IBM SPSS Modeler Text Analytics 17.1 사용자 안내서

참고

이 정보와 이 정보가 지원하는 제품을 사용하기 전에, 반드시 247 페이지의 『주의사항』의 정보를 읽으십시오.

제품 정보

이 개정판은 새 개정판에 별도로 명시하지 않는 한, 버전 17.1, 릴리스 0, IBM SPSS Modeler Text Analytics 수정 0 및 모든 후속 릴리스와 수정에 적용됩니다.

목차

서론

IBM® SPSS® Modeler Text Analytics는 고급 언어학적 기술 및 자연어 처리(NLP)를 사용하여 다양한 비정형 텍스트 데이터를 빠르게 처리하고, 이 텍스트에서 주요 개념을 추출 및 구성하는 강력한 텍스트 분석 기능을 제공합니다. 게다가, IBM SPSS Modeler Text Analytics는 이러한 개념을 범주로 그룹화할 수 있습니다.

조직 내에 보유된 데이터의 약 80%는 텍스트 문서 양식(예: 보고서, 웹 페이지, 이메일 및 콜센터 노트)으로 되어 있습니다. 텍스트는 조직이 해당 고객의 행동을 잘 이해할 수 있도록 할 때 핵심 요인입니다. NLP를 통합하는 시스템은 복합 구문을 포함하여 개념을 지능적으로 추출할 수 있습니다. 또한 의미와 컨텍스트를 사용하여, 기본적인 언어에 대한 지식을 통해 제품, 조직, 또는 사람과 같은, 관련 그룹으로 용어를 분류할 수 있습니다. 결과적으로, 신속하게 필요성에 대한 정보의 관련성을 판별할 수 있습니다. 추출된 개념과 범주는 인구통계와 같은 기존의 구조화된 데이터와 결합할 수 있고 보다 나은 집중적인 의사결정을 내리기 위해 전체 IBM SPSS Modeler 데이터 마이닝 도구 세트에서 모델링에 적용할 수 있습니다.

언어학적 시스템은 지식에 민감하며 사전에 더 많은 정보가 포함될수록 결과의 품질이 높아집니다. IBM SPSS Modeler Text Analytics는 용어 및 동의어, 라이브러리 및 템플리트와 같은 언어학적 자원 세트와 함께 전달됩니다. 이 제품은 추가적으로 컨텍스트에 대해 이러한 언어학적 자원을 개발하고 세분화할 수 있도록 합니다. 언어학적 자원의 미세한 조정은 종종 반복적 프로세스로, 정확한 개념 검색 및 범주화에 필요합니다. CRM 및 유전체학과 같은, 사용자 정의 템플리트, 라이브러리 및 특정 도메인용 사전도 포함됩니다.

IBM Business Analytics 소개

IBM Business Analytics 소프트웨어는 의사 결정자가 비즈니스 성능을 개선하기 위해 신뢰하는 완벽하고 일관되며 정확한 정보를 제공합니다. 비즈니스 지능, 예측 분석, 금융 성과와 전략 관리 및 분석 응용 프로그램의 종합 포트폴리오는 현재 성과와 앞으로의 결과를 예측하는 능력에 분명하고 즉각적이면서 실행 가능한 통찰력을 제공합니다. 풍부한 업계 솔루션, 입증된 사례 및 전문 서비스가 결합되어 어떠한 크기의 조직이라도 생산성을 극대화하고 자신있는 자동 결정을 내릴 수 있으며 더 나은 결과를 가져올 수 있습니다.

이 포트폴리오의 일부인 IBM SPSS Predictive Analytics 소프트웨어를 통해 조직은 미래의 사건을 예측하고 더 나은 비즈니스 결과를 얻기 위한 통찰력에 대해 적극적인 조치를 할 수 있습니다. 전 세계의 기업, 정부 및 학계 고객들은 고객을 매료시키고 유지하며 성장하게 만드는 동시에 불공정 행위를 줄이고 위험을 낮추는 IBM SPSS 기술의 경쟁 이점을 활용합니다. 일상 업무에서 IBM SPSS 소프트웨어를 활용한다면 예측형 기업으로 거듭날 수 있습니다. 즉 비즈니스 목표 달성을 위해 의사 결정의 방향을 정하고 이를 자동화하며 측정 가능한 경쟁 우위를 달성할 수 있습니다. 자세한 내용을 보거나 담당자에게 문의하려면 http://www.ibm.com/spss 사이트를 방문하십시오.

