale categoria.

Le tecniche di raggruppamento automatico selezionate tenteranno di identificare i concetti, i modelli TLA e le regole di categoria correlate ai descrittori della categoria esistente. Questi nuovi concetti, modelli e regole di categoria vengono quindi aggiunti come nuovi descrittori o aggiunti ai descrittori esistenti. Le tecniche di raggruppamento per l'estensione includono *concetto di derivazione principale* (non disponibile per giapponese), *concetto di inclusione*, *reti semantiche* (solo inglese) e *regole di ricorrenza*. Il metodo **Estendi categorie vuote con i descrittori generati dal nome categoria** genera descrittori utilizzando le parole nei nomi categoria, quindi, più i nomi categoria sono descrittivi, tanto migliori sono i risultati.

Nota: le tecniche di frequenza non sono disponibili quando si estendono le categorie.

L'estensione è un ottimo modo per migliorare le categorie in modo interattivo. Di seguito sono riportati alcuni esempi di quando si può estendere una categoria:
- Dopo aver trascinato/rilasciato modelli di concetto per creare le categorie nel riquadro Categorie
- Dopo la creazione di categorie manualmente e l'aggiunta di regole semplici di categoria e di descrittori
- Dopo l'importazione di un file di categoria predefinito in cui le categorie hanno nomi molto descrittivi
- Dopo aver perfezionato le categorie che provengono dal TAP scelto

È possibile estendere la categoria più volte. Ad esempio, se è stato importato un file di categoria predefinito con nomi molto descrittivi, è possibile estendere utilizzando l'opzione **Estendi categorie vuote con i descrittori generati dal nome categoria** per ottenere un primo gruppo di descrittori, e quindi estendere di nuovo tali categorie. Tuttavia, in altri casi, estendendo più volte, il risultato potrebbe essere una categoria troppo generica se i descrittori vengono estesi in modo sempre più ampio. Poiché la creazione e l'estensione di tecniche di raggruppamento utilizzano algoritmi simili sottostanti, è improbabile che l'estensione diretta dopo la creazione di categorie produca risultati più interessanti.

Suggerimenti:

- Se si tenta di estendere e non si desidera utilizzare i risultati, è sempre possibile annullare l'operazione (**Modifica > Annulla**) subito dopo aver eseguito l'estensione.
- L'estensione può produrre due o più regole di categoria in una categoria che corrispondono esattamente alla stessa serie di documenti poiché le regole vengono create in modo indipendente durante il processo. Se si desidera, è possibile esaminare le categorie ed eliminare ridondanze modificando manualmente la descrizione della categoria. Consultare la sezione “Modifica dei descrittori di categoria” a pagina 144 per ulteriori informazioni.

Per estendere le categorie

1. Nel riquadro Categorie, selezionare la categoria che si desidera estendere.
2. Dal menu, scegliere **Categorie > Estendi categorie**. Se non si è scelto di non visualizzare richieste, viene visualizzata una casella di messaggio.
3. Scegliere se si desidera creare ora o modificare prima le impostazioni.

- Fare clic su **Estendi ora** per iniziare la creazione di categorie utilizzando le impostazioni correnti. Il processo di creazione categoria inizia e viene visualizzata una finestra di dialogo di avanzamento.
- Fare clic su **Modifica** per rivedere e modificare le impostazioni.

Dopo aver tentato l'estensione, tutte le categorie per le quali vengono rilevati nuovi descrittori vengono segnalate con la parola **Esteso** nel riquadro Categorie, in modo da poterli identificare rapidamente. Il testo Esteso resta finché non si estende di nuovo, non si modifica la categoria in un altro modo o non si deselezionano queste categorie tramite il menu di scelta rapida.

Nota: il numero massimo di categorie che può essere visualizzato è 10.000. Viene visualizzato un avviso se questo numero viene raggiunto o superato. In tal caso è necessario modificare le opzioni di creazione o estensione delle categorie per ridurre il numero di categorie create.

Ciascuna delle tecniche disponibili durante la creazione o l'estensione delle categorie è adatto ad alcuni tipi di dati e situazioni, ma spesso è utile per combinare le tecniche nella stessa analisi e acquisire la gamma completa di documenti o record. Nel workbench interattivo i concetti e i tipi che sono stati raggruppati in una categoria sono ancora disponibili la volta successiva in cui si desidera creare le categorie. Ciò significa che è possibile visualizzare un concetto in più categorie o trovare categorie ridondanti.

All'interno di Estendi categorie, finestra di dialogo Impostazioni, sono disponibili i seguenti campi e aree:

Estendi con. Selezionare l'input che verrà utilizzato per estendere le categorie:

- **Risultati di estrazione inutilizzati**. Questa opzione consente di creare categorie dai risultati di estrazione che non vengono utilizzati in altre categorie esistenti. Ciò riduce la tendenza dei record di corrispondere più categorie e limita il numero di categorie prodotte.
- **Tutti i risultati di estrazione**. Questa opzione consente di creare categorie da tutti i risultati di estrazione. Questo è particolarmente utile quando non esistono già alcune o poche categorie.

Tecniche di raggruppamento

Per le descrizioni brevi di ciascuna di queste tecniche, consultare “Impostazioni linguistiche avanzate” a pagina 114. Queste tecniche includono:

- **Derivazione principale di concetto** *(non disponibile per giapponese)*
- **Reti semantiche** (*solo testo inglese e non utilizzato se viene selezionata solo l'opzione Generalizza.*)
- **Inclusione di concetti**
- Sottopzione **Ricorrenza** e **Numero minimo di documenti.**.

Un numero di tipi sono permanentemente esclusi dalla tecnica di reti semantiche poiché questi tipi non producono risultati pertinenti. Essi includono `<Positivo>`, `<Negativo>`, `<IP>`, altri tipi non linguistici, ecc.

Distanza massima di ricerca. Selezionare quanto si desidera che le tecniche effettuino la ricerca prima di produrre le categorie. Più basso è il valore, meno risultati vengono prodotti; e comunque questi risultati saranno meno clamorosi e avranno maggiori probabilità di essere significativamente associati o collegati tra loro. Maggiore è il valore, più risultati si ottengono; tuttavia questi risultati possono essere meno affidabili o pertinenti. Sebbene questa opzione venga applicata globalmente a tutte le tecniche, il suo effetto è maggiore sulle ricorrenze e reti semantiche.

Previeni l'abbinamento di concetti specifici. Selezionare questa casella di spunta per arrestare il processo di raggruppamento o di accoppiamento di due concetti insieme nell'output. Per creare o gestire coppie di concetto, fare clic su **Gestisci Coppie...**. Consultare la sezione “Gestione delle coppie di eccezione di collegamento” a pagina 116 per ulteriori informazioni.

Dove possibile: scegliere se estendere semplicemente, generalizzare i descrittori utilizzando caratteri jolly o entrambi.

- **Estendi e generalizza.** Questa opzione consente di estendere le categorie selezionate e poi generalizzare i descrittori. Quando si sceglie di generalizzare, il prodotto crea le regole di categoria generiche nelle categorie utilizzando il carattere jolly asterisco. Ad esempio, invece di produrre più descrittori come `[torta di mele + .]` e `[succo di mela + .]`, utilizzando caratteri jolly potrebbe produrre `[mela * + .]`. Se è necessario generalizzare con caratteri jolly, sarà spesso necessario ottenere esattamente lo stesso numero di record o documenti come è stato fatto in precedenza. Tuttavia, questa opzione ha il vantaggio di ridurre il numero e semplificare i descrittori di categoria. Inoltre, questa opzione aumenta la possibilità di classificare più record o documenti mediante queste categorie su nuovi dati di testo (ad esempio, in studi longitudinale/onda).
- **Estendi solo**. Questa opzione estende le categorie senza generalizzare. Può essere utile prima scegliere l'opzione **Estendi solo** per creare categorie ed estendere poi le stesse categorie di nuovo utilizzando l'opzione **Estendi e generalizza**.
- **Generalizza solo.** Questa opzione consente di generalizzare i descrittori senza estendere le categorie in qualsiasi altro modo.

  Nota: selezionando questa opzione si disabilita l'opzione **rete semantica**; ciò avviene perché l'opzione **rete semantica** è disponibile solo quando una descrizione deve essere estesa.

Altre opzioni per l'estensione di categorie

Oltre a selezionare le tecniche da applicare, è possibile modificare una qualsiasi delle seguenti opzioni:

Numero massimo di elementi con cui estendere un descrittore. Quando si estende un descrittore con elementi (concetti, tipi e altre espressioni), definire il numero massimo di elementi che possono essere aggiunti a un singolo descrittore. Se si imposta questo limite su 10, non possono essere aggiunti più di 10 nuovi elementi ad un descrittore esistente. Se esistono più di 10 voci da aggiungere, le tecniche smettono di aggiungere nuovi elementi dopo che è stato aggiunto il decimo. In tal modo è possibile creare un elenco di descrittori più breve ma ciò non garantisce che le voci più interessanti vengano utilizzate per prime. È possibile preferire di ridurre la dimensione dell'estensione senza penalizzare la qualità utilizzando l'opzione **Generalizza con caratteri jolly dove possibile**. Questa opzione è valida solo per i descrittori che contengono i valori booleani `&` (AND) o `!` (NOT).

Estendi anche sottocategorie. Questa opzione consente di estendere anche tutte le sottocategorie appartenenti alle categorie selezionate.

Estendi categorie vuote con i descrittori generati dal nome categoria. Questo metodo si applica solo alle categorie vuote, che hanno 0 descrittori. Se una categoria già contiene descrittori, essa non verrà estesa in questo modo. Questa opzione tenta di creare automaticamente i descrittori per ciascuna categoria basati sulle parole che costituiscono il nome della categoria. Il nome della categoria viene sottoposto a scansione per vedere se le parole nel nome corrispondono ad uno qualsiasi dei concetti estratti. Se un concetto viene riconosciuto, esso viene utilizzato per trovare modelli di concetto corrispondenti ed entrambi questi vengono utilizzati per formare i descrittori per la categoria. Questa opzione produce i migliori risultati quando i nomi categoria o le annotazioni sono entrambi lunghi e descrittivi. Si tratta di un metodo veloce per la generazione dei descrittori di categoria che consentono a turno alla categoria di catturare i record che contengono quei descrittori. Questa opzione è molto utile quando si desidera importare le categorie da altre parti o quando si creano manualmente le categorie con nomi lunghi descrittivi.

Genera descrittori con nome. Questa opzione si applica soltanto se è selezionata l'opzione precedente.

- **Concetti.** Scegliere questa opzione per produrre i descrittori nella forma di concetti, indipendentemente dal fatto che siano stati estratti dal testo di origine.
- **Modelli.** Scegliere questa opzione per produrre i descrittori nella forma di modelli, indipendentemente dal fatto che questi modelli siano stati estratti.

Creazione manuale di categorie

Oltre a creare categorie utilizzando le tecniche di creazione categorie automatizzata, e l'editor delle regole, è possibile creare categorie anche manualmente. Sono possibili i seguenti metodi manuali:

- Creazione di una categoria vuota in cui si aggiungeranno gli elementi uno alla volta. Consultare la sezione “Creazione o ridenominazione di categorie” per ulteriori informazioni.
- Trascinando termini, tipi e modelli nel riquadro delle categorie. Consultare la sezione “Creazione di categorie mediante trascinamento e rilascio” a pagina 126 per ulteriori informazioni.

Creazione o ridenominazione di categorie

È possibile creare categorie vuote a cui aggiungere concetti e tipi. È inoltre possibile ridenominare le categorie.

Per creare una nuova categoria vuota

1. Andare al riquadro Categorie.
2. Dai menu, scegliere **Categorie > Crea categoria vuota**. Viene visualizzata la finestra di dialogo Proprietà della categoria.
3. Immettere un nome per questa categoria nel campo Nome.
4. Fare clic su **OK** per accettare il nome e chiudere la finestra di dialogo. La finestra di dialogo viene chiusa e un nuovo nome categoria viene visualizzato nel riquadro.

È ora possibile iniziare ad aggiungere a questa categoria. Consultare la sezione “Aggiunta di descrittori alle categorie” a pagina 144 per ulteriori informazioni.

Per ridenominare una categoria

1. Selezionare una categoria e scegliere **Categorie > Rinomina categoria** . Viene visualizzata la finestra di dialogo Proprietà della categoria.
2. Immettere un nuovo nome per questa categoria nel campo Nome.
3. Fare clic su **OK** per accettare il nome e chiudere la finestra di dialogo. La finestra di dialogo viene chiusa e un nuovo nome categoria viene visualizzato nel riquadro.

Creazione di categorie mediante trascinamento e rilascio

La tecnica di rilascio-trascinamento è manuale e non è basata su algoritmi. È possibile creare categorie nel riquadro Categorie trascinando:

- Concetti estratti, tipi o modelli dal riquadro Risultati di estrazione nel riquadro Categorie.
- Concetti estratti dal pannello Dati nel riquadro Categorie.
- Righe intere dal pannello Dati nel riquadro Categorie. In questo modo verrà creata una categoria composta da tutti i concetti e i modelli estratti contenuti in tale riga.

Nota: il riquadro Risultati di estrazione supporta più selezioni per facilitare le operazioni di trascinamento e rilascio di più elementi.

Importante! Non è possibile trascinare e rilasciare i concetti dal riquadro dati che non sono stati estratti dal testo. Se si desidera forzare l'estrazione di un concetto che si trova nei dati, è necessario aggiungere questo concetto a un tipo. Quindi rieseguire l'estrazione. I nuovi risultati dell'estrazione conterranno il concetto appena aggiunto. È possibile quindi utilizzarlo nella propria categoria. Consultare la sezione "Aggiunta di concetti ai tipi" a pagina 97 per ulteriori informazioni.

Per creare le categorie utilizzando il trascinamento e rilascio:

1. Dal riquadro Risultati di estrazione o Dati, selezionare uno o più concetti, modelli, tipi, record o record parziali.
2. Tenendo premuto il pulsante del mouse, trascinare l'elemento da una categoria esistente o sull'area del riquadro per creare una nuova categoria.
3. Quando si raggiunge l'area in cui si desidera lasciare l'elemento, rilasciare il pulsante del mouse. L'elemento viene aggiunto al riquadro Categorie. Le categorie che sono stati modificate vengono visualizzate con un colore di sfondo speciale. Questo colore viene chiamato **Sfondo feedback di categoria**. Per ulteriori informazioni, consultare "Impostazione delle opzioni" a pagina 82.

Nota: la categoria che ne risulta viene denominata automaticamente. Se si desidera si può cambiare il nome. Consultare la sezione "Creazione o ridenominazione di categorie" a pagina 125 per ulteriori informazioni.

Se si desidera vedere quali record sono assegnati ad una categoria, selezionare tale categoria nel riquadro Categorie. Il pannello dei dati viene automaticamente aggiornato e visualizza tutti i record per tale categoria.

Uso delle regole di categoria

È possibile creare categorie in molti modi. Uno di questi è quella di definire le regole di categoria per esprimere idee. Le regole di categoria sono istruzioni che vengono automaticamente classificate come documenti o record in una categoria basata su un'espressione logica utilizzando concetti estratti, tipi e modelli e operatori booleani. Ad esempio, è possibile scrivere un'espressione che significa *includi tutti i record che contengono i concetti estratti* `ambasciata` *ma non* `argentina` *in questa categoria*.

Mentre alcune regole di categoria vengono generate automaticamente quando si creano le categorie utilizzando tecniche di raggruppamento come *ricorrenza* e *derivazione principale di concetto* (**Categorie > Impostazioni di creazione > Impostazioni avanzate: linguistica**), è possibile anche creare regole di categoria manualmente nell'editor di regole utilizzando la conoscenza di categoria dei dati e del contesto. Ogni regola viene collegata ad una singola categoria in modo che ogni documento o record che corrisponde alla regola viene poi calcolato in tale categoria.

Le regole di categoria di regole aiutano a migliorare la qualità e la produttività dei risultati di estrazione testo e ulteriore analisi quantitativa più approfondita che consente di classificare le risposte con maggiore specificità. L'esperienza e le conoscenze di business potrebbero fornire una conoscenza specifica dei dati e

del contesto. È possibile alzare il livello di questa comprensione per tradurre tale conoscenza in regole di categoria per classificare i documenti o i record in modo ancora più efficiente e preciso combinando elementi estratti con logica booleana.

La capacità di creare tali regole di codifica migliora la precisione, l'efficienza e la produttività permettendo di alzare il livello delle proprie conoscenze di business sulla tecnologia di estrazione del prodotto.

Nota: per esempi di come le regole corrispondono al testo, consultare “Esempi di regole di categoria” a pagina 133

Sintassi di regole di categoria

Mentre alcune regole di categoria vengono generate automaticamente quando si creano le categorie utilizzando tecniche di raggruppamento come *ricorrenza* e *derivazione principale di concetto* (**Categorie > Impostazioni di creazione > Impostazioni avanzate: linguistica**), è possibile anche creare regole di categoria manualmente nell'editor di regole. Ogni regola è un descrittore di una singola categoria; ogni documento o record che corrisponde alla regola viene automaticamente conteggiato in tale categoria.

Nota: per esempi di come le regole corrispondono al testo, consultare “Esempi di regole di categoria” a pagina 133

Quando si sta creando o modificando una regola, è necessario che la regola venga aperta nell'editor delle regole. È possibile aggiungere concetti, tipi o modelli o utilizzare i caratteri jolly per ampliare le corrispondenze. Quando si utilizzano concetti, tipi e modelli estratti, è possibile trarre vantaggio dal trovare tutti i concetti correlati.

Importante! Per evitare errori comuni, si consiglia di trascinare e rilasciare i concetti direttamente dal riquadro Risultati di estrazione, pannelli di Analisi di collegamento del testo o il pannello Dati nell'editor di regole o di aggiungerli mediante il menu di scelta rapida quando possibile.

Quando i concetti, i tipi e modelli vengono riconosciuti, viene visualizzata un'icona accanto al testo.

Tabella 18. Icone di estrazione

Icona	Descrizione
	Concetto estratto
	Tipo estratto
	Modello estratto

Sintassi e operatori di regola

La seguente tabella contiene i caratteri con cui è possibile definire la sintassi della regola. Utilizzare questi caratteri insieme ai concetti, i tipi e modelli per creare la regola.

Tabella 19. Sintassi supportata

Carattere	Descrizione

`&`	L'"e" booleano. Ad esempio `a & b` contiene entrambi `a` *e* `b` come:
	`- invasione & stati uniti`
	`- 2016 & olimpiadi`
	`- mela & buona`

Tabella 19. Sintassi supportata (Continua)

Carattere	Descrizione

|	L'"o" booleano è inclusivo, il che significa che se vengono trovati uno o tutti gli elementi, è realizzata una corrispondenza. Ad esempio, `a | b` contiene `a` *o* `b` come:
	`- attacco | francia`
	`- condominio | appartamento`

!()	Il "no" booleano. Ad esempio, `!(a)` non contiene `a` come, `!(buon & hotel), assassinio & !(austria)` o `!(oro) & !(rame)`
*	Un carattere jolly che rappresenta qualsiasi cosa da un singolo carattere ad una parola intera a seconda di come viene utilizzato. Consultare la sezione "Uso dei caratteri jolly nelle regole di categoria" a pagina 131 per ulteriori informazioni.
()	Delimitatore di espressione. Qualsiasi espressione in parentesi viene valutata per prima.
+	Il connettore del modello utilizzato per formare un modello specifico dell'ordine. Quando presente, devono essere utilizzate le parentesi quadre. Consultare la sezione "Uso dei modelli TLA nelle regole di categoria" per ulteriori informazioni.
[]	Il delimitatore di modello è necessario se si stanno cercando corrispondenze basate su un modello TLA estratto all'interno di una regola di categoria. Il contenuto all'interno delle parentesi fa riferimento a modelli TLA e non avrà mai corrispondenza con concetti o tipi basati su ricorrenza semplice. Se non è stato estratto questo modello TLA, non sarà possibile alcuna corrispondenza. Consultare la sezione "Uso dei modelli TLA nelle regole di categoria" per ulteriori informazioni. Non utilizzare parentesi quadre se si sta cercando corrispondenza con concetti e tipi e non con modelli. *Nota*: in precedenti versioni, le regole di ricorrenza e di sinonimo generate da tecniche di creazione categoria venivano racchiuse tra parentesi quadre. In tutte le nuove versioni, le parentesi quadre indicano ora la presenza di un modello TLA. Al contrario, le regole prodotte dalle tecniche e sinonimi di ricorrenza verranno racchiuse in parentesi, ad esempio `(sistemi di altoparlanti|altoparlanti)`.

Gli operatori `&` e `|` sono commutativi in modo tale che `a & b = b & a` e `a | b = b | a`.

Caratteri escape con la barra retroversa

Se si dispone di un concetto che contiene qualsiasi carattere che è anche un carattere di sintassi, è necessario posizionare una barra retroversa davanti quel carattere in modo che la regola venga interpretata correttamente. Il carattere barra retroversa (\) viene utilizzato per i caratteri escape che altrimenti hanno un significato speciale. Quando si trascina e si rilascia nell'editor, la barra retroversa viene inserita automaticamente.

I seguenti caratteri di sintassi di regola devono essere preceduti da una barra retroversa se si desidera che vengano trattati per quello che sono piuttosto che come sintassi di regola:
`& ! | + < > ( ) [ ] *`

Ad esempio, poiché il concetto `r&d` contiene l'operatore "and" (`&`), la barra retroversa viene richiesta quando viene immessa nell'editor di regole, ad esempio: `r\&d`.

Uso dei modelli TLA nelle regole di categoria

I modelli di analisi di collegamento del testo possono essere definiti esplicitamente nelle regole di categoria in modo da ottenere risultati ancor più specifici e contestuali. Quando si definisce un modello in una regola di categoria, si ignorano i risultati di estrazione concetto più semplici e solo i documenti e record corrispondenti basati sui risultati del modello di analisi di collegamento del testo.

Importante! Al fine di corrispondenza documenti tramite i modelli TLA nelle proprie regole di corrispondenza, è necessario avere eseguito un'estrazione con l'analisi di collegamento del testo abilitata.

La regola di categoria ricercherà le corrispondenze rilevate durante tale processo. Se non è stato scelto di esplorare i risultati TLA nella scheda Modello del nodo di estrazione testo, è possibile scegliere di abilitare l'estrazione TLA nelle impostazioni di estrazione all'interno della sessione interattiva e poi rieseguire l'estrazione. Consultare la sezione “Estrazione di dati” a pagina 88 per ulteriori informazioni.

Delimitazione con parentesi quadre. Un modello TLA deve essere racchiuso tra parentesi quadre `[ ]` se si utilizzando all'interno di una regola di categoria. Il delimitatore di modello è necessario se si stanno cercando corrispondenze basate su un modello TLA estratto. Poiché le regole di categoria possono contenere tipi, concetti o modelli, le parentesi chiariscono alla regola che il contenuto all'interno delle parentesi fa riferimento al modello TLA estratto. Se non è stato estratto questo modello TLA, non sarà possibile alcuna corrispondenza. Se si desidera visualizzare un modello senza parentesi come `mela + buono` nel riquadro Categorie, ciò probabilmente indica che il modello è stato aggiunto direttamente alla categoria al di fuori dell'editor della regola di categoria. Ad esempio, se si aggiunge un modello di concetto direttamente alla categoria dalla vista dell'analisi di collegamento del testo , esso non verrà visualizzato con parentesi quadre. Tuttavia, quando si utilizza un modello all'interno di una regola di categoria, è necessario racchiudere il modello tra parentesi quadre all'interno della regola di categoria, ad esempio `[banana + !(buona)]`.

Uso del segno + nei modelli. In IBM SPSS Modeler Text Analytics, è possibile un modello a 6 parti o slot. Per indicare che l'ordine è importante, utilizzare il simbolo + per connettere ogni elemento, ad esempio `[azienda1 + acquisisce + azienda2]`. Qui l'ordine è importante poiché potrebbe modificare l'identità dell'azienda che è stata acquisita. L'ordine non è determinato dalla struttura della frase ma piuttosto dal modo in cui l'output del modello TLA è strutturato. Ad esempio, se si ha il testo "*amo Parigi*" e si desidera estrarre questa idea, il modello TLA dovrebbe essere `[parigi + amo]` o `[<Ubicazione> + <Positivo>]` invece di `[<Positivo> + <Ubicazione>]` poiché le risorse di opinione predefinite generalmente inseriscono le opinioni al secondo posto dei modelli a 2 parti. Quindi può essere utile impiegare il modello direttamente come descrittore nella propria categoria per evitare problemi. Tuttavia, se è necessario utilizzare un modello come parte di un'istruzione più complessa, prestare particolare attenzione all'ordine degli elementi all'interno dei modelli presentati nella vista Analisi di collegamento del testo in quanto l'ordine svolge un ruolo importante nel trovare una corrispondenza.

Ad esempio, si supponga di avere i due seguenti testi campione: le espressioni "Io *amo l'ananas*" e "io odio *l'ananas*". Comunque io *amo* le fragole". L'espressione `amo & ananas` corrispondono a entrambi i testi poiché è un'espressione di concetto e non una regola di collegamento di testo (non racchiusi tra parentesi). L'espressione `ananas + amo` corrisponde solo a "Io *amo l'ananas*" poiché nel secondo testo, la parola *amo* è associata a *fragole*.

Raggruppamento con modelli. È possibile semplificare le regole con i propri modelli. Si supponga di voler catturare le seguenti tre espressioni, `pepe + amo`, `peperoncino + amo`, e `peperoni + amo`. È possibile raggruppare il tutto in una singola regola di categoria come `[* peperoni & amo]`. Se si avesse un'altra espressione `peperoni forti + buoni`, è possibile raggruppare questi quattro con una regola come `[* peperoni + <Positivo>]`.

Ordine nei modelli. Per meglio organizzare l'output, le regole di analisi di collegamento del testo fornite nei modelli installati con il prodotto tentano di emettere modelli di base nello stesso ordine indipendentemente dall'ordine delle parole nella frase. Ad esempio, se si dispone di un record contenente il testo "*Una bella presentazione.*" e un altro record che contiene "*la presentazione è stata buona*", entrambi i testi corrispondono alla stessa regola e output nello stesso ordine come `presentazione + buona` nei risultati del modello di concetto piuttosto che `presentazione + buona` e anche `buona + presentazione`. Nel modello a due slot come quelli nell'esempio, i concetti assegnati a tipi nella libreria Opinioni verrà visualizzato per ultimo nell'output per impostazione predefinita come `mela + cattiva`.

Tabella 20. Sintassi del modello e uso booleano

Espressione	Corrisponde ad un documento o record che
[ ]	Contiene qualsiasi modello TLA. Il delimitatore di modello è necessario *nelle regole di categoria* se si stanno cercando corrispondenze basate su un modello TLA estratto. Il contenuto all'interno delle parentesi quadre fa riferimento ai modelli TLA e non a concetti e tipi semplici. Se non è stato estratto questo modello TLA, non sarà possibile alcuna corrispondenza. Se si volesse creare una regola che non include modelli, è possibile utilizzare `!([ ])`.
[a]	Contiene un modello di cui almeno un elemento è `a` a prescindere dalla sua posizione nel modello. Ad esempio, `[accordo]` può corrispondere a `[accordo + buon]` o solo `[accordo + .]`
[a + b]	Contiene un modello di concetto. Ad esempio, `[accordo + buon]`. *Nota:* se si desidera solo catturare questo modello senza aggiungere altri elementi, si consiglia di aggiungere il modello direttamente alla propria categoria piuttosto che creare con esso una regola.
[a + b + c]	Contiene un modello di concetto. Il simbolo + indica che l'ordine degli elementi corrispondenti è importante. Ad esempio, `[azienda1 + ha acquistato + azienda2]`.
[<A> + <B>]	Contiene qualsiasi modello con tipo `<A>` nel primo slot e tipo `<B>` nel secondo slot e ci sono così esattamente due slot. Il simbolo + indica che l'ordine degli elementi corrispondenti è importante. Ad esempio, `[<Bilancio> + <Negativo>]`. *Nota:* se si desidera solo catturare questo modello senza aggiungere altri elementi, si consiglia di aggiungere il modello direttamente alla propria categoria piuttosto che creare con esso una regola.
[<A> & <B>]	Contiene qualsiasi tipo di modello con tipo `<A>` e tipo `<B>`. Ad esempio, `[<Bilancio> & + <Negativo>]`. Questo modello TLA non verrà mai estratto; tuttavia, scritto in questo modo è realmente uguale a `[<Bilancio> + <Negativo>]|[<Negativo> + <Bilancio>]`. L'ordine degli elementi corrispondenti non è importante. Inoltre, altri elementi potrebbero essere contenuti nel modello ma questo deve contenere almeno `<Bilancio>` e `<Negativo>`.
[a + .]	Contiene un modello dove `a` è il solo concetto e non esiste nulla in altri slot per quel modello. Per esempio, `[accordo + .]` corrisponde al modello di concetto in cui l'unico output è il concetto di `accordo`. Se è stato aggiunto il concetto `accordo` come descrittore di categoria, si potrebbero ottenere tutti i record con accordo come concetto che include dichiarazioni positive su un accordo. Tuttavia, utilizzando `[accordo + .]` corrisponderà solo a quei risultati di modello di record che rappresentano `accordo` e nessun'altra relazione o opinione e non corrisponderà a `accordo + fantastico`. *Nota:* se si desidera solo catturare questo modello senza aggiungere altri elementi, si consiglia di aggiungere il modello direttamente alla propria categoria piuttosto che creare con esso una regola.
[<A> + <>]	Contiene un modello dove `<A>` è l'unico tipo. Ad esempio, `[<Bilancio> + <>]` corrisponde al modello in cui l'unico output è un concetto del tipo `<Bilancio>`. *Nota:* è possibile utilizzare `<>` per indicare un tipo vuoto solo quando è inserito dopo il simbolo di modello + nel modello di tipo come `[<Bilancio> + <>]` ma non come `[prezzo + <>]`. *Nota:* se si desidera solo catturare questo modello senza aggiungere altri elementi, si consiglia di aggiungere il modello direttamente alla propria categoria piuttosto che creare con esso una regola.
[a + !(b)]	Contiene almeno un modello che include il concetto `a` ma non include il concetto `b`. Deve includere almeno un modello. Ad esempio, `[prezzo + !(alto)]` o per i tipi, `[!(<Frutta>|<Verdura>) + <Positivo>]`
!([<A> & <B>])	Non contiene un modello specifico. Ad esempio, `!([<Bilancio> & <Negativo>])`.

Nota: per esempi di come le regole corrispondono al testo, consultare "Esempi di regole di categoria" a pagina 133

Uso dei caratteri jolly nelle regole di categoria

I caratteri jolly possono essere aggiunti ai concetti nelle regole per estendere le capacità corrispondenti. Il carattere jolly asterisco * può essere posizionato prima e/o dopo una parola per indicare come i concetti possono corrispondere. Il carattere jolly può essere utilizzato in due modi:

- **Caratteri jolly come affisso**. Questi caratteri jolly devono essere prefisso o suffisso immediato senza spazio che separa la stringa e l'asterisco. Ad esempio, `opera*` può corrispondere a *operato*, *opera*, *operativo*, *operazioni*, *operatore* e così via.
- **Caratteri jolly di parole**. Questi caratteri jolly sono prefisso o suffisso di un concetto con uno spazio tra il concetto e l'asterisco. Ad esempio, `* operazione` potrebbe corrispondere a *operazione*, *operazione chirurgica*, *post-operazione*, e così via. Inoltre un carattere jolly di parola può essere utilizzato insieme a un carattere jolly di affisso come, ad esempio `* opera* *`, che può corrispondere a *operazione*, *operazione chirurgica*, *operatore del telefono*, *aria di opera* e così via. Come è possibile vedere in questo ultimo esempio, si consiglia di utilizzare i caratteri jolly con attenzione in modo da non eseguire ricerche troppo estese sulla rete e non catturare corrispondenze non desiderate.

Eccezioni!

- Un carattere jolly può mai stare da solo. Ad esempio, `(mela | * )` non verrebbe accettato.
- Un carattere jolly non può mai essere utilizzato per stabilire una corrispondenza con i nomi di tipo. `<Negativo*>` non corrisponde ad alcun nome tipo.
- Non è possibile filtrare determinati tipi con l'essere in corrispondenza con i concetti trovati tramite caratteri jolly. Il tipo a cui il concetto è assegnato viene utilizzato automaticamente.
- Un carattere jolly non può mai essere posizionato in mezzo ad una sequenza di parole, se è fine o inizio di una parola (conto* aperto) o un componente autonomo (conto * aperto). Non è possibile utilizzare caratteri jolly nemmeno nei nomi tipo. Ad esempio, `parola* parola`, come `mela* ricetta`, non corrisponderanno a ricetta di torta di mele o a qualsiasi altra cosa. Tuttavia, `mela* *` potrebbe corrispondere a *ricetta di torta di mele*, *torta di mele*, *mele* e così via. In un altro esempio, `parola* parola`, ad esempio `mela * toast`, non restituirà *mele cannella toast* o altro poiché l'asterisco viene visualizzato tra altre due parole. Tuttavia, `mela *` corrisponde a *mela cannella toast*, *mele*, *torta di mele* e così via.

Tabella 21. Uso di caratteri jolly

Espressione	Corrisponde ad un documento o record che
`*mela`	Contiene un concetto che termina con la lettera scritta ma può avere qualsiasi numero di lettere come prefisso. Ad esempio: `*mela` termina con le lettere *mela* , ma può richiedere un prefisso come ad esempio: - `mela` - `ananas` -`mela selvatica`

Tabella 21. Uso di caratteri jolly (Continua)

Espressione	Corrisponde ad un documento o record che
`mela*`	Contiene un concetto che inizia con la lettera scritta ma può avere qualsiasi numero di lettere come suffisso. Ad esempio: `*mela` inizia con le lettere *mela* , ma può contenere un suffisso o non contenerne affatto, ad esempio: - `mela` - `torta di mele` - `acquavite di mele` Ad esempio, `mela* & !(pera* | mela cotogna)`, che contiene un concetto che inizia con le lettere `mela` ma non un concetto che inizia con le lettere *pera* o il concetto `mela cotogna`, NON corrisponde a: `mela & mela cotogna` ma potrebbe corrispondere a: - `torta di mele` - `mela & arancia`
`*prodotto*`	Contiene un concetto che contiene le lettere scritte `prodotto`, ma può avere qualsiasi numero di lettere come un prefisso o un suffisso o entrambi. Ad esempio: `*prodotto*` potrebbe corrispondere a: - `prodotto` - `sottoprodotto` - `improduttivo`
`* prestito`	Contiene un concetto che contiene la parola `prestito` ma può essere un composto con un'altra parola inserita prima. Ad esempio `* prestito` potrebbe corrispondere a: - `prestito` - `prestito auto` - `prestito casa` Ad esempio, `[* spedizione + <Negativo>]` contiene un concetto che termina con la parola `spedizione` nella prima posizione e contiene un tipo `<Negativo>` nella seconda posizione che potrebbe corrispondere ai seguenti modelli di concetto: - `spedizione pacco + normale` - `spedizione rapida + ritardo`
`evento *`	Contiene un concetto che contiene la parola `evento` ma può essere un composto seguito da un'altra parola. Ad esempio `evento *` potrebbe corrispondere a: - `evento` - `ubicazione evento` - `commissione di pianificazione eventi`
`* mela *`	Contiene un concetto che potrebbe iniziare con qualsiasi parola seguito dalla parola `mela` seguita eventualmente da un'altra parola. `*` significa 0 o n, quindi corrisponde anche a `mela`. Ad esempio, `* mela *` potrebbe corrispondere a: - `torta di mele esclusiva` - `croccante di mela barilla` - `torta di mele famosa` - `mela` Ad esempio, `[* prenotazione* * + <Positivo>]`, che contiene un concetto con la parola `prenotazione` (indipendentemente da dove si trova nel concetto) nella prima posizione e contiene un tipo `<Positivo>` nella seconda posizione potrebbe corrispondere ai modelli di concetto: - `sistema di prenotazione + buono` - `prenotazioni online + buono`

Nota: per esempi di come le regole corrispondono al testo, consultare “Esempi di regole di categoria” a pagina 133

Esempi di regole di categoria

Per contribuire a dimostrare il modo in cui le regole vengono confrontate con i record in modo diverso in base alla sintassi utilizzata per esprimerli, considerare il seguente esempio.

Esempio di record

Immaginiamo di avere due record:

- **Record A**: *"quando ho controllato il mio portafoglio, ho visto che mancavano 5 euro."*
- **Record B**: *"nell'area picnic ho trovato 5 euro, ma non ho trovato la coperta."*

Le seguenti due tabelle mostrano cosa potrebbe essere estratto come concetti e tipi o anche come modelli di concetto e modelli di tipo.

Esempio di concetti e tipi estratti

Tabella 22. Esempio di concetti e tipi estratti

Concetto estratto	Concetti immessi come
portafoglio	<Sconosciuto>
mancante	<Negativo>
€ 5	<Valuta>
coperta	<Sconosciuto>
area picnic	<Sconosciuto>

Esempio di modelli TLA estratti

Tabella 23. Esempio di output di modello TLA estratto

Modelli di concetto estratti	Modelli di tipo estratti	Da record
area picnic + .	<Sconosciuto> + <>	Record B
portafoglio + .	<Sconosciuto> + <>	Record A
coperta + mancante	<Sconosciuto> + <Negativo>	Record B
€ 5 + .	<Valuta> + <>	Record B
€ 5 + mancante	<Valuta> + <Negativo>	Record A

Possibili corrispondenze di regole di categoria

La seguente tabella contiene sintassi che può essere immessa nell'editor di regole di categoria. Non tutte le regole funzionano e non tutte corrispondono agli stessi record. Ecco come sintassi differente influenza le corrispondenze di record.

Tabella 24. Esempi di regole

Sintassi di regola	Risultato
€ 5 & mancante	Corrisponde a entrambi i record A e B in quanto entrambi contengono il concetto estratto mancante e il concetto di € 5 estratto. Ciò equivale a: (€ 5 & mancante)
mancante & € 5	Corrisponde a entrambi i record A e B in quanto entrambi contengono il concetto estratto mancante e il concetto di € 5 estratto. Ciò equivale a: (mancante & € 5)

Tabella 24. Esempi di regole (Continua)

Sintassi di regola	Risultato
`mancante & <Valuta>`	Corrisponde a entrambi i record A e B in quanto entrambi contengono il concetto estratto `mancante` e un concetto che corrisponde al tipo `<Valuta>`. Ciò equivale a: `(mancante & <Valuta>)`
`<Valuta> & mancante`	Corrisponde a entrambi i record A e B in quanto entrambi contengono il concetto estratto `mancante` e un concetto che corrisponde al tipo `<Valuta>`. Ciò equivale a: `(<Valuta> & mancante)`
`[€ 5 + mancante]`	Corrisponde ad A ma **non** a B poiché non ha prodotto alcun output di modello TLA che contiene `€ 5 + mancante` (vedi tabella precedente). Ciò equivale all'output di modello TLA: `€ 5 + mancante`
`[mancante + € 5]`	Non corrisponde né ad A né a B poiché nessun modello TLA estratto (vedi tabella precedente) corrisponde all'ordine qui espresso con `mancante` nella prima posizione. Ciò equivale all'output di modello TLA: `€ 5 + mancante`
`[mancante & € 5]`	Sono presenti corrispondenze per A, ma **non** per B poiché non è stato estratto nessun modello TLA dal record B. Utilizzando il carattere `&` si indica che l'ordine non è importante per la corrispondenza; quindi questa regola ricerca una corrispondenza del modello per `[mancante + USD5]` o `[USD5 + mancante]`. Solo `[€ 5 + mancante]` dal record A ha una corrispondenza.
`[mancante + <Valuta>]`	Non corrisponde né ad A né a B poiché nessun modello TLA estratto corrisponde a tale ordine. Ciò non ha alcun equivalente, poiché un output TLA è basato solo sui termini `(€ 5 + mancante)` o sui tipi di `(<Valuta> + <Negativo>)`, ma non mischia concetti e tipi.
`[<Valuta> + <Negativo>]`	Sono presenti corrispondenze per A, ma non per B poiché non è stato estratto nessun modello TLA dal record B. Ciò equivale all'output TLA: `<Valuta> + <Negativo>`
`[<Negativo> + <Valuta>]`	Non corrisponde né ad A né a B poiché nessun modello TLA estratto corrisponde a tale ordine. Nel modello `Opinioni`, per impostazione predefinita, quando un *argomento* viene rilevato con una *opinione*, l'*argomento* (`<Valuta>`) occupa la prima posizione in casella e *opinione* (`<Negativo>`) occupa la seconda posizione.

Creazione di regole di categoria

Quando si sta creando o modificando una regola, è necessario che la regola venga aperta nell'editor delle regole. È possibile aggiungere concetti, tipi o modelli o utilizzare i caratteri jolly per ampliare le corrispondenze. Quando si utilizzano concetti riconosciuti, tipi e modelli, è possibile trarre vantaggio dal momento in cui si troveranno tutti i concetti correlati. Ad esempio, quando si utilizza un concetto anche tutti i suoi termini associati, forme plurali e sinonimi vengono associati alla regola. Allo stesso modo, quando si utilizza un tipo, anche tutti i suoi concetti vengono catturati dalla regola.

È possibile aprire l'editor delle regole per modificare una regola esistente o facendo clic con il tasto destro del mouse sul nome della categoria e scegliere **Crea regola**.

È possibile utilizzare i menu di scelta rapida, trascinare e rilasciare o immettere manualmente i concetti, i tipi e i modelli nell'editor. Combinare quindi questi valori con operatori booleani (`&`, `!()`, `|`) e parentesi per formare le espressioni della regola. Per evitare errori comuni, si consiglia di trascinare e rilasciare i

concetti direttamente dal riquadro Risultati di estrazione o il pannello Dati nell'editor di regole. Prestare attenzione alla sintassi delle regole per evitare errori. Consultare la sezione “Sintassi di regole di categoria” a pagina 127 per ulteriori informazioni.

Nota: per esempi di come le regole corrispondono al testo, consultare “Esempi di regole di categoria” a pagina 133.

Per creare una regola

1. Se non sono stati ancora estratti eventuali dati o l'estrazione non è aggiornata, farlo ora. Consultare la sezione “Estrazione di dati” a pagina 88 per ulteriori informazioni.

   Nota: se si desidera filtrare un'estrazione in modo tale che non ci siano più concetti visibili, viene visualizzato un messaggio di errore quando si tenta di creare o modificare una regola di categoria. Per evitare ciò, modificare il filtro di estrazione in modo che quei concetti siano disponibili.
2. Nel riquadro Categorie, selezionare la categoria in cui si desidera aggiungere la regola.
3. Dai menu scegliere **Categorie > Crea regola**. Il riquadro dell'editor delle regole di categoria si apre nella finestra.
4. Nel campo Nome regola, immettere un nome per la regola. Se non viene fornito alcun nome, l'espressione verrà utilizzata automaticamente come nome. È possibile ridenominare questa regola successivamente.
5. Nel campo di testo dell'espressione più grande, è possibile:
   - Immettere il testo direttamente nel campo o trascinare e rilasciare da un altro riquadro. Utilizzare solo i concetti, tipi e modelli estratti. Ad esempio, se si immette la parola `gatti`, ma solo la forma singolare, `gatto`, viene visualizzata nel riquadro Risultati di estrazione, l'editor non sarà in grado di riconoscere `gatti`. In quest'ultimo caso, la forma singolare potrebbe includere automaticamente il plurale, altrimenti si potrebbe utilizzare un carattere jolly. Consultare la sezione “Sintassi di regole di categoria” a pagina 127 per ulteriori informazioni.
   - Selezionare i concetti, i tipi o i modelli che si desidera aggiungere alle regole e utilizzare i menu.
   - Aggiungere gli operatori booleani per collegare insieme gli elementi nella regola. Utilizzare i pulsanti della barra degli strumenti per aggiungere l'"and" booleano **&**, l'"or" booleano **|**, il "not" booleano **!()**, le parentesi **()** e le parentesi quadre **[ ]** alla regola.
6. Fare clic sul pulsante **Verifica regola** per verificare che la regola sia formata correttamente. Consultare la sezione “Sintassi di regole di categoria” a pagina 127 per ulteriori informazioni. Il numero di documenti o record trovati viene visualizzato tra parentesi accanto al testo **risultati di testo**. Alla destra di questo testo, è possibile visualizzare gli elementi nella regola che sono stati riconosciuti o eventuali messaggi di errore. Se il grafico accanto al tipo, modello o concetto viene visualizzato con un punto interrogativo rosso, si vuole indicare che l'elemento non corrisponde ad alcuna delle estrazioni. Se non corrispondono, la regola non trova alcun record.
7. Per verificare una parte della propria regola, selezionare tale parte e fare clic su **Verifica selezione**.
8. Apportare tutte le modifiche necessarie e riverificare la regola se vengono riscontrati problemi.
9. Una volta terminato, fare clic su **Salva e chiudi** per salvare la regola di nuovo e chiudere l'editor. Nella categoria viene riportato il nome della nuova regola.

Modifica ed eliminazione delle regole

Dopo aver creato e salvato una regola, è possibile modificare tale regola in qualsiasi momento. Consultare la sezione “Sintassi di regole di categoria” a pagina 127 per ulteriori informazioni.

Se non si desidera più conservare una regola, è possibile eliminarla.

Per modificare le regole

1. Nella tabella dei descrittori nella finestra di dialogo Definizioni di categoria, selezionare la regola.

2. Dai menu scegliere **Categorie > Modifica regola** oppure fare doppio clic sul nome della regola. L'editor si apre con la regola selezionata.
3. Apportare eventuali modifiche alla regola utilizzando i risultati di estrazione e i pulsanti della barra degli strumenti.
4. Provare di nuovo la regola per assicurarsi che restituisca i risultati previsti.
5. Fare clic su **Salva e chiudi** per salvare la regola di nuovo e chiudere l'editor.

Per eliminare una regola

1. Nella tabella dei descrittori nella finestra di dialogo Definizioni di categoria, selezionare la regola.
2. Dal menu scegliere **Modifica > Elimina**. La regola viene eliminata dalla categoria.

Importazione ed esportazione di categorie predefinite

Se le proprie categorie sono memorizzate in un file Microsoft Excel (*.*xls*, *.*xlsx*), è possibile importarle in IBM SPSS Modeler Text Analytics .

È anche possibile esportare le categorie disponibili in una sessione workbench interattiva su un file Microsoft Excel (*.*xls*, *.*xlsx*). Quando si esportano le categorie, è possibile scegliere di includere o escludere alcune informazioni aggiuntive quali descrittori e punteggi. Consultare la sezione “Esportazione di categorie” a pagina 140 per ulteriori informazioni.

Se le categorie predefinite non dispongono di codici o si desidera codici nuovi, è possibile generare automaticamente un nuovo insieme di codici per la serie di categorie nel riquadro delle categorie, scegliendo **Categorie > Gestisci categorie > Crea codici automaticamente** dai menu. In questo modo vengono automaticamente eliminati e rinumerati tutti i codici esistenti.

Importazione di categorie predefinite

In IBM SPSS Modeler Text Analytics è possibile importare le categorie predefinite. Prima di importare, assicurarsi che il file di categoria predefinito è in un file di Microsoft Excel (*.*xls*, *.*xlsx*) strutturato in uno dei formati di sostegno. È inoltre possibile scegliere che il prodotto rilevi automaticamente il formato per l'utente. Sono supportati i seguenti formati:

- **Formato di elenco semplice**: consultare la sezione “Formato elenco semplice” a pagina 137 per ulteriori informazioni.
- **Formato compatto**: per ulteriori informazioni consultare la sezione “Formato compatto” a pagina 138.
- **Formato impresso**: per ulteriori informazioni consultare la sezione “Formato impresso” a pagina 139.