기술 지원

기술 지원은 유지 관리 고객에게 제공됩니다. IBM Corp. 제품 사용 및 지원된 하드웨어 환경 중 하나에 대해 설치하는 데 도움이 필요한 경우 기술 지원부로 문의하십시오. 기술 지원에 문의하려면 http://www.ibm.com/support의 IBM Corp. 웹 사이트를 참고하십시오. 지원을 요청하려면 본인의 신상과 소속 조직(회사) 및 지원 동의서를 제시해야 합니다.

제 1 장 IBM SPSS Modeler Text Analytics 정보

IBM SPSS Modeler Text Analytics는 고급 언어학적 기술 및 자연어 처리(NLP)를 사용하여 다양한 비정형 텍스트 데이터를 빠르게 처리하고, 이 텍스트에서 주요 개념을 추출 및 구성하는 강력한 텍스트 분석 기능을 제공합니다. 게다가, IBM SPSS Modeler Text Analytics는 이러한 개념을 범주로 그룹화할 수 있습니다.

조직 내에 보유된 데이터의 약 80%는 텍스트 문서 양식(예: 보고서, 웹 페이지, 이메일 및 콜센터 노트)으로 되어 있습니다. 텍스트는 조직이 해당 고객의 행동을 잘 이해할 수 있도록 할 때 핵심 요인입니다. NLP를 통합하는 시스템은 복합 구문을 포함하여 개념을 지능적으로 추출할 수 있습니다. 또한 의미와 컨텍스트를 사용하여, 기본적인 언어에 대한 지식을 통해 제품, 조직, 또는 사람과 같은, 관련 그룹으로 용어를 분류할 수 있습니다. 결과적으로, 신속하게 필요성에 대한 정보의 관련성을 판별할 수 있습니다. 추출된 개념과 범주는 인구통계와 같은 기존의 구조화된 데이터와 결합할 수 있고 보다 나은 집중적인 의사결정을 내리기 위해 전체 IBM SPSS Modeler 데이터 마이닝 도구 세트에서 모델링에 적용할 수 있습니다.

언어학적 시스템은 지식에 민감하며 사전에 더 많은 정보가 포함될수록 결과의 품질이 높아집니다. IBM SPSS Modeler Text Analytics는 용어 및 동의어, 라이브러리 및 템플리트와 같은 언어학적 자원 세트와 함께 전달됩니다. 이 제품은 추가적으로 컨텍스트에 대해 이러한 언어학적 자원을 개발하고 세분화할 수 있도록 합니다. 언어학적 자원의 미세한 조정은 종종 반복적 프로세스로, 정확한 개념 검색 및 범주화에 필요합니다. CRM 및 유전체학과 같은, 사용자 정의 템플리트, 라이브러리 및 특정 도메인용 사전도 포함됩니다.

배포. 비정형 데이터의 실시간 스코어링을 위해 IBM SPSS Modeler Solution Publisher를 사용하여 텍스트 마이닝 스트림을 배치할 수 있습니다. 이들 스트림을 배치하는 기능은 성공적인 폐쇄 루프 텍스트 마이닝 구현을 보장합니다. 예를 들어, 사용자 조직이 이제 예측 모형을 적용하여 마케팅 메시지의 정확도를 실시간으로 늘려서 인바운드 또는 아웃바운드 호출자의 메모철을 분석할 수 있습니다.

참고: IBM SPSS Modeler Solution Publisher와 함께 IBM SPSS Modeler Text Analytics를 실행하려면 `<install_directory>/ext/bin/spss.TMWBServer` 디렉토리를 `$LD_LIBRARY_PATH` 환경 변수에 추가하십시오.

지원되는 언어의 자동화된 변환. IBM SPSS Modeler Text Analytics는 SDL의 SaaS(Software as a Service)와 함께 사용하여 아랍어, 중국어 및 페르시아어를 포함한 지원되는 언어의 목록에서 영어로 텍스트를 변환할 수 있습니다. 그런 다음 변환된 텍스트에 대해 텍스트 분석을 수행하고 이들 결과를 소스 언어의 내용을 이해할 수 없었던 사람에게 배치할 수 있습니다. 텍스트 마이닝 결과가 자동으로 다시 대응하는 외국어 텍스트에 링크되므로, 조직은 훨씬 필요한 모국어 발표자 자원을 분석의 더 중요한 결과에만 집중할 수 있습니다. SDL은 연인원 20명의 고급 변환 연구의 결과인 통계적 변환 알고리즘을 사용한 자동 언어 변환을 제공합니다.

IBM SPSS Modeler Text Analytics 버전 17.1로 업그레이드

Clementine용 텍스트 마이닝 또는 PASW 텍스트 분석의 이전 버전에서 업그레이드합니다.

IBM SPSS Modeler Text Analytics 버전 17.1을 설치하기 전에, 새 버전에서 사용할 TAP, 템플리트 및 라이브러리를 현재 버전에서 저장하고 내보내야 합니다. 이러한 파일은 최신 버전을 설치할 때 삭제되거나 덮어쓰지 않는 디렉토리에 저장하는 것이 좋습니다.

IBM SPSS Modeler Text Analytics의 최신 버전을 설치한 후 저장한 TAP 파일을 로드하거나, 저장한 라이브러리를 추가하거나, 저장한 템플리트를 가져오고 로드하여 최신 버전에서 사용할 수 있습니다.

중요사항: 먼저 필요한 파일을 저장하고 내보내지 않고 현재 버전을 제거할 경우 이전 버전에서 수행한 모든 TAP, 템플리트 및 공용 라이브러리 작업은 손실되며 IBM SPSS Modeler Text Analytics 버전 17.1에서 사용할 수 없습니다.

텍스트 마이닝 정보

오늘날 점점 늘어나는 정보의 양이 구조화되지 않은 준구조화된 형식으로 보존되어 있습니다. 예를 들어, 고객 이메일, 콜 센터 메모, 개방형 설문 반응, 뉴스 피드, 웹 양식 등입니다. 이러한 정보의 풍요는 많은 조직에게 다음과 같은 질문을 던지는 문제점을 야기합니다. "이 정보를 어떻게 수집, 탐색하고 활용할 수 있습니까?"

*텍스트 마이닝*은 작성자가 개념을 표현하기 위해 사용한 정확한 단어와 용어를 모르더라도 주요 개념과 테마를 캡처하고 숨겨진 관계와 경향을 발견하기 위해 텍스트 자료 콜렉션을 분석하는 프로세스입니다. 텍스트 마이닝은 정보 검색과는 상당히 다르기는 하지만 종종 혼동되기도 합니다. 정확한 검색과 정보 저장은 엄청난 도전인 반면에 이러한 정보에 포함된 품질 컨텐츠, 용어 및 관계의 추출과 관리는 결정적이고 중요한 프로세스입니다.

텍스트 마이닝 및 데이터 마이닝

텍스트의 각 기사마다 언어학적 기반 텍스트 마이닝은 개념 지수뿐만 아니라 이러한 개념에 대한 정보를 리턴합니다. 이 순화된 구조화된 정보는 다른 데이터 소스와 결합되어 다음과 같은 질문을 처리할 수 있습니다.

- 어떤 개념이 함께 발생합니까?
- 그 밖에 어디에 링크되어 있습니까?
- 추출된 정보로부터 어떤 상위 수준 범주를 작성할 수 있습니까?
- 개념 또는 범주가 예상하는 것은 무엇입니까?
- 개념 또는 범주가 작동을 어떻게 예상합니까?

텍스트 마이닝을 데이터 마이닝과 결합하면 구조화된 또는 구조화되지 않은 데이터에서만 사용 가능한 것보다 더 많은 통찰력을 제공합니다. 이 프로세스는 일반적으로 다음 단계를 포함합니다.

1. **마이닝할 텍스트를 식별하십시오.**텍스트 마이닝을 준비하십시오. 텍스트가 여러 파일에 존재하면 파일을 한 위치에 저장하십시오. 데이터베이스의 경우 텍스트를 포함하는 필드를 판별하십시오.
2. **텍스트를 마이닝하고 구조화된 데이터를 추출하십시오.** 텍스트 마이닝 알고리즘을 소스 텍스트에 적용하십시오.
3. **개념 및 범주 모델을 작성하십시오.** 주요 개념을 식별하고 범주를 작성하십시오. 구조화되지 않은 데이터로부터 리턴된 개념 수는 일반적으로 매우 큽니다. 스코어링을 위해 최상의 개념과 범주를 식별하십시오.

4. **구조화된 데이터를 분석하십시오.** 군집, 분류 및 예측 모델링과 같은 일반적인 데이터 마이닝 기술을 사용하여 개념 간의 관계를 발견하십시오. 추출된 개념을 다른 구조화된 데이터와 병합하여 개념을 기반으로 추가로 동작을 예측하십시오.

텍스트 분석 및 범주화

질적 분석의 양식으로 된 텍스트 분석은 이 텍스트에 포함된 주요 아이디어와 개념이 적합한 개수의 범주로 그룹화될 수 있도록 텍스트로부터의 유용한 정보의 추출입니다. 텍스트 분석은 분석의 접근 방법은 다소 다르더라도 모든 유형과 텍스트 길이에서 수행될 수 있습니다.