Per importare categorie predefinite

1. Dai menu del workbench interattivo, scegliere **Categorie > Gestisci categorie > Importa categorie predefinite**. Viene visualizzata una procedura guidata di importazione di categorie predefinite.
2. Dall'elenco a discesa Cerca in, selezionare l'unità e la cartella in cui è ubicato il file.
3. Selezionare il file dall'elenco. Il nome del file viene visualizzato nella casella di testo Nome file.
4. Selezionare il foglio di lavoro contenente le categorie predefinite dall'elenco. Il nome foglio di lavoro viene visualizzato nel campo Foglio di lavoro.
5. Per iniziare a scegliere il formato dei dati, fare clic su **Avanti**.
6. Scegliere il formato del file o l'opzione che consente al prodotto di tentare di rilevare automaticamente il formato. Il rilevamento automatico funziona meglio per i formati più comuni.
   - **Formato di elenco semplice**: consultare la sezione “Formato elenco semplice” a pagina 137 per ulteriori informazioni.
   - **Formato compatto**: per ulteriori informazioni consultare la sezione “Formato compatto” a pagina 138.

- **Formato impresso**: per ulteriori informazioni consultare la sezione “Formato impresso” a pagina 139.

7. Per definire altre opzioni di importazione, fare clic su **Avanti**. Se si sceglie che il formato venga rilevato automaticamente, si viene indirizzati alla fase finale.
8. Se una o più righe contengono intestazioni di colonna o altre informazioni estranee, selezionare il numero di riga da cui si desidera avviare l'importazione nell'opzione **Avvia importazione alla riga**. Ad esempio, se i nomi di categoria iniziano sulla riga 7, è necessario immettere il numero 7 per questa opzione per importare il file correttamente.
9. Se il file contiene i codici delle categorie, scegliere l'opzione **Contiene codici di categoria**. In questo modo, si consente alla procedura guidata di riconoscere correttamente i dati.
10. Esaminare le cellule a colori e la legenda per assicurarsi che i dati siano stati identificati correttamente. Eventuali errori rilevati nel file sono mostrati in rosso con riferimenti al di sotto della tabella di anteprima del formato. Se è stato scelto il formato errato, tornare indietro e sceglierne un altro. Se è necessario apportare delle modifiche al file, effettuare tali modifiche e riavviare la procedura guidata selezionando di nuovo il file. È necessario correggere tutti gli errori prima di poter terminare la procedura guidata.
11. Per esaminare la serie di categorie e sottocategorie che verrà importata e per definire il modo in cui creare descrittori per queste categorie, fare clic su **Avanti**.
12. Esaminare l'insieme delle categorie che verrà importato nella tabella. Se non vengono visualizzate le parole chiave che si aspetta di vedere come descrittori, può essere che non sono state riconosciute durante l'importazione. Assicurarsi che siano corrette e che siano visualizzate nella cella corretta.
13. Scegliere come si desidera gestire qualsiasi categoria preesistente nella sessione.
    - **Sostituisci tutte le categorie esistenti**. Questa opzione elimina tutte le categorie esistenti e poi solo le categorie appena importate vengono utilizzate al loro posto.
    - **Accoda a categorie esistenti**. Questa opzione consente di importare le categorie e unire categorie comuni alle categorie esistenti. Quando si aggiunge a categorie esistenti, è necessario determinare in che modo si desidera gestire eventuali duplicati. Una scelta (opzione: **Unisci**) è quella di unire qualsiasi categoria importata con le categorie esistenti se condividono un nome categoria. Un'altra scelta (opzione: **Escludi da importazione**) è quello di vietare l'importazione di categorie se ne esiste già una con lo stesso nome.
14. **Importa parole chiave come descrittori** è un'opzione per importare le parole chiave identificate nei propri dati come descrittori per la categoria associata.
15. **Estendi categorie da descrittori derivanti** è un'opzione che genera descrittori dalle parole che rappresentano il nome della categoria o categoria secondaria e/o dalle parole che costituiscono l'annotazione. Se le parole corrispondono ai risultati estratti, questi vengono aggiunti come descrittori alla categoria. Questa opzione produce i migliori risultati quando i nomi categoria o le annotazioni sono entrambi lunghi e descrittivi. Si tratta di un metodo veloce per la generazione dei descrittori di categoria che consentono alla categoria di catturare i record che contengono quei descrittori.
    - Il campo **Da** consente di selezionare la derivazione dei descrittori, i nomi o categorie e sottocategorie, le parole nelle annotazioni o entrambi.
    - Il campo **Come** consente di scegliere di creare questi descrittori nella forma di concetti o modelli TLA. Se l'estrazione TLA non ha avuto luogo, le opzioni di **modelli** vengono disabilitate in questa procedura guidata.
16. Per importare le categorie predefinite nel riquadro Categorie, fare clic su **Fine**.

Formato elenco semplice

Nel formato elenco semplice, è presente un solo livello superiore delle categorie senza alcuna gerarchia, ovvero nessuna sottocategoria o sottorete. I nomi di categoria sono in una singola colonna.

Le seguenti informazioni possono essere contenute in un file di questo formato:

- La colonna facoltativa **codici** contiene valori che identificano in modo univoco ciascuna categoria. Se si specifica che il file di dati non contiene codici (opzione **Contiene codici di categoria** nel passo **Impostazioni del contenuto**), deve essere presente una colonna contenente i codici univoci per ciascuna categoria nella cella direttamente alla sinistra del nome categoria. Se i dati non contengono codici ma si desidera creare alcuni codici successivamente, è sempre possibile generare codici in (**Categorie> Gestisci categorie > Crea codici automaticamente**).
- Una colonna **nomi categoria** *obbligatoria* contiene tutti i nomi delle categorie. Questa colonna è richiesta per importare utilizzando questo formato.
- **Annotazioni** facoltative nella cella immediatamente alla destra del nome categoria. Questa annotazione consiste in un testo che descrive le categorie e le sottocategorie.
- **Parole chiave** facoltative possono essere importate come descrittori per le categorie. Per essere riconosciute, queste parole chiave devono essere presenti nella cella direttamente al di sotto del nome categoria o sottocategoria associato e l'elenco di parole chiave deve essere preceduto dal carattere di sottolineatura (_) come `_armi da fuoco, missili / fucili`. La cella di parola chiave può contenere una o più parole utilizzate per descrivere ciascuna categoria. Queste parole saranno importate come descrittori o ignorate a seconda di quanto specificato nell'ultimo passo della procedura guidata. Successivamente, i descrittori vengono confrontati con i risultati estratti dal testo. Se viene trovata una corrispondenza, tale record o documento viene calcolato nella categoria contenente questo descrittore.

Tabella 25. Formato di elenco semplice con codici, parole chiave e annotazioni

Colonna A	Colonna B	Colonna C
Codice di categoria (*facoltativo*)	Nome categoria	Annotazione
	Elenco _Descrittore/Paola chiave (*facoltativo*)	

Formato compatto

Il formato compatto è strutturato in modo simile al formato elenco semplice ad eccezione del fatto che il formato compatto viene utilizzato con categorie gerarchiche. Pertanto, è necessaria una colonna di livello di codice per definire il livello gerarchico di ogni categoria e sottocategoria.

Le seguenti informazioni possono essere contenute in un file di questo formato:

- Una colonna **di livello di codice** *obbligatoria* contiene numeri che indicano nella riga la posizione gerarchica per le informazioni successive. Ad esempio, se sono specificati i valori 1, 2 o 3 e si dispone di entrambe le categorie e sottocategorie, 1 è per categorie, 2 è per sottocategorie, 3 è per sottocategorie secondarie. Se si dispone solo di categorie e sottocategorie, 1 è per le categorie e 2 è per le sottocategorie. E così via, fino alla profondità della categoria desiderata.
- La colonna facoltativa **codici** contiene valori che identificano in modo univoco ciascuna categoria. Se si specifica che il file di dati non contiene codici (opzione **Contiene codici di categoria** nel passo **Impostazioni del contenuto**), deve essere presente una colonna contenente i codici univoci per ciascuna categoria nella cella direttamente alla sinistra del nome categoria. Se i dati non contengono codici ma si desidera creare alcuni codici successivamente, è sempre possibile generare codici in (**Categorie> Gestisci categorie > Crea codici automaticamente**).
- Una colonna **nomi categoria** *obbligatoria* contiene tutti i nomi delle categorie e delle sottocategorie. Questa colonna è richiesta per importare utilizzando questo formato.
- **Annotazioni** facoltative nella cella immediatamente alla destra del nome categoria. Questa annotazione consiste in un testo che descrive le categorie e le sottocategorie.
- **Parole chiave** facoltative possono essere importate come descrittori per le categorie. Per essere riconosciute, queste parole chiave devono essere presenti nella cella direttamente al di sotto del nome categoria o sottocategoria associato e l'elenco di parole chiave deve essere preceduto dal carattere di sottolineatura (_) come `_armi da fuoco, missili / fucili`. La cella di parola chiave può contenere una o più parole utilizzate per descrivere ciascuna categoria. Queste parole saranno importate come descrittori o ignorate a seconda di quanto specificato nell'ultimo passo della procedura guidata.

Successivamente, i descrittori vengono confrontati con i risultati estratti dal testo. Se viene trovata una corrispondenza, tale record o documento viene calcolato nella categoria contenente questo descrittore.

Tabella 26. Esempio di formato compatto con i codici

Colonna A	**Colonna B**	**Colonna C**
Livello gerarchico di codice	Codice di categoria (*facoltativo*)	Nome categoria
Livello gerarchico di codice	Codice di sottocategoria (*facoltativo*)	Nome sottocategoria

Tabella 27. Esempio di formato compatto senza codici

Colonna A	**Colonna B**
Livello gerarchico di codice	Nome categoria
Livello gerarchico di codice	Nome sottocategoria

Formato impresso

Nel formato file Impresso, il contenuto è gerarchico, il che significa che contiene categorie e uno o più livelli di sottocategorie. Inoltre, la sua struttura è impressa per denotare questa gerarchia. Ogni riga nel file contiene una categoria o una sottocategoria, ma le sottocategorie sono impresse dalle categorie e le sottocategorie sono impresse dalle sottocategorie secondarie e così via. È possibile creare manualmente questa struttura in Microsoft Excel o utilizzarne una che è stato esportata da un altro prodotto e salvata in un formato Microsoft Excel.

- I **codici e i nomi categoria di livello superiore** occupano rispettivamente le colonne A e B. Oppure, se non sono presenti codici, il nome della categoria è nella colonna A.
- I **codici e i nomi di sottocategoria** occupano rispettivamente le colonne B e C. Oppure, se non sono presenti codici, il nome della sottocategoria è nella colonna B. La sottocategoria è un membro di una categoria. Non è possibile avere sottocategorie se non si dispone di categorie di livello superiore.

Tabella 28. Struttura compatta con codici

Colonna A	**Colonna B**	**Colonna C**	**Colonna D**
Codice categoria (*facoltativo*)	Nome categoria		
	Codice di sottocategoria (*facoltativo*)	Nome sottocategoria	
		Codice di sottocategoria secondaria (*facoltativo*)	Nome sottocategoria secondaria

Tabella 29. Struttura compatta senza codici

Colonna A	**Colonna B**	**Colonna C**
Nome categoria		
	Nome sottocategoria	
		Nome sottocategoria secondaria

Le seguenti informazioni possono essere contenute in un file di questo formato:

- I **codici** facoltativi contengono valori che identificano in modo univoco ciascuna categoria o sottocategoria. Se si specifica che il file di dati non contiene codici (opzione **Contiene codici di categoria** nel passo **Impostazioni del contenuto**), deve essere presente un codice univoco per ciascuna categoria o sottocategoria nella cella direttamente alla sinistra del nome categoria/sottocategoria. Se i dati non contengono codici ma si desidera creare alcuni codici successivamente, è sempre possibile generare codici in (**Categorie> Gestisci categorie > Crea codici automaticamente**).

- Un **nome** *richiesto* per ogni categoria e sottocategoria. Le sottocategorie devono rientrare dalle categorie da una cella a destra in una riga separata.
- **Annotazioni** facoltative nella cella immediatamente a destra del nome categoria. Questa annotazione consiste in un testo che descrive le categorie e le sottocategorie.
- **Parole chiave** facoltative possono essere importate come descrittori per le categorie. Per essere riconosciute, queste parole chiave devono essere presenti nella cella direttamente al di sotto del nome categoria o sottocategoria associato e l'elenco di parole chiave deve essere preceduto dal carattere di sottolineatura (_) come `_armi da fuoco, missili / fucili`. La cella di parola chiave può contenere una o più parole utilizzate per descrivere ciascuna categoria. Queste parole saranno importate come descrittori o ignorate a seconda di quanto specificato nell'ultimo passo della procedura guidata. Successivamente, i descrittori vengono confrontati con i risultati estratti dal testo. Se viene trovata una corrispondenza, tale record o documento viene calcolato nella categoria contenente questo descrittore.

Importante! Se si utilizza un codice in un livello, è necessario includere un codice per ciascuna categoria e sottocategoria. In caso contrario, il processo di importazione non riesce.

Esportazione di categorie

È anche possibile esportare le categorie disponibili in una sessione workbench interattiva in un formato di file Microsoft Excel (**.xls*, **.xlsx*). I dati che verranno esportati provengono in gran parte dall'attuale contenuto del riquadro Categorie o dalle proprietà della categoria. Pertanto, si consiglia di eseguire il punteggio di nuovo se si pensa di esportare anche il punteggio **Docs.**.

Tabella 30. Opzioni di esportazione di categoria

Esportare sempre...	Esportare facoltativamente...
• Codici di categoria, se presenti • Nomi di categoria (e sottocategoria) • Livelli di codice, se presenti (formato *Semplice/Compatto*) • Intestazioni di colonna (formato *Semplice/Compatto*)	• **Documenti.** punteggi • Annotazioni di categoria • Nomi descrittore • Conteggi descrittori

Importante! Quando si esportano i descrittori, vengono convertiti in stringhe di testo e preceduti da un carattere di sottolineatura. Se si reimporta in questo prodotto, di perde la capacità di distinguere tra i descrittori che sono modelli, quelli che sono regole di categoria e quelli che sono concetti normali. Se si intende riutilizzare queste categorie in questo prodotto, si consiglia vivamente di creare un pacchetto di analisi del testo (TAP) poiché il formato TAP mantiene tutti i descrittori così come sono attualmente definiti e anche tutte le categorie, codici, ed anche le risorse linguistiche utilizzate. I file TAP possono essere utilizzati in IBM SPSS Modeler Text Analytics e IBM SPSS Text Analytics for Surveys . Per ulteriori informazioni, consultare la sezione “Uso dei pacchetti di analisi del testo (TAP)” a pagina 141.

Per esportare categorie predefinite

1. Dai menu del workbench interattivo, scegliere **Categorie Categorie > Gestisci categorie > Esporta categorie**. Viene visualizzata una procedura guidata Esporta categorie.
2. Scegliere la posizione e immettere il nome del file che verrà esportato.
3. Inserire un nome per il file di output nella casella di testo Nome file.
4. Per scegliere il formato in cui si desidera esportare i dati della categoria, fare clic su **Avanti**.
5. Scegliere il formato tra questi:
   - **Formato di elenco semplice o compatto**: consultare la sezione “Formato elenco semplice” a pagina 137 per ulteriori informazioni. L'elenco semplice non contiene sottocategorie. Consultare la sezione “Formato compatto” a pagina 138 per ulteriori informazioni. Il formato di elenco compatto contiene categorie gerarchiche.
   - **Formato impresso**: per ulteriori informazioni consultare la sezione “Formato impresso” a pagina 139.

6. Per iniziare a scegliere il contenuto da esportare e per esaminare i dati proposti, fare clic su **Avanti**.
7. Rivedere il contenuto per il file esportato.
8. Selezionare o deselezionare le impostazioni del contenuto aggiuntivo per essere esportate come **Annotazioni** o **Nomi descrittore**.
9. Per esportare le categorie, fare clic su **Fine**.

Uso dei pacchetti di analisi del testo (TAP)

Un pacchetto di analisi del testo, chiamato anche TAP, è un modello per la categorizzazione della risposta del testo. L'uso di un TAP è un modo facile per categorizzare i dati di testo con un intervento minimo poiché esso contiene serie di categoria precostituite e le risorse linguistiche necessarie per codificare un vasto numero di record in modo rapido e automatico. Utilizzando le risorse linguistiche, i dati di testo vengono analizzati e scavati per estrarre concetti chiave. In base ai concetti chiave e modelli trovati nel testo, i record possono essere classificati nella serie di categorie selezionata nella TAP. È possibile creare il proprio TAP o aggiornare uno.

Un TAP è composto dai seguenti elementi:

- **Serie di categorie**. Una serie di categorie è composta essenzialmente da categorie predefinite, codici di categoria, descrittori per ciascuna categoria ed, infine, da un nome per una serie intera di categorie. I descrittori sono elementi linguistici (concetti, tipi, modelli e regole), come il termine *a buon mercato* o il modello *buon prezzo*. I descrittori vengono utilizzati per definire una categoria in modo tale che quando il testo corrisponde ad un descrittore di categoria, il documento o record viene inserito nella categoria.
- **Risorse linguistiche**. Le risorse linguistiche sono una serie di librerie e risorse avanzate che vengono ottimizzate per estrarre concetti chiave e modelli. Questi concetti e modelli di estrazione, a loro volta, sono utilizzati come i descrittori che consentono che i record vengano inseriti in una categoria nella serie di categorie.

È possibile creare il proprio TAP, aggiornarne uno o caricare i pacchetti di analisi del testo.

Dopo aver selezionato il TAP e un insieme di categorie, IBM SPSS Modeler Text Analytics può estrarre e classificare i record in categorie.

Nota: i TAP possono essere creati ed utilizzati in modo interscambiabile tra IBM SPSS Text Analytics for Surveys e IBM SPSS Modeler Text Analytics.

Creazione dei pacchetti di analisi del testo

Ogni qualvolta si ha un sessione con almeno una categoria e alcune risorse, è possibile creare un pacchetto di analisi del testo (TAP) dal contenuto della sessione workbench interattiva. La serie di categorie e i descrittori (concetti, tipi, regole o output del modello TLA) può essere creata in un TAP insieme a tutte le risorse linguistiche aperte nell'editor della risorsa.

È possibile visualizzare la lingua per cui sono state create le risorse. La lingua viene impostata nella scheda Risorse avanzate di Editor di modelli o di Editor risorse.

Per creare un pacchetto di analisi del testo

1. Dai menu, scegliere **File > Pacchetti di analisi del testo > Crea pacchetto**. Viene visualizzata la finestra di dialogo Crea pacchetto.
2. Individuare la directory in cui si desidera salvare il TAP. Per impostazione predefinita, i TAP vengono salvati nella sottodirectory `\TAP` della directory di installazione del prodotto.
3. Immettere un nome per il TAP nel campo **Nome file**.
4. Immettere un'etichetta nel campo **Etichetta pacchetto**. Quando si immette un nome file, questo nome viene visualizzato automaticamente come etichetta ma è possibile modificare questa etichetta.

5. Per escludere una categoria impostata dal TAP, deselezionare la casella di spunta **Includi**. In questo modo si garantisce che non viene aggiunta al pacchetto. Per impostazione predefinita, nel TAP è inclusa una serie di categorie per ogni domanda. Nel TAP deve essere sempre presente almeno una serie di categorie.
6. Rinominare tutti gli insiemi di categorie. La colonna **Nuova serie di categorie** contiene, per impostazione predefinita, nomi generici che sono generati aggiungendo il prefisso `Cat_` al nome variabile del testo. Un singolo clic nella cella rende il nome modificabile. Immettere o fare clic altrove per applicare la ridenominazione. Se si ridenomina una serie di categorie, solo il nome cambia in TAP e non si modifica il nome della variabile nella sessione aperta.
7. Se si desidera riordinare le serie di categorie utilizzare i tasti freccia a destra della tabella della serie di categorie.
8. Fare clic su **Salva** per creare il pacchetto di analisi del testo. La casella di dialogo viene chiusa.

Caricamento dei pacchetti di analisi del testo (TAP)

Quando si configura un nodo di modellazione testo, è necessario specificare le risorse che verranno utilizzate durante l'estrazione. Invece di scegliere un modello di risorsa, è possibile selezionare un pacchetto di analisi del testo (TAP) per copiare non solo le proprie risorse, ma anche una categoria impostata nel nodo.

I TAP sono più interessanti quando si crea un modello interattivo di categoria poiché è possibile utilizzare l'insieme di categorie come punto di partenza per la categorizzazione. Quando si esegue il flusso, la sessione workbench interattiva viene avviata e questa serie di categorie viene visualizzata nel riquadro Categorie. In questo modo, è possibile calcolare il punteggio dei propri documenti e record utilizzando queste categorie e quindi continuare a perfezionare, creare ed estendere tali categorie finché non soddisfano le proprie esigenze. Consultare la sezione “Metodi e strategie per la Creazione di categorie” a pagina 104 per ulteriori informazioni.

A partire dalla versione 14, è possibile anche visualizzare la lingua per la quale le risorse in questo TAP sono state definite quando si fa clic su **Carica** e scegliere il TAP.

Per caricare un pacchetto di analisi del testo

1. Aprire il nodo di modellazione dell'estrazione di testo.
2. Nella scheda Modelli, scegliere *Pacchetto di analisi del testo* nella sezione **Copia risorse da**.
3. Fare clic su **Carica**. Viene visualizzata la finestra di dialogo Carica pacchetto di analisi del testo.
4. Passare al percorso TAP contenente le risorse e impostare la categoria che si desidera copiare nel nodo. Per impostazione predefinita, i TAP vengono salvati nella sottodirectory `\TAP` della directory di installazione del prodotto.
5. Immettere un nome per il TAP nel campo **Nome file**. L'etichetta viene visualizzata automaticamente.
6. Selezionare la categoria che si desidera utilizzare. Questa è la serie di categorie che verranno visualizzate nella sessione workbench interattiva. È possibile quindi selezionare e migliorare queste categorie manualmente o utilizzando l'opzione di creazione o di estensione delle categorie.
7. Fare clic su **Carica** per copiare il contenuto del pacchetto di analisi del testo nel nodo. La casella di dialogo viene chiusa. Quando viene caricato un TAP, nel nodo viene inserita una copia del TAP; di conseguenza, tutte le modifiche operate sulle risorse e sulle categorie non si rifletteranno nel TAP a meno che non viene esplicitamente aggiornato e ricaricato.

Aggiornamento dei pacchetti di analisi del testo

Se si apportano miglioramenti a un insieme di categorie, alle risorse linguistiche o si crea una nuova categoria per intero, è possibile aggiornare un pacchetto di analisi del testo (TAP) per facilitare il riutilizzo di tali miglioramenti. Per eseguire questa operazione, è necessario trovarsi nella sessione aperta

contenente le informazioni che si desidera inserire nel TAP. Quando si aggiorna, è possibile scegliere di accodare insiemi di categorie, sostituire le risorse, modificare l'etichetta del pacchetto, oppure ridenominare/riordinare gli insiemi di categoria.

Per aggiornare un pacchetto di analisi del testo

1. Dai menu, scegliere **File > Pacchetti di analisi del testo > Aggiorna pacchetto**. Viene visualizzata la finestra di dialogo Aggiorna pacchetto.
2. Andare alla directory che contiene il pacchetto di analisi del testo che si desidera aggiornare.
3. Immettere un nome per il TAP nel campo **Nome file**.
4. Per sostituire le risorse linguistiche all'interno del TAP con quelle della sessione corrente, selezionare l'opzione **Sostituisci le risorse in questo pacchetto con quelle nella sessione aperta**. In genere è sensato aggiornare le risorse linguistiche poiché sono state utilizzate per estrarre i concetti chiave e i modelli utilizzati per creare le definizioni di categoria. Disporre delle risorse linguistiche più recenti garantisce di ottenere i migliori risultati nella categorizzazione dei record. Se non si seleziona questa opzione, le risorse linguistiche che erano già nel pacchetto rimangono invariate.
5. Per aggiornare solo le risorse linguistiche, assicurarsi di selezionare l'opzione **Sostituisci le risorse in questo pacchetto con quelle nella sessione aperta** e selezionare solo le serie di categorie correnti che erano già nel TAP.
6. Per includere la nuova serie di categorie dalla sessione aperta nel TAP, selezionare la casella di spunta per ogni serie di categorie da aggiungere. È possibile aggiungere una, diverse o nessuna delle serie di categorie.
7. Per rimuovere una serie di categorie dal TAP, deselezionare la corrispondente casella di spunta **Includi**. È possibile scegliere di rimuovere una serie di categorie che era già nel TAP poiché se ne sta aggiungendo una migliore. A tal fine, deselezionare la casella di spunta **Includi** per la serie di categorie corrispondente impostata nella colonna Serie di categorie corrente. Nel TAP deve essere sempre presente almeno una serie di categorie.
8. Ridenominare gli insiemi di categoria se necessario. Un singolo clic nella cella rende il nome modificabile. Immettere o fare clic altrove per applicare la ridenominazione. Se si ridenomina una serie di categorie, solo il nome cambia nel TAP e non si modifica il nome della variabile nella sessione aperta. Se due serie di categorie hanno lo stesso nome, il nome verrà visualizzato in rosso fino a quando non si corregge il duplicato.
9. Per creare un nuovo pacchetto con il contenuto della sessione integrato con il contenuto del TAP selezionato, fare clic su **Salva con nome nuovo**. Viene visualizzata la finestra di dialogo Salva con nome pacchetto di analisi del testo. Consultare le istruzioni seguenti.
10. Fare clic su **Aggiorna** per salvare le modifiche apportate al TAP selezionato.

Per salvare un pacchetto di analisi del testo

1. Individuare la directory in cui si desidera salvare il file TAP. Per impostazione predefinita, i file TAP vengono salvati nella sottodirectory TAP della directory di installazione.
2. Immettere un nome per il file TAP nel campo Nome file.
3. Immettere un'etichetta nel campo Etichetta pacchetto. Quando si immette un nome file, tale nome viene utilizzato automaticamente come etichetta. Tuttavia, è possibile ridenominare questa etichetta. È necessario disporre di un'etichetta.
4. Fare clic su **Salva** per creare il nuovo pacchetto.

Modifica e perfezionamento delle categorie

Dopo aver creato alcune categorie, spesso si desidera esaminare e apportare alcune modifiche. Oltre a perfezionare le risorse linguistiche, è necessario esaminare le categorie, cercando modi per combinarle o per ripulire le loro definizioni e controllare alcuni documenti o record inseriti nelle categorie. È anche possibile consultare i documenti o record in una categoria ed apportare modifiche in modo che le categorie sono definite in modo da catturare sfumature e differenze.

È possibile utilizzare le tecniche integrate, automatiche, di creazione categoria per creare le categorie; tuttavia, è probabile che si desidera eseguire alcune modifiche a tali categorie. Dopo avere utilizzato uno o più tecniche, nella finestra viene visualizzata una serie di nuove categorie. È possibile quindi esaminare i dati in una categoria e apportare modifiche fino a che non si è soddisfatti delle definizioni di categoria. Consultare la sezione “Informazioni sulle categorie” a pagina 109 per ulteriori informazioni.

Di seguito sono riportate alcune opzioni per perfezionare le categorie , la maggior parte delle quali sono descritte nelle seguenti pagine:

Aggiunta di descrittori alle categorie

Dopo aver utilizzato tecniche automatizzate, è molto probabile che si hanno ancora risultati di estrazione che non sono stati utilizzati in tutte le definizioni di categoria. È necessario controllare questo elenco nel riquadro Risultati di estrazione. Se si desidera trovare gli elementi che si desidera spostare in una categoria, è possibile aggiungerli a una categoria esistente o nuova.

Per aggiungere un concetto o tipo ad una categoria

1. Dall'interno dei riquadri Risultati di estrazione e Dati, selezionare gli elementi che si desidera aggiungere a una categoria esistente.
2. Dai menu, scegliere **Categorie > Aggiungi a categoria**. La finestra di dialogo Tutte le categorie visualizza la serie di categorie. Selezionare la categoria a cui si desidera aggiungere gli elementi selezionati. Se si desidera aggiungere elementi ad una nuova categoria, selezionare **Nuova categoria**. Una nuova categoria viene visualizzata nel riquadro Categorie utilizzando il nome del primo elemento selezionato.

Modifica dei descrittori di categoria

Dopo aver creato alcune categorie, è possibile aprire ogni categoria per visualizzare tutti i descrittori che costituiscono la sua definizione. All'interno della finestra di dialogo Definizioni di categoria, è possibile apportare una serie di modifiche ai descrittori di categoria. Inoltre, se le categorie sono visualizzate nella struttura ad albero della categoria, è possibile gestirle nella struttura.

Per modificare una categoria

1. Selezionare la categoria che si desidera modificare nel riquadro Categorie.
2. Dal menu, scegliere **Visualizza > Definizioni di categoria**. Viene visualizzata la finestra di dialogo Definizioni di categoria.
3. Selezionare il descrittore che si desidera modificare e fare clic sul pulsante della barra degli strumenti corrispondente.

La seguente tabella descrive ciascun pulsante della barra degli strumenti che è possibile utilizzare per modificare le definizioni di categoria.

Tabella 31. Pulsanti e descrizioni della barra degli strumenti.

Icone	Descrizione
	Rimuove dalla categoria i descrittori selezionati.
	Sposta i descrittori selezionati a una categoria nuova o esistente.
	Sposta i descrittori selezionati nella forma di una regola di categoria & su una categoria. Consultare la sezione “Uso delle regole di categoria” a pagina 126 per ulteriori informazioni.
	Sposta ognuno dei descrittori selezionati sulla propria nuova categoria

Tabella 31. Pulsanti e descrizioni della barra degli strumenti (Continua).

Icone	Descrizione
 **Visualizza**	Aggiorna ciò che viene visualizzato nel pannello di dati e il riquadro di visualizzazione in base ai descrittori selezionati

Spostamento di categorie

Se si desidera inserire una categoria in un'altra categoria esistente o spostare descrittori in un'altra categoria, è possibile eseguire questo cambiamento.

Per spostare una categoria

1. Nel riquadro Categorie, selezionare le categorie che si desidera spostare in un'altra categoria.
2. Dai menu, scegliere **Categorie > Sposta su categoria**. Il menu presenta una serie di categorie con in cima all'elenco quella creata più recentemente. Selezionare il nome della categoria su cui si desidera spostare i concetti selezionati.

- Se si desidera visualizzare il nome che si sta cercando, selezionarlo e gli elementi selezionati vengono aggiunti a tale categoria.
- Se non viene visualizzato, selezionare **Altro** per visualizzare la finestra di dialogo Tutte le categorie e selezionare la categoria dall'elenco.

Livellamento delle categorie

Quando si dispone di una struttura gerarchica di categoria con categorie e sottocategorie, è possibile livellare la struttura. Quando si esegue il livellamento di una categoria, tutti i descrittori nelle sottocategorie di tale categoria vengono spostati nella categoria selezionata e le categorie secondarie ora vuote vengono eliminate. In questo modo, tutti i documenti che corrispondevano alle sottocategorie sono ora inseriti nella categoria selezionata.

Per livellare una categoria

1. Nel riquadro Categorie, selezionare una categoria (di livello superiore o sottocategoria) che si desidera livellare.
2. Dal menu, scegliere **Categorie > Livella categorie**. Le sottocategorie vengono rimosse e i descrittori vengono uniti nella categoria selezionata.

Unione o combinazione di categorie

Se si desidera combinare due o più categorie esistenti in una nuova categoria, è possibile unirle. Quando si uniscono le categorie, viene creata una nuova categoria con un nome generico. Tutti i concetti, tipi e modelli utilizzati nei descrittori di categoria vengono spostati in questa nuova categoria. È possibile ridenominare questa categoria successivamente modificando le proprietà della categoria.

Per unire una categoria o parte di una categoria

1. Nel riquadro Categorie, selezionare gli elementi che si desidera unire insieme.
2. Dal menu, scegliere **Categorie > Unisci categorie**. Viene visualizzato la finestra di dialogo Proprietà di categoria in cui è possibile immettere un nome per la categoria appena creata. Le categorie selezionate vengono unite nella nuova categoria come sottocategorie.

Eliminazione di categorie

Se non si desidera più conservare una categoria, è possibile eliminarla.

Per eliminare una categoria

1. Nel pannello Categorie, selezionare la categoria o le categorie che si desidera eliminare.

2. Dal menu scegliere **Modifica > Elimina**.

Capitolo 11. Analisi dei cluster

È possibile creare e analizzare cluster di concetti nella vista Cluster (**Visualizza > Cluster**). Un **cluster** è un raggruppamento di concetti correlati generati da algoritmi di cluster in base alla frequenza in cui questi concetti ricorrono nella serie di documenti/record e quanto spesso vengono visualizzati insieme nello stesso documento, cosa conosciuta anche come **ricorrenza**. Ogni concetto all'interno di un cluster ricorre con almeno un altro concetto nel cluster. L'obiettivo dei cluster è di raggruppare concetti che ricorrono insieme in base a come il testo che contengono corrisponda ai descrittori (concetti, regole, modelli) per ogni categoria.

Un buon cluster è quello con concetti che sono fortemente collegati e ricorrono spesso e con pochi collegamenti a concetti in altri cluster. Quando si lavora con insiemi di dati più grandi, questa tecnica può risultare in tempi di elaborazione significativamente maggiori.

Nota: utilizzare l'opzione **Numero massimo di documenti da utilizzare per il calcolo dei cluster** nella finestra di dialogo Crea cluster per creare con un solo sottoinsieme di tutti i documenti o record.

Il raggruppamento in cluster è un processo che inizia analizzando una serie di concetti e ricercando una serie di concetti che ricorrono spesso nei documenti. Due concetti che ricorrono in un documento vengono considerati come una coppia di concetti. Successivamente, il processo di raggruppamento in cluster valuta **il valore di similitudine** di ogni concetto, confrontando il numero di documenti in cui la coppia compare insieme con il numero di documenti in cui ogni concetto ricorre. Consultare la sezione “Calcolo dei valori di collegamento di similitudine” a pagina 150 per ulteriori informazioni.

Infine, il processo di raggruppamento in cluster raggruppa concetti simili per aggregazione nei cluster e prende in considerazione i valori di collegamento e le impostazioni definiti nella finestra di dialogo Crea cluster. Per aggregazione si intende che concetti vengono aggiunti o cluster più piccoli vengono uniti in un cluster più grande fino a che il cluster non è saturo. Un cluster è **saturo** quando l'unione di concetti cluster più piccoli potrebbe causare che il cluster superi il limite delle impostazioni nella finestra di dialogo Crea cluster (numero di concetti, collegamenti interni o link esterni). Un cluster prende il nome del concetto all'interno del cluster che ha il numero più alto complessivo di collegamenti ad altri concetti all'interno del cluster.

Alla fine, non tutte le coppie concetto finiscono insieme nello stesso cluster poiché potrebbe sussistere un legame più forte in un altro cluster o la saturazione potrebbe impedire l'unione dei cluster in cui esse ricorrono. Per questo motivo, esistono entrambi i collegamenti interni ed esterni.

- I **collegamenti interni** sono collegamenti tra coppie concetto all'interno di un cluster. Non tutti i concetti sono collegati tra loro in un cluster. Tuttavia, ogni concetto viene collegato ad almeno un altro concetto all'interno del cluster.
- I **collegamenti esterni** sono collegamenti tra coppie concetto in cluster separati (un concetto all'interno di un cluster e un concetto esterno in un altro cluster).

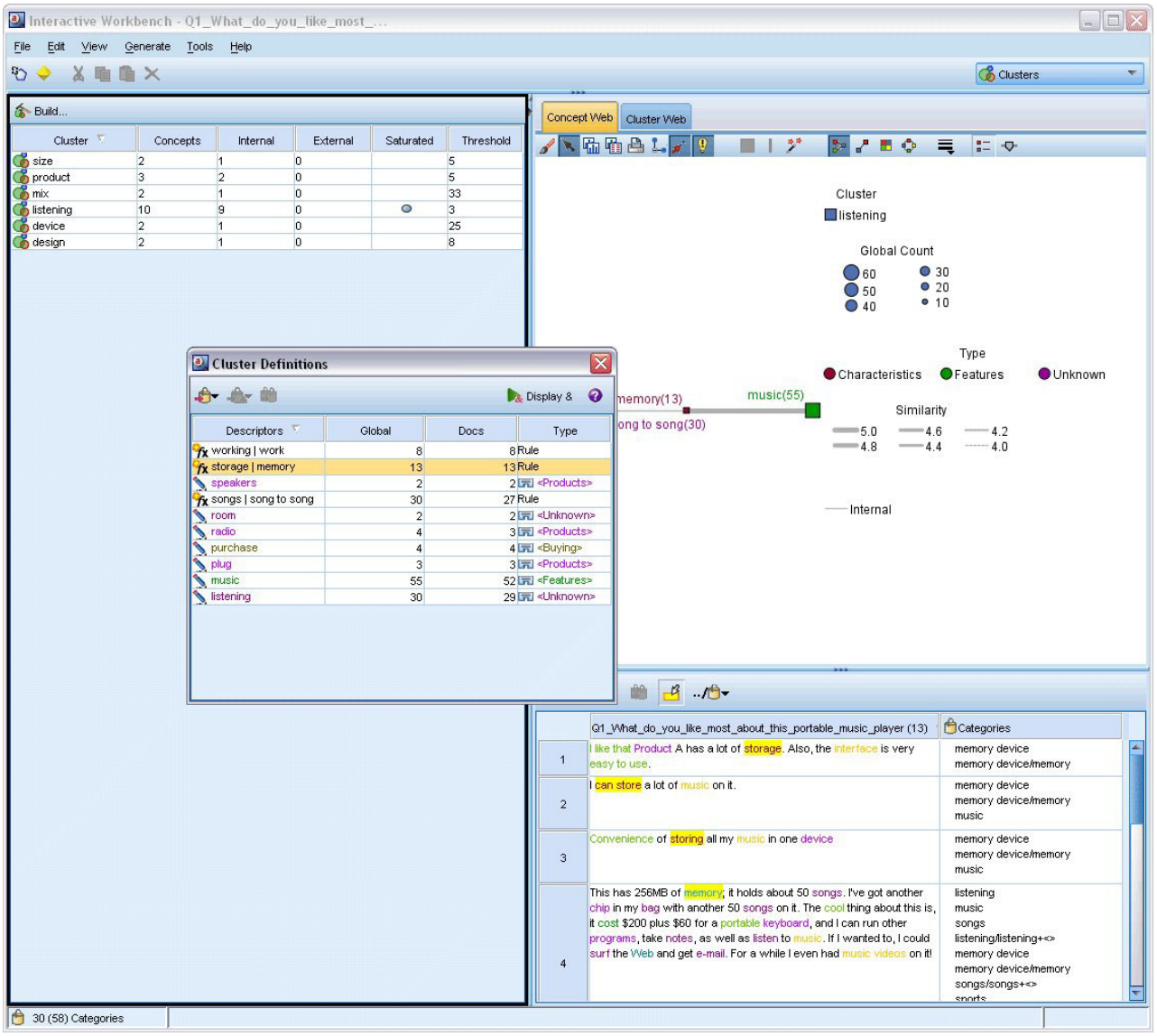


Figura 30. Vista Cluster

La vista Cluster è organizzata in tre riquadri, ognuno dei quali può essere nascosto o mostrato selezionandone il nome dal menu Visualizza:

- **Riquadro Cluster.** È possibile creare e gestire i cluster in questo riquadro. Consultare la sezione "Esplorazione dei cluster" a pagina 151 per ulteriori informazioni.
- **Riquadro Visualizzazione.** In questo riquadro è possibile esplorare visivamente i cluster e il modo in cui essi interagisco. Consultare la sezione "Grafici del cluster" a pagina 161 per ulteriori informazioni.
- **Riquadro Dati.** È possibile esplorare e rivedere il testo contenuto all'interno di documenti e record che corrisponde alle selezioni nella finestra di dialogo Definizioni cluster. Consultare la sezione "Definizioni di cluster" a pagina 151 per ulteriori informazioni.

Creazione di cluster

Quando si accede per la prima volta alla vista Cluster, nessun cluster è visibile. È possibile creare i cluster attraverso il menu (**Strumenti > Creazione di cluster**) o facendo clic sul pulsante **Crea...** sulla barra degli strumenti. Questa azione apre la finestra di dialogo Crea cluster in cui è possibile definire le impostazioni e i limiti per la creazione dei cluster.

Nota: ogni volta che i risultati di estrazione non corrispondono alle risorse, questo riquadro diventa giallo, come nel pannello Risultati di estrazione. È possibile riestrarre per ottenere gli ultimi risultati di estrazione e il colore giallo scompare. Tuttavia, ogni volta che un'estrazione viene effettuata, il riquadro Cluster è deselezionato e si dovranno ricreare i cluster. Allo stesso modo i cluster non vengono salvati da una sessione all'altra.

Le seguenti aree e campi disponibili nella finestra di dialogo Crea cluster:

Input

Tabella degli **input**. I cluster sono costruiti da descrittori derivati da determinati tipi. Nella tabella è possibile selezionare i tipi da includere nel processo di creazione. Per impostazione predefinita, sono preselezionati solo i tipi che acquisiscono la maggior parte dei record o documenti .

Concetti per cluster: Selezionare il metodo di selezione dei concetti che si desidera utilizzare per il raggruppamento in cluster. Riducendo il numero di concetti, è possibile accelerare il processo di raggruppamento in cluster. È possibile raggruppare in cluster utilizzando una serie di concetti più importanti, una percentuale di concetti più importanti o utilizzando tutti i concetti:

- **Numero basato su conteggio documenti.** Quando si seleziona **Numero maggiore di concetti**, immettere il numero di concetti da considerare per il raggruppamento in cluster. I concetti sono scelti in base a quelle che hanno il valore più alto di conteggio documenti. Il conteggio documenti è il numero di documenti o record in cui il concetto viene visualizzato.
- **Percentuale basata sul conteggio del documento.** Quando si seleziona **Percentuale maggiore di concetti**, immettere la percentuale di concetti da considerare per il raggruppamento in cluster. I concetti sono scelti in base a questa percentuale di concetti che hanno il valore più alto di conteggio documenti.

Numero massimo di documenti da utilizzare per calcolare i cluster. Per impostazione predefinita, i valori di collegamento vengono calcolati utilizzando l'intera serie di documenti o record. Tuttavia, in alcuni casi, è possibile che si desideri accelerare il processo di raggruppamento in cluster limitando il numero di documenti o record utilizzati per calcolare i collegamenti. Limitare i documenti può diminuire la qualità dei cluster. Per utilizzare questa opzione, selezionare la casella di spunta alla sua sinistra e immettere il numero massimo di documenti o record da utilizzare.

Limiti di output

Numero massimo di cluster da creare. Questo valore rappresenta il numero massimo di cluster da generare e visualizzare nel riquadro Cluster. Durante il processo di raggruppamento in cluster, i cluster saturi sono presentati prima di quelli non saturi, e quindi, molti dei cluster risultanti saranno saturi. Per visualizzare più cluster non saturi, è possibile modificare questa impostazione su un valore maggiore del numero di cluster saturi.

Numero massimo di concetti in un cluster. Questo valore è il numero massimo di concetti che un cluster può contenere.

Numero minimo di concetti in un cluster. Questo valore è il numero minimo di concetti che deve essere collegato per creare un cluster.

Numero massimo di collegamenti interni. Questo valore è il numero massimo di collegamenti interni che un cluster può contenere. I collegamenti interni sono collegamenti tra coppie concetto all'interno di un cluster.

Numero massimo di collegamenti esterni. Questo valore rappresenta il numero massimo di collegamenti ai concetti al di fuori del cluster. I collegamenti esterni sono collegamenti tra coppie concetto in cluster separati.

Valore minimo di collegamento. Questo valore è il valore minimo di collegamento accettato per una coppia concetto da considerare per il cluster. Il valore di collegamento è calcolato utilizzando una formula di similarità. Consultare la sezione "Calcolo dei valori di collegamento di similitudine" per ulteriori informazioni.

Previeni l'abbinamento di concetti specifici. Selezionare questa casella di spunta per arrestare il processo di raggruppamento o di accoppiamento di due concetti insieme nell'output. Per creare o gestire coppie di concetti, fare clic su **Gestisci Coppie**. Per ulteriori informazioni, consultare la sezione "Gestione delle coppie di eccezione di collegamento" a pagina 116.

Calcolo dei valori di collegamento di similitudine

Il conoscere il numero di documenti in cui una coppia concetto ricorre non indica di per sé come sono simili i concetti. In questi casi, il valore di similitudine può essere utile. Il valore del collegamento di similitudine viene misurato utilizzando il conteggio del documento di ricorrenza confrontato con i conteggi del singolo documento per ogni concetto nella relazione. Durante il calcolo della similarità, l'unità di misura è il numero di documenti (conteggio doc) in cui un concetto o coppia concetto viene trovato. Un concetto o coppia concetto viene "trovato" in un documento se si verifica *almeno* una volta nel documento. È possibile scegliere che lo spessore di riga nel grafico Concetto rappresenti il valore di collegamento di similitudine nei grafici.

L'algoritmo rivela quelle relazioni che sono più forti, ossia che la tendenza per i concetti di apparire insieme nei dati di testo è molto più elevata rispetto alla loro tendenza a verificarsi in modo indipendente. Internamente, l'algoritmo fornisce un coefficiente di similitudine compreso tra 0 e 1, dove un valore di 1 indica che i due concetti vengono visualizzati sempre insieme e mai separatamente. Il risultato del coefficiente di similitudine viene quindi moltiplicato per 100 ed arrotondato al numero intero più vicino. Il coefficiente di similitudine è calcolato utilizzando la formula riportata nella seguente figura.

$$\text{similarity coefficient} = \frac{(C_{IJ})^2}{(C_I \times C_J)}$$

Figura 31. Formula del coefficiente di similitudine

In cui:

- C_I è il numero di documenti o record in cui si verifica il concetto `I`.
- C_J è il numero di documenti o record in cui si verifica il concetto `J`.
- C_{IJ} è il numero di documenti o record in cui la coppia concetto `I` e `J` ricorre nella serie di documenti.

Ad esempio, si supponga di avere 5.000 documenti. `I` e `J` devono essere i concetti estratti e `IJ` è una ricorrenza di coppia concetto di `I` e `J`. La seguente tabella propone due scenari per illustrare come sono calcolati il valore del coefficiente e del collegamento.

Tabella 32. Esempio di frequenze di concetto

Concetto/Coppia	Scenario A	Scenario B
Concetto: `I`	Si verifica in 20 documenti	Si verifica in 30 documenti
Concetto: `J`	Si verifica in 20 documenti	Si verifica in 60 documenti
Coppia concetto: `IJ`	Ricorre in 20 documenti	Ricorre in 20 documenti
Coefficiente di similitudine	ERROR! SEGMENT DATA CORRUPTED, SEGDATA=1	0.22222
Valore di collegamento di similitudine	100	22

Nello scenario A, i concetti I e J come la coppia IJ ricorrono in 20 documenti, fornendo un coefficiente di similitudine 1, a indicare che i concetti ricorrono sempre insieme. Il valore di collegamento di similitudine per questa coppia sarà 100.

Nello scenario B, il concetto I si verifica in 30 documenti e il concetto J in 60 documenti, ma la coppia IJ ricorre solo in 20 documenti. Di conseguenza, il coefficiente di similitudine è 0,22222. Il valore di collegamento di similitudine per questa coppia sarà arrotondato a 22.