짧은 레코드 또는 문서가 가장 쉽게 범주화됩니다. 이들은 복잡하지 않고 일반적으로 애매한 단어나 반응이 적기 때문입니다. 예를 들어, 짧은 개방형 설문 질문에서 사람들에게 좋아하는 세 가지의 휴가 활동을 꼽으라고 물으면 *해변에 가기*, *국립공원 방문* 또는 *아무것도 안하기* 등과 같은 짧은 답변을 여러 개 예상할 수 있습니다. 반면 더 긴 개방형 반응은 복잡하고 길 수 있으며 반응자가 교육을 많이 받았고, 동기가 있고, 설문지를 작성할 시간이 충분한 경우에는 특히 그렇습니다. 설문에서 사람들에게 자신의 정치적 신념에 대해 얘기해 달라고 묻거나 정치에 대한 블로그 피드를 가지도록 요청하면 모든 종류의 문제와 위치에 대해 다소 긴 설명을 예측할 수 있습니다.

단시간에 주요 개념을 추출하고 이러한 긴 텍스트 소스로부터 통찰력있는 범주를 작성하는 기능은 IBM SPSS Modeler Text Analytics 사용의 주요 장점입니다. 이 장점은 텍스트 분석 프로세스의 각 단계마다 가장 안정적인 결과를 내기 위해 자동화된 언어학적 및 통계적 기술의 결합을 통해 획득됩니다.

언어학적 처리와 NLP

이 구조화되지 않은 모든 텍스트 데이터 관리의 주요 문제점은 컴퓨터가 이해할 수 있도록 텍스트를 쓰기 위한 표준 규칙이 없다는 점입니다. 언어, 따라서 의미는 모든 문서 및 모든 텍스트마다 다릅니다. 이러한 구조화되지 않은 데이터를 정확하게 검색하고 조직하는 유일한 방법은 언어를 분석하고 해당 의미를 발견하는 것입니다. 구조화되지 않은 정보로부터 개념 추출을 위한 여러 개의 자동화된 방법이 있습니다. 이러한 접근 방법은 언어학적 및 비언어학적의 두 가지 종류로 구분할 수 있습니다.

몇몇 조직에서는 통계 및 신경망을 기반으로 자동화된 비언어학적 솔루션을 사용하려고 시도했습니다. 컴퓨터 기술을 사용하면 이러한 솔루션은 사람보다 더 빨리 주요 개념을 스캔하고 범주화할 수 있습니다. 불행하게도 이러한 솔루션의 정확도는 매우 낮습니다. 대부분의 통계 기반 시스템은 단순히 단어가 발생한 횟수를 세고 관련 개념에 대한 통계적 인접성을 계산합니다. 이는 관련되지 않은 결과 또는 잡음을 생성하고, 반드시 있어야 하는 결과가 누락되고 침묵으로 처리됩니다.

정확도의 한계를 보충하기 위해서 몇몇 솔루션은 관련 결과와 비관련 결과를 구분하는 데 도움이 되는 복잡한 비언어 규칙을 사용합니다. 이를 *규칙 기반 텍스트 마이닝*이라고 부릅니다.

반면, *언어학적 기반 텍스트 마이닝*은 자연어 처리(NLP)-인간 언어의 컴퓨터 지원 분석-의 원칙을 텍스트의 단어, 구문 및 명령문 또는 구조에 적용합니다. NLP를 통합하는 시스템은 복합 구문을 포함하여 개념을 지능적으로 추출할 수 있습니다. 게다가, 기본 언어 지식을 사용하면 의미 및 컨텍스트를 사용하여 개념을 제품, 조직 또는 사람 등과 같은 관련 그룹으로 분류할 수 있습니다.

언어학적 기반 텍스트 마이닝은 방대한 단어 양식을 유사한 의미가 있는 것으로 인식하고 문장 구조를 분석하여 텍스트 이해를 위한 프레임워크를 제공하여 사람들이 하는 방법으로 텍스트에서 많은 의미를 찾아냅니다. 이 방법은 통계 기반 시스템의 속도와 비용 효율성을 제공하지만 사람의 개입은 덜 요구하면서 훨씬 더 높은 수준의 정확도를 제공합니다.

일본어를 제외한 모든 언어 텍스트에서 추출 프로세스 중에 통계 기반과 언어학적 기반 접근 방식 간의 차이를 설명하려면 reproduction of documents에 대한 쿼리에 대해 반응하는 방법을 고려하십시오. 통계 기반 및 언어학적 기반 솔루션 둘 모두는 reproduction 단어를 copy 및 duplication 등과 같은 동의어를 포함하기 위해 확장해야 합니다. 그렇지 않으면 관련 정보를 빠뜨리게 됩니다. 그러나 통계 기반 솔루션이 의미가 같은 다른 용어에 대해 이 유형의 동의어 검색을 하려고 시도하면 birth 용어 또한 포함하려고 하므로 관련이 없는 결과가 많이 생성됩니다. 다시 말해서 언어의 이해는 텍스트의 모호성을 극복하여 언어학적 기반 텍스트 마이닝을 보다 믿을 만한 방법으로 만들어 줍니다.

정서 분석기를 통해 언어학적 기반 기술을 사용하면 보다 의미있는 표현식을 추출할 수 있습니다. 감정의 분석과 캡처는 텍스트의 모호성을 극복하고, 다시 말해서 언어학적 기반 텍스트 마이닝을 보다 안정적인 방법으로 만들어 줍니다.

추출 프로세스의 작동 방법을 이해하면 언어학적 자원(라이브러리, 유형, 동의어 등)을 세부 조정할 때 중요한 결정을 내리는 데 도움이 됩니다. 추출 프로세스의 단계는 다음을 포함합니다.

- 소스 데이터를 표준 형식으로 변환
- 후보 항 식별
- 동의어의 동등 클래스 및 통합 식별
- 유형 지정
- 색인화 및 요청 시에 2차 분석기와 패턴 매치

1단계. 소스 데이터를 표준 형식으로 변환

이 첫 번째 단계에서, 사용자가 가져오는 데이터가 추가 분석에 사용될 수 있는 균일한 형식으로 변환됩니다. 이 변환은 내부적으로 수행되므로 원래 데이터를 변경하지 않습니다.

2단계. 후보 항 식별

언어학적 추출 중에 후보 항의 식별에서 언어학적 자원의 역할을 이해하는 것이 중요합니다. 언어학적 자원은 추출이 실행될 때마다 사용됩니다. 이들은 템플리트, 라이브러리 및 컴파일된 자원의 양식으로 존재합니다. 라이브러리에는 단어 목록, 관계 및 추출을 지정하거나 조정하는 데 사용되는 기타 정보가 포함됩니다. 컴파일된 자원은 보거나 편집할 수 없습니다. 그러나 나머지 자원은 템플리트 편집기에서나 대화식 워크벤치 세션에 있는 경우에는 자원 편집기에서 편집할 수 있습니다.

컴파일된 자원은 IBM SPSS Modeler Text Analytics 내에서 추출 엔진의 핵심적인 내부 구성요소입니다. 이러한 자원에는 품사 코드(명사, 동사, 형용사 등)가 있는 기본 양식 목록을 포함하는 일반 사전이 포함됩니다.

컴파일된 자원 외에, 여러 개의 라이브러리가 제품과 함께 제공되며 컴파일된 자원에서 유형 및 개념 정의를 보완하고 동의어를 제공하는 데 사용될 수 있습니다. 이러한 라이브러리 및 사용자가 작성하는 사용자 정의 라이브러리는 몇몇 사전으로 구성됩니다. 여기에는 유형 사전, 동의어 사전 및 제외 사전이 포함됩니다.

데이터를 가져와서 변환한 후 추출 엔진이 추출을 위한 후보 항 식별을 시작합니다. 후보 항은 텍스트에서 개념을 식별하는 데 사용되는 단어나 단어 그룹입니다. 텍스트를 처리하는 동안 단일 단어(**단일어**) 및 복합어(**다항어**)가 품사 패턴 추출기를 사용하여 식별됩니다. 그런 다음, 후보 정서 키워드는 정서 텍스트 링크 분석을 사용하여 식별됩니다.

참고: 앞서 언급한 컴파일된 일반 사전에 있는 용어는 관심이 없거나 언어학적으로 단일어로서는 애매한 모든 단어 목록을 나타냅니다. 이러한 단어는 단일어를 식별할 때 추출에서 제외됩니다. 그러나, 품사를 판별할 때나 더 긴 후보 복합 단어(다항어)를 찾을 때 다시 평가됩니다.

3단계. 동의어의 동등 클래스 및 통합 식별

후보 단일어 및 다항어가 식별된 후 소프트웨어는 정규화 사전을 사용하여 동등 클래스를 식별합니다. 동등 클래스는 한 구문의 기본 양식이거나 동일 구문에 대한 두 개의 변형이 있는 단일 양식입니다. 동등 클래스에 사용할 개념(을 판별하기 위해 추출 엔진은 다음 규칙을 나열된 순서대로 적용합니다.

- 라이브러리의 사용자 지정 양식.
- 사전에 컴파일된 자원으로 정의되는 최대 빈도 양식.