Esplorazione dei cluster

Dopo aver creato i cluster, è possibile visualizzare una serie di risultati nel riquadro Cluster. Per ogni cluster, sono disponibili le seguenti informazioni nella tabella:

- **Cluster.** Questo è il nome del cluster. I cluster vengono denominati dopo il concetto con il più alto numero di collegamenti interni.
- **Concetti.** Questo è il numero di concetti nel cluster. Consultare la sezione "Definizioni di cluster" per ulteriori informazioni.
- **Interno.** Questo è il numero di collegamenti interni nel cluster. I collegamenti interni sono collegamenti tra coppie concetto all'interno di un cluster.
- **Esterno.** Questo è il numero di collegamenti esterni nel cluster. I collegamenti esterni sono collegamenti tra coppie concetto quando un concetto è in un cluster e l'altro concetto si trova in un altro cluster.
- **Saturo.** Se è presente un simbolo, questo indica che questo cluster potrebbe essere più grande ma uno o più limiti sono stati superati e, quindi, il processo di raggruppamento è terminato per quel cluster che viene considerato *saturo*. Alla fine del processo di raggruppamento in cluster, i cluster saturi sono presentati prima di quelli non saturi, e quindi, molti dei cluster risultanti saranno saturi. Per visualizzare più cluster non saturi, è possibile modificare l'impostazione **numero massimo di cluster da creare** su un valore maggiore del numero di cluster saturi o diminuire il **valore minimo di collegamento**. Consultare la sezione "Creazione di cluster" a pagina 148 per ulteriori informazioni.
- **Soglia.** Per tutte le coppie di concetti che ricorrono nel cluster, questo è il valore di collegamento di similitudine più bassa di tutti nel cluster. Consultare la sezione "Calcolo dei valori di collegamento di similitudine" a pagina 150 per ulteriori informazioni. Un cluster con un valore di soglia superiore indica che i concetti contenuti in quel cluster presentano similitudine complessiva superiore e sono più strettamente correlati a quelli in un cluster il cui valore soglia è inferiore.

Per saperne di più su un determinato cluster, è possibile selezionarlo e il riquadro di visualizzazione sulla destra mostrerà due grafici che consentono di esplorare il cluster. Consultare la sezione "Grafici del cluster" a pagina 161 per ulteriori informazioni. È anche possibile eseguire taglia e incolla del contenuto della tabella in un'altra applicazione.

Ogni volta che i risultati di estrazione non corrispondono alle risorse, questo riquadro diventa giallo, come nel pannello Risultati di estrazione. È possibile riestrarre per ottenere gli ultimi risultati di estrazione e il colore giallo scompare. Tuttavia, ogni volta che un'estrazione viene effettuata, il riquadro Cluster è deselezionato e si dovranno ricreare i cluster. Allo stesso modo i cluster non vengono salvati da una sessione all'altra.

Definizioni di cluster

È possibile visualizzare tutti i concetti all'interno di un cluster selezionandolo nel riquadro Cluster e aprendo la finestra di dialogo Definizioni cluster (**Visualizza > Definizioni cluster**).

Tutti i concetti nel cluster selezionato vengono visualizzati nella finestra di dialogo Definizioni cluster. Se si selezionano uno o più concetti nella finestra di dialogo Definizioni cluster e si fa clic su **Visualizza &**, il pannello Dati visualizza tutti i record o documenti in cui *tutti i concetti selezionati vengono visualizzati insieme*. Tuttavia, il pannello Dati non visualizza alcun record o documento di testo quando si seleziona un cluster nel riquadro Cluster. Per informazioni generali sul pannello Dati, consultare in.

La selezione dei concetti in questa finestra di dialogo modifica anche il grafico web del concetto. Consultare la sezione “Grafici del cluster” a pagina 161 per ulteriori informazioni. Allo stesso modo, quando si selezionano uno o più concetti nella finestra di dialogo Definizioni cluster, il riquadro di visualizzazione riporta tutti i collegamenti esterni e interni da questi concetti.

Descrizioni delle colonne

Le icone vengono visualizzate in modo da poter facilmente identificare ciascun descrittore.

Tabella 33. Colonne e icone del descrittore

Colonne	Descrizione
Descrittori	Il nome del concetto.
 **Globali**	Mostra il numero di volte in cui questo descrittore viene visualizzato nell'intero dataset, cioè la frequenza globale.
 **Documenti**	Mostra il numero di documenti o record in cui questo descrittore viene visualizzato, cioè la frequenza del documento.
Tipo	Mostra il tipo o i tipi a cui il descrittore appartiene. Se il descrittore è una regola di categoria, nessun nome tipo viene visualizzato in questa colonna.

Azioni della barra degli strumenti

Da questa casella di dialogo, è possibile anche selezionare uno o più concetti da utilizzare in una categoria. Vi sono diversi modi per farlo, ma è molto interessante per selezionare i concetti che ricorrono in un cluster e aggiungerli come regola di categoria. Consultare la sezione “Regole di ricorrenza” a pagina 120 per ulteriori informazioni. È possibile utilizzare i pulsanti della barra degli strumenti per aggiungere i concetti alle categorie.

Tabella 34. Pulsanti della barra degli strumenti per aggiungere i concetti alle categorie

Icone	Descrizione
	Aggiunge i concetti selezionati a una categoria nuova o esistente.
	Aggiunge i concetti selezionati nella forma di una regola di categoria & a una categoria nuova o esistente. Consultare la sezione “Uso delle regole di categoria” a pagina 126 per ulteriori informazioni.
	Aggiunge ognuno dei concetti selezionati alla propria nuova categoria
	Aggiorna ciò che viene visualizzato nel pannello di dati e il riquadro di visualizzazione in base ai descrittori selezionati

Nota: è anche possibile aggiungere concetti a un tipo, come sinonimi o come elemento di esclusione, utilizzando il menu di scelta rapida.

Capitolo 12. Esplorazione di analisi di collegamento del testo

Nel campo TLA della vista, è possibile esplorare i risultati dei modelli di analisi di collegamento del testo. L'analisi di collegamento del testo è una tecnologia di corrispondenza modello che consente di definire le regole del modello e confrontarle con i concetti estratti e le relazioni trovati nel testo.

Ad esempio l'estrazione di idee su un'organizzazione potrebbe non risultare interessante. Utilizzando TLA, è possibile avere anche informazioni sui collegamenti tra questa organizzazione e altre organizzazioni o le persone all'interno di un'organizzazione. È inoltre possibile utilizzare TLA per estrarre le opinioni sui prodotti o, per alcune lingue, le relazioni tra geni.

Una volta estratti alcuni risultati di modello TLA, è possibile rivedere nei modelli di Tipo e Concetto pannelli della vista Analisi di collegamento del testo. Consultare la sezione “Modelli di tipo e concetto” a pagina 155 per ulteriori informazioni. È possibile esplorare ulteriormente nei pannelli Dati e Visualizzazione in questa vista. E cosa più importante, è possibile aggiungerli alle categorie.

Se non è già stato scelto di farlo, è possibile fare clic su **Estrai** e scegliere **Abilita estrazione del modello di analisi di collegamento del testo** nella finestra di dialogo Impostazioni di estrazione. Consultare la sezione “Estrazione dei risultati di modello TLA” a pagina 154 per ulteriori informazioni.

Sono presenti alcune regole di modello TLA le regole definite nel modello di risorsa o nelle librerie che si sta utilizzando per estrarre i risultati del modello TLA. È possibile utilizzare i modelli TLA in alcuni modelli di risorsa forniti con IBM SPSS Modeler Text Analytics. Il tipo di relazioni e modelli è possibile estrarre dipendono interamente sulle regole definite nelle proprie risorse TLA. È possibile definire le proprie regole TLA per tutte le lingue del testo *tranne* giapponese. I modelli sono costituite da macro, elenchi di parole e differenze di parole per formare una interrogazione booleana o una regola, che vengono confrontate con il testo di input. Consultare la sezione Capitolo 19, “Informazioni sulle regole di collegamento del testo”, a pagina 215 per ulteriori informazioni.

Ogni volta che una regola di modello TLA corrisponde al testo, questo testo può essere estratto come modello e ristrutturato come dati di output. I risultati sono quindi visibili nei riquadri della vista analisi di collegamento del testo. Ogni riquadro può essere nascosta o mostrata selezionandone il nome dal menu Visualizza:

- **Riquadri dei modelli Tipo e Concetto.** È possibile creare ed esplorare i modelli in questi due riquadri. Consultare la sezione “Modelli di tipo e concetto” a pagina 155 per ulteriori informazioni.
- **Riquadro Visualizzazione.** È possibile esplorare visivamente il modo in cui i concetti e i tipi nei propri modelli interagiscono in questo riquadro. Consultare la sezione “Grafici di analisi di collegamento del testo” a pagina 162 per ulteriori informazioni.
- **Riquadro Dati.** È possibile esplorare e rivedere il testo contenuto all'interno di documenti e record che corrispondono alle selezioni in un altro riquadro. Consultare la sezione “Riquadro Dati” a pagina 157 per ulteriori informazioni.

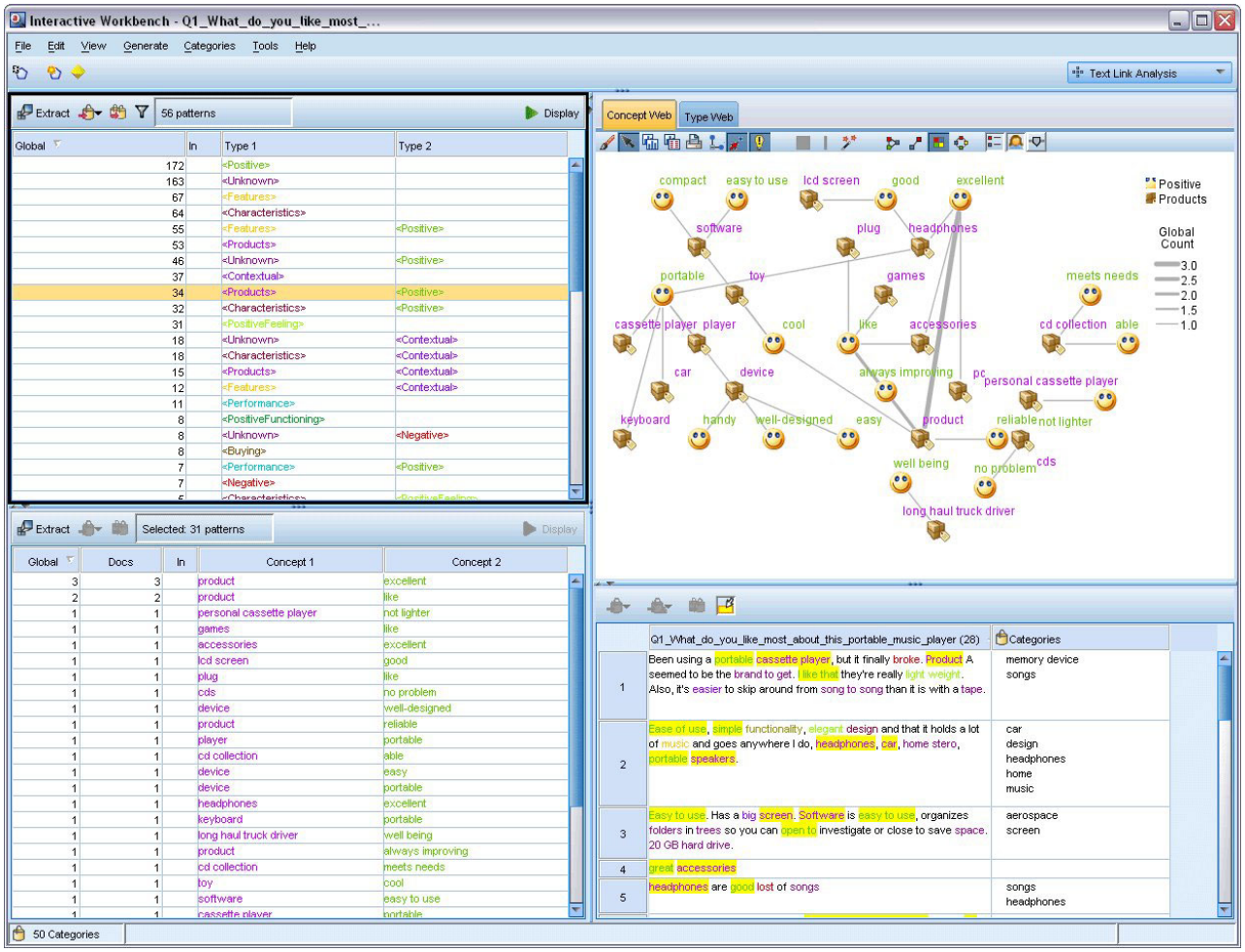


Figura 32. Visualizzazione Analisi di collegamento del testo

Estrazione dei risultati di modello TLA

Il processo di estrazione ha come risultato una serie di concetti e di tipi, come i modelli TLA, se abilitati. Se sono stati estratti i modelli TLA, è possibile visualizzarli nella vista Analisi di collegamento del testo. Quando i risultati dell'estrazione non sono sincronizzati con le risorse, i riquadri dei modelli diventano di colore giallo che indica che una nuova estrazione potrebbe produrre risultati diversi.

È inoltre necessario scegliere di estrarre questi modelli nell'impostazione nodo della finestra di dialogo Estrai utilizzando l'opzione **Abilita estrazione di modello di analisi di collegamento del testo**. Consultare la sezione “Estrazione di dati” a pagina 88 per ulteriori informazioni.

Nota: si tratta di una relazione tra la dimensione del proprio dataset e il tempo impiegato per completare il processo di estrazione. Consultare le istruzioni di installazione per le statistiche sulle prestazioni e i consigli. È possibile sempre considerare l'inserimento di un flusso a monte di nodo campione o ottimizzare la configurazione della macchina.

Per il data mining

1. Dai menu scegliere **Strumenti > Estrai**. In alternativa, fare clic sul pulsante della barra degli strumenti **Estrai**.

2. Modificare una qualsiasi delle opzioni che si desidera utilizzare. Tenere presente che l'opzione **Abilita modello di estrazione di analisi collegamenti del testo** deve essere selezionata in questa scheda e inoltre è necessario avere regole TLA nel modello per estrarre i risultati del modello TLA. Consultare la sezione "Estrazione di dati" a pagina 88 per ulteriori informazioni.
3. Fare clic su **Estrai** per avviare il processo di estrazione.

Una volta avviata l'estrazione, viene visualizzata la finestra di dialogo di avanzamento. Se si desidera interrompere l'estrazione, fare clic su **Annulla**. Una volta completata l'estrazione, la finestra di dialogo si chiude e i risultati vengono visualizzati nel riquadro. Consultare la sezione "Modelli di tipo e concetto" per ulteriori informazioni.

Modelli di tipo e concetto

I modelli sono costituiti da due parti, una combinazione di concetti e di tipi. I modelli sono molto utili quando si tenta di rilevare le opinioni su un particolare argomento o le relazioni tra concetti. Ad esempio l'estrazione del nome prodotto del proprio concorrente potrebbe non risultare interessante. In questo caso, è possibile esaminare i modelli estratti per vedere se è possibile trovare esempi in cui un record o un documento in cui gli intervistati trovano il prodotto buono, cattivo o costoso.

I modelli possono contenere fino a sei tipi o sei concetti. Per questo motivo, le righe in entrambi i pannelli di modelli contengono fino a sei slot o posizioni. Ogni slot corrisponde alla posizione specifica di un elemento nella regola modello TLA come definito nelle risorse linguistiche. Nel workbench interattivo, se uno slot non contiene valori, non viene visualizzato nella tabella. Ad esempio se i risultati di modello più lunghi non contengono più di quattro slot, gli ultimi due non vengono mostrati. Consultare la sezione Capitolo 19, "Informazioni sulle regole di collegamento del testo", a pagina 215 per ulteriori informazioni.

Quando si estraggono i risultati del modello, vengono prima raggruppati a livello di tipo e quindi suddivisi in modelli di concetto. Per questo motivo, vi sono due pannelli di risultati diversi: **Modelli di tipo** (in alto a sinistra) e **Modelli Concetto** (in basso a sinistra). Per visualizzare tutti i modelli concetto restituiti, selezionare tutti i modelli di tipo. Il riquadro inferiore dei modelli di concetto visualizza tutti i modelli di concetto fino al valore massimo di classificazione (come definito nella finestra di dialogo Filtro).

Modelli di tipo. Questo pannello visualizza i risultati del modello costituito da uno o più tipi correlati corrispondente ad una regola di modello TLA. I modelli di tipo vengono visualizzati come `<Organizzazione>` + `<Ubicazione>` + `<Positivo>`, che potrebbe fornire un feedback positivo su un'organizzazione in una posizione specifica. La sintassi è la seguente:

```
  <Tipo1> + <Tipo2> + <Tipo3> + <Tipo4> +
<Tipo5> + <Tipo6>
```

Modelli di concetto. Questo pannello presenta i risultati del modello a livello di concezione per tutti i modelli di tipo attualmente selezionati nel riquadro Modelli di tipo di cui sopra. I modelli di concetto seguono una struttura come `hotel + parigi + bellissimo`. La sintassi è la seguente:

```
  concetto1 + concetto2 + concetto3 + concetto4 + concetto5 + concetto6
```

Quando i risultati del modello utilizzano meno del massimo di sei slot, solo il numero di slot necessari (o colonne) viene visualizzato. Tutti gli slot vuoti trovati tra due slot riempiti vengono eliminati in modo tale che il modello di `<Tipo1>+<>+<Tipo2>+<>+<>+<>` può essere rappresentata da `<Tipo1>+<Tipo3>`. Per un modello di concetto, questa sarebbe `concetto1+.+concetto2` (dove `.` rappresenta un valore nullo.

Così come con i risultati di estrazione nella vista Categorie e concetti, è possibile anche qui visualizzare i risultati. Se si desidera visualizzare tutti i miglioramenti si desidera apportare ai tipi e concetti che costituiscono questi modelli, è possibile farlo nel riquadro Risultati di estrazione nella vista Categorie e concetti o direttamente nell'Editor di risorsa ed estrarre di nuovo i modelli. Ogni volta che un concetto,

tipo o modello viene utilizzato in una definizione di categoria come è o come parte di una regola, un'icona di categoria o di regola viene visualizzata nella colonna **interna** nella tabella dei risultati del modello o di estrazione.

Filtro dei risultati TLA

Quando si lavora con insiemi di dati di grandi dimensioni, il processo di estrazione potrebbe produrre milioni di risultati. Per molti utenti, tale quantità può rendere più difficile esaminare i risultati effettivi. È possibile, tuttavia, filtrare tali risultati per eseguire lo zoom su quelli che sono più interessanti. È possibile modificare le impostazioni nella finestra di dialogo Filtra per limitare i modelli visualizzati. Tutte queste impostazioni sono utilizzate insieme.

Nella vista TLA, la finestra di dialogo Filtro contiene le seguenti aree e campi.

Filtro per frequenza. È possibile filtrare in modo da visualizzare solo i risultati con un certo valore di frequenza globale o del documento.

- La **frequenza globale** è il numero totale di volte in cui un modello viene visualizzato nell'intera serie di documenti o record e viene visualizzata nella colonna **Globale**.
- La **frequenza del documento** è il numero totale di documenti o record in cui un modello viene visualizzato e mostrato nella colonna **Documenti**.

Ad esempio, se un modello è apparso 300 volte in 500 record, questo concetto avrebbe una frequenza globale di 300 e una frequenza di documento di 500.

E per testo corrispondente. È anche possibile filtrare per visualizzare solo i risultati che corrispondono alla regola definita qui. Immettere la serie di caratteri corrispondenti nel campo **Corrispondenza con testo** e selezionare se ricercare questo testo all'interno dei nomi concetto o tipo identificando il numero di slot o tutti. Quindi selezionare la condizione nella quale applicare la corrispondenza (non è necessario utilizzare le virgolette semplici per denotare l'inizio o la fine di un nome tipo). Selezionare **And** o **Or** dall'elenco a discesa in modo che la regola corrispondi ad entrambe le istruzioni o ad almeno una e definire la seconda istruzione di corrispondenza di testo nella stessa maniera della prima.

Tabella 35. Condizioni del testo di corrispondenza

Condizione	Descrizione
Contiene	Il testo è esatto se la stringa ricorre in qualsiasi punto. (Scelta predefinita)
Inizia con	Il testo è esatto solo se il concetto o il tipo inizia con il testo specificato.
Termina con	Il testo è esatto solo se il concetto o il tipo termina con il testo specificato.
Corrispondenza esatta	L'intera stringa deve corrispondere al nome tipo o nome concetto.

E per punteggio. È anche possibile filtrare per visualizzare solo i concetti più in alto nei modelli in base alla frequenza globale (**Globale**) o frequenza documento (**Documenti**) in ordine crescente o decrescente. Questo valore massimo limita il numero totale di modelli restituiti per la visualizzazione.

Quando il filtro viene applicato, il prodotto aggiunge i modelli di tipo fino a superare il numero massimo totale di modelli di concetto (massimo). Si comincia esaminando il tipo di modello con il livello superiore e prendendo la somma dei modelli di concetto corrispondenti. Se questa somma non supera il massimo, i modelli vengono visualizzati nella vista. Poi viene sommato il numero di modelli di concetto per il modello del tipo successivo. Se tale numero oltre al numero totale di modelli concetto nel modello tipo precedente è minore del numero massimo di livelli, tali modelli vengono visualizzati anche nella vista. Ciò continua per più modelli visualizzati possibili, senza superare il massimo.

Risultati visualizzati nel riquadro Modelli

Se si sta utilizzando una versione inglese del software, di seguito sono riportati alcuni esempi di come i risultati potrebbero essere visualizzati sulla barra degli strumenti del riquadro Modelli in base ai filtri.

Figura 33. Risultati di filtro esempio 1

In questo esempio, la barra degli strumenti mostra che il numero di matrici restituito era limitato a causa del massimo specificato nel filtro. Se è presente un'icona viola significa che è stato raggiunto il numero massimo di modelli. Passare con il mouse sull'icona per ulteriori informazioni. Consultare la spiegazione precedente del filtro **Classifica**.

Figura 34. Risultati di filtro esempio 2

In questo esempio, la barra degli strumenti mostra che i risultati sono stati limitati utilizzando un filtro di testo (icona lente di ingrandimento). È possibile passare il mouse sopra l'icona per visualizzare quale è il testo corrispondente.

Per filtrare i risultati

1. Dai menu scegliere **Strumenti > Filtro**. Si apre la finestra di dialogo Filtro.
2. Selezionare e perfezionare i filtri che si desidera utilizzare.
3. Fare clic su **OK** per applicare i filtri e vedere i nuovi risultati.

Riquadro Dati

Quando si estrae e si esplora i modelli di analisi di collegamento del testo, è possibile che si desideri rivedere alcuni dei dati che si stanno gestendo. Ad esempio, si potrebbe voler visualizzare i record effettivi nel quale un gruppo di modelli sono stati rilevati. È possibile rivedere i record o i documenti nel riquadro Dati che si trova in basso a destra. Se non è visibile per impostazione predefinita, scegliere **Visualizza > Riquadri > Dati** dal menu.

Il pannello presenta una riga per documento o record di dati corrispondente ad una selezione nella vista, fino a un certo limite di visualizzazione. Per impostazione predefinita, il numero di documenti o record visualizzati nel riquadro Dati è limitata per consentire all'utente di visualizzare i dati più rapidamente. Tuttavia, è possibile regolare questo valore nella finestra Opzioni. Consultare la sezione “Opzioni: scheda Sessione” a pagina 82 per ulteriori informazioni.

Visualizzazione e aggiornamento del riquadro Dati

Il pannello Dati non viene aggiornato automaticamente in quanto con grandi insiemi di dati l'aggiornamento automatico potrebbe richiedere del tempo. Pertanto, ogni volta che si effettua una selezione di modelli di concetto e tipo in questa vista, fare clic su **Visualizza** per aggiornare il contenuto del riquadro Dati.

Documenti di testo o record

Se i dati di testo è sotto forma di record e il testo è relativamente breve in lunghezza, il campo di testo nel pannello Dati visualizza i dati di testo nella loro interezza. Tuttavia, quando si lavora con record e grandi insiemi, la colonna campo di testo visualizza un breve del testo e apre un riquadro Anteprima testo a destra per visualizzare la maggior parte o tutto il testo del record selezionato nella tabella. Se i

dati di testo sono sotto forma di documenti singoli, il riquadro Dati visualizza il nome file del documento. Quando si seleziona un documento, il riquadro Anteprima testo si apre con il testo del documento selezionato.

Colori ed evidenziazione

Ogni volta che si visualizzano i dati, i concetti e i descrittori che si trovano in quei documenti o record vengono evidenziati con un colore per aiutare l'utente a identificarli facilmente nel testo. Il codice colore corrisponde ai tipi a cui i concetti appartengono. È inoltre possibile passare il mouse su elementi di colore per visualizzare il concetto da cui è stato estratto e il tipo a cui era assegnato. Tutto il testo non estratto appare in nero. Di solito, queste parole non estratte sono spesso connettori (*e* o *con*), pronomi (*me* o *essi*), e verbi (*è*, *ha* o *dai*).

Colonne del riquadro Dati

Mentre la colonna del campo di testo è sempre visibile, è possibile anche visualizzare altre colonne. Per visualizzare altre colonne, scegliere **Visualizza > riquadro Dati** dal menu e quindi selezionare la colonna che si desidera visualizzare nel riquadro Dati. È possibile visualizzare le seguenti colonne:

- **"Nome campo testo" (#)/Documenti.** Aggiunge una colonna per i dati di testo da cui i concetti e il tipo sono stati estratti. Se i dati sono nei documenti, la colonna viene denominata Documenti e solo il nome file documento o percorso completo è visibile. Per vedere il testo per tali documenti è necessario cercare nel riquadro Anteprima testo. Il numero di righe nel riquadro Dati viene visualizzato tra parentesi dopo il nome della colonna. Potrebbero esserci momenti in cui non tutti i documenti o record vengono visualizzati a causa di un limite nella finestra di dialogo Opzioni utilizzato per aumentare la velocità di caricamento. Se si raggiunge il massimo, il numero sarà seguito da **- Max**. Consultare la sezione “Opzioni: scheda Sessione” a pagina 82 per ulteriori informazioni.
- **Categorie.** Elenca ognuna delle categorie a cui il record appartiene. Quando questa colonna viene visualizzata, l'aggiornamento del riquadro dei dati potrebbe richiedere un po' più di tempo per poter visualizzare le informazioni più aggiornate.
- **Classifica di attinenza.** Fornisce una classificazione per ogni record in una categoria singola. Questa classificazione di attinenza indica quanto bene il record si inserisce nella categoria confrontata rispetto agli altri record di tale categoria. Selezionare una categoria nel riquadro Categorie (riquadro in alto a sinistra) per visualizzare la classificazione. Consultare la sezione “Attinenza tra categorie” a pagina 111 per ulteriori informazioni.
- **Conteggio di categoria.** Elenca il numero di categorie a cui il record appartiene.

Capitolo 13. Visualizzazione di grafici

La vista Categorie e concetti, vista Cluster e vista Analisi collegamento del testo hanno tutte un riquadro di visualizzazione nell'angolo in alto a destra della finestra. È possibile utilizzare questo riquadro per esplorare i dati. Sono disponibili i seguenti grafici e diagrammi.

- **Vista Categorie e concetti.** Questa vista contiene tre grafici e diagrammi: *Barra di categoria*, *Web di categoria* e *Tabella Web di categoria*. In questa vista, i grafici vengono aggiornati solo quando si fa clic su **Visualizza**. Consultare la sezione "Grafici e diagrammi di categoria" per ulteriori informazioni.
- **Vista Cluster.** Questa vista presenta due grafici web: *Grafico Web di concetto* e *Grafico Web di cluster*. Consultare la sezione "Grafici del cluster" a pagina 161 per ulteriori informazioni.
- **Vista Analisi di collegamento del testo.** Questa vista presenta due grafici web: *Grafico Web di concetto* e *Grafico Web di tipo*. Consultare la sezione "Grafici di analisi di collegamento del testo" a pagina 162 per ulteriori informazioni.

Per ulteriori informazioni su tutte le barre degli strumenti e tavolozze utilizzate per la modifica di grafici, consultare la sezione sulla modifica dei grafici nella guida in linea o nel file *modeler_nodes_general_book.pdf*, disponibile nella cartella *\Documentazione\it* in IBM SPSS Modeler DVD.

Grafici e diagrammi di categoria

Durante la creazione di categorie, è importante prendere il tempo necessario per esaminare le definizioni di categoria, i documenti o record che contengono e il modo in cui le categorie si sovrappongono. Il riquadro di visualizzazione offre varie prospettive sulle proprie categorie. Il pannello Visualizzazione è situato nell'angolo superiore destro della vista Categorie e Concetti . Se non è già visibile, è possibile accedere a questo pannello dal menu Visualizza (**Visualizza > Riquadri > Visualizzazione**).

In questa vista, il riquadro di visualizzazione offre tre prospettive sulle accomunanze nella categorizzazione del documento o record . I diagrammi e i grafici in questo riquadro possono essere utilizzati per analizzare i risultati di categorizzazione e come aiuto per regolare le categorie o la notifica. Quando si regolano le categorie, è possibile utilizzare questo riquadro per esaminare le definizioni di categoria per scoprire le categorie troppo simili (ad esempio, quando condividono oltre il 75% dei loro documenti o record) o troppo distinte. Se due categorie sono troppo simili, è utile decidere di combinarle. In alternativa, è possibile decidere di perfezionare le definizioni di categoria per rimuovere alcuni descrittori da una categoria o dall'altra.

In base a quanto è selezionato nel riquadro Risultati di estrazione , nel riquadro Categorie, o nella finestra di dialogo Definizioni di categoria, è possibile visualizzare le relative interazioni tra documenti/record e categorie su ciascuna delle schede di questo riquadro. Ciascuna presenta informazioni simili ma in modo differente o con un livello di dettaglio differente. Tuttavia, per aggiornare un grafico per la selezione corrente, fare clic su **Visualizza** sulla barra degli strumenti del riquadro o sulla finestra di dialogo in cui è stata effettuata la selezione.

Il riquadro Visualizzazione nella vista Categorie e Concetti fornisce i seguenti grafici e diagrammi:

- **Grafico a barre Categoria.** Una tabella e un grafico a barre mostrano la sovrapposizione tra documenti/record corrispondenti alla propria selezione e le categorie associate. Il grafico a barre presenta anche le proporzioni di documenti/record nelle categorie rispetto al numero totale di documenti/record. Consultare la sezione "Grafico a barre di categoria" a pagina 160 per ulteriori informazioni.
- **Grafico Web Categoria.** Questo grafico mostra la sovrapposizione di documento/record per le categorie a cui appartengono i documenti/record secondo la selezione negli altri riquadri. Consultare la sezione "Grafico Web di categoria" a pagina 160 per ulteriori informazioni.

- **Tabella Web Categoria.** Questa tabella visualizza le stesse informazioni della scheda Web di categoria ma in un formato tabella. La tabella contiene tre colonne che possono essere ordinate facendo clic sulle intestazioni delle colonne. Consultare la sezione “Tabella Web di categoria” per ulteriori informazioni.

Consultare la sezione Capitolo 10, “Categorizzazione dei dati di testo”, a pagina 101 per ulteriori informazioni.

Grafico a barre di categoria

Questa scheda visualizza una tabella e un grafico a barre che mostrano la sovrapposizione tra documenti/record corrispondenti alla propria selezione e le categorie associate. Il grafico a barre presenta anche le proporzioni di documenti/record nelle categorie rispetto al numero totale di documenti o record . Non è possibile modificare il layout di questo grafico. È possibile, tuttavia, ordinare le colonne facendo clic sulle intestazioni di colonna.

Il grafico contiene le seguenti colonne:

- **Categoria.** Questa colonna presenta il nome della categoria nella selezione. Per impostazione predefinita, la categoria più comune nella selezione viene elencata per prima.
- **Barra.** Questa colonna presenta, in modo visivo, il rapporto tra i documenti o record in una categoria specificata per il numero totale di documenti o record.
- **% di selezione.** Questa colonna presenta una percentuale basata sul rapporto tra il numero totale di documenti o record di una categoria per il numero totale di documenti o record rappresentato nella selezione.
- **Documenti.** Questa colonna riporta il numero di documenti o record in una selezione per una certa categoria.

Grafico Web di categoria

Questa scheda visualizza un grafico Web di categoria. Questo grafico mostra la sovrapposizione di documenti o record per le categorie a cui appartengono i documenti o record secondo la selezione negli altri riquadri. Se esistono etichette di categoria, queste etichette sono visualizzate nel grafico. È possibile scegliere un layout grafico (rete, cerchio, orientato o griglia) utilizzando i pulsanti della barra degli strumenti in questo riquadro.

Nel web, ogni nodo rappresenta una categoria. Utilizzando il mouse è possibile selezionare e spostare i nodi all'interno del riquadro. La dimensione del nodo rappresenta la dimensione relativa in base al numero di documenti o record per tale categoria nella selezione. Lo spessore e colore della linea tra due categorie denota il numero di documenti o record che sono comuni. Se si passa il mouse su un nodo in modalità Esplora, un suggerimento visualizza il nome (o etichetta) della categoria e il numero complessivo di documenti o record nella categoria.

Nota: per impostazione predefinita, la modalità di esplorazione è abilitata per i grafici su cui è possibile spostare i nodi. Tuttavia, è possibile passare alla modalità Modifica per modificare i layout grafici inclusi colori, font, legende e altro ancora. Consultare la sezione “Uso delle barre degli strumenti e tavolozze dei grafici” a pagina 163 per ulteriori informazioni.

Tabella Web di categoria

Questa scheda visualizza le stesse informazioni della scheda Web di categoria ma in un formato tabella. La tabella contiene tre colonne che possono essere ordinate facendo clic sull'intestazione di colonna:

- **Conteggio.** Questa colonna presenta il numero di documenti o record condivisi o comuni tra le due categorie.
- **Categoria 1.** Questa colonna presenta il nome della prima categoria seguito dal numero totale di documenti o record in essa contenuti, riportati tra parentesi.
- **Categoria 2.** Questa colonna presenta il nome della seconda categoria seguito dal numero totale di documenti o record in essa contenuti, riportati tra parentesi.

Grafici del cluster

Dopo la creazione dei cluster, è possibile esplorare visivamente nei grafici Web nel riquadro di visualizzazione. Il riquadro di visualizzazione offre due prospettive sui cluster: un grafico Web del concetto e di un grafico Web del cluster. I grafici Web in questo riquadro possono essere utilizzati per analizzare i risultati del raggruppamento in cluster e come aiuto a rilevare alcuni concetti e regole che si vorrebbero aggiungere alle proprie categorie. Il riquadro Visualizzazione si trova nell'angolo in alto a destra della vista Cluster. Se non è già visibile, è possibile accedere a questo riquadro dal menu Visualizza (**Visualizza > Riquadri > Visualizzazione**). Selezionando un cluster nel riquadro Cluster, è possibile visualizzare automaticamente i grafici corrispondenti nel riquadro Visualizzazione.

Nota: per impostazione predefinita, i grafici sono in modalità interattiva/selezione in cui è possibile spostare i nodi. Tuttavia, è possibile modificare i layout del grafico in modalità Modifica, inclusi colori e font, legende e altro. Consultare la sezione "Uso delle barre degli strumenti e tavolozze dei grafici" a pagina 163 per ulteriori informazioni.

La vista Cluster ha due grafici Web.

- **Grafico Web Concetto.** Questo grafico presenta tutti i concetti all'interno del cluster selezionato oltre a concetti collegati al di fuori del cluster. Questo grafico può aiutare a scoprire come i concetti all'interno di un cluster sono collegati e gli eventuali collegamenti esterni. Consultare la sezione "Grafico Web di concetto" per ulteriori informazioni.
- **Grafico Web Cluster.** Questo grafico presenta il cluster selezionato (i) con tutti i collegamenti esterni tra i cluster selezionati visualizzati come linee tratteggiate. Consultare la sezione "Grafico Web Cluster" per ulteriori informazioni.

Consultare la sezione Capitolo 11, "Analisi dei cluster", a pagina 147 per ulteriori informazioni.

Grafico Web di concetto

Questa tabella presenta tutti i concetti all'interno del cluster selezionato oltre a concetti collegati al di fuori del cluster. Questo grafico può aiutare a scoprire come i concetti all'interno di un cluster sono collegati e gli eventuali collegamenti esterni. Ogni concetto in un cluster è rappresentato come un nodo, che è codificato come colore in base al tipo di colore. Consultare la sezione "Creazione di tipi" a pagina 191 per ulteriori informazioni.

Vengono tracciati i collegamenti interni tra i concetti all'interno di un cluster e lo spessore della linea di ciascun collegamento è direttamente collegato al conteggio di documenti per ogni ricorrenza della coppia di concetti o al valore di collegamento di similitudine, in base alla propria scelta sulla barra degli strumenti del grafico. Vengono anche visualizzati i collegamenti esterni tra concetti di un cluster e quei concetti al di fuori del cluster.

Se i concetti vengono selezionati nella finestra di dialogo Definizioni cluster, il grafico Web Concetto visualizzerà tali concetti e gli eventuali collegamenti interni e esterni associati a questi concetti. Tutti i collegamenti tra gli altri concetti che non includono uno dei concetti selezionati non apparirà sul grafico.

Nota: per impostazione predefinita, i grafici sono in modalità interattiva/selezione in cui è possibile spostare i nodi. Tuttavia, è possibile modificare i layout del grafico in modalità Modifica, inclusi colori e font, legende e altro. Consultare la sezione "Uso delle barre degli strumenti e tavolozze dei grafici" a pagina 163 per ulteriori informazioni.

Grafico Web Cluster

Questa scheda visualizza un grafico Web che mostra i cluster selezionati. I collegamenti esterni tra i cluster selezionati e anche i collegamenti tra altri cluster sono tutti riportati come linee con punti. In un grafico Web del cluster, ogni nodo rappresenta un intero cluster e lo spessore delle linee tracciate tra loro rappresenta il numero di collegamenti esterni tra due cluster.

Importante! Per visualizzare il grafico Web Cluster, è necessario avere già creato i cluster con i collegamenti esterni. I collegamenti esterni sono collegamenti tra coppie concetto in cluster separati (un concetto all'interno di un cluster e un concetto esterno in un altro cluster).

Ad esempio, si supponga di avere due cluster. `Cluster A` ha tre concetti: `A1`, `A2` e `A3`. `Cluster B` ha due concetti: `B1` e `B2`. I seguenti concetti sono collegati: `A1-A2`, `A1-A3`, `A2-B1` (Esterna), `A2-B2` (Esterna), `A1-B2` (Esterna) e `B1-B2`. Ciò significa che nel grafico Web cluster, lo spessore della linea rappresenta i tre collegamenti esterni.

Nota: per impostazione predefinita, i grafici sono in modalità interattiva/selezione in cui è possibile spostare i nodi. Tuttavia, è possibile modificare i layout del grafico in modalità Modifica, inclusi colori e font, legende e altro. Consultare la sezione "Uso delle barre degli strumenti e tavolozze dei grafici" a pagina 163 per ulteriori informazioni.

Grafici di analisi di collegamento del testo

Dopo l'estrazione dei modelli TLA (Text Link Analysis), è possibile esplorarli visivamente nei grafici Web nel riquadro di visualizzazione. Il riquadro di visualizzazione offre due prospettive sui modelli TLA: un grafico web di concetto (modello) e un grafico Web di tipo (modello). I grafici Web in questo riquadro possono essere utilizzati per rappresentare visivamente i modelli. Il riquadro Visualizzazione si trova nell'angolo in alto a destra della vista Analisi di collegamento del testo. Se non è già visibile, è possibile accedere a questo riquadro dal menu Visualizza (**Visualizza > Riquadri > Visualizzazione**). Se non vi è alcuna selezione, l'area del grafico è vuota.

Nota: per impostazione predefinita, i grafici sono in modalità interattiva/selezione in cui è possibile spostare i nodi. Tuttavia, è possibile modificare i layout del grafico in modalità Modifica, inclusi colori e font, legende e altro. Consultare la sezione "Uso delle barre degli strumenti e tavolozze dei grafici" a pagina 163 per ulteriori informazioni.

La vista di analisi di collegamento del testo ha due grafici Web.

- **Grafico Web del concetto.** Questo grafico presenta tutti i concetti nel modello selezionato. La larghezza della riga e le dimensioni del nodo (se le icone tipo non vengono visualizzate) in un grafico concetto mostrano il numero di ricorrenze globali nella tabella selezionata. Consultare la sezione "Grafico Web di concetto" per ulteriori informazioni.
- **Grafico Web del tipo.** Questo grafico presenta tutti i tipi nel modello selezionato. La larghezza della riga e le dimensioni del nodo (se le icone tipo non vengono visualizzate) in un grafico mostrano il numero di ricorrenze globali nella tabella selezionata. I nodi sono rappresentati da un colore di tipo da un'icona. Consultare la sezione "Grafico Web del tipo" per ulteriori informazioni.

Consultare la sezione Capitolo 12, "Esplorazione di analisi di collegamento del testo", a pagina 153 per ulteriori informazioni.

Grafico Web di concetto

Questo grafico Web presenta tutti i concetti rappresentati nella selezione corrente. Ad esempio, se è stato selezionato un modello di tipo che ha tre modelli di concetto corrispondenti, questo grafico mostrerà tre serie di concetti collegati. La larghezza della riga e le dimensioni del nodo in un grafico di concetto rappresentano i conteggi di frequenza globale. Il grafico rappresenta visivamente le stesse informazioni di quanto selezionato nei riquadri dei modelli. I tipi di ogni concetto sono rappresentati con un colore o con un'icona a seconda di cosa è stato selezionato sulla barra degli strumenti del grafico. Consultare la sezione "Uso delle barre degli strumenti e tavolozze dei grafici" a pagina 163 per ulteriori informazioni.

Grafico Web del tipo

Questo grafico Web presenta ogni modello di tipo per la selezione corrente. Ad esempio, se si selezionano due modelli di concetto, questo grafico mostra un nodo per tipo nei modelli selezionati e i collegamenti

tra quelli trovati nello stesso modello. La larghezza della riga e le dimensioni nodo rappresentano i conteggi di frequenza globale per il gruppo. Il grafico rappresenta visivamente le stesse informazioni di quanto selezionato nei riquadri dei modelli. Oltre ai nomi tipo, sono anche rappresentati i tipi con un colore o con un'icona di tipo a seconda di cosa è stato selezionato sulla barra degli strumenti del grafico. Consultare la sezione “Uso delle barre degli strumenti e tavolozze dei grafici” per ulteriori informazioni.

Uso delle barre degli strumenti e tavolozze dei grafici

Per ciascun grafico, c'è una barra degli strumenti che consente di accedere rapidamente ad alcune delle tavolozze comuni da cui è possibile eseguire una serie di azioni con i propri grafici. Ogni vista (Categorie e concetti, Cluster e Analisi di collegamento del testo) dispone di una barra degli strumenti leggermente diversa. È possibile scegliere tra le modalità di visualizzazione *Esplora* o *Modifica*.

Mentre la modalità Esplora consente di esaminare in modo analitico i dati e i valori presenti nella visualizzazione, con la modalità Modifica è possibile intervenire sul layout e sull'aspetto della visualizzazione. Per esempio, è possibile modificare caratteri e colori in base alla guida di stile della propria organizzazione. Per selezionare questa modalità, scegliere **Visualizza > riquadro Visualizzazione > Modo Modifica** dai menu (o fare clic sull'icona nella barra degli strumenti).

In modalità Modifica sono disponibili numerose barre degli strumenti che riguardano diversi aspetti del layout della visualizzazione. Se lo si desidera, è possibile nascondere le barre degli strumenti che non vengono utilizzate in modo da incrementare lo spazio della finestra di dialogo in cui il grafico viene visualizzato. Per selezionare o deselezionare le barre degli strumenti, fare clic sul relativo nome di barra degli strumenti o tavolozza nel menu Visualizza.

Per ulteriori informazioni su tutte le barre degli strumenti e tavolozze utilizzate per la modifica di grafici, consultare la sezione sulla modifica dei grafici nella guida in linea o nel file *modeler_nodes_general_book.pdf*, disponibile nella cartella *\Documentazione\it* in IBM SPSS Modeler DVD.

Tabella 36. Pulsanti della barra degli strumenti di analisi del testo.

Pulsante/Elenco	Descrizione
	Attiva la modalità di modifica. Passare alla modalità di modifica per cambiare l'aspetto del grafico, come l'estensione del font, la modifica dei colori per adattarli al proprio stile aziendale o la rimozione di etichette e legende.
	Attiva la modalità di esplorazione. Per impostazione predefinita, la modalità di esplorazione è attivata, il che significa che è possibile spostare e trascinare i nodi nel grafico o anche sorvolare con il mouse su oggetti del grafico per visualizzare ulteriori informazioni e suggerimenti.
	Selezionare un tipo di visualizzazione Web per i graficinella vista Categorie e concetti o nella vista Analisi di collegamento del testo . • **Layout circolare.** Un layout generale che può essere applicato a qualsiasi grafico. Esso prevede un grafico in base al presupposto che i collegamenti siano indiretti e considera in egual modo tutti i nodi. I nodi vengono inseriti solo intorno al perimetro di un cerchio. • **Layout di rete.** Un layout generale che può essere applicato a qualsiasi grafico. Esso prevede un grafico in base al presupposto che i collegamenti siano indiretti e considera in egual modo tutti i nodi. Nel layout i nodi sono posizionati ovunque. • **Layout diretto.** Un layout che deve essere utilizzato solo per grafici diretti. Questo layout produce strutture ad albero dai nodi principali fino ai nodi di estremità organizzate in base ai colori. I dati gerarchici vengono ben visualizzati con questo tipo di layout. • **Layout di griglia.** Un layout generale che può essere applicato a qualsiasi grafico. Esso prevede un grafico in base al presupposto che i collegamenti siano indiretti e considera in egual modo tutti i nodi. I nodi sono collocati solo in punti della griglia all'interno dello spazio.

Tabella 36. Pulsanti della barra degli strumenti di analisi del testo (Continua).

Pulsante/Elenco	Descrizione
	Rappresentazione della dimensione del collegamento. Scegliere cosa rappresenta lo spessore della linea nel grafico. Si applica solo alla vista Cluster. Il grafico Web di cluster visualizza solo il numero di collegamenti esterni tra i cluster. È possibile scegliere tra: • **Similarità.** Lo spessore indica il numero di collegamenti esterni tra due cluster • **Ricorrenza.** Lo spessore indica il numero di documenti in cui si verifica una ricorrenza di descrittori.
	Un pulsante di attivazione/disattivazione che, quando selezionato, visualizza la legenda. Quando il pulsante non è premuto, la legenda non viene visualizzata.
	Un pulsante di attivazione/disattivazione che, quando selezionato, visualizza le icone di tipo nel grafico piuttosto che i colori di tipo. Ciò si applica solo alla vista Analisi di collegamento del testo .
	Un pulsante di attivazione/disattivazione che, quando selezionato, visualizza un pannello di scorrimento dei collegamenti al di sotto del grafico. È possibile filtrare i risultati facendo scorrere la freccia.
	Visualizzerà il grafico per il più alto livello di categorie selezionate piuttosto che per loro sottocategorie.
	Visualizzerà il grafico per il livello più basso di categorie selezionate.
	Questa opzione controlla come i nomi delle sottocategorie vengono visualizzati nell'output. • **Percorso completo di categoria.** Questa opzione emette il nome della categoria e il percorso completo di categorie principali, se applicabile, utilizzando barre per separare i nomi categoria dai nomi sottocategoria. • **Percorso abbreviato di categoria.** Questa opzione emette solo il nome della categoria, ma utilizza le ellissi per visualizzare il numero di categorie principali per la categoria in questione. • **Categoria di livello inferiore.** Questa opzione emette solo il nome della categoria senza il percorso completo o senza mostrare le categorie principali.

Capitolo 14. Editor di risorsa della sessione

IBM SPSS Modeler Text Analytics rapidamente e accuratamente cattura e estrae concetti chiave dai dati di testo. Questo processo di estrazione si basa pesantemente sulle risorse linguistiche per indicare come estrarre informazioni da dati di testo. Per impostazione predefinita, queste risorse provengono dai modelli di risorsa.