4단계. 유형 지정

다음으로 유형은 추출된 개념에 지정됩니다. 유형은 개념의 시맨틱 그룹입니다. 이 단계에서는 컴파일된 자원과 라이브러리 둘 모두가 사용됩니다. 유형은 상위 레벨 개념, 긍정적 및 부정적 단어, 이름, 장소, 조직 등과 같은 것을 포함합니다. 자세한 정보는 197 페이지의 『유형 사전』의 내용을 참조하십시오.

일본어 자원에는 특징적인 유형 세트가 있음을 유의하십시오.

언어학적 시스템은 지식에 민감하며 사전에 더 많은 정보가 포함될수록 결과의 품질이 높아집니다. 동의어 정의 등과 같이 사전 컨텐츠의 수정은 결과로 나오는 정보를 단순화할 수 있습니다. 이는 종종 반복적인 프로세스이며 정확한 개념 검색을 위해 필요합니다. NLP는 IBM SPSS Modeler Text Analytics의 코어 요소입니다.

추출 작동 방법

반응으로부터 주요 개념과 아이디어를 추출하는 동안 IBM SPSS Modeler Text Analytics 은 언어학적 기반 텍스트 분석에 의존합니다. 이 접근 방법은 통계 기반 시스템의 속도와 비용 효율성을 제공합니다. 그러나 인간의 개입은 덜 요구하면서 훨씬 더 높은 수준의 정확도를 제공합니다. 언어 기반 텍스트 분석은 자연어 처리로 알려지고 계산 언어학으로도 알려진 연구 분야를 기반으로 합니다.

중요사항: 일본어 텍스트의 경우 추출 프로세스는 다른 단계 세트를 따릅니다.

추출 프로세스의 작동 방법을 이해하면 언어학적 자원(라이브러리, 유형, 동의어 등)을 세부 조정할 때 중요한 결정을 내리는 데 도움이 됩니다. 추출 프로세스의 단계는 다음을 포함합니다.

- 소스 데이터를 표준 형식으로 변환
- 후보 항 식별
- 동의어의 동등 클래스 및 통합 식별
- 유형 지정
- 색인화
- 패턴 및 이벤트 추출 매치

1단계. 소스 데이터를 표준 형식으로 변환

이 첫 번째 단계에서, 사용자가 가져오는 데이터가 추가 분석에 사용될 수 있는 균일한 형식으로 변환됩니다. 이 변환은 내부적으로 수행되므로 원래 데이터를 변경하지 않습니다.

2단계. 후보 항 식별

언어학적 추출 중에 후보 항의 식별에서 언어학적 자원의 역할을 이해하는 것이 중요합니다. 언어학적 자원은 추출이 실행될 때마다 사용됩니다. 이들은 템플리트, 라이브러리 및 컴파일된 자원의 양식으로 존재합니다. 라이브러리에는 단어 목록, 관계 및 추출을 지정하거나 조정하는 데 사용되는 기타 정보가 포함됩니다. 컴파일된 자원은 보거나 편집할 수 없습니다. 그러나 나머지 자원(템플리트)은 템플리트 편집기에서나 대화식 워크벤치 세션에 있는 경우에는 자원 편집기에서 편집할 수 있습니다.

컴파일된 자원은 IBM SPSS Modeler Text Analytics 내에서 추출 엔진의 핵심 내부 구성요소입니다. 이러한 자원에는 품사 코드(명사, 동사, 형용사, 부사, 분사, 등위 접속사, 관사 또는 전치사)의 기본 양식 목록을 포함하는 일반 사전을 포함합니다.자원은 또한 다음과 같은 `<Location>`, `<Organization>` 또는 `<Person>` 유형에 많은 추출 항을 지정하는 데 사용되는 예약된 내장된 유형을 포함합니다. 자세한 정보는 198 페이지의 『내장 유형』의 내용을 참조하십시오.

컴파일된 자원 외에, 여러 개의 라이브러리가 제품과 함께 제공되며 컴파일된 자원에서 유형 및 개념 정의를 보완하고 동의어를 제공하는 데 사용될 수 있습니다. 이러한 라이브러리 및 사용자가 작성하는 사용자 정의 라이브러리는 몇몇 사전으로 구성됩니다. 여기에는 유형 사전, 대체 사전(동의어 및 선택적 요소) 및 제외 사전이 포함됩니다. 자세한 정보는 187 페이지의 제 16 장 『라이브러리에 대한 작업』의 내용을 참조하십시오.

데이터를 가져와서 변환한 후 추출 엔진이 추출을 위한 후보 항 식별을 시작합니다. 후보 항은 텍스트에서 개념을 식별하는 데 사용되는 단어나 단어 그룹입니다. 텍스트 처리 동안에는 컴파일된 자원에 있지 않은 단일 단어(*단일어*)는 후보 항 추출로 간주됩니다. 후보 복합어(*다항어*)는 품사 패턴 추출기를 사용하여 식별됩니다. 예를 들어, "형용사 명사" 품사 패턴을 따르는 다항어 `sports car`에는 두 개의 구성요소가 있습니다. "형용사 형용사 명사" 품사 패턴을 따르는 다항어 `fast sports car`에는 세 개의 구성요소가 있습니다.

참고: 앞서 언급한 컴파일된 일반 사전에 있는 용어는 관심이 없거나 언어학적으로 단일어로서는 애매한 모든 단어 목록을 나타냅니다. 이러한 단어는 단일어를 식별할 때 추출에서 제외됩니다. 그러나, 품사를 판별할 때나 더 긴 후보 복합 단어(다항어)를 찾을 때 다시 평가됩니다.

마지막으로, 작업 제목 등과 같은 대문자 글자 문자열을 처리할 때는 이러한 특수 패턴을 추출할 수 있도록 특수 알고리즘이 사용됩니다.

3단계. 동의어의 동등 클래스 및 통합 식별

후보 단일어 및 다항어가 식별된 후에는 소프트웨어는 알고리즘 세트를 사용하여 이를 비교하고 동등 클래스를 식별합니다. 동등 클래스는 한 구문으로 된 기본 양식이거나 동일한 구문의 두 개의 변형이 있는 단일 양식입니다. 구문을 동등 클래스에 지정하는 목적은 예를 들어, `president of the company` 및 `company president`가 별개의 개념으로 처리되지 않도록 하기 위한 것입니다. 동등 클래스에 사용할 개념을 판별하기 위해서 즉, `president of the company` 또는 `company president`가 리드 용어로 사용되는지 여부를 판별하기 위해서 추출 엔진은 다음 규칙을 나열된 순서대로 적용합니다.

- 라이브러리의 사용자 지정 양식.
- 텍스트의 전체 본문에서 가장 자주 사용되는 양식.
- 텍스트의 전체 본문에서 가장 짧은 양식(일반적으로 기본 양식에 해당함).

4단계. 유형 지정

다음으로 유형은 추출된 개념에 지정됩니다. 유형은 개념의 시맨틱 그룹입니다. 이 단계에서는 컴파일된 자원과 라이브러리 둘 모두가 사용됩니다. 유형은 상위 레벨 개념, 긍정적 및 부정적 단어, 이름, 장소, 조직 등과 같은 것을 포함합니다. 추가 유형은 사용자가 정의할 수 있습니다. 자세한 정보는 197 페이지의 『유형 사전』의 내용을 참조하십시오.

5단계. 색인화

레코드 또는 문서의 전체 세트는 텍스트 위치와 각 동등 클래스의 대표 용어 간에 포인터를 설정하여 색인화됩니다. 여기에서는 후보 개념의 모든 굴절된 양식 인스턴스가 후보 기본 양식으로서 색인화되는 것으로 가정합니다. 각 기본 양식에 대해 글로벌 빈도가 계산됩니다.

6단계. 패턴 및 이벤트 추출 매치

IBM SPSS Modeler Text Analytics에서는 유형과 개념뿐만 아니라 이들 간의 관계를 찾아낼 수 있습니다. 이 제품에서는 몇몇 알고리즘과 라이브러리를 사용할 수 있고 유형과 개념 간의 관계 패턴을 추출하는 기능을 제공합니다. 이들은 특정 의견(예: 제품 반응) 또는 사람이나 개체 간의 관계 링크(예: 정치적 그룹과 게놈 사이의 링크)를 찾아내려고 시도할 때 특히 유용합니다.

범주화 작동 방법

IBM SPSS Modeler Text Analytics에서 범주 모델을 작성할 때 범주를 작성하기 위해 선택할 수 있는 몇몇 기술이 있습니다. 모든 데이터 세트는 고유하므로 기술의 수와 이를 적용하는 순서는 변경될 수 있습니다. 사용자의 결과 해석이 다른 사람의 해석과 다를 수 있으므로 어떤 기술이 텍스트 데이터에 대해 최상의 결과를 내는지를 보려면 여러 기술을 실험해야 할 수도 있습니다. IBM SPSS Modeler Text Analytics에서, 워크벤치 세션에서 범주 모델을 작성하고 여기에서 추가로 범주를 탐색하고 세부 조정할 수 있습니다.

이 안내서에서 **범주 작성**은 하나 이상의 내장된 기술을 사용하여 범주 정의 및 분류의 생성을 가리키고, **범주화**는 각 레코드 또는 문서마다 고유 식별자(이름/ID/값)를 범주 정의에 지정하는 기준이 되는 스코어링 또는 레이블, 프로세스를 가리킵니다.