IBM SPSS Modeler Text Analytics viene fornito con una serie **di modelli di risorse** specialistici che contengono una serie di risorse linguistiche e non linguistiche, sotto forma di librerie e risorse avanzate, per aiutare a definire il modo in cui i dati verranno gestiti ed estratti. Consultare la sezione Capitolo 15, “Modelli e risorse”, a pagina 169 per ulteriori informazioni.

Nella finestra di dialogo del nodo, è possibile caricare una copia di risorse del modello nel nodo. Una volta all'interno di una sessione workbench interattiva, è possibile, se si desidera, personalizzare tali risorse in modo specifico per i dati di questo nodo. Durante una sessione workbench interattiva, è possibile gestire le proprie risorse nella vista di Editor risorse. Ogni volta che una sessione interattiva viene avviata, l'estrazione viene eseguita utilizzando le risorse caricate nella finestra di dialogo del nodo, a meno che non siano stati memorizzati nella cache i dati e i risultati di estrazione del nodo.

Modifica delle risorse nell'Editor di risorsa

Editor risorse fornisce l'accesso alla serie di risorse utilizzate per produrre i risultati di estrazione (concetti, tipi e modelli) per una sessione workbench interattiva. Questo editor è molto simile a Editor di modelli ad eccezione del fatto che in Editor risorse si stanno modificando le risorse per questa sessione. Dopo aver finito di lavorare sulle proprie risorse e qualsiasi altra attività è stata eseguita, è possibile aggiornare il nodo di modellazione per salvare questo lavoro in modo che possa essere ripristinato in una successiva sessione workbench interattiva. Consultare la sezione “Aggiornamento dei nodi di modellazione e salvataggio” a pagina 84 per ulteriori informazioni.

Se si desidera lavorare direttamente sui modelli utilizzati per caricare delle risorse nei nodi, si consiglia di utilizzare Editor di modelli. Molte delle attività che è possibile eseguire all'interno di Editor risorse vengono eseguite esattamente come sono in Editor di modelli, come:

- **Gestione di librerie**. Consultare la sezione Capitolo 16, “Gestione delle librerie”, a pagina 179 per ulteriori informazioni.
- **Creazione di dizionari di tipo**. Consultare la sezione “Creazione di tipi” a pagina 191 per ulteriori informazioni.
- **Aggiunta di termini a dizionari**. Consultare la sezione “Aggiunta di termini” a pagina 192 per ulteriori informazioni.
- **Creazione di sinonimi**. Consultare la sezione “Definizione dei sinonimi” a pagina 197 per ulteriori informazioni.
- **Importazione ed esportazione di modelli**. Consultare la sezione “Importazione ed esportazione di modelli” a pagina 177 per ulteriori informazioni.
- **Pubblicazione di librerie**. Consultare la sezione “Pubblicazione delle librerie” a pagina 185 per ulteriori informazioni.

Per testo olandese, inglese, francese, tedesco, italiano, portoghese e spagnolo

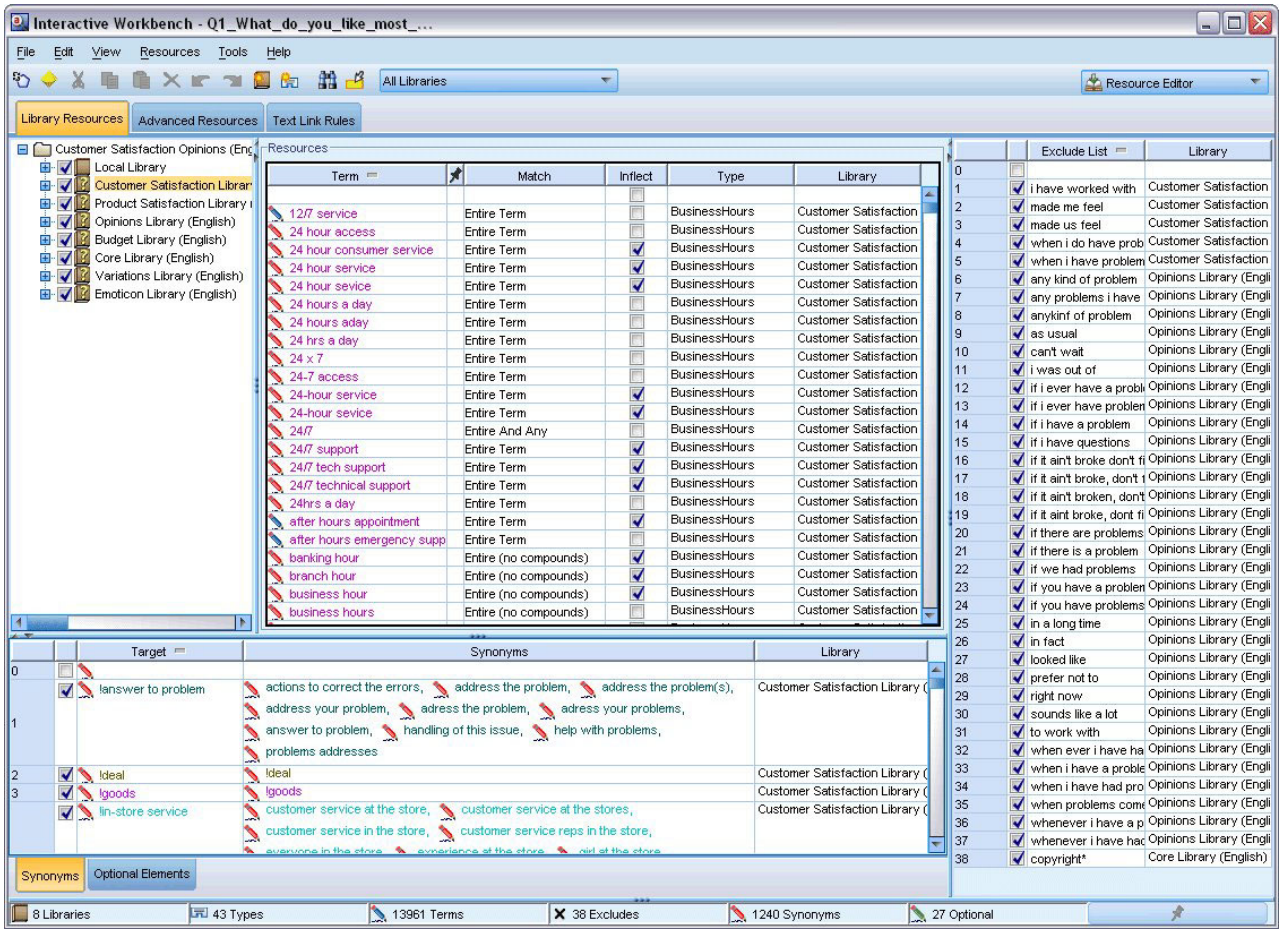


Figura 35. Vista Editor di risorsa per le lingue non giapponese

Per il testo giapponese

L'interfaccia di editor per la lingua del testo in giapponese è diversa da altre lingue di testo.

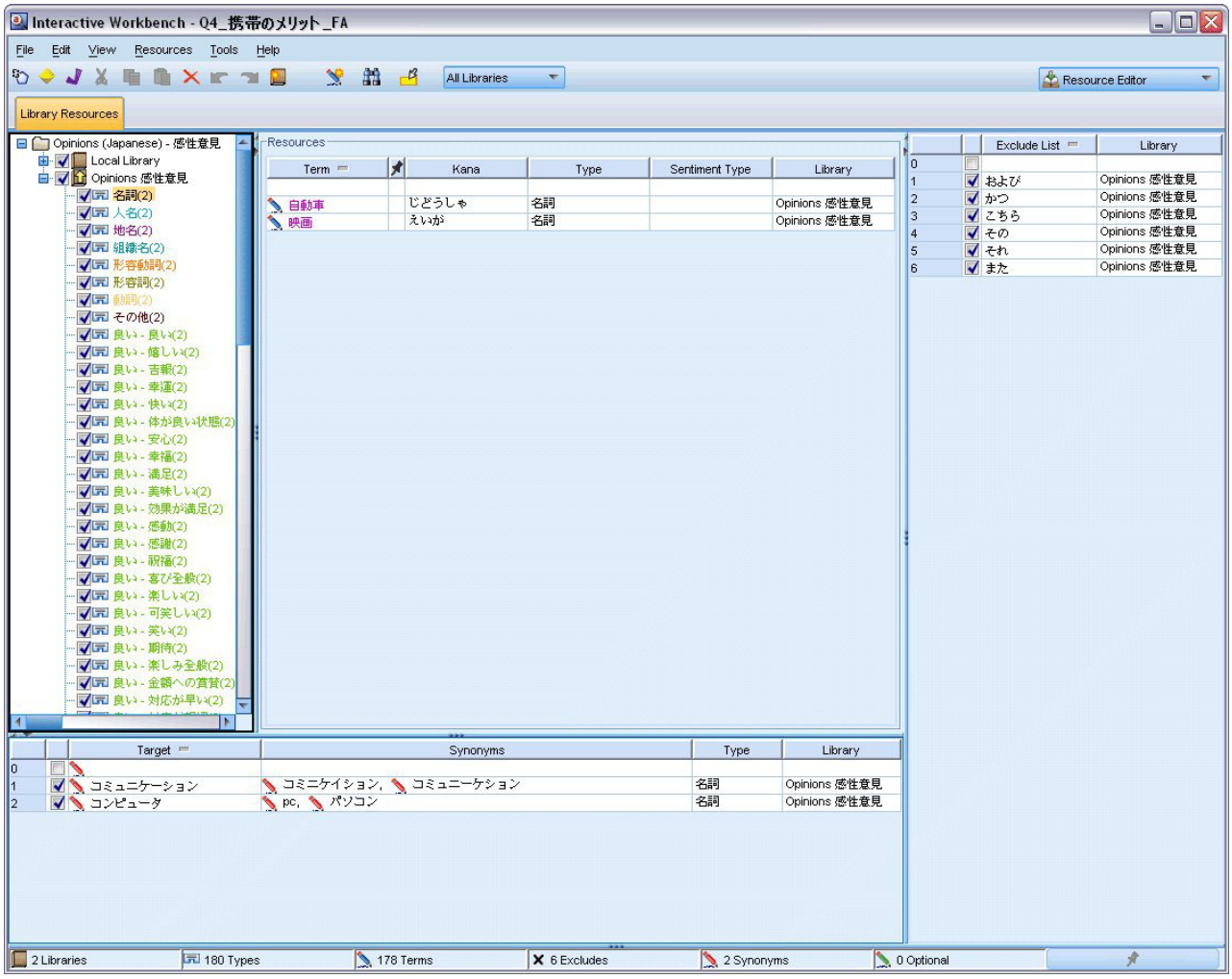


Figura 36. Vista Editor di risorsa per testo giapponese

Creazione ed aggiornamento di modelli

Ogni volta che si apportano delle modifiche alle risorse e si desidera riutilizzarle in futuro, è possibile salvare le risorse come modello. Quando si esegue questa operazione, è possibile scegliere di salvare utilizzando un nome modello esistente o fornendo un nuovo nome. Quindi, ogni volta che si desidera caricare questo modello, in futuro si potranno ottenere le stesse risorse. Per ulteriori informazioni, consultare la sezione “Copia risorse da modelli e TAP” a pagina 27.

Nota: è anche possibile pubblicare e condividere le proprie librerie. Consultare la sezione “Condivisione librerie” a pagina 184 per ulteriori informazioni.

Per eseguire (o aggiornare) un modello

1. Dal menu nella vista di Editor risorse, selezionare **Risorse > Crea modello di risorsa**. Viene aperta la finestra di dialogo Crea modello di risorsa.
2. Immettere un nuovo nome nel campo Nome modello, se si desidera creare un nuovo modello. Selezionare un modello nella tabella, se si desidera sovrascrivere un modello esistente con le risorse attualmente caricate.
3. Fare clic su **Salva** per creare il modello.

Importante! Poiché i modelli vengono caricati quando li si seleziona nel nodo e non quando viene eseguito il flusso, assicurarsi di ricaricare il modello di risorsa in tutti gli altri nodi in cui viene utilizzato se si desidera ricevere le ultime modifiche. Consultare la sezione “Aggiornamento delle risorse del nodo dopo il caricamento” a pagina 175 per ulteriori informazioni.

Scambio di modelli di risorsa

Se si desidera sostituire le risorse attualmente caricate nella sessione con una loro copia da un altro modello, è possibile passare a tali risorse. In questo modo verranno sovrascritte tutte le risorse attualmente caricate nella sessione. Se si stanno scambiando le risorse per ottenere regole di modello TLA (Text Link Analysis), assicurarsi di selezionare un modello che li presenta contrassegnati nella colonna TLA.

Importante! Non è possibile passare da un modello giapponese ad un modello non giapponese o viceversa.

Lo scambio delle risorse è particolarmente utile quando si desidera ripristinare il lavoro della sessione (categorie, modelli e risorse) ma si desidera caricare una copia aggiornata delle risorse da un modello senza perdere il lavoro delle altre sessioni. È possibile selezionare il modello di cui si desidera copiare il contenuto nel Editor risorse e fare clic su **OK.** Questo sostituisce le risorse contenute in questa sessione. Assicurarsi di aggiornare il nodo di modellazione alla fine della sessione se si desidera conservare tali modifiche la volta successiva che si avvia la sessione workbench interattiva.

Nota: se si passa al contenuto di un altro modello durante una sessione interattiva, il nome del modello elencato nel nodo sarà ancora il nome dell'ultimo modello caricato e copiato. Per poter beneficiare di queste risorse o di altro lavoro di sessione, aggiornare il nodo di modellazione prima di uscire dalla sessione e selezionare l'opzione **Usa lavoro di sessione** nel nodo. Consultare la sezione “Aggiornamento dei nodi di modellazione e salvataggio” a pagina 84 per ulteriori informazioni.

Scambio tra le risorse

1. Dal menu nella vista di Editor risorse, selezionare **Risorse > Scambia modelli di risorsa**. Viene aperta la finestra di dialogo Scambia risorse.
2. Selezionare il modello che si desidera utilizzare tra quelli visualizzati nella tabella.
3. Fare clic su **OK** per abbandonare le risorse attualmente caricate e caricare al loro posto una copia di quelle del modello selezionato. Se sono state apportate modifiche alle proprie risorse e si desidera salvare le librerie per un uso futuro, è possibile pubblicarle, aggiornarle e condividerle prima dello scambio. Consultare la sezione “Condivisione librerie” a pagina 184 per ulteriori informazioni.

Capitolo 15. Modelli e risorse

IBM SPSS Modeler Text Analytics rapidamente e accuratamente cattura e estrae concetti chiave dai dati di testo. Questo processo di estrazione si basa pesantemente sulle risorse linguistiche per indicare come estrarre informazioni da dati di testo. Consultare la sezione “Come funziona il processo di estrazione” a pagina 5 per ulteriori informazioni. È possibile regolare tali risorse in Editor risorse.

Quando si installa il software, è possibile anche ottenere una serie di risorse specializzate. Queste risorse fornite consentono di trarre vantaggio da anni di ricerche e di ottimizzazioni per specifiche lingue e specifiche applicazioni. Poiché le risorse fornite potrebbero non essere sempre perfettamente adattate al contesto dei dati, è possibile modificare tali modelli di risorsa o addirittura creare e utilizzare le librerie personalizzate ottimizzate univocamente per i dati della propria organizzazione. Queste risorse provengono in varie forme e ciascuna può essere utilizzata nella propria sessione. Le risorse possono essere trovate in:

- **Modelli di risorsa**. I modelli sono costituiti da una serie di librerie, tipi, e alcune risorse avanzate che insieme formano un serie specializzata di risorse adeguate per un particolare dominio o contesto come ad esempio i pareri sul prodotto.
- **TAP (Text analysis package)**. Oltre alle risorse memorizzate in una modello, anche i TAP riuniscono insieme una o più categorie specializzate generate utilizzando questi insiemi di risorse in modo che entrambe le categorie e le risorse vengano memorizzate insieme e riutilizzate. Consultare la sezione “Uso dei pacchetti di analisi del testo (TAP)” a pagina 141 per ulteriori informazioni.
- **Librerie**. Le librerie sono utilizzate come blocchi di creazione per entrambi i TAP e i modelli. Possono anche essere aggiunti individualmente a risorse nella propria sessione. Ogni libreria è costituita da diversi dizionari utilizzati per definire e gestire i tipi, i sinonimi e gli elenchi di esclusione. Mentre le librerie vengono fornite anche singolarmente esse sono anche precompresse in modelli e TAP. Consultare la sezione Capitolo 16, “Gestione delle librerie”, a pagina 179 per ulteriori informazioni.

Nota: durante l'estrazione, vengono anche utilizzate alcune risorse interne compilate. Queste risorse compilate contengono un gran numero di definizioni integrando i tipi nella libreria principale. Queste risorse compilate non possono essere modificate.

Editor risorse fornisce l'accesso alla serie di risorse utilizzate per produrre i risultati di estrazione (concetti, tipi e modelli). Vi sono alcune attività che è possibile eseguire in Editor risorse e includono:

- **Gestione di librerie**. Consultare la sezione Capitolo 16, “Gestione delle librerie”, a pagina 179 per ulteriori informazioni.
- **Creazione di dizionari di tipo**. Consultare la sezione “Creazione di tipi” a pagina 191 per ulteriori informazioni.
- **Aggiunta di termini a dizionari**. Consultare la sezione “Aggiunta di termini” a pagina 192 per ulteriori informazioni.
- **Creazione di sinonimi**. Consultare la sezione “Definizione dei sinonimi” a pagina 197 per ulteriori informazioni.
- **Aggiornamento di risorse nei TAP**. Consultare la sezione “Aggiornamento dei pacchetti di analisi del testo” a pagina 142 per ulteriori informazioni.
- **Creare modelli**. Consultare la sezione “Creazione ed aggiornamento di modelli” a pagina 167 per ulteriori informazioni.
- **Importazione ed esportazione di modelli**. Consultare la sezione “Importazione ed esportazione di modelli” a pagina 177 per ulteriori informazioni.
- **Pubblicazione di librerie**. Consultare la sezione “Pubblicazione delle librerie” a pagina 185 per ulteriori informazioni.

Editor di modello e Editor di risorsa

Vi sono due metodi principali per la gestione e la modifica di modelli, le librerie e le relative risorse. È possibile lavorare sulle risorse linguistiche in Editor di modelli o Editor risorse.

Editor di modelli

Editor di modelli consente di creare e modificare i modelli di risorse senza una sessione workbench interattiva e indipendente da un determinato nodo o flusso. È possibile utilizzare questo editor per creare o modificare i modelli di risorsa prima di caricarli nel nodo di analisi collegamento del testo e nel nodo di modellazione di estrazione testo.

Editor di modelli è accessibile tramite la barra degli strumenti principale di IBM SPSS Modeler dal menu **Strumenti > Editor di modello di analisi del testo**.

Editor risorse

Editor risorse, che è accessibile all'interno di una sessione workbench interattiva, consente di lavorare con le risorse nel contesto di un nodo specifico e di dataset. Quando si aggiunge un nodo di modellazione di estrazione testo a un flusso è possibile caricare una copia del contenuto di un modello di risorsa o una copia di un pacchetto di analisi del testo (serie di categoria *e* risorse) per controllare il modo in cui viene estratto il testo per l'estrazione testo. Quando viene avviata una sessione workbench interattiva, oltre a creare categorie, estrarre modelli di analisi di collegamento del testo e creare modelli di categoria, è possibile anche ottimizzare le risorse per i dati di sessione nella vista di Editor risorse integrata. Consultare la sezione “Modifica delle risorse nell'Editor di risorsa” a pagina 165 per ulteriori informazioni.

Ogni volta che si lavora sulle risorse in una sessione workbench interattiva, tali modifiche si applicano solo a tale sessione. Se si desidera salvare il lavoro (risorse, categorie, modelli, ecc.) così da continuare in una sessione successiva, è necessario aggiornare il nodo di modellazione. Consultare la sezione “Aggiornamento dei nodi di modellazione e salvataggio” a pagina 84 per ulteriori informazioni.

Se si desidera salvare le modifiche e riportarle al modello originale, i cui contenuti sono stati copiati nel nodo di modellazione, in modo che questo modello aggiornato può essere caricato in altri nodi, è possibile creare un modello dalle risorse. Consultare la sezione “Creazione ed aggiornamento di modelli” a pagina 167 per ulteriori informazioni.

L'interfaccia dell'Editor

Le operazioni che è possibile eseguire in Editor di modello o Editor risorse si articolano intorno alla gestione e ottimizzazione delle risorse linguistiche. Queste risorse sono memorizzate sotto forma di modelli e librerie. Consultare la sezione “Dizionari di tipo” a pagina 189 per ulteriori informazioni.

Scheda Risorse della libreria

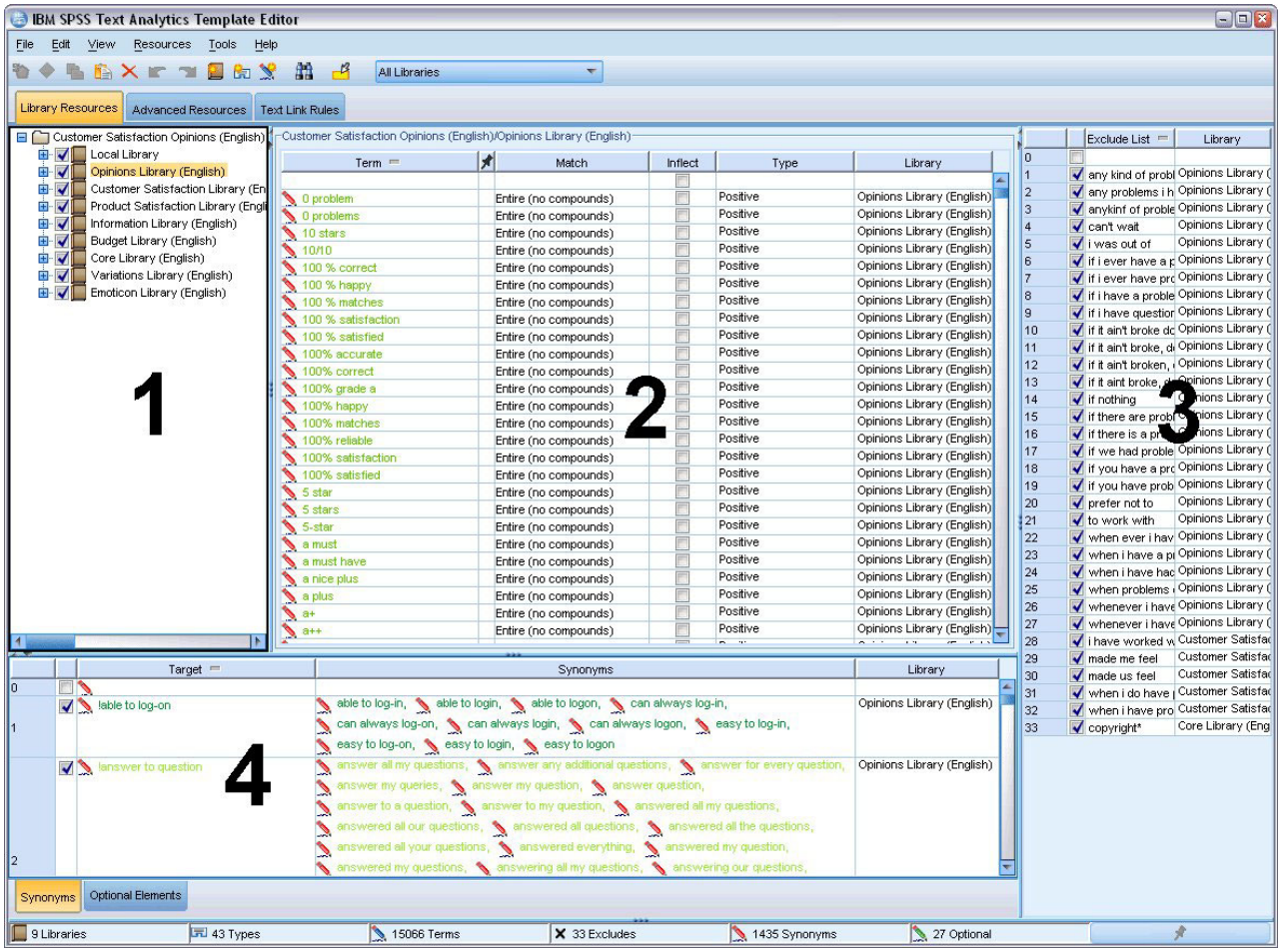


Figura 37. Editor di modello di estrazione testo

L'interfaccia è organizzata in quattro parti:

1. Riquadro ad albero Libreria. Ubicato nell'angolo in alto a sinistra, questo piano visualizza una struttura ad albero delle librerie. In questa struttura ad albero è possibile abilitare e disabilitare librerie e anche filtrare le viste negli altri riquadri selezionando una libreria nella struttura ad albero. È possibile eseguire molte operazioni in questa struttura ad albero utilizzando il menu di scelta rapida. Se si espande una libreria nella struttura ad albero, è possibile visualizzare la serie di tipi che contiene. È anche possibile filtrare questo elenco attraverso il menu **Visualizza** se si desidera concentrarsi su una sola particolare libreria.

2. Elenco dei termini dal riquadro Dizionari di tipo. Situato a destra della struttura ad albero delle librerie, questo riquadro visualizza gli elenchi di termini dei dizionari di tipo per le librerie selezionate nella struttura ad albero. Un **dizionario di tipo** è una raccolta di termini da raggruppare sotto un'etichetta, un tipo o un nome. Quando il motore di estrazione legge i dati di testo, esso confronta le parole trovate nel testo con i termini dei dizionari di tipo. Se un concetto estratto si presenta come un termine in un dizionario di tipo, viene assegnato tale nome tipo. Il dizionario di tipo può essere considerato come un dizionario distinto di termini che hanno qualcosa in comune. Ad esempio il tipo `<Ubicazione>` nella libreria Centrale contiene concetti quali `new orleans`, `great britain` e `new york`. Questi termini rappresentano tutti posizioni geografiche. Una libreria può contenere uno o più dizionari di tipo. Per ulteriori informazioni, consultare la sezione “Dizionari di tipo” a pagina 189.

3. Riquadro Dizionario di esclusione. Ubicato sul lato destro, questo riquadro visualizza la raccolta di termini che verranno esclusi dai risultati finali dell'estrazione. I termini che compaiono in questo

dizionario di esclusione non vengono visualizzati nel riquadro Risultati di estrazione. I termini esclusi possono essere memorizzati in una libreria di propria scelta. Tuttavia, il riquadro Dizionario di esclusione visualizza tutti i termini esclusi per tutte le librerie visibili nella struttura ad albero delle librerie. Consultare la sezione “Dizionari di esclusione” a pagina 200 per ulteriori informazioni.

4. Riquadro Dizionario di sostituzione. Situato in basso a sinistra, questo riquadro visualizza i sinonimi e gli elementi facoltativi, ciascuno nella propria scheda. Sinonimi e gli elementi facoltativi aiutano a raggruppare termini simili in un unico concetto guida o di destinazione nei risultati dell'estrazione finale. Questo dizionario può contenere sinonimi conosciuti e sinonimi ed elementi definiti dall'utente, come anche errori di ortografia accoppiati all'ortografia corretta. Le definizioni dei sinonimi e gli elementi facoltativi possono essere memorizzati in una libreria di propria scelta. Tuttavia, il riquadro dizionario di sostituzione visualizza tutti i contenuti per tutte le librerie visibili nella struttura ad albero della libreria. Mentre questo riquadro visualizza tutti i sinonimi o elementi facoltativi da tutte le librerie, in questo riquadro vengono visualizzate insieme le sostituzioni per tutte le librerie della struttura ad albero. Una libreria può contenere un solo dizionario di sostituzione. Per ulteriori informazioni, consultare la sezione “Dizionari dei sinonimi/di sostituzione” a pagina 196. Notare che la scheda Elementi facoltativi non si applica alle risorse della lingua di testo giapponese.

Notas:

- Se si desidera filtrare in modo da visualizzare solo le informazioni relative ad una singola libreria, è possibile modificare la vista Libreria utilizzando l'elenco a discesa sulla barra degli strumenti. Essa contiene una voce di livello superiore denominata **Tutte le librerie** oltre ad una voce aggiuntiva per ogni singola libreria. Per ulteriori informazioni, consultare la sezione “Visualizzazione librerie” a pagina 182.
- L'interfaccia di editor per la lingua del testo in giapponese è diversa da altre lingue di testo.

Scheda Risorse avanzate

Le risorse avanzate sono disponibili nella seconda scheda della vista editor. È possibile esaminare e modificare le risorse avanzate in questa scheda. Consultare la sezione Capitolo 18, “Informazioni su Risorse avanzate”, a pagina 203 per ulteriori informazioni.

Importante! Questa scheda non è disponibile per le risorse regolate per il testo giapponese.

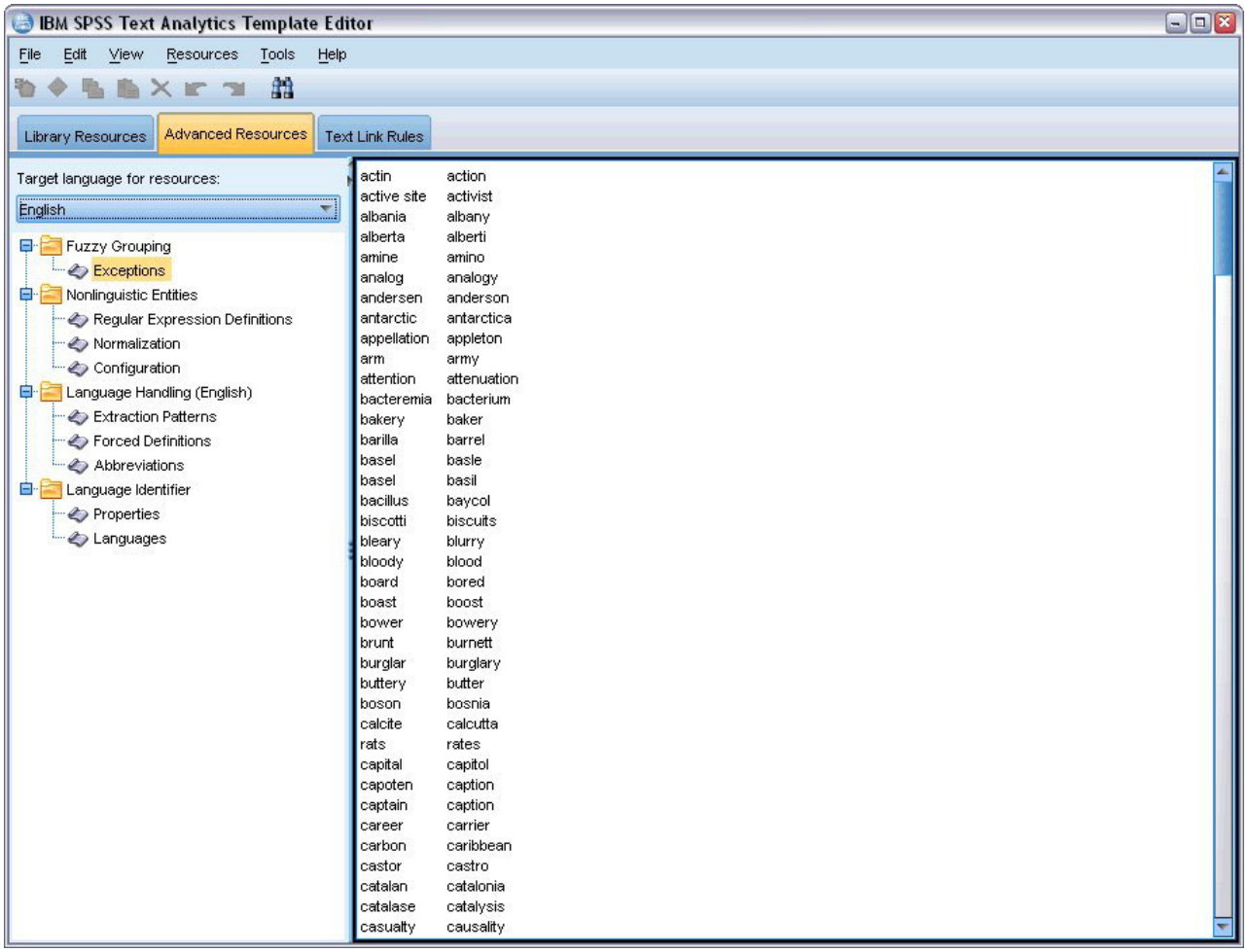


Figura 38. Editor di modello di estrazione testo - scheda Risorse avanzate

Scheda Regole di collegamento del testo

Dalla versione 14, le regole di analisi di collegamento del testo sono modificabili nella propria scheda della vista editor. È possibile lavorare nell'editor di regole, creare regole proprie e persino eseguire le simulazioni per vedere in che modo le regole influenzano i risultati TLA. Consultare la sezione Capitolo 19, “Informazioni sulle regole di collegamento del testo”, a pagina 215 per ulteriori informazioni.

Importante! Questa scheda non è disponibile per le risorse regolate per il testo giapponese.

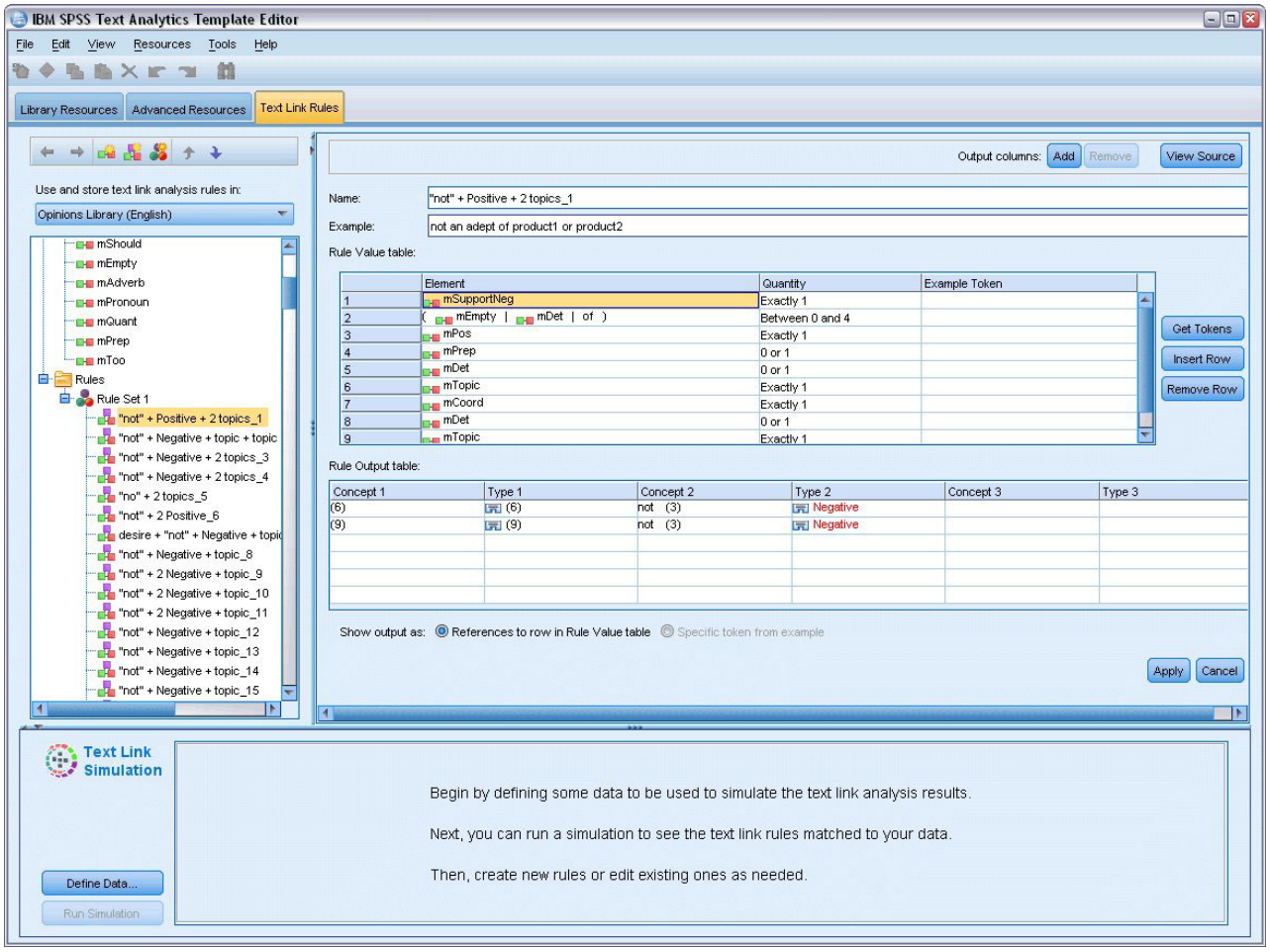


Figura 39. Editor di modello di estrazione testo - scheda Regole di collegamento del testo

Apertura di modelli

Quando si avvia Editor di modelli, viene richiesto di aprire un modello. Allo stesso modo, è possibile aprire un modello dal menu File. Se si desidera un modello che contiene alcune regole di analisi di collegamento del testo (TLA), assicurarsi di selezionare un modello che dispone di un'icona nella colonna TLA. La lingua per la quale un modello è stato creato viene visualizzata nella colonna Lingua.

Se si desidera importare un modello che non è riportato nella tabella o se si desidera esportare un modello, è possibile utilizzare i pulsanti nella finestra di dialogo Apri modello. Consultare la sezione “Importazione ed esportazione di modelli” a pagina 177 per ulteriori informazioni.

Per aprire un modello

1. Dal menu in Editor di modelli, selezionare **File > Apri modello di risorsa**. Viene aperta la finestra di dialogo Apri modello di risorsa.
2. Selezionare il modello che si desidera utilizzare tra quelli visualizzati nella tabella.
3. Fare clic su **OK** per aprire questo modello. Se si dispone di un altro modello aperto al momento nell'editor, facendo clic su OK questo modello verrà abbandonato e verrà visualizzato il modello selezionato. Se sono state apportate modifiche alle proprie risorse e si desidera salvare le librerie per un uso futuro, è possibile pubblicarle, aggiornarle e condividerle prima di aprirne un'altra. Consultare la sezione “Condivisione librerie” a pagina 184 per ulteriori informazioni.

Salvataggio dei modelli

In Editor di modelli, è possibile salvare le modifiche apportate ad un modello. Se si desidera, è possibile scegliere di salvare utilizzando un nome modello esistente o fornendo un nuovo nome.

Se si apportano modifiche a un modello che già caricato in un nodo la volta precedente, è necessario ricaricare il contenuto del modello nel nodo per ottenere le ultime modifiche. Consultare la sezione “Copia risorse da modelli e TAP” a pagina 27 per ulteriori informazioni.

Oppure, se si sta utilizzando l'opzione **Usa lavoro interattivo salvato** nella scheda Modello del nodo di estrazione testo, cioè si stanno utilizzando le risorse da una sessione workbench interattiva precedente, è necessario passare alle risorse di questo modello dall'interno della sessione workbench interattiva. Consultare la sezione “Scambio di modelli di risorsa” a pagina 168 per ulteriori informazioni.

Nota: è anche possibile pubblicare e condividere le proprie librerie. Consultare la sezione “Condivisione librerie” a pagina 184 per ulteriori informazioni.

Per salvare un modello

1. Dal menu in Editor di modelli, selezionare **File > Salva modello di risorsa**. Viene aperta la finestra di dialogo Salva modello di risorsa.
2. Immettere un nuovo nome nel campo Nome modello, se si desidera salvare questo modello come un nuovo modello. Selezionare un modello nella tabella, se si desidera sovrascrivere un modello esistente con le risorse attualmente caricate.
3. Se lo si desidera, immettere una descrizione per visualizzare un commento o un'annotazione nella tabella.
4. Fare clic su **Salva** per salvare il modello.

Importante! Poiché le risorse da modelli o TAP vengono caricate/copiate nel nodo, è necessario aggiornare le risorse per ricaricarle se si apportano modifiche a un modello e si desidera trarre vantaggio da queste modifiche in un flusso esistente. Consultare la sezione “Aggiornamento delle risorse del nodo dopo il caricamento” per ulteriori informazioni.

Aggiornamento delle risorse del nodo dopo il caricamento

Per impostazione predefinita, quando si aggiunge un nodo a un flusso, una serie di risorse da un modello predefinito vengono caricate e integrate nel proprio nodo. E se si modificano i modelli o si utilizza un TAP, quando si caricano, una copia di tali risorse sovrascrive le risorse. Poiché i modelli e i TAP non sono collegati al nodo direttamente, le modifiche apportate ad un modello o TAP non sono automaticamente disponibili nel nodo preesistente. Per poter beneficiare di tali modifiche, è necessario aggiornare le risorse in tale nodo. Le risorse possono essere aggiornate in uno dei due modi.

Metodo 1: Ricaricamento delle risorse nella scheda Modello

Se si desidera aggiornare le risorse nel nodo utilizzando un modello o TAP nuovi o aggiornati, è possibile ricaricarle nella scheda Modello del nodo. Per il nuovo caricamento, verrà sostituita la copia delle risorse nel nodo con una copia più attuale. Per comodità, la data e l'ora aggiornate verranno visualizzate nella scheda Modello insieme con il nome del modello di origine. Consultare la sezione “Copia risorse da modelli e TAP” a pagina 27 per ulteriori informazioni.

Tuttavia, se si sta lavorando con dati della sessione interattiva in un nodo di modellazione di estrazione testo ed è stata selezionata l'opzione **Usa lavoro di sessione** nella scheda Modello, verranno utilizzati il lavoro della sessione salvata e le risorse e il pulsante **Carica** è disabilitato. Esso viene disabilitato poiché, in una sola volta, durante una sessione workbench interattiva, è stata scelta l'opzione **Aggiorna nodo di**

modellazione e mantenute le categorie, le risorse e altre sessioni di lavoro. In tal caso, se si desidera modificare o aggiornare tali risorse, è possibile provare il metodo successivo di passare le risorse in Editor risorse.

Metodo 2: Passaggio delle risorse in Editor risorse

In qualsiasi momento si desidera utilizzare risorse differenti durante una sessione interattiva, è possibile scambiare le risorse utilizzando la finestra di dialogo Scambia risorse. Ciò è particolarmente utile quando si desidera riutilizzare lavoro di categoria esistente ma sostituire le risorse. In questo caso, selezionare l'opzione **Usa lavoro di sessione** nella scheda Modello di un nodo di modellazione estrazione testo. In questo modo si disabilita la possibilità di ricaricare un modello tramite la finestra di dialogo del nodo e invece mantenere le impostazioni e le modifiche apportate durante la sessione. Quindi, è possibile avviare la sessione workbench interattiva per eseguire il flusso ed alternare le risorse in Editor risorse. Consultare la sezione “Scambio di modelli di risorsa” a pagina 168 per ulteriori informazioni.

Per mantenere il lavoro della sessione per le sessioni successive, incluse le risorse, è necessario aggiornare il nodo di modellazione dall'interno della sessione workbench interattiva in modo che le risorse (e altri dati) vengano salvati sul nodo. Consultare la sezione “Aggiornamento dei nodi di modellazione e salvataggio” a pagina 84 per ulteriori informazioni.

Nota: se si passa al contenuto di un altro modello durante una sessione interattiva, il nome del modello elencato nel nodo sarà ancora il nome dell'ultimo modello caricato e copiato. Per poter beneficiare di queste risorse o altro lavoro di sessione, aggiornare il nodo di modellazione prima di uscire dalla sessione.

Gestione modelli

Esistono alcune attività di gestione di base che si potrebbero voler eseguire di volta in volta sui modelli, quali la ridenominazione di modelli, l'importazione e l'esportazione dei modelli o l'eliminazione di modelli obsoleti. Queste attività vengono eseguite nella finestra di dialogo Gestisci modelli. L'importazione ed esportazione di modelli consente di condividere i modelli con altri utenti. Consultare la sezione “Importazione ed esportazione di modelli” a pagina 177 per ulteriori informazioni.

Nota: non è possibile ridenominare o eliminare i modelli che sono installati (o forniti) con questo prodotto. Invece, se si desidera ridenominare, è possibile aprire il modello installato e farne uno nuovo con il nome di propria scelta. È possibile eliminare i modelli personalizzati; tuttavia, se si tenta di eliminare un modello fornito, verrà reimpostato alla versione originariamente installata.

Per ridenominare un modello

1. Dai menu, scegliere **Risorse > Gestisci modelli di risorsa**. Viene visualizzata la finestra di dialogo Gestisci modelli.
2. Selezionare il modello da ridenominare e fare clic su **Rinomina**. La casella del nome diventa un campo modificabile nella tabella.
3. Immettere un nuovo nome e premere il tasto Invio. Viene aperta una finestra di dialogo di conferma.
4. Se si è soddisfatti delle modifiche del nome, fare clic su **Sì**. In caso contrario, fare clic su **No**.

Per eliminare un modello

1. Dai menu, scegliere **Risorse > Gestisci modelli di risorsa**. Viene visualizzata la finestra di dialogo Gestisci modelli.
2. Nella finestra di dialogo Gestisci modelli, selezionare il modello che si desidera eliminare.
3. Fare clic su **Elimina**. Viene aperta una finestra di dialogo di conferma.
4. Fare clic su **Sì** per eliminare o fare clic su **No** per annullare la richiesta. Se si fa clic su **Sì**, il modello viene eliminato.

Importazione ed esportazione di modelli

È possibile condividere i modelli con altri utenti o macchine per l'importazione e l'esportazione. I modelli sono memorizzati in un database interno, ma possono essere esportati come file *.*lrt* sull'unità disco fisso.

Poiché vi sono circostanze in cui potrebbe essere necessario importare o esportare i modelli, vi sono diverse finestre di dialogo che offrono tali funzionalità.

- Finestra di dialogo Apri modello in Editor di modelli
- Finestra di dialogo Carica risorse nel nodo di modellazione di estrazione testo e il nodo di analisi del collegamento del testo.
- Finestra di dialogo Gestisci modelli in Editor di modelli e Editor risorse.

Per importare un modello

1. Nella casella di dialogo, fare clic su **Importa**. Viene visualizzata la casella di dialogo Importa modello.
2. Selezionare il file di modello di risorsa (*.*lrt*) da importare e fare clic su **Importa**. È possibile salvare il modello che si sta importando con un altro nome o sovrascrivere quello esistente. La finestra di dialogo si chiude e il modello viene ora visualizzato nella tabella.

Per esportare un modello

1. Nella finestra di dialogo, selezionare il modello che si desidera esportare e fare clic su **Esporta**. Viene aperta la finestra di dialogo Seleziona directory.
2. Selezionare la directory in cui si desidera esportare e fare clic su **Esporta**. Questa finestra di dialogo viene chiusa e il modello viene esportato con estensione del file (*.*lrt*)

Uscita da Editor di modelli

Una volta terminato l'uso di Editor di modelli, è possibile salvare il proprio lavoro ed uscire dall'editor.

Per uscire da Editor di modelli

1. Dal menu scegliere **File > Chiudi**. Viene visualizzata la finestra di dialogo Salva e chiudi.
2. Selezionare **Salva modifiche su modello** per salvare il modello aperto prima di chiudere l'editor.
3. Selezionare **Pubblica librerie** se si desidera pubblicare le librerie nel modello aperto prima di chiudere l'editor. Se si seleziona questa opzione, verrà richiesto di selezionare le librerie da pubblicare. Per ulteriori informazioni, consultare la sezione “Pubblicazione delle librerie” a pagina 185.

Backup delle risorse

È possibile eseguire il backup delle risorse di volta in volta come misura di sicurezza.

Importante! Quando si ripristina, l'intero contenuto delle proprie risorse verranno cancellate pulito e solo il contenuto del file di backup sarà accessibile al prodotto. Ciò include qualsiasi lavoro aperto.

Nota: è solo possibile eseguire il backup e il ripristino alla versione principale del proprio software. Ad esempio, se si esegue il backup dalla versione 15, non è possibile ripristinare quel backup alla versione 16.

Eseguire la copia di riserva delle risorse

1. Dai menu scegliere **Risorse > Strumenti di backup > Backup delle risorse**. Viene aperta la finestra di dialogo Backup.
2. Immettere un nome per il file di backup e fare clic su **Salva**. La finestra di dialogo si chiude e il file di backup viene creato.

Per ripristinare le risorse

1. Dai menu scegliere **Risorse > Strumenti di backup > Ripristina risorse**. Un avviso segnala che il ripristino sovrascriverà il contenuto corrente del database.
2. Fare clic su **Sì** per procedere. Si apre la finestra di dialogo.
3. Selezionare il file che si desidera ripristinare e fare clic su **Apri**. La finestra di dialogo si chiude e le risorse vengono ripristinate nell'applicazione.

Importazione di file di risorsa

Se sono state apportate modifiche direttamente in file di risorsa al di fuori di questo prodotto, è possibile importarli in una libreria selezionata selezionando tale libreria e procedendo con l'importazione. Quando si importa una directory, è possibile importare tutti i file supportati in una specifica libreria aperta. È possibile importare solo file **.txt*.

Importante! Per i file di lingua giapponese, i file .txt che si desidera importare devono essere codificati in UTF8. Inoltre, non è possibile importare elenchi di esclusione per la lingua giapponese.

Ogni file importato deve contenere una sola voce per riga e se il contenuto è strutturato come:

- Un elenco di parole o frasi (una per riga). Il file viene importato come un elenco di termini per un dizionario di tipo, dove il dizionario di tipo prende il nome del file meno l'estensione.
- Un elenco di voci quali *term1* `<TAB>` *term2* viene importato come un elenco di sinonimi, dove *term1* è l'insieme dei termini sottostanti e *term2* è il termine di destinazione.

Per importare un singolo file di risorse

1. Dai menu, scegliere **Risorse > Importa file > Importa file singolo**. Viene visualizzata la finestra di dialogo Importa file.
2. Selezionare il file che si desidera importare e fare clic su **Importa**. Il contenuto del file viene trasformato in un formato interno e aggiunto alla libreria.

Per importare tutti i file in una directory

1. Dai menu, scegliere **Risorse > Importa file > Importa intera directory**. Viene visualizzata la finestra di dialogo Importa directory.
2. Selezionare la libreria in cui si desidera importare tutti i file di risorsa dall'elenco **Importa**. Se si seleziona l'opzione **Predefinito**, una nuova libreria verrà creata utilizzando il nome della directory come suo nome.
3. Selezionare la directory dalla quale importare i file. Le sottodirectory non verranno lette.
4. Fare clic su **Importa**. La finestra di dialogo viene chiusa e il contenuto di questi file di risorsa importati viene ora visualizzato nell'editor nella forma di dizionari e file di risorse avanzate.

Capitolo 16. Gestione delle librerie

Le risorse utilizzate dal motore di estrazione per estrarre e raggruppare termini dai propri dati di testo contengono sempre una o più librerie. È possibile visualizzare la serie di librerie nella struttura ad albero delle librerie che si trova nella parte superiore sinistra della finestra Editor di modelli e Editor risorse. Le librerie sono composte da tre tipi di dizionari: Tipo, Sostituzione e Esclusione. Per ulteriori informazioni, consultare la sezione Capitolo 17, “Informazioni sui dizionari di libreria”, a pagina 189.

Il modello di risorsa o le risorse dal TAP scelte includono diverse librerie per consentire all'utente di iniziare immediatamente l'estrazione di concetti dai propri dati di testo. Tuttavia, è possibile creare le proprie librerie come pure pubblicarle anche in modo che sia possibile riutilizzarle. Consultare la sezione “Pubblicazione delle librerie” a pagina 185 per ulteriori informazioni.

Ad esempio, si supponga che si gestiscono frequentemente dati di testo correlati all'industria automobilistica. Dopo aver analizzato i dati, si decide di voler creare alcune risorse personalizzate per gestire gergo o vocabolario specifico per quel tipo di industria. Utilizzando Editor di modelli, è possibile creare un nuovo modello e in esso una libreria per estrarre e raggruppare termini automobilistici. Poiché sono di nuovo necessarie le informazioni in questa libreria, è possibile pubblicare la libreria in un repository centrale, accessibile nella finestra di dialogo **Gestisci librerie**, in modo che possa essere riutilizzata in maniera indipendente in diversi sessioni di flusso .

Si supponga che si desidera anche raggruppare termini che sono specifici di diverse industrie del settore, come i dispositivi elettronici, i motori, i sistemi di raffreddamento o addirittura un determinato costruttore o mercato. È possibile creare una libreria per ogni gruppo e quindi pubblicare le librerie in modo che possano essere utilizzate con più serie di dati di testo. In questo modo, è possibile aggiungere le librerie che meglio corrispondono al contesto dei propri dati di testo.

Nota: le risorse supplementari possono essere configurate e gestite nella scheda Risorse avanzate. Alcune si applicano a tutte le librerie e gestiscono entità non linguistiche, eccezioni di raggruppamento confuso e così via. Inoltre, è anche possibile modificare le regole del modello di analisi di collegamento del testo, che sono specifiche della libreria, nella scheda Regole di collegamento del testo. Consultare la sezione Capitolo 18, “Informazioni su Risorse avanzate”, a pagina 203 per ulteriori informazioni.

Librerie fornite

Per impostazione predefinita, diverse librerie vengono installate con IBM SPSS Modeler Text Analytics. È possibile utilizzare queste librerie preformattate per accedere a migliaia di termini predefiniti e sinonimi, oltre a molti tipi diversi. Queste librerie fornite vengono regolate per diversi domini e sono disponibili in diverse lingue.

Esiste un certo numero di librerie ma le più comunemente utilizzate sono le seguenti:

- **Libreria locale.** Utilizzata per memorizzare dizionari definiti dall'utente. Si tratta di una libreria vuota aggiunta per impostazione predefinita a tutte le risorse. Essa contiene un dizionario di tipo vuoto. Risulta molto utile quando si apportano modifiche o aggiustamenti alle risorse (come l'aggiunta di una parola a un tipo) direttamente nella vista Categorie e concetti, vista Cluster e la vista Analisi di collegamento del testo . In questo caso, quelle modifiche e aggiustamenti verranno automaticamente memorizzati nella prima libreria elencata nella struttura ad albero della libreria in Editor risorse: per impostazione predefinita, questa è la *Libreria locale*. Non è possibile pubblicare questa libreria perché è specifica per i dati di sessione . Se si desidera pubblicare i suoi contenuti, è necessario ridenominare prima la libreria.
- **Libreria principale.** Utilizzata nella maggior parte dei casi, poiché include cinque tipi integrati di base che rappresentano persone, ubicazioni, organizzazioni, prodotti e sconosciuto. Mentre è possibile

visualizzare solo pochi termini in uno dei suoi dizionari di tipo, i tipi rappresentati nella libreria principale sono effettivamente complementari ai tipi più solidi trovati nelle risorse interne, compilate fornite con il prodotto di estrazione testo. Tali risorse interne, compilate contengono migliaia di termini per ciascun tipo. Per questo motivo, mentre si potrebbe non visualizzare un termine nell'elenco dei termini del dizionario di tipo, può essere estratto e immesso con un tipo principale. Ciò spiega come i nomi come *Giorgio* possono essere estratti e immessi come `<Persona>` quando solo *Giovanni* appare nel dizionario di tipo `<Persona>` nella libreria principale. Allo stesso modo, se non si include la libreria principale, è comunque possibile visualizzare tali tipi nei propri risultati di estrazione, poiché le risorse compilate contenenti questi tipi verranno ancora utilizzate dal motore di estrazione.

- **Libreria di pareri.** Utilizzata più comunemente per estrarre i pareri e opinioni da dati di testo. Questa libreria include migliaia di parole che rappresentano atteggiamenti, i qualificativi e le preferenze che, quando utilizzato insieme ad altri termini, indicano un parere su un argomento. Questa libreria include un numero di tipi incorporati, sinonimi ed esclusioni. Inoltre, include una serie estesa di regole di modello utilizzata per l'analisi di collegamento del testo. Per trarre vantaggio dalle regole di analisi di collegamento del testo in questa libreria e i risultati di modello prodotti, questa libreria deve essere specificata nella scheda Regole di collegamento del testo. Per ulteriori informazioni, consultare la sezione Capitolo 19, “Informazioni sulle regole di collegamento del testo”, a pagina 215.
- **Libreria di bilancio.** Utilizzata per estrarre i termini con riferimento al costo di qualcosa. Questa libreria include molte parole e frasi che rappresentano aggettivi, qualificativi e giudizi il prezzo o la qualità di qualcosa.
- **Libreria di variazioni.** Utilizzata per includere i casi in cui determinate variazioni della lingua appropriata richiedono delle definizioni di sinonimi per raggrupparli. Questa libreria include solo le definizioni dei sinonimi.

Sebbene alcune delle librerie fornite al di fuori dei modelli somigliano al contenuto di alcuni modelli, i modelli sono stati specificatamente regolati per particolari applicazioni e contengono altre risorse avanzate. È consigliabile tentare di utilizzare un modello progettato per il tipo di dati di testo che si sta gestendo e apportare le modifiche alle risorse piuttosto che solo aggiungere librerie individuali a un modello più generico.

Le risorse compilate vengono anche fornite con IBM SPSS Modeler Text Analytics. Esse sono sempre utilizzate durante il processo di estrazione e contengono un numero elevato di definizioni complementari a dizionari di tipo integrati nelle librerie predefinite. Poiché queste risorse vengono compilate, esse non possono essere visualizzate o modificate. È possibile, tuttavia, forzare un termine che è stato immesso da queste risorse compilate in qualsiasi altro dizionario. Consultare la sezione “Forzatura di termini” a pagina 195 per ulteriori informazioni.

Creazione di librerie

È possibile creare qualsiasi numero di librerie. Dopo avere creato una nuova libreria, è possibile iniziare a creare dizionari di tipo in questa libreria e immettere termini, sinonimi ed esclusioni.

Per creare una libreria

1. Dai menu scegliere **Risorse > Nuova libreria**. Viene aperta la finestra di dialogo Proprietà della libreria .
2. Immettere un nome per la libreria nella casella di testo Nome.
3. Se lo si desidera, immettere un commento nella casella di testo Annotazione.
4. Fare clic su **Pubblica** se si desidera pubblicare questa libreria ora prima di immettere qualcosa nella libreria. Consultare la sezione “Condivisione librerie” a pagina 184 per ulteriori informazioni. È possibile anche pubblicare successivamente in qualsiasi momento.
5. Fare clic su **OK** per creare la libreria. La finestra di dialogo viene chiusa e il nome della libreria viene visualizzato nella vista ad albero. Se si espandono le librerie nella struttura ad albero, è possibile che

un dizionario di tipo vuoto sia stato automaticamente incluso nella libreria. Nel dizionario è possibile iniziare immediatamente l'aggiunta di termini. Consultare la sezione “Aggiunta di termini” a pagina 192 per ulteriori informazioni.

Aggiunta di librerie pubbliche

Se si desidera riutilizzare una libreria da un'altra sessione dei dati, è possibile aggiungerlo alle risorse attuali fin tanto che è una biblioteca pubblica. Un **libreria pubblica** è una libreria che è stata pubblicata. Consultare la sezione “Pubblicazione delle librerie” a pagina 185 per ulteriori informazioni.

Importante! Non è possibile aggiungere una libreria di risorse in giapponese a risorse non in giapponese o viceversa.

Quando si aggiunge una libreria pubblica, una copia **locale** è integrata nei dati di sessione . È possibile apportare modifiche a questa libreria; tuttavia, è necessario ripubblicare la versione pubblica della libreria se si desidera condividere le modifiche.

Quando si aggiunge una libreria pubblica, una finestra di dialogo Risolvi conflitti può apparire se vengono rilevate eventuali conflitti tra i termini e i tipi in una libreria e le altre librerie locali. È necessario risolvere questi conflitti o accettare le risoluzioni proposte per completare questa operazione. Consultare la sezione “Risoluzione dei conflitti” a pagina 186 per ulteriori informazioni.

Nota: se si desidera aggiornare le librerie quando viene avviata una sessione workbench interattiva o si pubblica quando se ne chiude una, è meno probabile che le librerie siano non sincronizzate. Consultare la sezione “Condivisione librerie” a pagina 184 per ulteriori informazioni.

Per aggiungere una libreria

1. Dai menu scegliere **Risorse > Aggiungi libreria**. Viene visualizzata la finestra di dialogo Aggiungi libreria.
2. Selezionare la libreria o le librerie nell'elenco.
3. Fare clic su **Aggiungi**. Se si verificano conflitti tra le librerie appena aggiunte e tutte le librerie che già esistono, verrà richiesto di verificare le risoluzioni di conflitto o modificarle prima di completare l'operazione. Consultare la sezione “Risoluzione dei conflitti” a pagina 186 per ulteriori informazioni.

Ricerca termini e tipi

È possibile ricercare nei diversi pannelli nell'editor utilizzando la funzione Trova. Nell'editor, è possibile scegliere **Modifica > Trova** dal menu e viene visualizzata la barra degli strumenti Trova. È possibile utilizzare questa barra degli strumenti per individuare una ricorrenza alla volta. Facendo clic su **Trova** nuovamente, è possibile trovare ricorrenze successive del proprio termine di ricerca.

Quando si ricerca, l'editor ricerca solo la libreria o librerie elencate nell'elenco a discesa sulla barra degli strumenti Trova. Se è selezionato **Tutte le librerie**, il programma ricercherà tutto nell'editor.

Quando si avvia una ricerca, si inizia dall'area che ha lo stato attivo. La ricerca continua per ogni sezione, ripercorrendo fino a quando non ritorna alla cella attiva. È possibile invertire l'ordine della ricerca utilizzando le frecce direzionali. È inoltre possibile scegliere se la ricerca è sensibile al maiuscolo/minuscolo.

Per trovare le stringhe nella vista

1. Dai menu scegliere **Modifica > Trova**. Viene visualizzata la barra degli strumenti Trova.
2. Immettere la stringa per cui si desidera eseguire la ricerca.
3. Fare clic sul pulsante **Trova** per avviare la ricerca. La ricorrenza successiva del termine o tipo viene quindi evidenziata.

4. Fare clic sul pulsante di nuovo per spostarsi di ricorrenza in ricorrenza.

Visualizzazione librerie

È possibile visualizzare il contenuto di una particolare libreria o di tutte le librerie. Ciò può essere utile quando si gestiscono molte librerie o quando si desidera esaminare il contenuto di una libreria specifica prima della pubblicazione. La modifica della vista influisce solo ciò che si vede in questa scheda Risorse libreria ma non disabilita tutte le librerie dall'essere utilizzate durante l'estrazione. Consultare la sezione "Disattivazione di librerie locali" per ulteriori informazioni.

La vista predefinita è **Tutte le librerie**, che mostra tutte le librerie nella struttura ad albero e il loro contenuto in altri riquadri. È possibile modificare questa selezione utilizzando l'elenco a discesa sulla barra degli strumenti o mediante una selezione di menu (**Visualizza > Librerie**). Quando una singola libreria viene visualizzata, tutti gli elementi presenti in altre librerie scompaiono dalla vista, ma sono ancora letti durante l'estrazione.

Per modificare la vista Libreria

1. Dai menu nella scheda Risorse della libreria, scegliere **Visualizza > Librerie**. Si apre un menu con tutte le librerie locali.
2. Selezionare la libreria che si desidera visualizzare o selezionare l'opzione **Tutte le librerie** per visualizzare il contenuto di tutte le librerie. Il contenuto della vista viene filtrato in base alla selezione.

Gestione delle librerie locali

Le librerie locali sono le librerie all'interno di sessione workbench interattiva o all'interno di un modello, in contrapposizione alle librerie pubbliche. Consultare la sezione "Gestione delle librerie pubbliche" a pagina 183 per ulteriori informazioni. Vi sono inoltre alcune attività di gestione libreria locale di base che si potrebbero voler eseguire, inclusi: la ridenominazione, la disabilitazione o l'eliminazione di una libreria locale.

Ridenominazione di librerie locali

È possibile ridenominare le librerie locali. Se una libreria locale viene ridenominata, essa verrà dissociata dalla versione pubblica, se ne esiste una. Ciò significa che le modifiche successive non possono essere più condivise con la versione pubblica. È possibile ripubblicare questa libreria locale con il suo nuovo nome. Ciò significa che non sarà possibile aggiornare la versione originale pubblica con le modifiche apportate a questa versione locale.

Nota: non è possibile ridenominare una libreria pubblica.

1. Dal menu scegliere **Modifica > Proprietà della libreria**. Viene visualizzata la finestra di dialogo Proprietà della libreria.

Per ridenominare una libreria locale

1. Nella vista ad albero della libreria, selezionare la libreria che si desidera ridenominare.
2. Immettere un nuovo nome per la libreria nella casella di testo Nome.
3. Fare clic su **OK** per accettare il nuovo nome per la libreria. La finestra di dialogo viene chiusa e il nome della libreria viene aggiornato nella vista ad albero.

Disattivazione di librerie locali

Se si desidera escludere temporaneamente una libreria dal processo di estrazione, è possibile deselezionare la casella di spunta alla sinistra del nome libreria nella vista ad albero. Ciò segnala che si desidera conservare la libreria, ma si desidera che il contenuto sia ignorato durante la verifica dei conflitti e durante l'estrazione.

Per disattivare una libreria

1. Nel riquadro ad albero della libreria, selezionare la libreria che si desidera disabilitare.
2. Fare clic sulla barra spaziatrice. La casella di spunta alla sinistra del nome è deselezionata.

Eliminazione di librerie locali

È possibile rimuovere una libreria senza eliminare la versione pubblica della libreria e viceversa. L'eliminazione di una libreria locale cancellerà la libreria e tutto il suo contenuto solo dalla sessione o dal . L'eliminazione di una versione locale di una libreria non rimuove tale libreria da altre sessioni, o dalla versione pubblica. Per ulteriori informazioni, consultare la sezione “Gestione delle librerie pubbliche”.

Per eliminare una libreria locale

1. Nella vista ad albero, selezionare la libreria che si desidera eliminare.
2. Dal menu scegliere **Modifica > Elimina** per eliminare la libreria. La libreria viene rimossa.
3. Se questa libreria non è mai stata pubblicata, viene visualizzato un messaggio che richiede se si desidera eliminare o mantenere questa libreria. Fare clic su **Elimina** per continuare o **Mantieni** se si desidera mantenere questa libreria.

Nota: deve sempre rimanere almeno una libreria.

Gestione delle librerie pubbliche

Per riutilizzare le librerie locali, è possibile pubblicarle e quindi lavorare con loro e visualizzarle attraverso la finestra di dialogo Gestione delle librerie (**Risorse > Gestisci librerie**). Consultare la sezione “Condivisione librerie” a pagina 184 per ulteriori informazioni. Alcune attività di gestione di libreria pubblica di base che si potrebbero voler eseguire includono l'importazione, l'esportazione o l'eliminazione di una libreria pubblica. Non è possibile ridenominare una libreria pubblica.

Importazione di librerie pubbliche

1. Nella casella di dialogo Gestisci librerie, fare clic su **Importa...**. Viene visualizzata la finestra di dialogo Importa libreria.
2. Selezionare il file della libreria (*.*lib*) che si desidera importare e se si desidera aggiungere questa libreria locale, selezionare **Aggiungi libreria al progetto corrente**.
3. Fare clic su **Importa**. La casella di dialogo viene chiusa. Se una libreria pubblica con lo stesso nome esiste già, verrà richiesto di ridenominare la libreria che si sta importando o di sovrascrivere la libreria pubblica corrente.

Esportazione delle librerie pubbliche

È possibile esportare le librerie pubbliche nel formato *.lib* in modo che sia possibile condividerle.

1. Nella finestra di dialogo Gestisci librerie, selezionare la libreria che si desidera esportare nell'elenco.
2. Fare clic su **Esporta**. Viene aperta la finestra di dialogo Seleziona directory.
3. Selezionare la directory in cui si desidera esportare e fare clic su **Esporta**. La finestra di dialogo viene chiusa e il file di libreria (*.*lib*) viene esportato.

Eliminazione di librerie pubbliche

È possibile rimuovere una libreria senza eliminare la versione pubblica della libreria e viceversa. Tuttavia, se la libreria viene eliminata da questa finestra di dialogo, non può più essere aggiunta a qualsiasi sessione risorse fino a non viene pubblicata di nuovo una versione locale.

Per eliminare una libreria installata con il prodotto, viene ripristinata la versione installata in origine.

1. Nella finestra di dialogo Gestisci librerie, selezionare la libreria che si desidera eliminare. E' possibile ordinare l'elenco facendo clic sull'intestazione appropriata.
2. Fare clic su **Elimina** per eliminare la libreria. IBM SPSS Modeler Text Analytics verifica se la versione locale della libreria è la stessa della libreria pubblica. In tal caso, la libreria viene rimossa senza alcun avviso. Se le versioni della libreria differiscono, una segnalazione viene visualizzata per chiedere se si desidera conservare o rimuovere la versione pubblica.

Condivisione librerie

Librerie consentono di gestire le risorse in un modo che sia facile da condividere tra più sessioni workbench interattive . Le librerie possono esistere in due stati o versioni. Le librerie che sono modificabili nell'editor e parte di una sessione workbench interattiva sono denominate **librerie locali**. Mentre si lavora all'interno di una sessione workbench interattiva, è possibile apportare molte modifiche nella libreria *Ortaggi*, ad esempio. Se le modifiche potrebbero essere utili con altri dati, è possibile apportare queste risorse disponibili creando una versione di **libreria pubblica** della libreria *Ortaggi*. Una libreria pubblica, come indica il nome, è disponibile per qualsiasi altra risorse in qualsiasi sessione workbench interattiva.

È possibile visualizzare le librerie pubbliche nella finestra di dialogo Gestisci librerie. Una volta che esiste la versione di libreria pubblica, è possibile aggiungerla alle risorse in altri contesti in modo che queste risorse linguistiche personalizzate possono essere condivise.

Le librerie fornite son inizialmente librerie pubbliche. È possibile modificare le risorse in queste librerie e quindi creare una nuova versione pubblica. Le nuove versioni dovrebbero quindi essere accessibili in altre sessioni workbench interattive.

Continuando a lavorare con le librerie e apportare le modifiche, le versioni della libreria saranno desincronizzate. In alcuni casi, una versione locale potrebbe essere più recente della versione pubblica e, in altri casi, la versione pubblica può essere più recente della versione locale. È inoltre possibile per le versioni pubbliche e locali contenere modifiche che gli altri non fanno se la versione pubblica è stata aggiornata dall'interno di un'altra sessione workbench interattiva. Se le versioni di libreria diventano desincronizzate, è possibile sincronizzarle di nuovo. La sincronizzazione delle versioni della libreria consiste nel ripubblicare e/o aggiornare le librerie locali.

Quando viene avviata una sessione workbench interattiva o se ne chiude una , verrà richiesto di sincronizzare tutte le librerie che devono essere aggiornate o ripubblicate. Inoltre, è possibile identificare facilmente lo stato di sincronizzazione di libreria locale dall'icona che appare accanto al nome libreria nella vista ad albero o visualizzando la finestra di dialogo Proprietà della libreria. È inoltre possibile scegliere di farlo in qualsiasi momento tramite le selezioni del menu. La seguente tabella descrive i cinque possibili stati e le icone associate.

Tabella 37. Stati di sincronizzazione della libreria locale.

Icona	Descrizione stato della libreria locale
	Non pubblicata — La libreria locale non è mai stata pubblicata.
	Sincronizzata - Le versioni di librerie locali e pubbliche sono identiche. Ciò vale anche per la *Libreria locale*, che non può essere pubblicata perché è destinata a contenere solo risorse specifiche per la sessione .
	Scaduta - La versione della libreria pubblica è più recente della versione locale. È possibile aggiornare la versione locale con le modifiche.
	Più recente - La versione della libreria locale è più recente della versione pubblica. È possibile ripubblicare la versione locale sulla versione pubblica.

Tabella 37. Stati di sincronizzazione della libreria locale (Continua).

Icona	Descrizione stato della libreria locale
	Fuori sincronizzazione — Entrambe le librerie locale e pubblica contengono le modifiche che le altre non contengono. È necessario decidere se aggiornare o pubblicare la libreria locale. Se si aggiorna, si perderanno le modifiche apportate dall'ultima volta in cui è stato aggiornata o pubblicata. Se si sceglie di pubblicare, sarà possibile sovrascrivere le modifiche nella versione pubblica.

Nota: se si desidera aggiornare le librerie quando viene avviata una sessione workbench interattiva o si pubblica se se ne chiude una, è meno probabile che le librerie siano non sincronizzate.

È possibile ripubblicare una libreria ogni volta che si pensa che le modifiche nella libreria potrebbero portare altri vantaggi a flussi che possono contenere anche questa libreria. Quindi, se le modifiche potrebbero portare vantaggio ad altri flussi, è possibile aggiornare le versioni locali in tali flussi. In questo modo, è possibile creare flussi per ogni contesto o dominio che si applica ai dati creando nuove librerie e/o aggiungendo qualsiasi numero di librerie pubbliche alle risorse.

Se una versione pubblica di una libreria è condivisa, c'è una maggiore probabilità che sorgeranno differenze tra le versioni locali e pubbliche. Quando si avvia o chiude e pubblica da una sessione workbench interattiva oppure si apre o chiude un modello da Editor di modelli, viene visualizzato un messaggio che abilita a pubblicare e/o aggiornare le librerie le cui versioni non sono sincronizzate con quelle nella finestra di dialogo Gestisci librerie. Se la versione della libreria pubblica è più recente della versione locale, si apre una finestra di dialogo che chiede se si desidera aggiornare. È possibile scegliere se conservare la versione locale come è invece di aggiornare con la versione pubblica o unire gli aggiornamenti nella libreria locale.

Pubblicazione delle librerie

Se una particolare libreria non è mai stata pubblicata, la pubblicazione comporta la creazione di una copia pubblica della propria libreria locale nel database. Se si sta eseguendo una nuova pubblicazione di una libreria, il contenuto della libreria locale sostituirà il contenuto della versione pubblica esistente. Dopo la pubblicazione, è possibile aggiornare questa libreria in qualsiasi altra sessione del flusso in modo che le loro versioni locali siano sincronizzate con la versione pubblica. Anche se è possibile pubblicare una libreria, una versione locale è sempre memorizzata nella sessione .

Importante! Se si apportano modifiche alla propria libreria locale e, nel frattempo, anche la versione pubblica della libreria è stata modificata, la libreria viene considerata essere fuori sincronizzazione. Si consiglia di iniziare aggiornando la versione locale con le modifiche pubbliche, apportare le modifiche desiderate e, quindi, di pubblicare di nuovo la versione locale in modo da rendere le due versioni identiche. Se si eseguono modifiche, sarà possibile sovrascrivere le modifiche nella versione pubblica.

Per pubblicare le librerie locali sul database

1. Dai menu scegliere **Risorse > Pubblica libreria**. La finestra di dialogo Pubblica librerie si apre con tutte le librerie che necessitano di pubblicazione selezionate per impostazione predefinita.
2. Selezionare la casella di spunta alla sinistra di ciascuna libreria che si desidera pubblicare o ripubblicare.
3. Fare clic su **Pubblica** per pubblicare le librerie sul database Gestisci librerie.

Aggiornamento delle librerie

Quando si avvia o chiude una sessione workbench interattiva, è possibile aggiornare o pubblicare tutte le librerie che non sono più sincronizzate con le versioni pubbliche. Se la versione della libreria pubblica è più recente della versione locale, si apre una finestra di dialogo che chiede se si desidera aggiornare. È possibile scegliere se conservare la versione locale invece di aggiornare la versione pubblica o sostituire la versione locale con quella pubblica. Se una versione pubblica di una libreria è più recente della versione

locale, è possibile aggiornare la versione locale per sincronizzare il suo contenuto con quella della versione pubblica. Per aggiornamento si intende incorporare le modifiche rilevate nella versione pubblica nella versione locale.

Nota: se si desidera aggiornare le librerie quando viene avviata una sessione workbench interattiva o si pubblica quando se ne chiude una, è meno probabile che le librerie siano non sincronizzate. Consultare la sezione "Condivisione librerie" a pagina 184 per ulteriori informazioni.

Per aggiornare le librerie locali

1. Dai menu scegliere **Risorse > Aggiorna librerie**. La finestra di dialogo Aggiorna librerie si apre con tutte le librerie che necessitano di aggiornamenti selezionati per impostazione predefinita.
2. Selezionare la casella di spunta alla sinistra di ciascuna libreria che si desidera pubblicare o ripubblicare.
3. Fare clic su **Aggiorna** per aggiornare le librerie locali.

Risoluzione dei conflitti

Conflitti tra librerie locali e pubbliche

Quando si avvia un flusso di sessione , IBM SPSS Modeler Text Analytics esegue un confronto tra le librerie locali e quelle elencate nella finestra di dialogo Gestisci librerie. Se tutte le librerie locali nella propria sessione non sono sincronizzate con le versioni pubblicate, viene visualizzata la finestra di dialogo Avvertimento di sincronizzazione libreria. È possibile scegliere tra le seguenti opzioni per selezionare le versioni della libreria che si desidera utilizzare:

- **Tutte le librerie locali su file.** Questa opzione conserva tutte le librerie locali come sono. È sempre possibile ripubblicarle o aggiornarle in seguito.
- **Tutte le librerie pubblicate su questa macchina.** Questa opzione consente di sostituire le librerie locali visualizzate con le versioni trovate nel database.
- **Tutte le librerie più recenti.** Questa opzione sostituirà tutte le vecchie librerie locali con le versioni pubbliche più recenti dal database.
- **Altro.** Questa opzione consente di selezionare manualmente la versione che si desidera scegliendo nella tabella.

Conflitti di termini forzati

Ogni volta che si aggiunge una libreria pubblica o aggiornare una libreria locale, i conflitti e le voci duplicate possono essere scoperte tra i termini e i tipi in questa libreria e i termini e i tipi nelle altre librerie delle proprie risorse. Se ciò si verifica, verrà richiesto di verificare le risoluzioni di conflitto proposte o di modificare i termini prima di completare l'operazione nella finestra di dialogo Modifica termini forzati. Consultare la sezione "Forzatura di termini" a pagina 195 per ulteriori informazioni.

La finestra di dialogo Modifica termini forzati contiene ogni coppia di termini o i tipi in conflitto. L'alternanza dei colori di sfondo serve a distinguere visivamente ogni coppia in conflitto. Tali colori possono essere modificati nella finestra Opzioni. Per ulteriori informazioni, consultare la sezione "Opzioni: scheda Visualizza" a pagina 82. La finestra di dialogo Modifica termini forzati contiene due schede:

- **Duplicati.** Questa scheda contiene i termini duplicati trovati nelle librerie. Se viene visualizzata un'icona, ciò significa che questa ricorrenza del termine è stata forzata. Se viene visualizzata un'icona X nera, ciò significa che questa ricorrenza del termine sarà ignorata durante l'estrazione in quanto è stata forzata altrove.
- **Definito dall'utente.** Questa scheda contiene un elenco di tutti i termini che sono stati forzati manualmente nel pannello termine di dizionario di tipo e non mediante conflitti.

Nota: la finestra di dialogo Modifica termini forzati si apre dopo aver aggiunto o aggiornato una libreria. Se si annulla questa finestra di dialogo, non sarà possibile annullare l'aggiornamento o aggiunta della libreria.

Per risolvere i conflitti

1. Nella finestra di dialogo Modifica termini forzati, selezionare il pallino nella colonna Usa per il termine da forzare.
2. Una volta terminato, fare clic su **OK** per applicare i termini forzati e chiudere la finestra di dialogo. Se si fa clic su **Annulla**, sarà necessario annullare le modifiche apportate in questa finestra di dialogo.

Capitolo 17. Informazioni sui dizionari di libreria

Le risorse utilizzate per estrarre i dati di testo vengono memorizzate sotto forma di modelli e librerie. Una libreria può essere composta da tre dizionari.

- Il **dizionario di tipo** è una raccolta di termini raggruppati sotto un'etichetta, un tipo o un nome. Quando il motore di estrazione legge i dati di testo, esso confronta le parole trovate nel testo con i termini definiti nei propri dizionari di tipo. Durante l'estrazione, le forme con inflessione di termini e sinonimi di un tipo sono raggruppati sotto un termine di destinazione denominato concetto. I concetti estratti vengono assegnati al dizionario di tipo in cui appaiono come termini. È possibile gestire i dizionari di tipo in alto a sinistra e nei pannelli centrali dell'editor-la struttura ad albero Libreria ed il riquadro del termine. Consultare la sezione “Dizionari di tipo” per ulteriori informazioni.
- Il **dizionario di sostituzione** contiene una raccolta di parole definite come sinonimio come elementi facoltativi utilizzato per raggruppare i termini simili in un termine di destinazione, chiamato concetto nei risultati dell'estrazione finale. È possibile gestire i dizionari di sostituzione nel pannello in basso a sinistra dell'editor utilizzando i sinonimi schedae la scheda Facoltativo. Consultare la sezione “Dizionari dei sinonimi/di sostituzione” a pagina 196 per ulteriori informazioni.
- Il **dizionario di esclusione** contiene una raccolta di termini e i tipi che verranno rimossi dai risultati dell'estrazione finale. È possibile gestire i dizionari di esclusione nel riquadro più a destra dell'editor. Consultare la sezione “Dizionari di esclusione” a pagina 200 per ulteriori informazioni.

Per ulteriori informazioni, consultare la sezione Capitolo 16, “Gestione delle librerie”, a pagina 179.

Dizionari di tipo

Un **dizionario di tipo** è costituito da un nome tipo o etichetta e da un elenco di termini. I dizionari di tipo vengono gestiti nei riquadri in alto a sinistra e al centro della scheda Risorse della libreria nell'editor. È possibile accedere a questa vista con **Vista > Editor delle risorse** nel menu , se ci si trova in una sessione di postazione interattiva. In caso contrario, è possibile revisionare i dizionari per un modello specifico in Editor di modelli.

Quando il motore di estrazione legge i dati di testo, esso confronta le parole trovate nel testo con i termini definiti nei propri dizionari di tipo. I termini sono parole o frasi dei dizionari di tipo nelle proprie risorse linguistiche.

Quando una parola corrisponde a un termine, essa viene assegnata al nome tipo per questo termine. Quando le risorse vengono lette durante l'estrazione, i termini trovati nel testo attraversano diversi passi di elaborazione prima di diventare concetti nel pannello dei risultati di estrazione. Se i termini appartenenti allo stesso dizionario di tipo sono stabiliti per essere sinonimi dal motore di estrazione, essi vengono raggruppati sotto il termine più frequente che diventa **concetto** nel riquadro Risultati di estrazione. Ad esempio, quando i termini `domanda` e `interrogazione` potrebbero apparire sotto il nome concetto `domanda` alla fine.

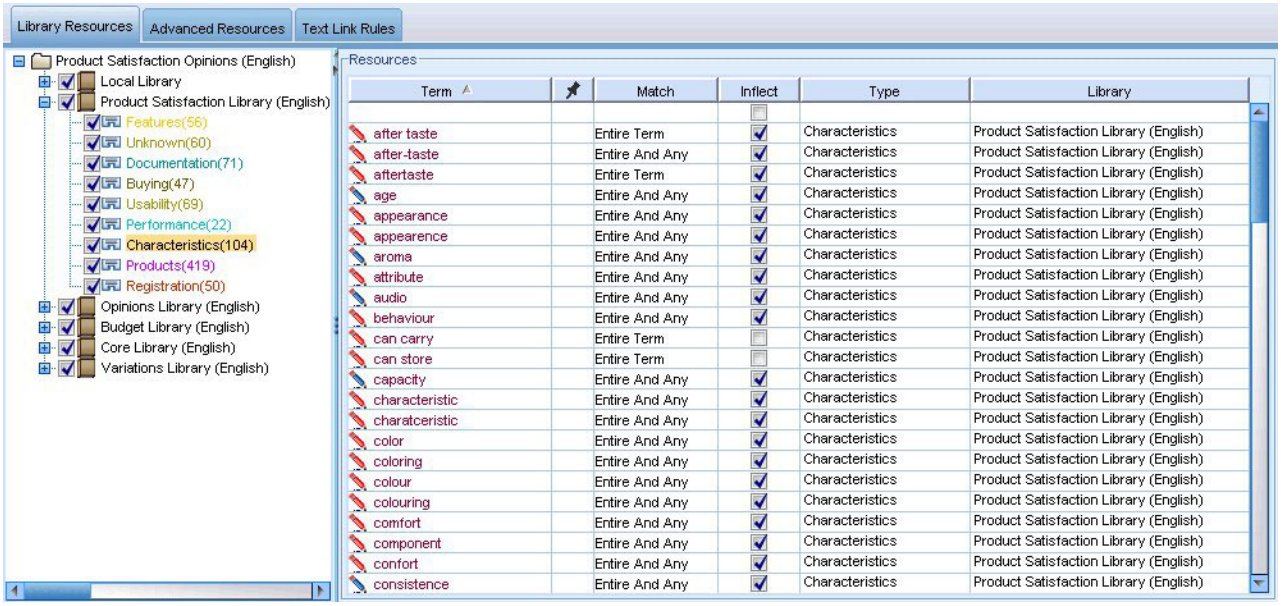


Figura 40. Struttura ad albero Libreria e riquadro dei termini

L'elenco dei dizionari di tipo viene visualizzato nel riquadro della struttura ad albero Libreria sulla sinistra. Il contenuto di ogni dizionario di tipo viene visualizzato nel pannello centrale. I dizionari di tipo consistono in più di un semplice elenco di termini. Il modo in cui le parole e le frasi nei dati di testo vengono confrontati con i termini definiti nei dizionari di tipo viene determinato dall'opzione di corrispondenza definita. Una **opzione di corrispondenza** specifica il modo in cui un termine è sancito rispetto a una parola o frase candidate nei dati di testo. Per ulteriori informazioni, consultare la sezione "Aggiunta di termini" a pagina 192.

Nota: non tutte le opzioni, come l'opzione di corrispondenza e forme con molte inflessioni, si applicano al testo giapponese.

Inoltre, è possibile estendere i termini nel dizionario di tipo per specificare se si desidera generare automaticamente ed aggiungere forme di inflessione dei termini al dizionario. Generando forme di inflessione, è possibile aggiungere automaticamente forme plurali di termini singolari, forme plurali di termini e aggettivi al dizionario di tipo. Consultare la sezione "Aggiunta di termini" a pagina 192 per ulteriori informazioni.

Nota: Per la maggior parte delle lingue, concetti che non sono presenti nei dizionari ma vengono estratti dal testo vengono automaticamente inserite come `<Sconosciuto>`

Tipi incorporati

IBM SPSS Modeler Text Analytics viene fornito con una serie di risorse linguistiche nella forma librerie fornite e risorse compilate. Le librerie fornite contengono una serie di dizionari di tipo incorporati come `<Ubicazione>`, `<Organizzazione>`, `<Persona>` e `<Prodotto>`.

Nota: la serie di tipi predefiniti e incorporati è differente per il testo giapponese.

Questi dizionari di tipo vengono utilizzati dal motore di estrazione per assegnare i tipi ai concetti che estrae, come è assegnato il tipo `<Ubicazione>` al concetto di `parigi`. Benché un gran numero di termini sono stati definiti nel dizionari di tipo incorporati, essi non coprono ogni possibilità. Pertanto, è possibile aggiungerne o crearne uno proprio. Per una descrizione del contenuto di un particolare dizionario di tipo fornito, leggere l'annotazione nella finestra di dialogo Proprietà tipo. Selezionare il tipo nella struttura ad albero e scegliere **Modifica > Proprietà** dal menu di scelta rapida.

Nota: oltre alle librerie fornite, le risorse compilate (utilizzate anche dal motore di estrazione) contengono un gran numero di definizioni complementari ai dizionari di tipo incorporati, ma il contenuto non è visibile nel prodotto. È possibile, tuttavia, forzare un termine che è stato immesso da questi dizionari compilati in qualsiasi altro dizionario. Consultare la sezione “Forzatura di termini” a pagina 195 per ulteriori informazioni.

Creazione di tipi

È possibile creare dizionari di tipo simili per facilitare il raggruppamento di termini simili. Quando i termini che compaiono in questo dizionario vengono rilevati durante il processo di estrazione, saranno assegnati a questo nome tipo ed estratti con un nome concetto. Quando si crea una libreria, una libreria di tipo vuota è sempre inclusa in modo da poter iniziare immettendo termini immediatamente.

Importante:: non è possibile creare nuovi tipi per risorse in giapponese.

Se si sta analizzando il testo sul cibo e si vogliono raggruppare i termini relativi alla verdura, è possibile creare il proprio dizionario di tipo `<Verdura>`. È possibile aggiungere termini come `carota`, `broccoli` e `spinaci` se si pensa che siano termini importanti da visualizzare nel testo. Quindi, durante l'estrazione, se qualcuno di questi termini viene rilevato, esso viene estratto come concetto ed assegnato al tipo `<Verdura>` .

Non è necessario definire ogni forma di una parola o espressione, perché è possibile scegliere di generare le forme di inflessione dei termini. Selezionando questa opzione, il motore di estrazione riconosce automaticamente forme plurali o singolari dei termini tra le altre forme come appartenenti a questo tipo. Questa opzione è particolarmente utile quando il tipo contiene più nomi, poiché è improbabile che si desiderano forme di inflessione di verbi o aggettivi.

La finestra di dialogo Proprietà di tipo contiene i seguenti campi.

Nome. Il nome che si dà al dizionario di tipo che si sta creando. Si consiglia di non utilizzare spazi nei nomi di tipo, soprattutto se due o più nomi di tipo iniziano con la stessa parola.

Nota: vi sono alcune restrizioni sui nomi di tipo e l'utilizzo di simboli. Non utilizzare, ad esempio, i simboli "@" o "!" all'interno del nome.

Corrispondenza predefinita. La corrispondenza predefinita indica al motore di estrazione come mettere in corrispondenza tale termine a dati di testo. Ogni volta che si aggiunge un termine a questo dizionario di tipo, questo è l'attributo di corrispondenza assegnato automaticamente ad esso. Si può sempre modificare la scelta manualmente nell'elenco dei termini. Le opzioni includono: **Termine intero**, **Avvia**, **Fine**, **Qualsiasi**, **Avvia o Fine**, **Intero e iniziale**, **Intero e finale**, **Intero e (iniziale o finale)**, e **Intero (non composti)**. Per ulteriori informazioni, consultare la sezione “Aggiunta di termini” a pagina 192. Questa opzione non si applica alle risorse in giapponese.

Aggiungi a. Questo campo indica la libreria in cui verrà creato il nuovo dizionario di tipo.

Genera forme di inflessione predefinite. Questa opzione indica al motore di estrazione la morfologia grammaticale da utilizzare per catturare e raggruppare forme simili dei termini che si aggiungono a questo dizionario, quali la forma singolare o plurale del termine. Questa opzione è particolarmente utile quando il tipo contiene maggiormente sostantivi. Quando si seleziona questa opzione, tutti i nuovi termini aggiunti a questo tipo conterranno automaticamente questa opzione anche se è possibile modificarla manualmente nell'elenco. Questa opzione non si applica alle risorse in giapponese.

Colore font. Questo campo consente di distinguere i risultati di questo tipo da altri nell'interfaccia. Se si seleziona **Usa colore principale**, il colore tipo predefinito viene utilizzato anche per questo dizionario di

tipo. Questo colore predefinito è impostato nella finestra di dialogo Opzioni. Per ulteriori informazioni, consultare “Opzioni: scheda Visualizza” a pagina 82. Se si seleziona **Personalizza**, selezionare un colore dall'elenco a discesa.

Annotazione. Questo campo è facoltativo e può essere utilizzato per qualsiasi commento o descrizione.

Per creare un dizionario di tipo

1. Selezionare la libreria in cui si desidera creare un nuovo dizionario di tipo.
2. Dai menu scegliere **Strumenti > Nuovo tipo**. Viene visualizzata la finestra di dialogo Proprietà del tipo.
3. Immettere il nome del dizionario di tipo nella casella di testo **Nome** e selezionare le opzioni desiderate.
4. Fare clic su **OK** per creare il dizionario di tipo. Il nuovo tipo è visibile nel riquadro della struttura ad albero delle librerie e viene visualizzato nel pannello centrale. È possibile iniziare ad aggiungere i termini immediatamente. Per ulteriori informazioni, vedere “Aggiunta di termini”.

Nota: queste istruzioni mostrano come eseguire le modifiche all'interno della Editor risorsevistao Editor di modelli. Tenere presente che è anche possibile effettuare questo tipo di ottimizzazione direttamente dal pannello Risultati di estrazione , Pannello dati, riquadro Categorie o dalla finestra di dialogo Definizioni cluster nelle altre viste. Per ulteriori informazioni, consultare la sezione “Perfezionamento dei risultati di estrazione” a pagina 95.

Aggiunta di termini

La struttura ad albero Libreria visualizza le librerie e può essere espansa per visualizzare i dizionari di tipo che essi contengono. Nel riquadro centrale, un elenco di termini visualizza i termini nella libreria selezionata o nel dizionario di tipo, a seconda della selezione nella struttura ad albero.

Importante! I termini sono definiti in modo diverso per le risorse in giapponese.

In Editor risorse, è possibile aggiungere termini al dizionario di tipo direttamente nel riquadro del termine o tramite la finestra di dialogo Aggiungi nuovi termini. I termini che è possibile aggiungere possono essere parole singole o composte. Si troverà sempre una riga vuota in cima all'elenco per consentire all'utente di aggiungere un nuovo termine.

Nota: queste istruzioni mostrano come eseguire le modifiche all'interno della Editor risorsevistao Editor di modelli. Tenere presente che è anche possibile effettuare questo tipo di ottimizzazione direttamente dal pannello Risultati di estrazione , Pannello dati, riquadro Categorie o dalla finestra di dialogo Definizioni cluster nelle altre viste. Per ulteriori informazioni, consultare la sezione “Perfezionamento dei risultati di estrazione” a pagina 95.

Colonna Termine

In questa colonna, immettere parole singole o composte nella cella. Il colore in cui il termine viene visualizzato dipende dal colore per il tipo in cui il termine viene memorizzato o forzato. È possibile modificare i colori di tipo nella finestra di dialogo Proprietà tipo. Consultare la sezione “Creazione di tipi” a pagina 191 per ulteriori informazioni.

Colonna Forza

In questa colonna, facendo clic e inserendo una icona in questa cella si indica al motore di estrazione di ignorare tutte le altre ricorrenze di questo stesso termine in altre librerie. Consultare la sezione “Forzatura di termini” a pagina 195 per ulteriori informazioni.

Colonna Corrispondenza

In questa colonna, selezionare un'opzione di corrispondenza per istruire il motore di estrazione su come mettere in corrispondenza tale termine a dati di testo. Consultare la tabella per gli esempi. È possibile modificare il valore predefinito modificando le proprietà di tipo. Consultare la sezione "Creazione di tipi" a pagina 191 per ulteriori informazioni. Dai menu scegliere **Modifica > Cambia corrispondenza**. Di seguito sono riportate le opzioni di corrispondenza di base poiché sono possibili le combinazioni di questi:

- **Avvia**. Se il termine nel dizionario corrisponde alla prima parola in un concetto estratto dal testo, questo tipo viene assegnato. Ad esempio, se si immette `mela`, la corrispondenza sarà `torta di mele`.
- **Fine**. Se il termine nel dizionario corrisponde all'ultima parola in un concetto estratto dal testo, viene assegnato questo tipo. Ad esempio, se si immette `mela`, la corrispondenza sarà anche `sidro di mele`.
- **Qualsiasi**. Se il termine nel dizionario corrisponde ad una parola qualsiasi in un concetto estratto dal testo, viene assegnato questo tipo. Ad esempio, se si immette `mela`, l'opzione **Qualsiasi** immetterà `torta di mele`, `sidro di melee torta di sidro di mele` allo stesso modo.
- **Termine intero**. Se l'intero concetto estratto dal testo corrisponde al termine esatto del dizionario, viene assegnato questo tipo. L'aggiunta di un termine come **Termine intero**, **Intero e iniziale**, **Intero e finale**, **Intero e qualsiasi**, oppure **intero (non composti)** obbligherà l'estrazione di un termine.

  Inoltre, poiché il tipo `<Persona>` estrae solo due nomi parte, come *piaf Edith* o *mohandas gandhi*, si potrebbe voler aggiungere esplicitamente i primi nomi per questo dizionario di tipo se si sta tentando di estrarre un nome proprio quando non viene menzionato il cognome. Ad esempio, se si desidera catturare tutte le istanze di *Edith* come nome, è necessario aggiungere `edith` al tipo `<Persona>` utilizzando **Termine intero** o **Intero e iniziale**.
- **Intero (non composti)**. Se il concetto intero estratto dal testo corrisponde al termine esatto nel dizionario, questo tipo viene assegnato e l'estrazione viene arrestata per proibire l'estrazione di corrispondenza del termine a un composti più lungo. Ad esempio, se si immette `mela`, l'opzione **Intero (non composto)** l'opzione di tipo `mela` e non si estrae il composto `succo di mela` a meno che non sia costretto altrove.

Nella seguente tabella, si supponga che il termine `mela` si trova in un dizionario di tipo. A seconda dell'opzione di corrispondenza, questa tabella mostra quali concetti sarebbero estratti e inseriti se sono stati rilevati nel testo.

Tabella 38. Esempi di corrispondenza.

Opzioni di corrispondenza per il termine:  **mela**	**Concetti estratti**			
	mela	**mela** *torta*	*matura* **mela**	*artigianali* **mela** *torta*
Termine intero	✔			
Avvia.		✔		
Fine.			✔	
Inizio o Fine		✔	✔	
Intero e iniziale	✔	✔		
Intero e finale	✔		✔	

Tabella 38. Esempi di corrispondenza (Continua).

Opzioni di corrispondenza per il termine:  **mela**	**Concetti estratti**			
	mela	**mela** *torta*	*matura* **mela**	*artigianali* **mela** *torta*
Intero e (iniziale o finale)	✔	✔	✔	
Qualsiasi.		✔	✔	✔
Intero e qualsiasi	✔	✔	✔	✔
Intero (non composti).	✔	*mai estratto*	*mai estratto*	*mai estratto*

Colonna Con inflessioni

In questa colonna, selezionare se il motore di estrazione deve generare forme di inflessione di questo termine durante l'estrazione in modo che siano tutte raggruppate insieme. Il valore predefinito per questa colonna è definito nella proprietà Tipo ma è possibile modificare questa opzione caso per caso direttamente nella colonna. Dai menu scegliere **Modifica > Cambia inflessione**.

Colonna Tipo

In questa colonna, selezionare un dizionario di tipo dall'elenco a discesa. L'elenco dei tipi è filtrato in base alla propria selezione nel riquadro della struttura ad albero delle librerie. Il primo tipo nell'elenco è sempre il tipo predefinito selezionato nel riquadro della struttura ad albero della libreria. Dai menu scegliere **Modifica > Cambia tipo**.

Colonna libreria

In questa colonna, viene visualizzata la libreria in cui è memorizzato il termine. È possibile trascinare e rilasciare un termine in un altro tipo nel riquadro della struttura ad albero per modificare la sua libreria.

Per aggiungere un termine singolo ad un dizionario di tipo

1. Nella struttura ad albero della libreria, selezionare il dizionario di tipo a cui si desidera aggiungere il termine.
2. Nell'elenco dei termini nel riquadro centrale, immettere il termine nella prima cella vuota disponibile e impostare tutte le opzioni che si desidera per questo termine.

Per aggiungere più termini ad un dizionario di tipo

1. Nella struttura ad albero della libreria, selezionare il dizionario di tipo a cui si desidera aggiungere il termine.
2. Dai menu scegliere **Strumenti > Nuovi termini**. Viene visualizzata la finestra di dialogo Aggiungi nuovi termini.
3. Immettere i termini che si desidera aggiungere al dizionario del tipo selezionato immettendo i termini o con copia e incolla di un gruppo di termini. Se si immettono più termini, è necessario separarli utilizzando il delimitatore che è definito nella finestra di dialogo Opzioni oppure aggiungere ogni termine a una nuova riga. Per ulteriori informazioni, consultare “Impostazione delle opzioni” a pagina 82.

4. Fare clic su **OK** per aggiungere i termini al dizionario. L'opzione di corrispondenza viene automaticamente impostata per l'opzione predefinita per questa libreria di tipo. La finestra di dialogo viene chiusa e i nuovi termini vengono visualizzati nel dizionario.

Forzatura di termini

Se si desidera che un termine venga assegnato ad un particolare tipo, è possibile aggiungerlo al dizionario di tipo corrispondente. Tuttavia, se vi sono più termini con lo stesso nome, il motore di estrazione deve sapere quale tipo deve essere utilizzato. Quindi, verrà richiesto di selezionare quale tipo deve essere utilizzato. Ciò viene detto **forzatura** di un termine in un tipo. Questa opzione è molto utile quando si sovrascrive l'assegnazione del tipo da un dizionario compilato (interno, non modificabile). In generale, si consiglia di evitare termini duplicati.

La forzatura non *rimuove* le altre ricorrenze di questo termine; piuttosto, questi verranno ignorati dal motore di estrazione. È possibile modificare successivamente quale ricorrenza deve essere utilizzata forzando o non forzando un termine. Potrebbe anche essere necessario forzare un termine in un dizionario di tipo quando si aggiunge o si aggiorna una libreria pubblica.

È possibile vedere quali termini sono forzati o ignorati nella colonna Forza, la seconda colonna nel riquadro del termine. Se viene visualizzata un'icona, ciò significa che questa ricorrenza del termine è stata forzata. Se viene visualizzata un'icona X nera, ciò significa che questa ricorrenza del termine sarà ignorata durante l'estrazione in quanto è stata forzata altrove. Inoltre, quando si forza un termine, esso verrà visualizzato nel colore del tipo per cui è stato forzato. Ciò significa che se si è forzato un termine che è in entrambi `Tipo 1` e `Tipo 2` in `Tipo 1`, ogni volta che si visualizza questo termine nella finestra, viene visualizzato nel colore font definito per `Tipo 1`.

È possibile fare doppio clic con il mouse sull'icona per modificare lo stato. Se il termine appare altrove, viene visualizzata una finestra di dialogo Risolvi conflitti per consentire di selezionare quale ricorrenza deve essere utilizzata.

Ridenominazione dei tipi

È possibile ridenominare un dizionario di tipo o modificare le altre impostazioni di dizionario modificando le proprietà di tipo.

Importante! Si consiglia di non utilizzare spazi nei nomi di tipo, soprattutto se due o più nomi di tipo iniziano con la stessa parola. Si consiglia inoltre di non ridenominare i tipi nelle librerie Centrale o Opinioni o modificare gli attributi di corrispondenza predefiniti.

Per ridenominare un tipo

1. Nel riquadro ad albero della libreria, selezionare il dizionario di tipo che si desidera ridenominare.
2. Fare clic con il mouse e scegliere **Proprietà del tipo** dal menu di scelta rapida. Viene visualizzata la finestra di dialogo Proprietà del tipo.
3. Immettere il nuovo nome per il dizionario di tipo nella casella di testo Nome.
4. Fare clic su **OK** per accettare il nuovo nome. Il nome del nuovo tipo è visualizzabile nel riquadro della struttura ad albero della libreria.

Spostamento dei tipi

È possibile trascinare un dizionario del tipo in un'altra ubicazione all'interno di una libreria o in un'altra libreria nella struttura ad albero.

Per riordinare un tipo all'interno di una libreria

1. Nel riquadro ad albero della libreria, selezionare il dizionario di tipo che si desidera spostare.

2. Dai menu, scegliere **Modifica > Sposta su** per spostare il tipo di dizionario in alto di una posizione nel riquadro della struttura ad albero della libreria o **Modifica > Sposta giù** per spostare in basso di una posizione.

Per spostare un tipo in un'altra libreria

1. Nel riquadro ad albero della libreria, selezionare il dizionario di tipo che si desidera spostare.
2. Fare clic con il mouse e scegliere **Proprietà del tipo** dal menu di scelta rapida. Viene visualizzata la finestra di dialogo Proprietà del tipo. (È inoltre possibile trascinare e rilasciare il tipo in un'altra libreria).
3. Nella finestra Aggiungi a elenco, selezionare la libreria in cui si desidera spostare il dizionario di tipo.
4. Fare clic su **OK**. La finestra di dialogo viene chiusa e il tipo è ora nella libreria selezionata.

Disattivazione ed eliminazione dei tipi

Se si desidera rimuovere temporaneamente un dizionario di tipo, è possibile disabilitarlo deselezionando la casella di spunta alla sinistra del nome dizionario nel riquadro della struttura ad albero della libreria. Ciò indica che si desidera conservare il dizionario nella libreria, ma si desidera che il contenuto venga ignorato durante il controllo dei conflitti e durante il processo di estrazione.

È anche possibile eliminare definitivamente dizionari di tipo da una libreria.

Per disabilitare un dizionario di tipo

1. Nel riquadro ad albero della libreria, selezionare il dizionario di tipo che si desidera disabilitare.
2. Fare clic sulla barra spaziatrice. La casella di spunta alla sinistra del nome è deselezionata.

Per eliminare un dizionario di tipo

1. Nel riquadro ad albero della libreria, selezionare il dizionario di tipo che si desidera eliminare.
2. Dal menu scegliere **Modifica > Elimina** per eliminare il dizionario di tipo.

Dizionari dei sinonimi/di sostituzione

Un **dizionario di sostituzione** è una raccolta di termini che contribuiscono a raggruppare termini analoghi sotto un termine di destinazione. I dizionari di sostituzione sono gestiti nel riquadro inferiore della scheda Risorse della libreria. È possibile accedere a questa vista con **Vista > Editor delle risorse** nel menu , se ci si trova in una sessione di postazione interattiva. In caso contrario, è possibile revisionare i dizionari per un modello specifico in Editor di modelli.

È possibile definire due forme di sostituzioni in questo dizionario: **sinonimi** e **elementi facoltativi**. È possibile fare clic sulle schede in questo riquadro per passare tra di loro.

Dopo aver eseguito un'estrazione di dati di testo, è possibile trovare diversi concetti che sono sinonimi o forme di inflessione di altri concetti. Identificando gli elementi facoltativi e sinonimi, è possibile forzare il motore di estrazione per associare questi ad un singolo termine di destinazione.

Quando si sostituisce utilizzando sinonimi ed elementi facoltativi si riduce il numero di concetti nel pannello Risultati di estrazione e si combinano insieme in concetti più significativi e più rappresentativi con maggiore conteggio Doc. di frequenza.

Nota: per risorse in giapponese, gli elementi facoltativi non si applicano e non sono disponibili. Inoltre, i sinonimi vengono gestiti in modo leggermente diverso per il testo giapponese.

Sinonimi

I sinonimi associano due o più parole che hanno lo stesso significato. I sinonimi sono spesso anche utilizzati per raggruppare i termini con le loro abbreviazioni o per raggruppare parole comunemente errate con l'ortografia corretta. È possibile definire questi sinonimi sulla scheda Sinonimi.

Una definizione di sinonimo è composta da due parti. La prima è un termine **di destinazione**, che è il termine sotto cui si desidera che il motore di estrazione raggruppi tutti i termini sinonimo. A meno che questo termine di destinazione non viene utilizzato come sinonimo di un altro termine di destinazione o è escluso, è probabile che diventi il concetto che viene visualizzato nel riquadro Risultati di estrazione. Il secondo è l'elenco dei sinonimi che saranno raggruppati sotto il termine di destinazione.

Ad esempio, se si desidera che `automobile` sia sostituito da `veicolo` `automobile` è il sinonimo e `veicolo` è il termine di destinazione.

È possibile immettere qualsiasi parola nella colonna **Sinonimo**, ma se la parola non viene trovata durante l'estrazione e il termine ha un'opzione di corrispondenza con l'opzione `Intero`, non può avvenire alcuna sostituzione. Tuttavia, il termine di destinazione non deve essere estratto per i sinonimi da raggruppare sotto questo termine.

Elementi facoltativi

Gli elementi facoltativi identificano le parole facoltative in un termine composto che possono essere ignorate durante l'estrazione in modo da mantenere termini simili, anche se appaiono leggermente diversi nel testo. Gli elementi facoltativi sono parole singole che, se rimosse da un composto, potrebbero creare una corrispondenza con un altro termine. Tali singole parole possono apparire dovunque all'interno del composto--all'inizio, nel mezzo o alla fine. È possibile definire gli elementi facoltativi sulla scheda Facoltativo.

Ad esempio per raggruppare insieme i termini `ibm` e `ibm corp` bisogna dichiarare che `corp` deve essere trattato, in questo caso, come elemento facoltativo. In un altro esempio, se si designa il termine `accesso` come un elemento facoltativo e durante l'estrazione vengono trovati `velocità di accesso a internet` e `velocità internet`, essi verranno raggruppati insieme sotto il termine che ricorre più frequentemente.

Nota: per le risorse di testo in giapponese, non esiste alcuna scheda Elementi facoltativi poiché gli elementi facoltativi non si applicano.

Definizione dei sinonimi

Nella scheda Sinonimi, è possibile immettere una definizione di sinonimo nella riga vuota nella parte superiore della tabella. Iniziare definendo il termine di destinazione e i suoi sinonimi. È inoltre possibile selezionare la libreria in cui si desidera memorizzare questa definizione. Durante l'estrazione, tutte le ricorrenze dei sinonimi verranno raggruppate sotto il termine di destinazione nell'estrazione finale. Consultare la sezione “Aggiunta di termini” a pagina 192 per ulteriori informazioni.

Ad esempio, se i dati di testo includono un sacco di informazioni sulle telecomunicazioni, è possibile avere questi termini: `cellulare`, `wireless`, e `portatile`. In questo esempio, si potrebbe voler definire `cellulare` e `portatile` come sinonimi di `wireless`. Se si definiscono tali sinonimi, ogni ricorrenza estratta di `cellulare` e `portatile` verranno trattati come `wireless` e verranno visualizzati insieme nell'elenco di termini.

Quando si creano i propri dizionari di tipo, è possibile immettere un termine e anche avere in mente tre o quattro sinonimi per questo termine. In tal caso, è possibile immettere tutti i termini e quindi il proprio termine di destinazione nel dizionario di sostituzione e quindi trascinare i sinonimi.

Nota: i sinonimi sono gestiti diversamente nel testo giapponese.

La sostituzione si sinonimo viene applicata anche ai moduli con inflessione (ad esempio la forma plurale) del sinonimo. A seconda del contesto, è possibile imporre limiti su come vengono sostituiti i termini. Alcuni caratteri possono essere utilizzati per inserire limiti sull'elaborazione:

- **Punto esclamativo (!).** Quando il punto esclamativo precede direttamente il sinonimo `!sinonimo` si indica che nessuna forma di inflessione del sinonimo verrà sostituita dal termine di destinazione. Tuttavia, un punto esclamativo che precede direttamente il termine di destinazione `!termine di destinazione` indica che non si desidera che alcuna parte del composto o varianti ricevano qualsiasi ulteriore sostituzione.
- **Asterisco (*).** Un asterisco posto direttamente dopo un sinonimo, ad esempio `sinonimo*`, significa che si desidera che questa parola sia sostituita dal termine di destinazione. Ad esempio, se viene definito `gestisci*` come sinonimo e `gestione` come destinazione, `gestori associati` verrà sostituito dal termine di destinazione `gestione di associati`. È inoltre possibile aggiungere uno spazio e un asterisco dopo la parola (`sinonimo *`) come `internet *`. Se è stata definita la destinazione come `internet` e i sinonimi come `internet * *` e `web *`, `scheda di accesso a internet` e `portale web` sarà sostituito con `internet`. In questo dizionario non è possibile iniziare una parola o una stringa con il carattere jolly asterisco.
- **Accento circonflesso (^).** Un accento circonflesso e uno spazio che precedono il sinonimo, come `^ sinonimo`, indica che il raggruppamento del sinonimo si applica solo quando il termine inizia con il sinonimo. Ad esempio, se si definisce `^ stipendio` come sinonimo e `reddito` come destinazione e entrambi i termini vengono estratti, essi saranno raggruppati sotto il termine `reddito`. Tuttavia, se vengono estratti `paga base` e `reddito`, essi non saranno raggruppati insieme, poiché `paga base` non inizia `stipendio`. Uno spazio deve essere posto tra questo simbolo e il sinonimo.
- **Simbolo del dollaro ($).** Uno spazio e un simbolo del dollaro che seguono il sinonimo, come `sinonimo $` indicano che il raggruppamento del sinonimo si applica solo quando il termine finisce con il sinonimo. Ad esempio, se si definisce `contanti $` come sinonimo e `soldi` come destinazione e entrambi i termini vengono estratti, essi saranno raggruppati insieme sotto il termine `soldi`. Tuttavia, se vengono estratti `vacca da mungere` e `soldi`, essi non saranno raggruppati insieme, poiché `vacca da mungere` non termina con `contanti`. Uno spazio deve essere posto tra questo simbolo e il sinonimo.
- **Accento circonflesso (^) e simbolo del dollaro ($).** Se il simbolo dell'accento circonflesso e dollaro vengono utilizzati insieme, ad esempio `^ sinonimo $`, un termine corrisponde al sinonimo solo se è una corrispondenza esatta. Ciò significa che nessuna parola può apparire prima o dopo il sinonimo nel termine estratto affinché abbia luogo il raggruppamento di sinonimi. Ad esempio, si potrebbe voler definire `^ van $` come sinonimo e `camion` come destinazione in modo che solo `van` viene raggruppato con `camion`, mentre `maria van guerin` verrà lasciato invariato. Inoltre, ogni volta che viene definito un sinonimo utilizzando il simbolo dell'accento circonflesso e il dollaro e questa parola viene visualizzata in qualsiasi punto del testo di origine, il sinonimo viene estratto automaticamente.

Nota: questi caratteri speciali e i caratteri jolly non sono supportati per il testo giapponese.

Per aggiungere una voce di sinonimo

1. Con il pannello di sostituzione visualizzato, fare clic sulla scheda **Sinonimi** nell'angolo inferiore sinistro.
2. Nella riga vuota nella parte superiore della tabella, immettere il termine di destinazione nella colonna Destinazione. Il termine di destinazione immesso viene visualizzato a colori. Questo colore rappresenta il tipo in cui il termine viene visualizzato o forzato, se questo è il caso. Se il termine appare in nero, questo significa che non compare in alcun dizionario di tipo.
3. Fare clic nella seconda cella a destra della destinazione e immettere la serie di sinonimi. Separare ciascuna voce utilizzando il delimitatore globale come definito nella finestra Opzioni. Per ulteriori informazioni, consultare “Impostazione delle opzioni” a pagina 82. I termini immessi vengono visualizzati a colori. Questo colore rappresenta il tipo in cui il termine appare. Se il termine appare in nero, questo significa che non compare in alcun dizionario di tipo.
4. Fare clic nell'ultima cella per selezionare la libreria in cui si desidera memorizzare la definizione di sinonimo.

Nota: queste istruzioni mostrano come eseguire le modifiche all'interno della Editor risorsevistao Editor di modelli. Tenere presente che è anche possibile effettuare questo tipo di ottimizzazione direttamente dal pannello Risultati di estrazione , Pannello dati, riquadro Categorie o dalla finestra di dialogo Definizioni cluster nelle altre viste. Per ulteriori informazioni, consultare la sezione "Perfezionamento dei risultati di estrazione" a pagina 95.

Definizione degli elementi facoltativi

Nella scheda Facoltativo, è possibile definire gli elementi facoltativi per qualsiasi libreria desiderata. Queste voci sono raggruppate insieme per ogni libreria. Non appena una libreria viene aggiunta al riquadro della struttura ad albero della libreria, un riga vuota di elemento facoltativo viene aggiunta alla scheda Facoltativo.

Tutte le voci vengono trasformate in parole minuscole automaticamente. Il motore di estrazione metterà in corrispondenza le voci con entrambe le parole in lettere minuscole e maiuscole nel testo.

Nota: per risorse in giapponese, gli elementi facoltativi non si applicano e non sono disponibili.

Nota: i termini vengono delimitati tramite il delimitatore definito nella finestra di dialogo Opzioni. Per ulteriori informazioni, consultare la sezione "Impostazione delle opzioni" a pagina 82. Se l'elemento facoltativo che si sta immettendo include lo stesso delimitatore come parte del termine, esso deve essere preceduto da una barra retroversa.

Per aggiungere una voce

1. Con il pannello di sostituzione visualizzato, fare clic sulla scheda Facoltativo nell'angolo inferiore sinistro.
2. Fare clic sulla cella nella colonna Elementi facoltativi per la libreria a cui si desidera aggiungere questa voce.
3. Immettere l'elemento facoltativo. Separare ciascuna voce utilizzando il delimitatore globale come definito nella finestra Opzioni. Per ulteriori informazioni, consultare la sezione "Impostazione delle opzioni" a pagina 82.

Disattivazione ed eliminazione delle sostituzioni

È possibile rimuovere una voce in modo temporaneo disabilitandola nel dizionario. Disabilitando una voce, la voce verrà ignorata durante l'estrazione.

È inoltre possibile eliminare tutte le voci obsolete nel dizionario di sostituzione.

Per disabilitare una voce

1. Nel dizionario, selezionare la voce da disabilitare.
2. Fare clic sulla barra spaziatrice. La casella di spunta alla sinistra del nome è deselezionata.

Nota: è anche possibile deselezionare la casella di spunta alla sinistra della voce per disabilitarla.

Per eliminare una voce di sinonimi

1. Nel dizionario, selezionare la voce da eliminare.
2. Dai menu, scegliere **Modifica > Elimina** o premere il tasto **Canc** sulla tastiera. La voce non è più nel dizionario.

Per eliminare una voce di elemento facoltativa

1. Nel dizionario, fare doppio clic sulla voce da eliminare.
2. Eliminare manualmente il termine.
3. Premere Invio per applicare la modifica.

Dizionari di esclusione

Un **dizionario di esclusione** è un elenco di parole, frasi o stringhe parziali. I termini corrispondenti o contenenti una voce nel dizionario di esclusione verranno ignorati o esclusi dall'estrazione. I dizionari di esclusione vengono gestiti nel riquadro a destra dell'editor. Di solito, i termini da aggiungere a questo elenco sono parole o frasi utilizzate per inserire nel testo la continuità ma che non aggiungono nulla di importante e possono intasare i risultati di estrazione. Aggiungendo questi termini al dizionario di esclusione, si è assicurato che non verranno mai estratti.

I dizionari di esclusione vengono gestiti nel riquadro in alto a destra della scheda Risorse di libreria nell'editor. È possibile accedere a questa vista con **Vista > Editor delle risorse** nel menu , se ci si trova in una sessione di postazione interattiva. In caso contrario, è possibile revisionare i dizionari per un modello specifico in Editor di modelli.

Nel dizionario di esclusione, è possibile immettere una parola, frase o stringa parziale nella riga vuota nella parte superiore della tabella. È possibile aggiungere stringhe di caratteri al dizionario di esclusione come una o più parole o addirittura parole parziali utilizzando l'asterisco come carattere jolly. Le voci dichiarate nel dizionario di esclusione verranno utilizzate per impedire l'estrazione di concetti. Se una voce viene inoltre dichiarata altrove nell'interfaccia, come in un dizionario di tipo, essa viene visualizzata con una barra negli altri dizionari, indicando che attualmente è esclusa. Questa stringa non deve apparire nei dati di testo o essere dichiarata come parte di ogni dizionario di tipo da applicare.

Nota: se si aggiunge un concetto al dizionario di esclusione che agisce anche come destinazione in una voce di sinonimo e quindi anche la destinazione e tutti i suoi sinonimi saranno esclusi. Consultare la sezione "Definizione dei sinonimi" a pagina 197 per ulteriori informazioni.

Uso dei caratteri jolly (*)

Per tutte le lingue a parte il giapponese, è possibile possibile utilizzare il carattere jolly asterisco per indicare che si desidera che la voce di esclusione sia trattata come una stringa parziale. Eventuali termini rilevati dal motore di estrazione che contengono una parola che inizia o termina con una stringa immessa nel dizionario di esclusione saranno esclusi dall'estrazione finale. Tuttavia, esistono due casi dove l'utilizzo di caratteri jolly non è consentito:

- Il carattere trattino (-) preceduto da un asterisco ad esempio *-
- L'apostrofo (') preceduto da un asterisco ad esempio *'*s*

Tabella 39. Esempi di voci di esclusione.

Inserimento	Esempio	Risultati
parola	*avanti*	Nessun concetto (o termine) verrà estratto se contiene la parola *avanti*.
frase	*ad esempio*	Nessun concetto (o termine) verrà estratto se contiene la frase *ad esempio*.
parziale	*copyright**	Si escludono tutti i concetti (o termini) corrispondenti o contenenti le varianti della parola *copyright*, ad esempio `note di copyright`, `norme di copyright`, `informazioni di copyright`, o `copyright 2010`.
parziale	**ware*	Si escludono tutti i concetti (o termini) corrispondenti o contenenti le varianti della parola *ware*, come `freeware`, `shareware`, `software`, `hardware`, `firmware` o `silverware`.

Per aggiungere voci

1. Nella riga vuota nella parte superiore della tabella, immettere un termine. Il termine immesso viene visualizzato a colori. Questo colore rappresenta il tipo in cui il termine appare. Se il termine appare in nero, questo significa che non compare in alcun dizionario di tipo.

Per disabilitare le voci

È possibile rimuovere temporaneamente una voce disabilitandola nel dizionario di esclusione. Disabilitando una voce, la voce verrà ignorata durante l'estrazione.

1. Nel dizionario, selezionare la voce da disabilitare.
2. Fare clic sulla barra spaziatrice. La casella di spunta alla sinistra del nome è deselezionata.

Nota: è anche possibile deselezionare la casella di spunta alla sinistra della voce per disabilitarla.

Per eliminare le voci

È possibile eliminare qualsiasi voce non necessaria nel dizionario di esclusione.

1. Nel dizionario di esclusione, selezionare la voce da eliminare.
2. Dal menu scegliere **Modifica > Elimina**. La voce non è più nel dizionario.

Capitolo 18. Informazioni su Risorse avanzate

Oltre al tipo, i dizionari di esclusione e sostituzione, è possibile anche gestire una varietà di impostazioni di risorsa avanzate come le impostazioni Raggruppamento confuso o le definizioni di tipo non linguistiche. È possibile gestire queste risorse nella scheda Risorse avanzate in Editor di modelli o Editor risorse.

Importante! Questa scheda non è disponibile per le risorse regolate per il testo giapponese.

Quando si passa alla scheda Risorse Avanzate, è possibile modificare le seguenti informazioni:

- **Lingua di destinazione per le risorse.** Utilizzata per selezionare la lingua per cui le risorse verranno create e ottimizzate. Consultare la sezione “Lingua di destinazione per le risorse” a pagina 205 per ulteriori informazioni.
- **Raggruppamento confuso (Eccezioni).** Utilizzata per escludere le coppie di parole dall'algoritmo di raggruppamento confuso (correzione di errori di ortografia). Consultare la sezione “Raggruppamento confuso” a pagina 206 per ulteriori informazioni.
- **Entità non linguistiche.** Utilizzato per abilitare e disabilitare quali entità non linguistiche possono essere estratte, così come le espressioni regolari e le regole di normalizzazione che vengono applicate durante la loro estrazione. Consultare la sezione “Entità non linguistiche” a pagina 207 per ulteriori informazioni.
- **Gestione lingua.** Utilizzato per dichiarare i modi speciali di strutturare le frasi (modelli di estrazione e definizioni forzate) e per utilizzare le abbreviazioni per la lingua selezionata. Consultare la sezione “Gestione lingua” a pagina 211 per ulteriori informazioni.
- **Identificativo lingua.** Utilizzato per configurare l'identificativo lingua automatico richiamato quando la lingua è impostata su **Tutti**. Consultare la sezione “Identificativo di lingua” a pagina 213 per ulteriori informazioni.

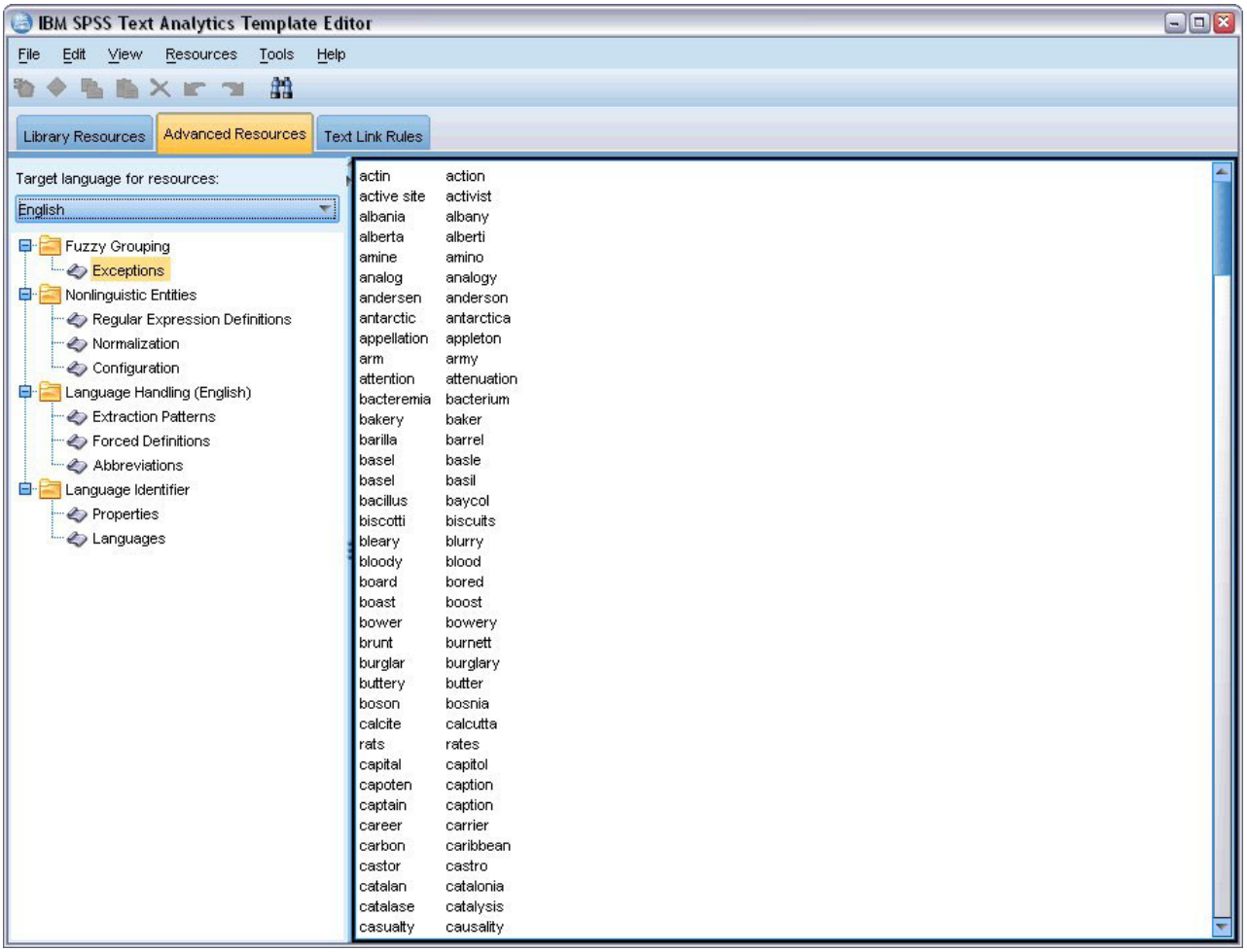


Figura 41. Editor di modello di estrazione testo - scheda Risorse avanzate

Nota: è possibile utilizzare la barra degli strumenti Trova/Sostituisci per ricercare rapidamente le informazioni o per apportare modifiche uniformi ad una sezione. Consultare la sezione "Sostituzione" a pagina 205 per ulteriori informazioni.

Per modificare le risorse avanzate

1. Individuare e selezionare la sezione di risorsa che si desidera modificare. Il contenuto viene visualizzato nel riquadro di destra.
2. Utilizzare il menu o i pulsanti della barra degli strumenti per tagliare, copiare o incollare il contenuto, se necessario.
3. Modificare i file che si desidera modificare utilizzando le regole di formattazione in questa sezione. Le modifiche vengono salvate non appena fatte. Utilizzare le frecce di annullamento o di riesecuzione sulla barra degli strumenti per ritornare alle modifiche precedenti.

Trova

In alcuni casi, potrebbe essere necessario individuare rapidamente le informazioni in una sezione particolare. Ad esempio, se si esegue l'analisi di collegamento del testo, è possibile avere centinaia di macro e definizioni di modello. Utilizzando la funzione Trova, è possibile trovare una regola specifica rapidamente. Per ricercare informazioni in una sezione, è possibile utilizzare la barra degli strumenti Trova.

Per utilizzare la funzione Trova

1. Individuare e selezionare la sezione di risorsa che si desidera ricercare. Il contenuto viene visualizzato nel riquadro di destra dell'editor.
2. Dai menu scegliere **Modifica > Trova**. La barra degli strumenti Trova viene visualizzata nella parte superiore destra della finestra di dialogo Modifica risorse avanzate.
3. Immettere la stringa di parole che si desidera ricercare nella casella di testo. È possibile utilizzare il pulsante della barra degli strumenti per controllare la sensibilità al maiuscolo / minuscolo, la corrispondenza parziale e la direzione della ricerca.
4. Fare clic su **Trova** per avviare la ricerca. Se viene individuata una corrispondenza, il testo viene evidenziato nella finestra.
5. Fare clic su **Trova** di nuovo per ricercare la corrispondenza successiva.

Nota: quando si lavora nella scheda Regole di collegamento del testo l'opzione Trova è disponibile solo quando si visualizza il codice di origine.

Sostituzione

In alcuni casi, potrebbe essere necessario apportare aggiornamenti più estesi alle proprie risorse avanzate. La funzione Sostituisci consente di effettuare aggiornamenti uniformi al proprio contenuto.

Per utilizzare la funzione Sostituisci

1. Individuare e selezionare la sezione di risorsa che si desidera ricercare e sostituire. Il contenuto viene visualizzato nel riquadro di destra dell'editor.
2. Dai menu scegliere **Modifica > Sostituisci**. Si apre la finestra di dialogo Sostituisci.
3. Nel casella di testo **Trova**, immettere la stringa di parole che si desidera ricercare.
4. Nella casella di testo **Sostituisci con**, immettere la stringa che si desidera utilizzare al posto del testo che è stato trovato.
5. Selezionare **Solo parole intere** se si desidera trovare o sostituire solo parole complete.
6. Selezionare **Maiuscole / minuscole** se si desidera trovare o sostituire solo le parole che corrispondono al caso.
7. Fare clic su **Trova successivo** per trovare una corrispondenza. Se viene individuata una corrispondenza, il testo viene evidenziato nella finestra. Se non si desidera sostituire questa corrispondenza, fare clic di nuovo su **Trova successivo** fino a che non si trova una corrispondenza che si desidera sostituire.
8. Fare clic su **Sostituisci** per sostituire la corrispondenza selezionata.
9. Fare clic su **Sostituisci** per sostituire tutte le corrispondenze nella sezione. Si apre un messaggio con il numero di sostituzioni effettuate.
10. Una volta terminato di apportare sostituzioni, fare clic su **Chiudi**. La casella di dialogo viene chiusa.

Nota: se viene commesso un errore di sostituzione, è possibile annullare la sostituzione chiudendo la finestra di dialogo e scegliendo **Modifica > Annulla** dal menu. È necessario eseguire questa operazione una volta per ogni modifica che si desidera annullare.

Lingua di destinazione per le risorse

Le risorse vengono create per una particolare lingua di testo. La lingua per la quale queste risorse vengono ottimizzate è definita nella scheda Avanzate. È possibile passare ad un'altra lingua, se necessario, selezionando tale lingua nella casella **Lingua di destinazione per le risorse**. Inoltre, la lingua riportata viene visualizzata come lingua per i pacchetti di analisi del testo creati con queste risorse.

Importante! Raramente sarà necessario modificare la lingua nelle proprie risorse. Se succede è possibile che si verifichino problemi quando le risorse non corrispondono più alla lingua di estrazione. Sebbene impiegato di rado, è possibile modificare una lingua se si prevede di utilizzare l'opzione di lingua `TUTTI`

durante l'estrazione perché si aspetta di avere testo in più di una lingua. Modificando la lingua, è possibile accedere, ad esempio, alle risorse di gestione di lingua per i modelli di estrazione, abbreviazioni e definizioni forzate per la lingua secondaria a cui si è interessati. Tuttavia, tenere presente che prima di pubblicare o salvare le modifiche apportate alle risorse o eseguire un'altra estrazione, impostare la lingua di nuovo su quella primaria a cui si è interessati per l'estrazione.

Raggruppamento confuso

Nel nodo di estrazione testo e nelle impostazioni di estrazione, se si seleziona **Correggi ortografia per limite minimo di carattere principale**, è stata abilitato l'algoritmo di raggruppamento confuso.

Il raggruppamento confuso consente di raggruppare le parole che hanno errori comuni di ortografia o parole quasi esatte rimuovendo tutte le vocali temporaneamente (ad eccezione della prima) e consonanti doppie o triple dalle parole estratte e quindi confrontandole per vedere se sono uguali. Durante il processo di estrazione, la funzione di raggruppamento confuso viene applicata ai termini estratti e i risultati vengono confrontati per determinare se vengono trovate corrispondenze. In tal caso, i termini originali vengono raggruppati insieme nell'elenco di estrazione finale. Essi sono raggruppati sotto il termine che ricorre più frequentemente nei dati.

Nota: se i due termini confrontati sono assegnati a tipi differenti, escluso il tipo `<Sconosciuto>`, quindi la tecnica di raggruppamento confuso non viene applicata a questa coppia. In altre parole, i termini devono appartenere allo stesso tipo o al tipo `<Sconosciuto>` per applicare la tecnica.

Se è stata abilitata questa funzione e sono state trovate due parole con ortografia simile che sono state correttamente raggruppate insieme, è possibile escluderle dal raggruppamento confuso. È possibile effettuare questa operazione immettendo le coppie in corrispondenza nella sezione Eccezioni nella scheda Risorse avanzate. Consultare la sezione Capitolo 18, “Informazioni su Risorse avanzate”, a pagina 203 per ulteriori informazioni.

L'esempio che segue dimostra il modo in cui viene eseguito il raggruppamento confuso. Se il raggruppamento confuso è abilitato, queste parole sembrano essere le stesse e sono associate nel modo seguente:

```
colore  -> colr                montagna -> montn
colore -> colr                montana  -> montn

modellazione -> modlng           mobili -> mbl
modellazione -> modlng           mobili -> mobl
```

Nell'esempio precedente, è più probabile che si desidera escludere `montagna` e `montana` dall'essere raggruppati insieme. Pertanto, è possibile immettere le parole nella sezione Eccezioni nel modo seguente:

```
montagna      montana
```

Importante! In alcuni casi, le eccezioni di raggruppamento confuso non impediscono a 2 parole di essere associate perché vengono applicate certe regole di sinonimi. In tal caso, è possibile provare a immettere sinonimi utilizzando il carattere jolly punto esclamativo (`!`) per vietare alle parole di diventare sinonimo nell'output. Consultare la sezione “Definizione dei sinonimi” a pagina 197 per ulteriori informazioni.

Regole per la formattazione delle eccezioni di raggruppamento confuso

- Definire solo una coppia di eccezione per riga.
- Utilizzare le parole semplici o composte.
- Utilizzare solo caratteri minuscoli per le parole. Le parole maiuscole verranno ignorate.
- Utilizzare un carattere `TAB` per separare ciascuna parola di una coppia.

Entità non linguistiche

Quando si lavora con certi tipi di dati, potrebbe risultare molto interessante estrarre date, numeri di codice fiscale, percentuali o altre entità non linguistiche. Queste entità sono dichiarate esplicitamente nel file di configurazione, in cui è possibile abilitare o disabilitare le entità. Consultare la sezione “Configurazione” a pagina 210 per ulteriori informazioni. Al fine di ottimizzare l'output dal motore di estrazione, l'input dall'elaborazione non linguistica è normalizzato per il gruppo come entità in base a formati predefiniti. Consultare la sezione “Normalizzazione” a pagina 210 per ulteriori informazioni.

Nota: è possibile attivare o disattivare l'estrazione di entità non linguistiche nelle impostazioni di estrazione.

Entità non linguistiche disponibili

Le entità non linguistiche nella seguente tabella possono essere estratte. Il nome tipo è in parentesi.

Tabella 40. Entità non linguistiche che possono essere estratte

Indirizzi	(<Address>)
Amminoacidi	(<Aminoacid>)
Valute	(<Currency>)
Date	(<Date>)
Ritardo	(<Delay>)
Cifre	(<Digit>)
Indirizzi e-mail	(<email>)
Indirizzi HTTP/URL	(<url>)
Indirizzo IP	(<IP>
Organizzazioni	(<Organization>)
Percentuali	(<Percent>)
Prodotti	(<Product>)
Proteine	(<Gene>)
Numeri di telefono	(<PhoneNumber>)
Ore	(<Time>)
Codice fiscale	(<SocialSecurityNumber>)
Pesi e misure	(<Weights-Measures>)

Ripulitura del testo per l'elaborazione

Prima che si verifichi l'estrazione di entità non linguistiche, il testo viene ripulito. Durante questa fase, le seguenti modifiche temporanee vengono effettuate in modo che le entità non linguistiche possono essere identificate ed estratte in questo modo:

- Qualsiasi sequenza di due o più spazi viene sostituita da un singolo spazio.
- Le tabulazioni sono sostituite da uno spazio.
- I caratteri o le sequenze di caratteri di fine riga vengono sostituiti da uno spazio, mentre le sequenze di fine riga multiple sono contrassegnate come fine di un paragrafo. La fine della riga può essere indicata da ritorni a capo (CR) e nuova riga (LF) o anche entrambi.
- Le tag HTML e XML vengono temporaneamente tolte e ignorate.

Definizioni delle espressioni regolari

Quando si estraggono entità non linguistiche, si potrebbe desiderare di modificare o aggiungere le definizioni di espressione regolare che vengono utilizzate per identificare le espressioni regolari. Tale operazione viene eseguita nella sezione **Definizioni di espressione regolare** nella scheda Risorse avanzate. Consultare la sezione Capitolo 18, “Informazioni su Risorse avanzate”, a pagina 203 per ulteriori informazioni.

Il file viene suddiviso in sezioni distinte. La prima sezione si chiama `[macros]`. Oltre a quella sezione, può sussistere una sezione aggiuntiva per ciascuna entità non linguistica. È possibile aggiungere sezioni a questo file. All'interno di ciascuna sezione, le regole sono numerate (*regexp1*, *regexp2* e così via). Queste regole devono essere numerate in sequenza da 1-*n*. Qualsiasi interruzione nella numerazione provocherà la sospensione dell'elaborazione di questo file.

In alcuni casi, un'entità dipende dalla lingua. Un'entità è considerata dipendente dalla lingua se prende un valore diverso da 0 per il parametro della lingua nel file di configurazione. Consultare la sezione “Configurazione” a pagina 210 per ulteriori informazioni. Quando un'entità è dipendente dalla lingua, la lingua deve essere utilizzata per inserire un prefisso nel nome sezione, ad esempio `[english/PhoneNumber]`. Questa sezione conterrà le regole che si applicano solo a numeri di telefono inglesi quando all'entità `PhoneNumber` viene assegnato un valore 2 per la lingua.

Importante! Se si eseguono delle modifiche a questo file o a qualsiasi altro nell'editor e il motore di estrazione non funziona come desiderato, utilizzare l'opzione **Reimposta su Originale** sulla barra degli strumenti per reimpostare il file con il contenuto originale fornito. Questo file richiede un determinato livello di familiarità con espressioni regolari. Se è necessaria ulteriore assistenza in questo settore, contattare IBM Corp. per la guida.

Caratteri speciali. [] {} () \ * + ? | ^ $

Tutti i caratteri corrispondono a se stessi tranne che per i seguenti caratteri speciali, che sono utilizzati per scopi specifici nelle espressioni: `.[{()\*+?|^$` Per utilizzare questi caratteri in questo modo, essi devono essere preceduti da una barra retroversa (\) nella definizione.

Ad esempio, se si stava tentando di estrarre gli indirizzi Web, il carattere completo stop è molto importante per l'entità; quindi è necessario barra retroversa, come ad esempio:

```
www\.[a-z]+\.[a-z]+
```

Operatori di ripetizione e quantificatori ? + * {}

Per abilitare le definizioni in modo da essere più flessibili, è possibile utilizzare caratteri jolly più volte; essi sono standard per le espressioni regolari. Essi sono `*? +`

- *L'asterisco* * indica che vi sono *zero o più* della precedente stringa. Ad esempio, `ab*c` corrisponde a "*ac*", "*abc*", "*abbbc*", e così via.
- *Il segno più* + indica che vi sono *uno o più* della precedente stringa. Ad esempio, `ab+c` corrisponde a "*abc*", "*abbc*", "*abbbc*" ma non a "*ac*".
- Il *punto interrogativo ?* indica che c'è *zero o uno* della precedente stringa. Ad esempio, `modell?azione` corrisponde sia a "*modella*" che a "*modellazione*".
- *Limitazione ripetizione con parentesi {}* indica i limiti della ripetizione. Per esempio,

  `[0-9]{n}` corrisponde ad una cifra ripetuta esattamente *n* volte. Ad esempio, `[0-9]{4}` corrisponde a "*1998*" ma non a "*33*" né a "*19983*".

  `[0-9]{n,}` corrisponde ad una cifra ripetuta esattamente *n o più* volte. Ad esempio, `[0-9]{3,}` corrisponde a "*199*" o a "*1998*", ma non a "*19*".

  `[0-9]{n,m}` corrisponde ad una cifra ripetuta tra *n ed m volte incluso*. Ad esempio, `[0-9]{3,5}` corrisponde a “*199*”, “*1998*” o “*19983*” ma non a “*19*” né a “*199835*”.

Spazi e trattini facoltativi

In alcuni casi, si desidera includere uno spazio facoltativo in una definizione. Ad esempio se si desidera estrarre valute come "*peso uruguaiano*", "*peso uruguaiano*", "*peso dell'uruguay*", "*peso di uruguay*", "*peso*" o "*peso*", sarà necessario affrontare il fatto che ci possono essere due parole separate da uno spazio. In questo caso, questa definizione deve essere scritta come `(uruguaiano |uruguay )?peso?`. Poiché *uruguaiano* o *uruguay* sono seguiti da uno spazio quando utilizzati con *peso/peso*, lo spazio facoltativo deve essere definito all'interno della sequenza `(uruguaiano |uruguay )` facoltativa. Se lo spazio non è nella sequenza facoltativa come `(uruguaiano|uruguay)? peso?`, ma non corrisponde a "*peso*" o "*peso*" poiché lo spazio sarebbe necessario.

Se si sta cercando una serie di cose tra cui un carattere trattino (-) in un elenco allora il trattino deve essere definito per ultimo. Ad esempio, f si sta cercando una virgola (,) o un trattino (-), utilizzare `[,-]` e non `[-,]` mai.

Ordine di stringhe in elenchi e macro

È opportuno definire sempre la più lunga sequenza prima di una più breve oppure la più lunga non verrà mai letta dal momento che la corrispondenza avverrà in quella più breve. Ad esempio, se si stavano cercando le stringhe "*miliardi*" o "*miglia*", quindi "*miliardi*" deve essere definito prima di "*miglia*". Quindi, ad esempio `(miliardo|miglia)` e non `(miglia|miliardo)`. Ciò vale anche per le macro, perché le macro sono elenchi di stringhe.

Ordine di regole nella sezione Definizione

Definire una regola per riga. All'interno di ciascuna sezione, le regole sono numerate (*regexp1*, *regexp2* e così via). Queste regole devono essere numerate in sequenza da 1-*n*. Qualsiasi interruzione nella numerazione provocherà la sospensione dell'elaborazione di questo file. Per disabilitare una voce, posizionare il simbolo # all'inizio di ogni riga utilizzata per definire l'espressione regolare. Per abilitare una voce, rimuovere il carattere # prima della riga.

In ogni sezione, le regole più specifiche devono essere definite prima di quelle più generiche al fine di garantire una corretta elaborazione. Ad esempio, se si stava cercando una data nel formato "*anno mese*" e nel formato "*mese*", il "*mese dell'anno*" la regola deve essere definita prima della regola "*mese*". Di seguito è riportato il modo in cui deve essere definito:

```
#@# gennaio 1932
regexp1=$(MONTH),? [0-9]{4}

#@# gennaio
regexp2=$(MONTH)
```

e non

```
#@# gennaio
regexp1=$(MONTH)

#@# gennaio 1932
regexp2=$(MONTH),? [0-9]{4}
```

Uso di macro nelle regole

Ogni volta che una sequenza specifica viene utilizzato in diverse regole, è possibile utilizzare una macro. Quindi, se è necessario modificare la definizione di questa sequenza, sarà necessario modificarla solo una volta e non in tutte le regole che fanno riferimento ad esso. Per esempio, si supponga di eseguire la seguente macro:

```
MONTH=((gennaio|febbraio|marzo|aprile|giugno|luglio|agosto|settembre|ottobre|
novembre|dicembre)|(gen|feb|mar|apr|mag|giu|lug|ago|set|ott|nov|dic)(\.)?)
```

Ogni volta che si fa riferimento al nome della macro, deve essere racchiuso in $(), ad esempio: regexp1=$(MONTH)

Tutte le macro devono essere definite nella sezione [macros].

Normalizzazione

Quando si estraggono entità non linguistiche, le entità rilevate vengono normalizzate in gruppo come entità in base a formati predefiniti. Ad esempio, i simboli di valuta e relativi equivalenti in termini vengono trattati allo stesso modo. Le voci di normalizzazione vengono memorizzate nella sezione **Normalizzazione** nella scheda Risorse avanzate. Consultare la sezione Capitolo 18, “Informazioni su Risorse avanzate”, a pagina 203 per ulteriori informazioni. Il file viene suddiviso in sezioni distinte.

Importante! Questo file è solo per utenti avanzati. E' altamente improbabile che sarà necessario modificare questo file. Se è necessaria ulteriore assistenza in questo settore, contattare IBM Corp. per la guida.

Regole di formattazione per la normalizzazione

- Aggiungere solo una voce di normalizzazione per riga.
- Rispettare strettamente le sezioni di questo file. Non è possibile aggiungere nuove sezioni.
- Per disabilitare una voce, inserire il simbolo # all'inizio della riga. Per abilitare una voce, rimuovere il carattere # prima della riga.

Date in inglese nella normalizzazione

Per impostazione predefinita le date nel modello inglese sono riconosciute nel formato data in stile americano; cioè: mese, giorno, anno. Per modificare nel formato giorno, mese, anno, disabilitare la riga "format:US" (aggiungendo # all'inizio della riga) e abilitare "format:UK" (rimuovendo il segno # da tale riga).

Configurazione

È possibile abilitare e disabilitare i tipi di entità non linguistiche che si desidera estrarre nel file di configurazione delle entità non linguistiche. Disabilitando le entità non necessarie, è possibile diminuire il tempo di elaborazione richiesto. Tale operazione viene eseguita nella sezione **Configurazione** nella scheda Risorse avanzate. Consultare la sezione Capitolo 18, “Informazioni su Risorse avanzate”, a pagina 203 per ulteriori informazioni. Se l'estrazione non linguistica è abilitata, il motore di estrazione legge questo file di configurazione durante il processo di estrazione per determinare quali tipi di entità non linguistiche devono essere estratti.

La sintassi per questo file è la seguente:

#nome<TAB>*Lingua*<TAB>*Codice*

Tabella 41. La sintassi per il file di configurazione.

Etichetta di colonna	Descrizione
#nome	La formulazione tramite cui le entità non linguistiche verranno indicate nei due altri file richiesti per l'estrazione delle entità non linguistiche. I nomi utilizzati qui sono sensibili al maiuscolo/minuscolo.
Lingua	La lingua dei documenti . Si consiglia di selezionare la lingua specifica; tuttavia, esiste un'opzione **Qualsiasi**. Le opzioni possibili sono: 0 = una qualsiasi che viene utilizzata ogni volta che un regexp non è specifico di una lingua e può essere utilizzato in diversi modelli con diverse lingue, ad esempio un indirizzo IP/URL/email; 1 = francese 2 = inglese; 4 = tedesco; 5 = spagnolo; 6 = olandese; 8 = portoghese/brasiliano 10 = italiano.

Tabella 41. La sintassi per il file di configurazione (Continua).

Etichetta di colonna	Descrizione
`Codice`	Codice parte del discorso. La maggior parte delle entità contengono un valore "s" tranne in alcuni casi. I valori possibili sono: `s` = parola; `a` = aggettivo; `n` = nome. Se abilitato, le entità non linguistiche vengono prima estratte e i modelli di estrazione vengono applicati per identificare il proprio ruolo in un contesto più ampio. Ad esempio, le percentuali sono un valore "a". Si supponga che il 30% viene estratto come un'entità non linguistica. Sarebbe identificata come un aggettivo. Quindi, se il testo contiene "aumento di stipendio del 30%", "30%" entità non linguistica si adatta al modello parte del discorso "ann" (nome nome aggettivo).

Ordine in definizione delle entità

L'ordine in cui le entità sono dichiarate in questo file è importante e influenza il modo in cui vengono estratte. Esse vengono applicate nell'ordine elencato. La modifica dell'ordine modifica i risultati. Le entità non linguistiche più specifiche devono essere definite prima di quelle più generiche.

Ad esempio, l'entità non linguistica "`Amminoacidi`" è definita da:

```
regexp1=($(AA)-?$(NUM))
```

dove `$(AA)` corrisponde a "`(ala|arg|asn|asp|cys|gln|glu|gly|his|ile|leu|lys|met|phe|pro|ser)`", che sono sequenze specifiche di 3 lettere che corrispondono ad amminoacidi particolari.

Dall'altra parte, l'entità non linguistica "`Gene`" è più generale ed è definita da:

```
regexp1=p[0-9]{2,3}
regexp2=[a-z]{2,4}-?[0-9]{1,3}-?[r]
regexp3=[a-z]{2,4}-?[0-9]{1,3}-?p?
```

Se "`Gene`" viene definita prima di "`Amminoacido`" nella sezione Configurazione, "`Amminoacido`" non sarà mai una corrispondenza, poiché `regexp3` da "`Gene`" sarà sempre prima corrispondenza.

Regole di formattazione per la configurazione

- Utilizzare un carattere `TAB` per separare ciascuna voce in una colonna.
- Non eliminare righe.
- Rispettare la sintassi mostrata nella tabella precedente.
- Per disabilitare una voce, inserire il simbolo # all'inizio di tale riga. Per abilitare un'entità, rimuovere il carattere # prima di tale riga.

Gestione lingua

Ogni lingua oggi usata ha modi speciali di esprimere idee, strutturare frasi e utilizzare abbreviazioni. Nella sezione Gestione lingua, è possibile modificare i modelli di estrazione, forzare le definizioni per tali modelli e dichiarare le abbreviazioni per la lingua che è stata selezionata nell'elenco a discesa Lingua.

- Modelli di estrazione
- Definizioni forzate
- Abbreviazioni

Modelli di estrazione

Quando si estraggono le informazioni dai documenti, il motore di estrazione applica una serie di modelli di estrazione parti del discorso a un "blocco" di parole presenti nel testo per identificare termini candidato (parole e frasi) per l'estrazione. È possibile aggiungere o modificare i modelli di estrazione.

Alcune parti del discorso includono elementi grammaticali, come nomi, aggettivi, participi passati, determinativi, preposizioni, coordinatori, nomi di persona, iniziali e particelle. Una serie di questi elementi costituiscono un modello di estrazione di parte del discorso. Nei prodotti di estrazione testo di IBM Corp., ciascuna parte del discorso è rappresentata da un singolo carattere per rendere più facile definire i propri modelli. Ad esempio, un aggettivo è rappresentato dalla lettera minuscola *a*. La serie di codici supportata viene visualizzata per impostazione predefinita nella parte superiore di ogni sezione di modello di estrazione con una serie di modelli ed esempi di ogni modello per aiutare l'utente a comprendere ogni codice che viene utilizzato.

Regole di formattazione per i modelli di estrazione

- Un modello per riga.
- Utilizzare # all'inizio di una riga per disabilitare il modello.

L'ordine in cui si desidera elencare i modelli di estrazione è molto importante perché una data sequenza di parole è letta solo una volta dal motore di estrazione e viene assegnata ai modelli di estrazione per cui il motore trova una corrispondenza.

Definizioni forzate

Quando si estraggono le informazioni da documenti, il motore di estrazione analizza il testo e identifica la parte di discorso per ogni parola che incontra. In alcuni casi, una parola può assumere ruoli diversi a seconda del contesto. Se si desidera forzare una parola ad assumere il ruolo particolare di una parte del discorso o per escludere la parola completamente dall'elaborazione, è possibile farlo nella sezione **Definizione forzata** della scheda Risorse avanzate. Consultare la sezione Capitolo 18, “Informazioni su Risorse avanzate”, a pagina 203 per ulteriori informazioni.

Per forzare un ruolo di parte del discorso per una determinata parola, è necessario aggiungere una riga a questa sezione utilizzando la seguente sintassi:

```
termine:codice
```

Tabella 42. Descrizione della sintassi.

Inserimento	Descrizione
`termine`	Un nome termine.
`codice`	Un codice di un singolo carattere che rappresenta il ruolo di parte del discorso. È possibile elencare fino a sei codici parte del discorso diversi per termine unico. Inoltre, è possibile arrestare una parola da essere estratta in parole/frasi composte utilizzando il codice in minuscolo `s`, ad esempio `aggiuntivo:s`.

Regole di formattazione per definizioni forzate

- Una riga per parola.
- I termini non possono contenere una virgola.
- Utilizzare il minuscolo `s` come codice parte del discorso per interrompere una parola dall'essere estratta.
- Utilizzare fino a sei codici parte del discorso per riga. I codici di parte del discorso sono visualizzati nella sezione Modelli di estrazione. Consultare la sezione “Modelli di estrazione” a pagina 211 per ulteriori informazioni.
- Utilizzare il carattere asterisco (*) come un carattere jolly alla fine di una stringa per corrispondenze parziali. Ad esempio, se si immette `agg*:s`, le parole come `aggregato`, `aggiuntivo`, `aggiunto`, `aggregazione` e `aggettivo` vengono mai estratti come un termine o come parte di un termine parola composta. Tuttavia, se una corrispondenza parola viene esplicitamente dichiarata come termine in un dizionario compilato o nelle definizioni forzate, sarà comunque estratta. Ad esempio, se si immette sia `agg*:s` che `aggettivo:n`, `aggettivo` sarà comunque estratto se trovato nel testo.

Abbreviazioni

Quando il motore di estrazione elabora testo, generalmente considera qualsiasi punto che trova come un'indicazione di una frase terminata. Ciò è in genere corretto; tuttavia, questa gestione dei caratteri punto non si applica quando nel testo sono contenute delle abbreviazioni.

Se dal testo vengono estratti dei termini e viene rilevato che alcune abbreviazioni risultano scorrette, è necessario dichiarare esplicitamente l'abbreviazione in questa sezione.

Nota: se l'abbreviazione appare già in una definizione di sinonimo o è definita come un termine in un dizionario, non è necessario aggiungere qui la voce dell'abbreviazione.

Regole di formattazione per le abbreviazioni

- Definire un'abbreviazione per riga.

Identificativo di lingua

Mentre è sempre meglio selezionare la lingua specifica per i dati di testo che si stanno analizzando, è anche possibile specificare l'opzione **Tutte** quando il testo potrebbe essere in lingue diverse o sconosciute. L'opzione di lingua **Tutte** utilizza un motore di riconoscimento automatico di lingua denominato Identificativo di lingua. L'identificativo di lingua analizza i documenti per identificare quelli che si trovano in una lingua supportata e applica automaticamente il migliore dizionario interno per ciascun file durante l'estrazione. L'opzione **Tutte** viene gestita dai parametri delle sezioni Proprietà.

Proprietà

L'identificativo della lingua viene configurato utilizzando i parametri in questa sezione. La seguente tabella descrive i parametri che è possibile impostare nella sezione **Identificativo lingua - Proprietà** nella scheda Risorse avanzate. Consultare la sezione Capitolo 18, “Informazioni su Risorse avanzate”, a pagina 203 per ulteriori informazioni.

Tabella 43. Descrizioni dei parametri

Parametro	Descrizione
NUM_CHARS	Specifica il numero di caratteri che devono essere letti dal motore di estrazione per determinare la lingua del testo. Più basso è il numero, più velocemente la lingua è identificata. Più alto è il numero, più accuratamente la lingua viene identificata. Se si imposta il valore su 0, verrà letto il testo completo del documento .
USE_FIRST_SUPPORTED _LANGUAGE	Specifica se il motore di estrazione deve utilizzare la prima lingua supportata trovata dall'identificativo della lingua. Se si imposta il valore su 1, viene utilizzata la prima lingua supportata. Se si imposta il valore su 0, viene utilizzato un valore di lingua di riserva.
FALLBACK_LANGUAGE	Specifica la lingua da utilizzare se la lingua restituita dall'identificativo non è supportata. I valori possibili sono `english`, `french`, `german`, `spanish`, `dutch`, `italian` e `ignore`. Se si imposta il valore su `ignore`, il documento senza lingua supportata verrà ignorato.

Lingue

L'identificativo di lingua supporta numerose lingue diverse. È possibile modificare l'elenco di lingue nella sezione **Identificativo lingua - Lingue** nella scheda Avanzate.

È possibile prendere in considerazione l'eliminazione delle lingue che hanno probabilità di non essere utilizzate da questo elenco perché più lingue sono rappresentate, maggiore è la possibilità di falsi e di rallentamento delle prestazioni. Non è comunque possibile aggiungere nuove lingue a questo file.

Prendere in considerazione di posizionare le lingue più probabili all'inizio dell'elenco per aiutare l'identificativo di lingua a trovare più rapidamente una corrispondenza con i documenti.

Capitolo 19. Informazioni sulle regole di collegamento del testo

TLA (Text link analysis) è un modello di tecnologia corrispondente che viene utilizzato per estrarre le relazioni trovate nel suo testo utilizzando una serie di regole. Quando l'analisi di collegamento del testo è abilitata per l'estrazione, i dati di testo vengono confrontati rispetto a queste regole. Quando viene trovata una corrispondenza, il modello di analisi di collegamento del testo viene estratto e presentato. Queste regole vengono definite nella scheda Regole di collegamento del testo.

Ad esempio, l'estrazione dei concetti che rappresentano idee semplici su un'organizzazione potrebbero non essere abbastanza interessanti, ma utilizzando TLA, è anche possibile avere informazioni sui collegamenti tra organizzazioni diverse e sulle persone associate all'organizzazione. TLA può essere utilizzati anche per estrarre i pareri su argomenti quali cosa i cittadini sentono su un prodotto o un'esperienza.

Per trarre vantaggio da TLA, è necessario avere risorse che contengono regole TLA (text link analisys). Quando si seleziona un modello, è possibile vedere quali sono i modelli di regole TLA o se hanno un'icona nella colonna TLA.

I modelli TLA vengono trovati nei dati di testo durante la fase di corrispondenza del modello del processo di estrazione. Durante questa fase, le regole vengono confrontate con i dati di testo e quando viene trovata una corrispondenza, queste informazioni vengono estratte come modello. Ci sono momenti in cui si potrebbe voler ottenere di più dalle analisi di collegamento del testo o modificare come procedere per la corrispondenza. In questi casi, è possibile ridefinire le regole per adattarle alle proprie esigenze specifiche. Ciò viene eseguito nella scheda Regole di collegamento del testo.

Nota: il supporto per le variabili è stato sospeso nella versione 13. Utilizzare invece le macro. Consultare la sezione “Gestione delle macro” a pagina 220 per ulteriori informazioni.

Dove lavorare sulle regole di collegamento del testo

È possibile modificare e creare regole direttamente nella vista Editor di modelli o Editor risorse. Per aiutare l'utente a vedere in che modo le regole potrebbero corrispondere al testo, è possibile eseguire una simulazione in questa scheda. Durante la simulazione, l'estrazione viene eseguita solo sui dati della simulazione di esempio e vengono applicate le regole di collegamento del testo per vedere se vi sono modelli che corrispondono. Le regole che corrispondono al testo vengono quindi visualizzate nel riquadro di simulazione. In base alle corrispondenze, è possibile scegliere di modificare le regole e le macro per modificare il modo in cui il testo viene messo in corrispondenza.

A differenza dalle altre risorse avanzate, le regole TLA sono specifiche per la libreria; di conseguenza, è possibile utilizzare le regole TLA solo da una libreria alla volta. Dall'interno di Editor di modelli o di Editor risorse, andare alla scheda **Regole di collegamento del testo**. In questa scheda è possibile specificare la libreria nel modello che contiene le regole TLA da utilizzare o modificare. Per questo motivo, si raccomanda di memorizzare tutte le regole in una libreria a meno che non vi sia un motivo grave o specifico per non farlo.

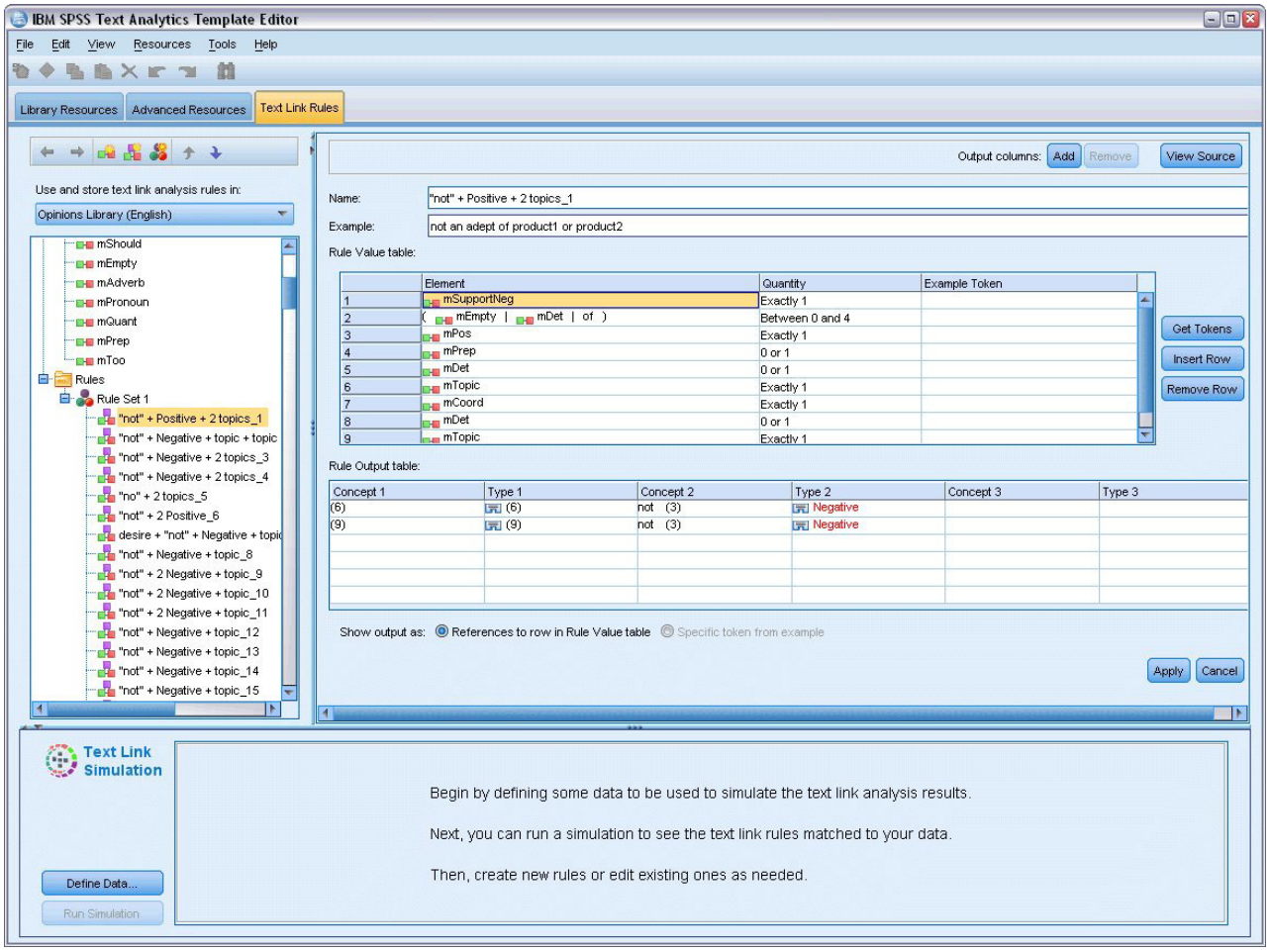


Figura 42. Scheda Regole di collegamento del testo

Importante! Questa scheda non è disponibile per le risorse in lingua giapponese.

Come iniziare

Esistono diversi modi per iniziare a lavorare nell'editor della scheda Regole di collegamento del testo:

- Iniziare simulando risultati con un testo di esempio e modificare o creare le regole di corrispondenza in base al modo in cui la serie corrente di modelli estrae le regole dai dati di simulazione.
- Creare una nuova regola da zero o modificare una regola esistente.
- Lavorare direttamente nella vista di origine.

Quando modificare o creare regole

Mentre le regole di analisi di collegamento del testo fornito con ogni modello vengono spesso adeguate per estrarre molte relazioni semplici o complesse dal testo, vi sono momenti in cui si vorrebbero apportare alcune modifiche a queste regole o creare alcune regole proprie. Ad esempio:

- Catturare un'idea o relazione che non è in corso di estrazione con le regole esistenti creando una nuova regola o macro.
- Modificare il comportamento predefinito di un tipo aggiunto alle risorse. Ciò richiede di solito di modificare una macro come `mTopic` o `mNonLingEntities`. Consultare la sezione “Macro speciali: mTopic, mNonLingEntities, SEP” a pagina 223 per ulteriori informazioni.

- Aggiungere nuovi tipi di testo alle regole e le macro di analisi di collegamento del testo esistenti. Ad esempio, se si pensa che il tipo <Organizzazione> è troppo ampio, è possibile creare nuovi tipi di organizzazioni in settori di business differenti come <Farmaceutici>, <Automobili>, <Finanza> e così via. In questo caso, è necessario modificare le regole di analisi di collegamento del testo e/o creare una macro per prendere in considerazione questi nuovi tipi ed elaborarli.
- Aggiungere i tipi ad una regola di analisi di collegamento del testo esistente. Ad esempio, si supponga di avere una regola che cattura il seguente testo `john doe ha chiamato jane doe`, ma si desidera che questa regola acquisisca oltre alle comunicazioni telefoniche anche scambi di email. È possibile aggiungere il tipo di entità non linguistica per email per la regola in modo da catturare anche testo come: `johndoe@ibm.com emailed janedoe@ibm.com`.
- Modificare lievemente una regola esistente, invece di crearne una nuova. Ad esempio, si supponga di avere una regola che corrisponde al seguente testo `xyz è buonissimo` ma si desidera che questa regola prenda anche `xyz è molto, molto buono`.

Simulazione dei risultati di analisi di collegamento del testo

Per definire le nuove regole di collegamento del testo o capire come certe frasi sono corrispondenti durante l'analisi di collegamento del testo, è spesso utile prendere un pezzo di testo ed eseguire una simulazione. Durante la simulazione, l'estrazione viene eseguita solo su dati di simulazione di esempio utilizzando la serie corrente di risorse linguistiche e delle impostazioni di estrazione correnti. L'obiettivo è quello di ottenere i risultati simulati e utilizzare questi risultati per migliorare le proprie regole, crearne di nuove o meglio comprendere come si verificano le corrispondenze. Per ogni pezzo di testo (una frase, una parola o una proposizione in base al contesto), un output di simulazione visualizza la raccolta di token e le regole TLA che hanno rilevato un modello in tale testo. Un **token** viene definito come qualsiasi parola o frase di parole identificate durante il processo di estrazione.

A differenza dalle altre risorse avanzate, le regole TLA sono specifiche per la libreria; di conseguenza, è possibile utilizzare le regole TLA solo da una libreria alla volta. Dall'interno di Editor di modelli o di Editor risorse, andare alla scheda **Regole di collegamento del testo**. In questa scheda è possibile specificare la libreria nel modello che contiene le regole TLA da utilizzare o modificare. Per questo motivo, si raccomanda di memorizzare tutte le regole in una libreria a meno che non vi sia un motivo grave o specifico per non farlo.

Importante! Si consiglia fortemente, se si utilizza un file di dati, di verificare che il testo contenuto sia breve per ridurre al minimo il tempo di elaborazione. L'obiettivo della simulazione è vedere come un blocco di testo viene interpretato e comprendere in che modo le regole corrispondono a questo testo. Queste informazioni consentono di scrivere e modificare le regole. Utilizzare il nodo analisi del collegamento del testo o eseguire un flusso con sessione interattiva con l'estrazione TLA abilitata per ottenere risultati per una serie di dati più completa. Questa simulazione è di prova e la regola è solo per scopi di sviluppo.

Definizione di dati per la simulazione

Per aiutare l'utente a vedere in che modo le regole potrebbero corrispondere al testo, è possibile eseguire una simulazione utilizzando i dati di esempio. Il primo passa consiste nel definire i dati.

Definizione dei dati

1. Fare clic su **Definisci dati** nel pannello di simulazione nella parte inferiore della scheda **Regole di collegamento del testo**. In alternativa, se i dati non sono stati precedentemente definiti, scegliere **Strumenti** > **Esegui simulazione** dal menu. Viene aperta la procedura guidata Dati di simulazione.
2. Specificare il tipo di dati selezionando una delle seguenti opzioni:
   - **Incolla o immetti il testo direttamente** Viene fornita una casella di testo per incollare un testo dal blocco appunti o per immettere manualmente il testo desiderato da elaborare. È possibile immettere una frase per riga o utilizzare punteggiatura per rompere la frase come punti o virgole. Dopo aver immesso il testo, è possibile iniziare la simulazione facendo clic su **Esegui simulazione**.

- **Specifica un'origine dati del file** Questa opzione indica che si desidera elaborare un file che contiene testo. Fare clic su **Avanti** per procedere al passo della procedura guidata in cui è possibile definire il file da elaborare. Una volta selezionato il file, è possibile iniziare la simulazione facendo clic su **Esegui simulazione**. Sono supportati i seguenti tipi di file: `.txt` e `.text`. Il file di dati selezionato viene letto "nello stato' durante la simulazione. L'intero file viene trattato nello stesso modo come se avesse connesso un nodo Elenco file per un nodo Estrazione testo.

Importante: Si consiglia fortemente, se si utilizza un file di dati, di verificare che il testo contenuto sia breve per ridurre al minimo il tempo di elaborazione. L'obiettivo della simulazione è vedere come un blocco di testo viene interpretato e comprendere in che modo le regole corrispondono a questo testo. Queste informazioni consentono di scrivere e modificare le regole. Utilizzare il nodo analisi di collegamento del testo o eseguire un flusso con sessione interattiva con l'estrazione TLA abilitata per ottenere risultati per una serie di dati più completa. Questa simulazione è di prova e la regola è solo per scopi di sviluppo.

3. Per iniziare il processo di simulazione, fare clic su **Esegui simulazione**. Viene visualizzata una finestra di dialogo di avanzamento. Se ci si trova in una sessione interattiva, le impostazioni di estrazione utilizzate durante la simulazione sono quelle attualmente selezionate nella sessione interattiva (consultare **Strumenti** > **Impostazioni di estrazione** nella vista Concetti e categorie). In Editor di modelli, le impostazioni utilizzate durante l'estrazione della simulazione sono le impostazioni di estrazione predefinite, che sono le stesse di quelle mostrate nella scheda di livello avanzato di un nodo di analisi di collegamento del testo. Per ulteriori informazioni, vedere “Comprensione dei risultati della simulazione”.

Comprensione dei risultati della simulazione

Per aiutare l'utente a vedere in che modo le regole potrebbero corrispondere al testo, è possibile eseguire una simulazione utilizzando i dati di esempio e analizzando i risultati. Da lì è possibile modificare la serie di regole per meglio adattare i dati. Quando il processo di estrazione e la simulazione sono stati completati, verranno presentati i risultati della simulazione.

Per ogni "frase" identificata durante l'estrazione, vengono visualizzate diverse informazioni inclusa la "frase" esatta, l'analisi dei token trovati in questa frase di testo di input e, infine, le regole che corrispondono al testo in questa frase. Per "**frase**", si intende una parola, una frase o una proposizione a seconda di come l'estrattore ha diviso il testo in blocchi leggibili.

Un **token** viene definito come qualsiasi parola o frase di parole identificate durante il processo di estrazione. Ad esempio, nella frase *Mio zio vive a New York*, potrebbero essere rilevati i seguenti token durante l'estrazione: *mio*, *zio*, *vive*, *a* e *new york*. Inoltre, *zio* potrebbe essere estratto come concetto e immesso come `<Sconosciuto>` e *new york* potrebbe anche essere estratto come concetto e immesso come `<Ubicazione>`. Tutti i concetti sono token ma non tutti i token sono concetti. I token possono anche essere altre macro, stringhe di costanti letterali e mancanze di parole. Solo le parole o frasi immesse possono essere concetti.

Quando si sta lavorando nella sessione interattiva o nell'editor di risorsa, si sta lavorando su un livello di concetto. Le regole TLA sono più granulari e i token singoli in una frase possono essere utilizzati nella definizione di una regola anche se non sono mai stati estratti e digitati. Essere in grado di utilizzare i token che non sono concetti offre anche una maggiore flessibilità nell'acquisizione di relazioni complesse nel testo.

Se si dispone di più di una frase nei dati di simulazione, è possibile spostarsi in avanti e indietro attraverso i risultati facendo clic su **Avanti** e **Precedente**.

Nei casi in cui una frase non corrisponde ad alcuna regola TLA nella libreria selezionata (vedi Nome libreria sulla struttura ad albero in questa scheda), i risultati vengono considerati non corrispondenti e i

pulsanti **Avanti non corrispondenti** e **Indietro non corrispondenti** sono abilitati per consentire di sapere se è presente testo per il quale nessuna regola ha trovato una corrispondenza e per consentire all'utente di accedere a questi casi rapidamente.

Dopo la creazione di nuove regole, la modifica delle proprie o aver modificato le risorse o impostazioni di estrazione, è possibile che si desideri eseguire una simulazione. Per rieseguire una simulazione, fare clic su **Esegui simulazione** nel riquadro di simulazione e verranno utilizzati di nuovo gli stessi dati di input.

I seguenti campi e tabelle vengono visualizzati nei risultati della simulazione:

Testo di input. La 'frase" effettiva individuata dal processo di estrazione dai dati di simulazione definiti nella procedura guidata. Per "frase", si intende una parola, una frase o una proposizione a seconda di come l'estrattore ha diviso il testo in blocchi leggibili.

Vista Sistema. Una raccolta di token che il processo di estrazione ha identificato.

- **Token di testo di input.** Ogni token trovato nel testo di input. I token sono stati definiti in precedenza in questa sezione.
- **Immesso come.** Se il token è stato identificato come un concetto e immesso, il nome tipo associato (ad esempio, `<Sconosciuto>`, `<Persona>`, `<Ubicazione>`) viene visualizzato in questa colonna.
- **Corrispondenza macro.** Se il token corrisponde a una macro esistente, il nome della macro associata viene visualizzato in questa colonna.

Regole corrispondenti al testo di input. Questa tabella mostra tutte le regole TLA corrispondenti al testo di input. Per ciascuna regola di corrispondenza, verrà visualizzato il nome della regola nella colonna **Output di regola** e i valori di output associati per tale regola (coppie Concetto + Tipo). È possibile fare doppio clic con il mouse sul nome regola corrispondente per aprire la regola nel riquadro editor sul riquadro di simulazione.

Pulsante **Genera regola**. Se si fa clic su questo pulsante nel pannello di simulazione, una nuova regola verrà aperta nel riquadro di editor sopra il pannello di simulazione. Il testo di input viene considerato come esempio. Allo stesso modo, tutti i token immessi o corrispondenti a una macro durante la simulazione vengono inseriti automaticamente nella colonna Elementi nella **tabella Valori di regola**. Se il token è stato immesso *e* corrisponde a una macro, il valore della macro è quello che verrà usato nella regola in modo da semplificare la regola. Ad esempio, la frase "*Mi piace la pizza*" potrebbe essere immessa durante la simulazione come `<Sconosciuto>` e corrispondere alla macro `mTopic` se si stanno utilizzando le risorse di inglese di base. In questo caso `mTopic` verrà utilizzato come elemento nella regola generata. Consultare la sezione “Gestione delle regole di collegamento del testo” a pagina 223 per ulteriori informazioni.

Navigazione di regole e macro nella struttura ad albero

Quando l'analisi di collegamento del testo viene eseguita durante l'estrazione, verranno utilizzate le regole di collegamento del testo memorizzate nella libreria selezionata nella scheda **Regole di collegamento del testo**.

A differenza dalle altre risorse avanzate, le regole TLA sono specifiche per la libreria; di conseguenza, è possibile solo utilizzare le regole TLA da una libreria alla volta. Dall'interno di Editor di modelli o di Editor risorse, andare alla scheda **Regole di collegamento del testo**. In questa scheda è possibile specificare la libreria nel modello che contiene le regole TLA da utilizzare o modificare. Per questo motivo, si raccomanda di memorizzare tutte le regole in una libreria a meno che non vi sia un motivo grave o specifico per non farlo.

È possibile specificare in quale libreria si desidera lavorare nella scheda delle regole di collegamento del testo selezionando tale libreria nel campo **Utilizza e memorizzare le regole di analisi del collegamento**

del testo in: elenco a discesa in questa scheda. Quando l'analisi di collegamento del testo viene eseguita durante l'estrazione, verranno utilizzate le regole di collegamento del testo memorizzate nella libreria selezionata nella scheda **Regole di collegamento del testo**. Pertanto, se sono state definite regole di collegamento di testo (regole TLA) in più di una libreria, solo la prima libreria in cui le regole TLA sono trovate verrà utilizzata per l'analisi del collegamento del testo. Per questo motivo, si raccomanda di memorizzare tutte le regole in una libreria a meno che non vi sia un motivo grave o specifico per non farlo.

Quando si seleziona una macro o regola nella struttura ad albero, il suo contenuto viene visualizzato nel riquadro dell'editor a destra. Se si seleziona con il tasto destro su qualsiasi elemento nella struttura ad albero, un menu di contesto verrà aperto per mostrare quali altre attività sono possibili, come:

- Creare una nuova macro nella struttura ad albero e aprirla nell'editor a destra.
- Creare una nuova regola nella struttura ad albero e aprirla nell'editor a destra.
- Creare una nuova regola impostata nella struttura ad albero.
- Taglia, Copia e Incolla per gli elementi per semplificare la modifica.
- Eliminare macro, regole e serie di regole per rimuoverle dalle risorse.
- Disabilitare macro, regole e serie di regole per indicare che devono essere ignorati durante l'elaborazione.
- Spostare le regole in alto o in basso per influenzare l'ordine di elaborazione.

Avvertenze nella struttura ad albero

Le avvertenze vengono visualizzate con un triangolo giallo nella struttura ad albero e sono lì per informare che potrebbe essersi verificato un problema. Spostare il puntatore del mouse sulla macro o sulla regola difettose per visualizzare una spiegazione a comparsa. Nella maggior parte dei casi, verrà visualizzato qualcosa come: **Avvertenza: nessun esempio fornito; immettere un esempio**; è necessario quindi immettere un esempio.

Se si manca un esempio o se l'esempio non corrisponde alla regola, non sarà possibile utilizzare la funzione Get Tokens e quindi si consiglia di immettere solo un esempio per regola.

Quando la regola viene evidenziata in giallo significa che un tipo o una macro è sconosciuto per l'editor TLA. Il messaggio sarà simile a: **Avvertenza: tipo sconosciuto o macro**. Questo è per informare l'utente che un elemento che sarebbe definito da `$something` nella vista di origine, ad esempio `$myType`, non è un tipo ereditato nella libreria, né è una macro.

Per aggiornare il controllo della sintassi è necessario passare ad un'altra regola o macro; non è necessario ricompilare. Pertanto, ad esempio, se la regola `A` visualizza un'avvertenza poiché l'esempio è mancante, è necessario aggiungere un esempio, fare clic su una regola superiore o inferiore e poi tornare di nuovo a regola `A` per controllare che sia stato ora corretto.

Gestione delle macro

Le macro possono semplificare la visualizzazione di regole di analisi di collegamento del testo, consentendo di raggruppare tipi, altre macro e stringhe di valori letterali (parola) insieme ad un operatore `OR` (`|`). Il vantaggio di utilizzare le macro è che non solo è possibile riutilizzare le macro in più regole di analisi di collegamento del testo per semplificarle, ma consente anche di effettuare gli aggiornamenti in una macro invece di dover apportare aggiornamenti in tutte le regole di analisi di collegamento del testo. La maggior parte delle regole TLA fornite contengono macro predefinite. Le macro vengono visualizzate nella parte superiore della struttura ad albero nel riquadro a sinistra della scheda Regole di collegamento del testo.

I seguenti campi e tabelle vengono visualizzati nei risultati della simulazione:

Nome. Un nome univoco di identificazione della macro. Si raccomanda di apporre un prefisso ai nomi di macro con una m minuscola per identificare rapidamente le macro nelle regole. Quando si fa riferimento alle macro nelle proprie regole manualmente (modificando in linea o nella vista di origine) è necessario utilizzare il carattere $ prefisso in modo che il processo di estrazione sa che deve ricercare tale nome speciale. Tuttavia se si trascina e si rilascia il nome macro o lo si aggiunge attraverso il menu di scelta rapida, il prodotto lo riconosce automaticamente come macro e non verrà aggiunto alcun segno $.

Tabella **Valore macro**.

- Il numero di righe che rappresenta tutti i possibili valori che questa macro può rappresentare. Questi valori sono sensibili al maiuscolo/minuscolo.
- Questi valori possono includere uno o una combinazione di tipi, stringhe di valori letterali, divari o le macro. Consultare la sezione “Elementi supportati per regole e macro” a pagina 230 per ulteriori informazioni.
- Per immettere un valore per un elemento in una macro, fare doppio clic sulla riga che si desidera utilizzare. Viene visualizzata una casella di testo modificabile in cui è possibile immettere un riferimento di tipo, un riferimento macro, una stringa letterale oppure una differenza di parole. In alternativa, fare clic con il tastino destro del mouse nella cella per visualizzare un menu contestuale contenente gli elenchi di macro comuni, i nomi tipo e i nomi di tipo non linguistico. Per fare riferimento ad un tipo o a una macro è necessario anteporre al nome macro o al tipo un carattere '$' come `$mTopic` per la macro `mTopic`. Quando si combinano gli argomenti, è necessario utilizzare le parentesi `( )` per raggruppare gli argomenti e il carattere `|` per indicare un valore booleano `OR`.
- È possibile aggiungere o rimuovere le righe nella tabella Valore macro utilizzando i pulsanti a destra.
- Immettere ciascun elemento nella propria riga. Ad esempio, se si desidera creare una macro che rappresenta una di 3 stringhe di valore letterali come `sono OR ero OR è`, immettere ciascuna stringa di valore letterale su una riga separata nella vista e la tabella delle macro conterrà 3 righe.

Creazione e modifica di macro

È possibile creare nuove macro o modificare quelle esistenti. Seguire le indicazioni e le descrizioni per l'editor di macro. Consultare la sezione “Gestione delle macro” a pagina 220 per ulteriori informazioni.

Creazione di nuove macro

1. Dai menu scegliere **Strumenti > Nuova macro**. In alternativa, fare clic sull'icona Nuova macro nella barra degli strumenti della struttura ad albero per aprire una nuova regola nell'editor.
2. Immettere un nome univoco e definire gli elementi di valore della macro.
3. Fare clic su **Applica** una volta terminato per verificare la presenza di errori.

Modifica macro

1. Fare clic sul nome macro nella struttura ad albero. La macro viene visualizzata nel riquadro dell'editor a destra.
2. Effettuare le modifiche.
3. Fare clic su **Applica** una volta terminato per verificare la presenza di errori.

Disattivazione ed eliminazione delle macro

Disabilitazione delle macro

Se si desidera che una macro venga ignorata durante l'elaborazione, è possibile disabilitarla. In questo modo potrebbero sorgere avvertenze o errori nelle regole che fanno ancora riferimento questa macro disabilitata. Procedere con attenzione quando si eliminano e si disabilitano delle macro.

1. Fare clic sul nome macro nella struttura ad albero. La macro viene visualizzata nel riquadro dell'editor a destra.
2. Fare clic con il pulsante destro del mouse sul nome.

3. Dal menu di scelta rapida, scegliere **Disabilita**. L'icona di macro diventa grigia e la macro stessa diventa non modificabile.

Eliminazione delle macro

Se si desidera sbarazzarsi di una macro, è possibile eliminarla. In questo modo potrebbero sorgere errori nelle regole che fanno ancora riferimento a questa macro. Procedere con attenzione quando si eliminano e si disabilitano delle macro.

1. Fare clic sul nome macro nella struttura ad albero. La macro viene visualizzata nel riquadro dell'editor a destra.
2. Fare clic con il pulsante destro del mouse sul nome.
3. Dal menu di scelta rapida, scegliere **Elimina**. La macro scompare dall'elenco.

Controllo errori, salvataggio e annullamento

Applicazione delle modifiche di macro

Se si fa clic al di fuori dell'editor della macro o se si fa clic su **Applica**, la macro viene analizzata automaticamente per rilevare eventuali errori. Se viene trovato un errore, è necessario correggerlo prima di passare a un'altra parte dell'applicazione.

Tuttavia, se vengono rilevati errori meno gravi, viene fornita solo un'avvertenza. Ad esempio, se la macro contiene definizioni incomplete o senza riferimenti a tipi o altre macro, viene visualizzato un messaggio di avvertenza. Dopo aver fatto clic su **Applica**, qualsiasi avvertenza di errore causa un'icona di avvertenza che viene visualizzata alla sinistra del nome macro nella struttura ad albero di regole e macro nel riquadro a sinistra.

L'applicazione di una macro non significa che la macro viene salvata in modo permanente. L'applicazione farà sì che il processo di convalida verifichi la presenza di errori e avvertenze.

Salvataggio delle risorse all'interno di una sessione workbench interattiva

1. Per salvare le modifiche apportate alle risorse durante una sessione workbench interattiva in modo che sia possibile ottenerle la volta successiva che si esegue il flusso, è necessario:
   - Aggiornare il nodo di modellazione per accertarsi che è possibile ottenere queste stesse risorse la volta successiva che si esegue il flusso. Consultare la sezione “Aggiornamento dei nodi di modellazione e salvataggio” a pagina 84 per ulteriori informazioni. Salvare poi il flusso. Per salvare il flusso, effettuare questa operazione nella finestra principale di IBM SPSS Modeler dopo aver aggiornato il nodo di modellazione.
2. Per salvare le modifiche apportate alle risorse durante una sessione workbench interattiva in modo che sia possibile utilizzarle in altri flussi, è necessario:

- Aggiornare il modello utilizzato oppure crearne uno nuovo. Consultare la sezione “Creazione ed aggiornamento di modelli” a pagina 167 per ulteriori informazioni. Ciò non salverà le modifiche per il nodo corrente (vedi passo precedente)
- In alternativa, aggiornare la TAP utilizzata. Consultare la sezione “Aggiornamento dei pacchetti di analisi del testo” a pagina 142 per ulteriori informazioni.

Salvataggio delle risorse all'interno di Editor di modelli

1. In primo luogo, pubblicare la libreria. Consultare la sezione “Pubblicazione delle librerie” a pagina 185 per ulteriori informazioni.
2. Salvare poi il modello mediante **File > Salva modello di risorsa** nel menu.

Annullamento delle modifiche di macro

1. Se si desidera annullare le modifiche, fare clic su **Annulla**.

Macro speciali: mTopic, mNonLingEntities, SEP

Il modello di pareri (come i modelli) così come i modelli di risorse di base vengono forniti con due macro speciali denominate `mTopic` e `mNonLingEntities`.

mTopic

Per impostazione predefinita, la macro `mTopic` raggruppa tutti i tipi forniti nel modello che sono suscettibili di essere collegati con un parere, come i seguenti tipi di libreria *Core* tipi libreria: `<Persona>`, `<Organizzazione>`, `<Ubicazione>` e così via, purché il tipo non sia un tipo di avviso (ad esempio, `<Negativo>` o `<Positivo>`) o un tipo definito come entità non linguistica nelle risorse avanzate.

Quando si crea un nuovo tipo in un modello Pareri (o simile), il prodotto presuppone che a meno che questo tipo non sia specificato in un'altra macro o nella sezione Entità non linguistica della scheda Risorse avanzate, esso sarà trattato allo stesso modo degli altri tipi definiti nella macro `mTopic`.

Si supponga di aver creato nuovi tipi nelle risorse da un modello Pareri: `<Verdura>` e `<Frutta>`. Senza dover apportare alcuna modifica, i nuovi tipi vengono trattati come tipi `mTopic` in modo che sia possibile scoprire automaticamente i pareri positivi, negativi, imparziali e contestuali sui nuovi tipi. Durante l'estrazione, ad esempio, la frase " *A me piacciono i broccoli, ma odio il pompelmo*" genera il seguente output: 2 modelli

```
broccoli <Verdura> + piace <Positivo>

pompelmo <Frutta> + non piace <Negativo>
```

Tuttavia, se si desidera elaborare questi tipi in modo diverso rispetto agli altri tipi in mTopic, è possibile aggiungere il nome tipo a una macro esistente come ad esempio `mPos`, che raggruppa tutti i tipi di parere positivo o creare una nuova macro a cui è possibile in seguito fare riferimento in una o più regole.

Importante! Se si crea un nuovo tipo come `<Verdura>`, questo nuovo tipo verrà incluso come un tipo in mTopic; tuttavia, questo nome tipo non sarà visibile esplicitamente nella definizione macro.

mNonLingEntities

Allo stesso modo, se si aggiungono nuove entità non linguistiche nella sezione **Entità non linguistiche** della scheda Risorse avanzate, queste verranno automaticamente elaborate come `mNonLingEntities` a meno che diversamente specificato. Consultare la sezione “Entità non linguistiche” a pagina 207 per ulteriori informazioni.

SEP

È inoltre possibile utilizzare la macro predefinita `SEP` che corrisponde al separatore globale definito sulla macchina locale, solitamente una virgola (,).

Gestione delle regole di collegamento del testo

Una regola di analisi del collegamento del testo è una interrogazione booleana che viene utilizzata per eseguire una corrispondenza su una frase. Le regole di analisi di collegamento del testo possono contenere uno o più dei seguenti argomenti: tipi, macro, stringhe di costanti letterali o divari tra parole. È necessario avere almeno una regola di analisi di collegamento del testo per estrarre i risultati TLA.

Le seguenti aree e i campi vengono visualizzati nella scheda Regole di collegamento del testo, Editor di regole:

Campo **Nome**. Il nome univoco per la regola di collegamento del testo.

Campo **Esempio**. Facoltativamente, è possibile includere una frase o parola di esempio che dovrebbero essere catturate da questa regola. Si consiglia di utilizzare gli esempi. In questo editor, si possono generare token da questo testo di esempio per vedere in che modo il testo corrisponde alla regola e quale sarà l'output. Un **token** viene definito come qualsiasi parola o frase di parole identificate durante il processo di estrazione. Ad esempio, nella frase *Mio zio vive a New York*, potrebbero essere rilevati i seguenti token durante l'estrazione: *mio*, *zio*, *vive*, *a* e *new york*. Inoltre, *zio* potrebbe essere estratto come concetto e immesso come `<Sconosciuto>` e *new york* potrebbe anche essere estratto come concetto e immesso come `<Ubicazione>`. Tutti i concetti sono token ma non tutti i token sono concetti. I token possono anche essere altre macro, stringhe di costanti letterali e mancanze di parole. Solo le parole o frasi immesse possono essere concetti.

Tabella di valore di regola. Questa tabella contiene gli elementi della regola che sono utilizzati per associare una regola ad una frase. È possibile aggiungere o rimuovere le righe nella tabella utilizzando i pulsanti a destra. La tabella consiste di 3 colonne:

- Colonna **Elemento**. Immettere i valori come unico o come una combinazione di tipi, stringhe di valori letterali, divari (`<Tutti i token>`) o macro. Consultare la sezione "Elementi supportati per regole e macro" a pagina 230 per ulteriori informazioni. Fare doppio clic sulla cella Elemento per immettere direttamente le informazioni. In alternativa, fare clic con il tastino destro del mouse nella cella per visualizzare un menu di scelta rapida contenente gli elenchi di macro comuni, i nomi tipo e i nomi di tipo non linguistico. Tenere presente che se si immettono le informazioni nella cella inserendole direttamente, anteporre al nome macro o tipo un carattere '`$`' come `$mTopic` per la macro `mTopic`. L'ordine in cui vengono create le righe di elemento è fondamentale per il modo di corrispondenza della regola al testo. Quando si combinano gli argomenti, è necessario utilizzare le parentesi `( )` per raggruppare gli argomenti e il carattere `|` per indicare un valore booleano `OR`. Tenere presente che i valori sono sensibili al maiuscolo/minuscolo.
- Colonna **Quantità**. Indica il numero minimo e massimo di volte per cui l'elemento deve essere trovato affinché si verifichi una corrispondenza. Ad esempio, se si desidera definire un vuoto, o una serie di parole, tra altri due elementi compresi tra 0 e 3 parole, è possibile scegliere **Tra 0 e 3** dall'elenco o immettere i numeri direttamente nella finestra di dialogo. Il valore predefinito è '**Esattamente 1**'. In alcuni casi si desidera rendere un elemento facoltativo. Se questo è il caso, allora avrà un quantitativo minimo di 0 e una quantità massima superiore a 0 (ad esempio 0 o 1, tra 0 e 2). Notare che il primo elemento in una regola non può essere facoltativo, ovvero non può avere una quantità pari a 0.
- Colonna **Token di esempio**. Se si fa clic su **Ottieni Token**, il programma interrompe il testo di **Esempio** nei token e utilizza tali token per riempire questa colonna con quelli che corrispondono agli elementi definiti. È possibile anche visualizzare questi token nella tabella di output.

Tabella di output della regola. Ogni riga di questa tabella definisce il modo in cui l'output di modello TLA apparirà nei risultati. L'output di regola può produrre modelli di fino a sei coppie di colonne Concetto/Tipo, ognuna rappresentante uno *slot*. Ad esempio, il modello di tipo `<Ubicazione> + <Positivo>` è un modello a due slot che significa che è composto da 2 coppie di colonna Concetto/Tipo.

Proprio come la lingua concede la libertà di esprimere le stesse idee fondamentali in molti modi diversi, così è possibile disporre di un numero di regole definite per catturare la stessa idea fondamentale. Ad esempio, il testo *"Parigi è un luogo che amo"* e il testo *"Sono molto, molto simili Parigi e Firenze"* rappresentano lo stesso fondamentale concetto -- che Parigi è piaciuta -- ma sono espressi diversamente e richiederebbero due regole diverse per entrambe essere catturate. Tuttavia, è più semplice gestire i risultati del modello se idee simili sono raggruppate insieme. Per questo motivo, mentre si potrebbero avere 2 regole diverse per catturare queste 2 frasi, è possibile definire lo stesso output per entrambe le regole, come il tipo di modello di `<Ubicazione> + <Positivo>` in modo che rappresenti entrambi i testi. E in questo modo, è possibile vedere che l'output non sempre imita la struttura o ordine delle parole trovate nel testo originale. Inoltre, un tipo di modello potrebbe corrispondere a altre diciture e potrebbero produrre modelli di concetto come ad esempio: `parigi + amo` e `tokyo + amo`.

Per aiutare l'utente a definire l'output rapidamente con pochi errori, è possibile utilizzare il menu di scelta rapida per selezionare l'elemento che si desidera visualizzare nell'output. In alternativa, è possibile

anche trascinare e rilasciare elementi dalla tabella Valore regola nell'output. Ad esempio, se si dispone di una regola che contiene un riferimento alla macro `mTopic` nella riga 2 della tabella Valore di regola e si desidera che il valore deve essere nell'output, è possibile semplicemente trascinare/rilasciare l'elemento per `mTopic` alla coppia della prima colonna nella tabella di output della regola. In questo modo verranno automaticamente popolati sia Concetto che Tipo per la coppia selezionata. In alternativa, se si desidera che l'output per iniziare con il tipo definito dal terzo elemento (riga 3) della tabella dei valori della regola, trascinare il tipo dalla tabella Valore di regola sulla cella **Tipo 1** nella tabella di output. La tabella verrà aggiornata per mostrare il riferimento di riga in parentesi (3).

In alternativa, è possibile immettere tali riferimenti manualmente nella tabella facendo doppio clic sulla cella in ciascuna colonna **Concetto** di cui si desidera l'output e immettendo il simbolo $ seguito dal numero di riga, ad esempio `$2` per fare riferimento all'elemento definito nella riga 2 della tabella Valore di regola. Quando si immettono le informazioni manualmente, è necessario anche definire la colonna **Tipo**, immettere il simbolo # seguito dal numero di riga, ad esempio `#2` per fare riferimento all'elemento definito nella riga 2 della tabella Valore di regola.

Inoltre, è possibile anche combinare i metodi. Si supponga di aver avuto il tipo `<Positivo>` nella riga 4 della tabella Valore di regola. È possibile trascinare nella colonna `Tipo 2` e quindi fare doppio clic sulla cella nella colonna `Concetto 2` e quindi immettere manualmente la parola "*non*' davanti ad essa. La colonna di output riporta `non (4)` nella tabella o, in modalità modifica o modalità origine, `non $4`. È quindi possibile fare clic nella colonna Tipo 1 e selezionare, ad esempio, la macro denominata `mTopic`. Poi questa emissione potrebbe risultare in un modello di concetto come: `auto + cattiva`.

La maggior parte delle regole sono solo una riga di output ma ci sono momenti in cui più di un output è possibile e desiderato. In questo caso, definire un output per riga nella tabella Output della regola.

Importante! Tenere presente che le altre operazioni di gestione linguistica vengono eseguite durante l'estrazione dei modelli TLA. Pertanto quando l'output legge `t$3\t#3`, questo significa che il modello visualizza il concetto finale per il terzo elemento e il tipo finale per il terzo elemento dopo che è stata applicata tutta l'elaborazione linguistica (sinonimi e altri raggruppamenti).

- **Mostra output come.** Per impostazione predefinita, l'opzione **Riferimenti a riga nella tabella Valore regola** è selezionata e l'output viene visualizzato utilizzando i riferimenti numerici alla riga come definito nella scheda Valore di regola. Se è stato precedentemente selezionato Ottieni token si hanno token nella colonna Token di esempio nella tabella Valore di regola, è possibile scegliere di visualizzare l'output per questi token specifici scegliendo l'opzione.

Nota: se non ci sono abbastanza coppie di output concetto/tipo visualizzate nella tabella di output, è possibile aggiungere un'altra coppia facendo clic sul pulsante Aggiungi nella barra degli strumenti dell'editor. Se 3 coppie sono attualmente visualizzate e si fa clic su Aggiungi, altre 2 colonne (Concetto 4 e Tipo 4) vengono aggiunte alla tabella. Ciò significa che è possibile ora vedere 4 coppie nella tabella di output per tutte le regole. È anche possibile rimuovere coppie non utilizzate fino a che nessuna altra regola nella serie di regole di questa libreria non utilizzi quella coppia.

Esempio di regola

Si supponga che le risorse contengono la seguente regola di analisi di collegamento del testo e che è stata abilitata l'estrazione dei risultati TLA:

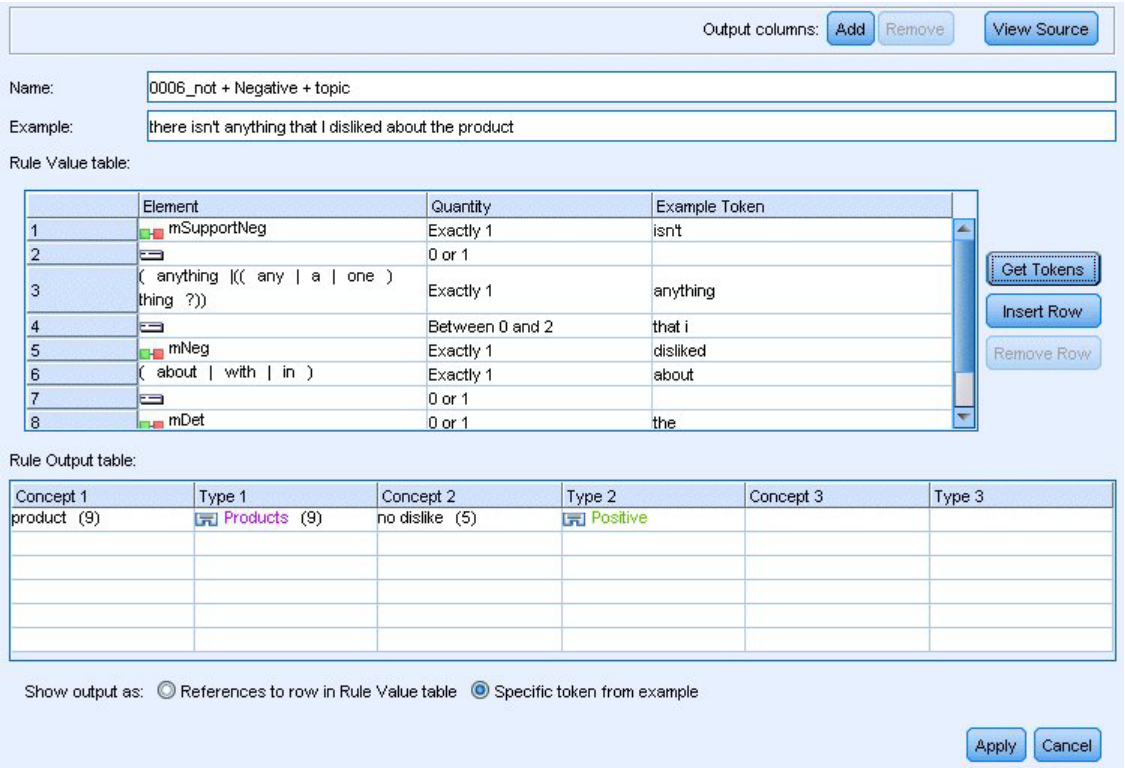


Figura 43. Scheda Regole di collegamento del testo: Editor di regole

Quando si estrae, il motore di estrazione legge ogni frase e cercherà la corrispondenza con la seguente sequenza:

Tabella 44. Esempio di sequenza di estrazione

Elemento (riga)	**Descrizione degli argomenti**
ERROR! SEGMENT DATA CORRUPTED, SEGDATA=1	Il concetto da uno dei tipi rappresentato dalla macro `mPos` o `mNeg` o dal tipo `<Uncertain>`.
2	Un concetto immesso come uno dei tipi rappresentati dalla macro `mTopic`.
3	Una delle parole rappresentate dalla macro `mBe`.
4	Un elemento facoltativo, 0 o 1 parole, anche noto come una differenza parola o `<Tutti i token>`
5	Un concetto immesso come uno dei tipi rappresentati dalla macro `mTopic`.

La tabella di output mostra che tutto ciò che è auspicato da questa regola è un modello in cui qualsiasi concetto o un tipo corrispondente alla macro `mTopic` che è stato definita nella riga 5 della tabella dei valori di regola + qualsiasi concetto o tipo corrispondente a `mPos`, `mNeg` o `<Uncertain>` come è stato definito nella riga 1 della tabella Valore di regola. Questo potrebbe essere `salsiccia + amo` o `<Sconosciuto> + <Positivo>`.

Creazione e modifica di regole

È possibile creare nuove regole o modificare quelle esistenti. Seguire le indicazioni e le descrizioni per l'editor di regole. Consultare la sezione “Gestione delle regole di collegamento del testo” a pagina 223 per ulteriori informazioni.

Creazione di nuove regole

1. Dai menu scegliere **Strumenti > Nuova regola**. In alternativa, fare clic sull'icona Nuova regola nella barra degli strumenti della struttura ad albero per aprire una nuova regola nell'editor.
2. Immettere un nome univoco e definire gli elementi di valore della regola.
3. Fare clic su **Applica** una volta terminato per verificare la presenza di errori.

Modifica regole

1. Fare clic sul nome regola nella struttura ad albero. La regola viene visualizzata nel riquadro dell'editor a destra.
2. Effettuare le modifiche.
3. Fare clic su **Applica** una volta terminato per verificare la presenza di errori.

Disattivazione ed eliminazione delle regole

Disattivazione delle regole

Se si desidera che una regola venga ignorata durante l'elaborazione, è possibile disabilitarla. Procedere con attenzione quando si eliminano e si disabilitano delle regole.

1. Fare clic sul nome regola nella struttura ad albero. La regola viene visualizzata nel riquadro dell'editor a destra.
2. Fare clic con il pulsante destro del mouse sul nome.
3. Dal menu di scelta rapida, scegliere **Disabilita**. L'icona di regola diventa grigia e la regola stessa diventa non modificabile.

Eliminazione delle regole

Se si desidera sbarazzarsi di una regola, è possibile eliminarla. Procedere con attenzione quando si eliminano e si disabilitano delle regole.

1. Fare clic sul nome regola nella struttura ad albero. La regola viene visualizzata nel riquadro dell'editor a destra.
2. Fare clic con il pulsante destro del mouse sul nome.
3. Dal menu di scelta rapida, scegliere **Elimina**. La regola scompare dall'elenco.

Controllo errori, salvataggio e annullamento

Applicazione delle modifiche della regola

Se si fa clic al di fuori dell'editor della regola o se si fa clic su **Applica**, la regola viene analizzata automaticamente per rilevare eventuali errori. Se viene trovato un errore, è necessario correggerlo prima di passare a un'altra parte dell'applicazione.

Tuttavia, se vengono rilevati errori meno gravi, viene fornita solo un'avvertenza. Ad esempio, se la regola contiene definizioni incomplete o senza riferimenti a tipi o altre macro, viene visualizzato un messaggio di avvertenza. Dopo aver fatto clic su **Applica**, qualsiasi avvertenza di errore causa un'icona di avvertenza che viene visualizzata alla sinistra del nome regola nella struttura ad albero nel riquadro a sinistra.

L'applicazione di una regola non significa che la regola viene salvata in modo permanente. L'applicazione farà sì che il processo di convalida verifichi la presenza di errori e avvertenze.

Salvataggio delle risorse all'interno di una sessione workbench interattiva

1. Per salvare le modifiche apportate alle risorse durante una sessione workbench interattiva in modo che sia possibile ottenerle la volta successiva che si esegue il flusso, è necessario:

- Aggiornare il nodo di modellazione per accertarsi che è possibile ottenere queste stesse risorse la volta successiva che si esegue il flusso. Consultare la sezione "Aggiornamento dei nodi di modellazione e salvataggio" a pagina 84 per ulteriori informazioni. Salvare poi il flusso. Per salvare il flusso, effettuare questa operazione nella finestra principale di IBM SPSS Modeler dopo aver aggiornato il nodo di modellazione.

2. Per salvare le modifiche apportate alle risorse durante una sessione workbench interattiva in modo che sia possibile utilizzarle in altri flussi, è possibile:
   - Aggiornare il modello utilizzato oppure crearne uno nuovo. Consultare la sezione "Creazione ed aggiornamento di modelli" a pagina 167 per ulteriori informazioni. Ciò non salverà le modifiche per il nodo corrente (vedi passo precedente)
   - In alternativa, aggiornare la TAP utilizzata. Consultare la sezione "Aggiornamento dei pacchetti di analisi del testo" a pagina 142 per ulteriori informazioni.

Salvataggio delle risorse all'interno di Editor di modelli

1. In primo luogo, pubblicare la libreria. Consultare la sezione "Pubblicazione delle librerie" a pagina 185 per ulteriori informazioni.
2. Salvare poi il modello mediante **File > Salva modello di risorsa** nel menu.

Annullamento delle modifiche di regola

1. Se si desidera annullare le modifiche, fare clic su **Annulla** nel riquadro dell'editor.

Ordine di elaborazione per le regole

Quando l'analisi di collegamento del testo viene eseguita durante l'estrazione, la "frase" (proposizione, parola, frase) verrà confrontata a turno con ogni regola fino a quando non viene trovata una corrispondenza o fino a che tutte le regole non sono state esaurite. La posizione nella struttura ad albero indica l'ordine in cui le regole vengono provate. La migliore prassi stabilisce che è necessario ordinare le regole dalla più specifica alla più generica. Quelle più specifiche devono trovarsi all'inizio della struttura ad albero. Per modificare l'ordine di una regola specifica o serie di regole, selezionare **Sposta su** o **Sposta giù** dal menu di scelta rapida della struttura ad albero regole e macro o le frecce su e giù nella barra degli strumenti.

Se si sta *nella vista di origine*, non è possibile modificare l'ordine delle regole spostandole nell'editor. Più in alto viene visualizzata la regola nella vista di origine, prima verrà elaborata. Si consiglia vivamente di riordinare solo nella struttura ad albero per evitare problemi di copia/incolla.

Importante! Nelle versioni precedenti di IBM SPSS Modeler Text Analytics, è stato richiesto di avere un ID regola univoco, numerico. A partire dalla versione di 17.1, è possibile indicare solo l'elaborazione dell'ordine spostando una regola in alto o in basso nella struttura ad albero o dalla loro posizione nella vista di origine.

Ad esempio, supponiamo che il testo contiene le seguenti due frasi:

Io amo le acciughe

Io amo le acciughe e i peperoni verdi

Inoltre, si supponga che esistano due regole di analisi di collegamento del testo con i seguenti valori:

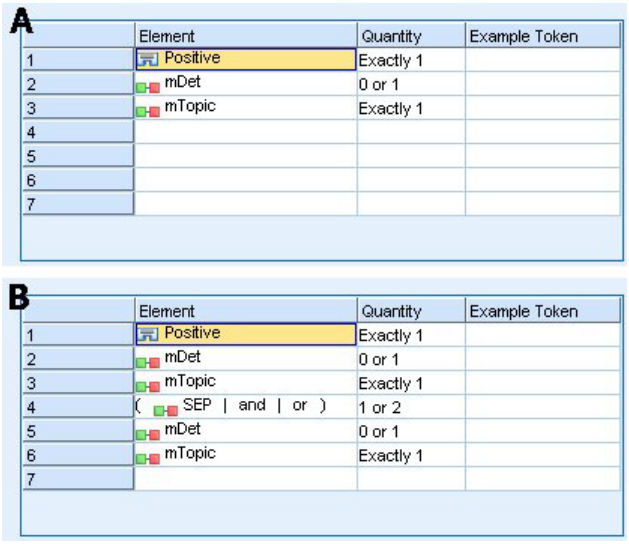


Figura 44. 2 Regole di esempio

Nella vista origine, i valori della regola potrebbero avere il seguente aspetto:

A: `value = $Positive $mDet? $mTopic`

B: `value = $Positive $mDet? $mTopic ($SEP|and|or){1,2} $mDet? $mTopic`

Se la regola **A** è più in alto nella struttura ad albero (più vicino alla cima) della regola **B** la regola **A** verrà elaborata per prima e la frase *Io amo le acciughe e il pepe verde* verrà prima accompagnata da `$Positive $mDet? $mTopic` e produrrà un output di modello incompleto (`acciughe + amo`) poiché era in corrispondenza con una regola che non cercava le corrispondenze 2 `$mTopic`.

Pertanto, per catturare l'essenza vera del testo, la regola più specifica, in questo caso **B** deve essere posizionata più in alto nella struttura ad albero rispetto a quella più generica, in questo caso la regola **A**.

Gestione delle serie di regole (più passi)

Una serie di regole è un modo utile per raggruppare una serie correlata di regole insieme nella struttura ad albero di regole e macro in modo da eseguire l'elaborazione a più passi. Una serie di regole non presenta nessuna definizione diversa da un nome ed è utilizzata per organizzare le proprie regole in gruppi significativi. In alcuni contesti, il testo è troppo ricco e variegato per essere elaborato in un solo passo. Ad esempio, quando si gestiscono i dati di intelligence sulla sicurezza, il testo potrebbe contenere collegamenti tra gli individui che vengono scoperti tramite metodi di contatto (*x chiama y*), tramite relazioni familiari (*y del cognato x*), mediante scambio di denaro (*x $100 collegati a y*), e così via. In questo caso, risulta utile creare serie specializzate di regole di analisi di collegamento del testo, ognuna delle quali è incentrata su un certo tipo di relazione come una per i contatti non protetti, un'altra per i membri della famiglia e così via.

Per creare una serie di regole, selezionare "Crea serie di regole" dal menu di scelta rapida della struttura ad albero regole e macro oppure dalla barra degli strumenti. È quindi possibile creare nuove regole direttamente in un nodo di serie di regole nella struttura ad albero o spostare le regole esistenti su una serie di regole.

Quando si esegue un'estrazione utilizzando le risorse in cui le regole vengono raggruppate in serie di regole, il motore di estrazione è costretto ad eseguire più passaggi attraverso il testo in modo da

corrispondere a diversi tipi di modelli in ogni passo. In questo modo, una "frase" può essere associata a una regola in ogni serie di regole mentre, senza una serie di regole, è possibile solo per una singola regola.

Nota: è possibile aggiungere fino a 512 regole per serie di regole.

Creazione di nuove serie di regole

1. Dai menu scegliere **Strumenti > Nuova serie di regole**. In alternativa, fare clic sull'icona Imposta nuova serie di regole nella barra degli strumenti della struttura ad albero. La serie di regole viene visualizzata nella struttura ad albero.
2. Aggiungere nuove regole per questa serie di regole o spostare le regole esistenti nella serie.

Disattivazione delle serie di regole

1. Fare clic con il tastino destro del mouse sul nome serie di regole nella struttura ad albero.
2. Dal menu di scelta rapida, scegliere **Disabilita**. L'icona della serie di regole diventa grigia e tutte le regole contenute all'interno di quella serie di regole sono anche disabilitate e ignorate durante l'elaborazione.

Eliminazione della serie di regole

1. Fare clic con il tastino destro del mouse sul nome serie di regole nella struttura ad albero.
2. Dal menu di scelta rapida, scegliere **Elimina**. La serie di regole e tutte le regole in essa contenute sono eliminate dalle risorse.

Elementi supportati per regole e macro

I seguenti argomenti sono accettati per i parametri di valore nelle regole di analisi del collegamento del testo e nelle macro:

Macro

È possibile utilizzare una macro direttamente in una regola di analisi del testo o all'interno di un'altra macro. Se si sta immettendo il nome macro manualmente o dall'interno della vista di origine (invece di selezionare il nome della macro da un menu di scelta rapida), assicurarsi di aggiungere un prefisso al nome con un carattere segno del dollaro ($), come `$mTopic`. Il nome macro è sensibile al maiuscolo/minuscolo. È possibile scegliere da qualsiasi macro definita nella scheda Regole di collegamento del testo corrente quando si selezionano le macro attraverso il menu di scelta rapida.

Tipi

È possibile utilizzare un tipo direttamente in una regola di analisi di collegamento del testo o all'interno di un'altra macro. Se si sta immettendo il nome tipo manualmente o dall'interno della vista di origine (invece di selezionare il nome tipo da un menu di scelta rapida), assicurarsi di aggiungere un prefisso al nome con un carattere segno del dollaro ($), come `$Persona`. Il nome tipo è sensibile al maiuscolo/minuscolo. Se si utilizza il menu di scelta rapida, è possibile scegliere da qualsiasi tipo dalla serie corrente di risorse utilizzate.

Se si fa riferimento a un tipo non riconosciuto, si riceverà un messaggio di avvertenza e la regola avrà un'icona di avvertenza nella struttura ad albero delle regole e macro fino a quando non verrà corretta.

Stringhe a valore letterale

Per includere le informazioni mai estratte, è possibile definire una stringa letterale per la quale il motore di estrazione effettuerà la ricerca. Tutte le parole o le frasi estratte sono state assegnate ad un tipo e per

questo motivo, non possono essere utilizzate in stringhe di costanti letterali. Se si utilizza una parola che è stata estratta essa sarà ignorata, anche se il suo tipo è `<Sconosciuto>`.

Una stringa letterale può essere una o più parole. Quando si definisce un elenco di stringhe a valore letterale, si applicano le eseguenti regole:

- Chiudere l'elenco di stringhe in parentesi come (suo). Se c'è una scelta di stringhe a valore letterale, ogni stringa deve essere separata dall'operatore OR, come `(a|da|il)` o `(suo|sua|loro)`.
- Utilizzare parole semplici o composte.
- Separare ciascuna parola nell'elenco con il carattere `|`, che è come un valore booleano `OR`.
- Immettere moduli singolari e plurali se si desidera la corrispondenza in entrambi. L'inflessione non viene generata automaticamente.
- Utilizzare solo caratteri minuscoli.
- Per riutilizzare le stringhe di a valore letterale, definirli come una macro e quindi utilizzare tale macro in altre macro e nelle regole dell'analisi del collegamento del testo.
- Se una stringa contiene punti o trattini, è necessario includerli. Ad esempio, per trovare `a.k.a` nel testo, immettere i punti con le lettere `a.k.a` come stringa a valore letterale.

Operatore di esclusione

Utilizzare `!` come operatore di esclusione per impedire a qualsiasi espressione della negazione di occupare un particolare slot. È possibile solo aggiungere un operatore di esclusione a mano attraverso la modifica della cella in linea (fare doppio clic sulla cella nella tabella del valore di regola o nella tabella del valore macro) o nella vista di origine. Ad esempio, se si aggiunge a `$mTopic @{0,2} !($Positivo) $Bilancio` alla propria regola di analisi di collegamento del testo, si sta cercando testo che contiene (1) un termine assegnato a uno qualsiasi dei tipi nella macro `mTopic`, (2) una differenza zero per due parole lunghe, (3) nessuna istanza di un termine assegnato al tipo `<Positivo>` e (4) un termine assegnato al tipo `<Bilancio>`. Ciò potrebbe rilevare "*le auto hanno un prezzo inflazionato*" ma ignorare "*il negozio offre sconti sorprendenti*".

Per utilizzare questo operatore, è necessario immettere il punto esclamativo e le parentesi manualmente nella cella elemento facendo doppio clic sulla cella.

Differenze tra parole (<tutti i token>)

Una differenza tra parole riferito anche come `<Tutti i token>`, definisce una serie numerica di token che possono essere presenti tra due elementi. Le distanze tra parole sono molto utili quando frasi molto simili corrispondenti possono differire solo leggermente a causa della presenza di determinativi aggiuntivi, frasi di preposizione, aggettivi o altre parole simili.

Tabella 45. Esempio degli elementi in una tabella di valore di regola senza la differenza tra parole

#	Elemento
ERROR! SEGMENT DATA CORRUPTED, SEGDATA=1	`Sconosciuto`
ERROR! SEGMENT DATA CORRUPTED, SEGDATA=2	`mBeHave`

Tabella 45. Esempio degli elementi in una tabella di valore di regola senza la differenza tra parole (Continua)

ERROR! SEGMENT DATA CORRUPTED, SEGDATA=3	 `Positivo`

Nota: nella vista origine questo valore è definito come: `$Unknown $mBeHave $Positive`

Questo valore corrisponde a frasi come "*il personale dell'hotel era simpatico*" in cui *il personale dell'hotel* appartiene al tipo `<Sconosciuto>`, *era* è nella macro `mBeHave` e *simpatico* è `<Positivo>`. Ma non restituirà "*il personale dell'hotel è stato molto simpatico*".

Tabella 46. Esempio degli elementi in una tabella di valore di regola con la differenza tra parole <Tutti i token>

#	Elemento
ERROR! SEGMENT DATA CORRUPTED, SEGDATA=1	 `Sconosciuto`
ERROR! SEGMENT DATA CORRUPTED, SEGDATA=2	 `mBeHave`
ERROR! SEGMENT DATA CORRUPTED, SEGDATA=3	
ERROR! SEGMENT DATA CORRUPTED, SEGDATA=4	 `Positivo`

Nota: nella vista origine questo valore è definito come: `$Unknown $mBeHave @{0,1} $Positive`

Se si aggiunge una differenza tra parole per il valore di regola, essa corrisponde a "*il personale dell'hotel è stato simpatico*" e "*il personale dell'hotel è stato molto simpatico*".

Nella vista origine o con la modifica in linea, la sintassi per una differenza tra parole è `@{#,#}`, dove `@` indica una differenza parole e `{#,#}` definisce il minimo e massimo di parole accettate tra l'elemento precedente e quello successivo. Ad esempio, `@{1,3}` indica che una corrispondenza può essere fatta tra due elementi definiti se esiste almeno una parola ma non più di tre parole che compare tra questi due elementi. `@{0,3}` indica che una corrispondenza può essere fatta tra due elementi definiti se c'è 0, 1, 2 o 3 parole presenti ma non più di tre parole.

Visualizzazione e gestione in modalità origine

Per ogni regola e macro l'editor TLA genera il codice di origine sottostante che viene utilizzato dall'Estrattore per la corrispondenza e la produzione di output TLA. Se si preferisce gestire il codice stesso, è possibile visualizzare questo codice origine e modificarlo direttamente facendo clic sul pulsante "Visualizza Origine" nella parte superiore dell'Editor. Viene riportata la vista Origine con evidenziata la regola o la macro attualmente selezionata. Tuttavia, si consiglia di utilizzare i pannelli dell'editor per ridurre la possibilità di errori.

Dopo aver visualizzato o modificato l'origine, fare clic su **Esci da origine**. Se genera sintassi non valida per una regola, verrà richiesto di correggere, prima di uscire dalla vista di origine.

Importante: Se si desidera modificare nella vista di origine, si consiglia di modificare le regole e le macro una alla volta. Dopo la modifica di una macro convalidare i risultati per l'estrazione. Se si è soddisfatti del risultato, si raccomanda di salvare il modello prima di apportare un'altra modifica. Se non si è soddisfatti del risultato o si verifica un errore, ripristinare le risorse salvate.

Macro nella vista Origine

```
[macro]
name = nome_macro
value = ([nome_tipo|nome_macro|stringa_letterale|divario_parole])
```

Tabella 47. Voci di macro

`[macro]`	Ciascuna macro deve iniziare con la riga contrassegnata [macro] per denotare l'inizio di una macro.
`name`	Il nome della definizione di macro. Il nome deve essere univoco.
`value`	Una combinazione di uno o più tipi, stringhe di costanti letterali, divari parola o macro. Consultare la sezione “Elementi supportati per regole e macro” a pagina 230 per ulteriori informazioni. Quando si combinano gli argomenti, è necessario utilizzare le parentesi ( ) per raggruppare gli argomenti e il carattere | per indicare un valore booleano OR.

Oltre alle linee guida e alla sintassi descritte nella sezione sulla macro, la vista di origine dispone di alcune indicazioni aggiuntive che non sono richieste quando si lavora nella vista dell'editor. Le macro devono anche rispettare le seguenti indicazioni quando si lavora in modalità origine:

- Ciascuna macro deve iniziare con la riga contrassegnata `[macro]` per denotare l'inizio di una macro.
- Per disabilitare un elemento, posizionare un indicatore di commento (#) prima di ogni riga.

Esempio. Questo esempio definisce una macro denominata `mTopic`. Il valore di `mTopic` è la presenza di un termine che corrisponde a *uno* dei seguenti tipi: `<Prodotto>`, `<Persona>`, `<Ubicazione>`, `<Organizzazione>`, `<Bilancio>` o `<Sconosciuto>`.

```
[macro]
name=mTopic
value=($Sconosciuto|$Prodotto|$Persona|$Ubicazione|$Organizzazione|$Bilancio|$Valuta)
```

Regole nella vista Origine

```
[pattern(ID)]
name = nome_modello
value = [$nome_tipo|nome_macro|divari_parole|stringhe_letterali]
output = $digit[\t]#digit[\t]$digit[\t]#digit[\t]$digit[\t]#digit[\t]
```

Tabella 48. Voci di regola

`[pattern (<ID>)]`	Indica l'inizio di una regola di analisi di collegamento del testo e fornisce un ID univoco numerico per stabilire l'ordine dell'elaborazione.

Tabella 48. Voci di regola (Continua)

`name`	Fornisce un nome univoco per questa regola di analisi di collegamento del testo.
`value`	Fornisce la sintassi e gli argomenti da associare al testo. Consultare la sezione "Elementi supportati per regole e macro" a pagina 230 per ulteriori informazioni.
`output`	Il formato di output per i modelli corrispondenti risultanti rilevati nel testo. L'output non sempre riprende l'esatta posizione originale degli elementi nel testo di origine. Inoltre, è possibile avere più righe di output per una regola di analisi di collegamento del testo inserendo ogni output su una riga separata. Sintassi per l'output: • Separare l'output con il codice scheda `\t`, ad esempio `$1\t#1\t$3\t#3` • `$` con un numero chiama il termine trovato corrispondente all'argomento definito nel parametro value in quella posizione. Pertanto `$1` indica il termine che corrisponde al primo argomento definiti per il valore. • `#` con un numero chiama il nome tipo dell'elemento in quella posizione. Se un elemento è un elenco di stringhe a valore letterale, verrà assegnato il tipo `<Sconosciuto>`. • Un valore `Null\tNull` non creerà alcun output.

Oltre alle indicazioni e alla sintassi descritte nella sezione sulle regole, la vista di origine dispone di alcune indicazioni aggiuntive che non sono richieste quando si lavora nella vista dell'editor. Le regole devono anche rispettare le seguenti indicazioni quando si lavora in modalità origine:

- Quando due o più elementi sono definiti, devono essere racchiusi tra parentesi se sono facoltativi (ad esempio, `($Negativo|$Positivo)` *o* `($mCoord|$SEP)?`). `$SEP` rappresenta una virgola.
- Il primo elemento in una regola di analisi di collegamento del testo non può essere un elemento facoltativo. Ad esempio, non è possibile iniziare con `value = $mTopic?` o `value = @{0,1}`.
- È possibile associare una quantità (o numero di istanza) a un token. Ciò è utile per scrivere una sola regola che includa tutti i casi invece di scrivere una regola separata per ciascun caso. Ad esempio, è possibile utilizzare la stringa di valore letterale `($SEP|and)` se si tenta di corrispondere a `,` (virgola) o `and`. Se si estende questa istruzione aggiungendo un quantitativo in modo che la stringa di valore letterale diventa `($SEP|and){1,2}`, essa corrisponde alle seguenti istanze: `","` `"and"` `", and"`.
- Gli spazi non sono supportati tra il nome macro e i caratteri `$` e `?` nel `valore` della regola di analisi di collegamento del testo.
- Gli spazi non sono supportati nell'`output` di analisi di collegamento del testo.
- Per disabilitare un elemento, posizionare un indicatore di commento (#) prima di ogni riga.

Esempio. Si supponga che le risorse contengono la seguente regola di analisi di collegamento del testo e che è stata abilitata l'estrazione dei risultati TLA:

```
## Jean Doe era l'ex direttore HR di IBM in Francia
[pattern(201)]
name= 1_201
value = $Persona ($SEP|$mDet|$mSupport|as|then){1,2} @{0,1} $Funzione
 (of|with|for|in|to|at) @{0,1} $Organizzazione @{0,2} $Ubicazione
output = $1\t#1\t$4\t#4\t$7\t#7\t$9\t#9
```

Quando si estrae, il motore di estrazione legge ogni frase e cerca la corrispondenza con la seguente sequenza:

Tabella 49. Esempio di sequenza di estrazione

Posizione	Descrizione degli argomenti
ERROR! SEGMENT DATA CORRUPTED, SEGDATA=1	Il nome di una persona ($Persona),
2	Uno dei seguenti: virgola ($SEP), determinativo ($mDet), verbo ausiliare ($mSupport), le stringhe "then" o "as",
3	0 o 1 parola (@{0,1})
4	Una funzione ($Funzione)
5	Una delle seguenti stringhe: "of", "with", "for", "in", "to" o "at",
6	0 o 1 parola (@{0,1})
7	Il nome di un'organizzazione ($Organizzazione)
8	0, 1 o 2 parole (@{0,2})
9	Il nome di un'ubicazione ($Ubicazione)

Questo esempio di regola di analisi di collegamento del testo può corrispondere a frasi come:

Jean Doe, il direttore HR di IBM in Francia

Jean Doe è stato il direttore HR di IBM in Francia

IBM ha nominato Jean Doe direttore HR di IBM in Francia

Questo esempio di regola di analisi del collegamento del testo produce il seguente output:

```
jean doe <Persona> direttore hr <Funzione> ibm
<Organizzazione> in francia <Ubicazione>
```

In cui:

- jean doe è il termine corrispondente a $1 (il primo elemento nella regola di analisi di collegamento del testo) e <Persona> è il tipo per jean doe (#1),
- direttore hr è il termine corrispondente a $4 (il quarto elemento nella regola di analisi di collegamento del testo) e <Funzione> è il tipo per direttore hr (#4),
- ibm è il termine corrispondente a $7 (il settimo elemento nella regola di analisi di collegamento del testo) e <Organizzazione> è il tipo per ibm. (#7),
- francia è il termine corrispondente a $9 (il primo elemento nella regola di analisi di collegamento del testo) e <Ubicazione> è il tipo per francia (#9)

Serie di regole nella vista Origine

```
[set(<ID>)]
```

[set (<ID>)] indica l'inizio di una serie di regole di analisi di collegamento del testo e fornisce un ID univoco numerico per stabilire l'ordine dell'elaborazione delle serie.

Esempio. La seguente frase contiene informazioni sugli individui, la loro funzione all'interno di un'azienda, ma anche le attività di fusione/acquisizione di società.

```
Org1 Inc ha stabilito un accordo di fusione definitiva con Org2 Ltd, ha dichiarato John Doe,
Amministratore delegato di Org2 Ltd.
```

È possibile scrivere una regola con più output per gestire tutti i possibili output come:

```
## Org1 Inc entered into a definitive merger agreement with Org2 Ltd, said
John Doe, CEO of Org2 Ltd.

[pattern(020)]
name=020
value = $Organization @{0,4} $ActionNouns @{0,6}  $mOrg @{1,2}
 $Person @{0,2} $Function @{0,1} $Organization
output = $1\t#1\t$3\t#3\t$5\t#5
output = $7\t#7\t$9\t#9\t$11\t#11
```

che produce i seguenti 2 modelli di output:

- `org1 inc<Organizzazione> + fusioni con <ActiveVerb> + org2 ltd<Organizzazione>`
- `john doe <Person> + ceo <Funzione> + org2 ltd<Organizzazione>`

Importante! Tenere presente che le altre operazioni di gestione linguistica vengono eseguite durante l'estrazione dei modelli TLA. In questo caso, `fusione` è raggruppato in `fusione` durante la fase di raggruppamento del processo di estrazione. E poiché `<fusione>` appartiene al tipo `ActiveVerb`, questo nome tipo è quello che viene visualizzato nell'output del modello TLA finale. Pertanto quando l'output legge `t$3\t#3` significa che il modello visualizza il concetto finale per il terzo elemento e il tipo finale per il terzo elemento dopo che è stata applicata tutta l'elaborazione linguistica (sinonimi e altri raggruppamenti).

Invece di scrivere le regole complesse come la precedente può essere più semplice gestire due regole. La prima è specializzata nello scoprire fusioni/acquisizioni fra società:

```
[set(1)]
## Org1 Inc has entered into a definitive merger agreement with Org2 Ltd
[pattern(44)]
name=firm + action + firm_0044
value=$mOrg @{0,20} $ActionNouns @{0,6} $mOrg
output(1)=$1\t#1\t$3\t#3\t$5\t#5
```

che dovrebbe produrre `org1 inc<Organizzazione> + fusioni con <ActiveVerb> + org2 ltd <Organizzazione>`

La seconda è specializzata in singolo/funzione/azienda:

```
[set(2)]
## said John Doe, CEO of Org2 Ltd
[pattern(52)]
name=individual + role + firm_0007
value=$Person @{0,3} $mFunction (at|of)? ($mOrg|$Media|$Unknown)
output(1)=$1\t#1\t$3\tFunction\t$5\t#5
```

che dovrebbe produrre `john doe <Persona> + ceo <Funzione> + org2 ltd <Organizzazione>`

Informazioni particolari

Queste informazioni sono sviluppate per prodotti e servizi offerti in tutto il mondo.

È possibile che IBM non offra i prodotti, servizi o funzioni illustrati in questa documentazione. Consultare il rappresentante locale IBM per le informazioni sui prodotti e servizi attualmente disponibili nella propria zona. Qualsiasi riferimento a un prodotto, programma o servizio IBM non implica o intende dichiarare che può essere utilizzato solo quel prodotto, programma o servizio IBM. In sostituzione a quelli forniti da IBM è possibile utilizzare qualsiasi prodotto, programma o servizio funzionalmente equivalente che non comporti la violazione dei diritti di proprietà intellettuale IBM o altri diritti. Tuttavia, è responsabilità dell'utente valutare e verificare il funzionamento di qualsiasi prodotto, programma o servizio non IBM.

IBM potrebbe avere brevetti o domande di brevetti in corso relativi ad argomenti discussi nella presente pubblicazione. La consegna del presente documento non conferisce alcuna licenza rispetto a questi brevetti. Rivolgere per iscritto i quesiti sulle licenze a:

IBM Director of Licensing
IBM Europe
Schoenaicher Str.220
D-7030 Boeblingen
Deutschland

Per richieste di licenze relative ad informazioni double-byte (DBCS) contattare il Dipartimento di Proprietà Intellettuale IBM nel proprio paese o inviare richieste per iscritto a:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502 Japan

Il seguente paragrafo non è valido nel Regno Unito o per tutti i paesi le cui leggi nazionali siano in contrasto con le disposizioni in esso contenute INTERNATIONAL BUSINESS MACHINES CORPORATION FORNISCE QUESTA PUBBLICAZIONE "NELLO STATO IN CUI ESSA SI TROVA" SENZA ALCUNA GARANZIA ESPLICITA O IMPLICITA IVI INCLUSE EVENTUALI GARANZIE DI COMMERCIABILITÀ ED IDONEITÀ AD UNO SCOPO PARTICOLARE Alcuni stati non consentono limitazioni di garanzie espresse o implicite in determinate transazioni, pertanto quanto sopra potrebbe non essere applicabile.

Le presenti informazioni possono includere imprecisioni tecniche o errori tipografici. Le modifiche periodiche apportate alle informazioni contenute in questa pubblicazione verranno inserite nelle nuove edizioni della pubblicazione. IBM si riserva il diritto di apportare miglioramenti e/o modifiche al prodotto o programma descritto in questa pubblicazione in qualsiasi momento e senza preavviso.

Qualsiasi riferimento nelle presenti informazioni a siti Web non IBM viene fornito esclusivamente per facilitare la consultazione e non rappresenta in alcun modo un'approvazione o sostegno da parte nostra di tali siti Web. I materiali disponibili sui siti Web non fanno parte di questo prodotto IBM e l'utilizzo di questi è a discrezione dell'utente.

IBM può utilizzare o distribuire qualsiasi informazione fornita dall'utente nel modo che ritiene più idoneo senza incorrere in alcun obbligo nei confronti dell'utente stesso.

Coloro che detengono la licenza di questo programma e che desiderano avere informazioni allo scopo di consentire: (i) lo scambio di informazioni tra programmi creati in modo indipendente e gli altri programmi (incluso questo) e (ii) l'utilizzo reciproco delle informazioni che sono state scambiate, devono contattare:

IBM Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
U.S.A.

Tali informazioni saranno fornite in conformità ai termini e alle condizioni in vigore e, in alcuni casi, dietro pagamento.

Il programma su licenza descritto in questa documentazione e tutto il materiale su licenza ad esso relativo vengono forniti da IBM nei termini del Customer Agreement IBM IBM International Program License Agreement o di eventuali accordi equivalenti intercorsi tra le parti.

Tutti i dati sulle prestazioni qui contenuti sono stati elaborati in ambiente controllato. Di conseguenza, i risultati ottenuti con sistemi operativi diversi possono variare in modo significativo. Alcune misurazioni potrebbero essere state effettuate su sistemi in corso di sviluppo e non c'è garanzia che tali misurazioni coincidano con quelle effettuate sui sistemi comunemente disponibili. Inoltre, alcune misurazioni potrebbero essere stime elaborate tramite l'estrapolazione. I risultati effettivi potrebbero variare. Gli utenti di questo documento devono verificare i dati relativi al proprio ambiente specifico.

le informazioni relative a prodotti non IBM sono state ottenute dai fornitori di tali prodotti, da loro annunci pubblicati e da altre fonti disponibili al pubblico. IBM non ha testato quei prodotti e non può garantire l'accuratezza delle prestazioni, la compatibilità o qualsiasi altra dichiarazione relativa a prodotti non-IBM. Eventuali domande in merito alle funzionalità dei prodotti non IBM vanno indirizzate ai fornitori di tali prodotti.

Qualsiasi affermazione relativa agli obiettivi e alla direzione futura di IBM è soggetta a modifica o revoca senza preavviso e concerne esclusivamente gli scopi dell'azienda.

Le presenti informazioni includono esempi di dati e report utilizzati in operazioni di business quotidiane. Per fornire una descrizione il più possibile esaustiva, gli esempi includono nomi di persone, società, marchi e prodotti. Tutti questi nomi sono fittizi e ogni somiglianza a nomi e indirizzi utilizzati da aziende reali è puramente casuale.

Per chi visualizza queste informazioni a video: le fotografie e le illustrazioni a colori potrebbero non essere disponibili.

Marchi

IBM, il logo IBM e ibm.com sono marchi o marchi registrati di International Business Machines Corp., registrati in molte giurisdizioni nel mondo. Altri nomi di prodotti e servizi possono essere marchi di IBM o altre società. Un elenco aggiornato di marchi IBM è disponibile sul web "Copyright and trademark information" all'indirizzo www.ibm.com/legal/copytrade.shtml.

Intel, il logo Intel, Intel Inside, il logo Intel Inside, Intel Centrino, il logo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium e Pentium sono marchi o marchi registrati di Intel Corporation o delle sue consociate negli Stati Uniti e in altri paesi.

Linux è un marchio registrato di Linus Torvalds negli Stati Uniti e/o negli altri paesi.

Microsoft, Windows, Windows NT e il logo Windows sono marchi di Microsoft Corporation negli Stati Uniti e/o negli altri paesi.

UNIX è un marchio registrato di The Open Group negli Stati Uniti e in altri paesi.

Java e tutti i marchi e logo basati su Java sono marchi o marchi registrati di Oracle e/o suoi affiliati.

Altri nomi di prodotti e servizi possono essere marchi commerciali di IBM o di altre aziende.

Indice analitico

Caratteri speciali

A

C

D

E

F

G

H

I

L

M

N

O

P

R

S

T

U

V

W

Stampato in Italia