범주 작성 동안에 추출된 개념 및 유형은 범주의 구성 요소로서 사용됩니다. 범주를 작성할 때 레코드 또는 문서는 범주의 정의 요소와 매치하는 텍스트를 포함하는 경우 자동으로 범주에 지정됩니다.

IBM SPSS Modeler Text Analytics 에서는 문서 또는 레코드를 빠르게 범주화할 수 있도록 몇몇 자동화된 범주 작성 기술을 제공합니다.

그룹화 기술

사용 가능한 각 기술은 특정 데이터 유형과 상황에 잘 맞지만, 종종 문서 또는 레코드의 전체 범위를 캡처하기 위해 동일한 분석으로 기술을 결합하는 것이 유용합니다. 다중 범주에서 개념을 확인하거나 중복 범주를 찾을 수 있습니다.

개념 루트 파생. 이 기술은 개념을 취하고 개념 구성요소가 형태소 분석으로 관련되거나 루트를 공유하는지 여부를 분석하여 관련되는 다른 개념을 찾아서 범주를 작성합니다. 이 기술은 동의 복합어 개념 식별에 아주 유용합니다. 생성된 각 범주의 개념은 동의어이거나 의미에서 거의 관련되기 때문입니다. 이는 다양한 길이의 데이터에 대해 작동하여 더 적은 수의 최소 범주를 생성합니다. 예를 들어, `opportunities to advance` 개념은 `opportunity for advancement` 및 `advancement opportunity` 개념을 사용하여 그룹화됩니다. 자세한 정보는 122 페이지의 『개념 루트 파생』의 내용을 참조하십시오. 이 옵션은 일본어 텍스트에는 사용할 수 없습니다.

시맨틱 네트워크. 이 기술은 광범위한 단어 색인 관계에서 각 개념의 가능한 의미를 식별하여 시작한 다음 관련된 개념을 그룹화하여 범주를 작성합니다. 이 기술은 개념이 시맨틱 네트워크에 알려져 있고 너무 애매하지 않을 경우에 가장 좋습니다. 텍스트에 네트워크에 알려지지 않은 용어나 특수화된 전문용어가 포함된 경우에는 덜 유용합니다. 하나의 예에서, 개념 `granny smith apple`은 `gala apple` 및 `winesap apple`과 그룹화될 수 있습니다. 이들은 granny smith의 형제어이기 때문입니다. 다른 예에서, 개념 `animal`은 `cat` 및 `kangaroo`와 그룹화될 수 있습니다. 이들은 `animal`의 하위어이기 때문입니다. 이 기술은 이 릴리스에서 영어 텍스트에만 사용할 수 있습니다. 자세한 정보는 124 페이지의 『시맨틱 네트워크』의 내용을 참조하십시오.

개념 포함. 이 기술은 다른 개념에서 단어의 서브세트 또는 수퍼세트인 단어를 포함하는지 여부를 기초로 다항어 개념(복합어)을 그룹화하여 범주를 작성합니다. 예를 들어, 개념 `seat`는 `safety seat`, `seat belt` 및 `seat belt buckle`과 함께 그룹화됩니다. 자세한 정보는 123 페이지의 『개념 포함』의 내용을 참조하십시오.

동시 발생. 이 기술은 텍스트에서 발견된 동시 발생에서 범주를 작성합니다. 개념 또는 개념 패턴이 종종 함께 문서 및 레코드에서 발견될 때, 동시 발생은 사용자 범주 정의의 값일 수 있는 기본적인 관계를 반영합니다. 단어가 현저하게 동시 발생하는 경우, 동시 발생 규칙이 작성되고 새 하위 범주에 대한 범주 디스크립터로 사용할 수 있습니다. 예를 들어, 많은 레코드에 단어 `price` 및 `availability`가 포함되어 있는 경우(그러나 몇 개의 레코드는 다른 하나 없이 하나만 포함함), 이 개념은 동시 발생 규칙으로 그룹화될 수 있고(`price & available`), 예를 들어 범주 `price`의 하위 범주에 지정됩니다. 자세한 정보는 125 페이지의 『동시 발생 규칙』의 내용을 참조하십시오.

최소 문서 수. 동시 발생 흥미 정도를 판별하기 위해, 범주에서 디스크립터로 사용되도록 지정된 동시 발생을 포함해야 하는 최소 문서 또는 레코드 수를 정의하십시오.

IBM SPSS Modeler Text Analytics 노드

IBM SPSS Modeler와 함께 제공되는 많은 표준 노드와 함께, 텍스트 마이닝 노드에 대해 작업하여 텍스트 분석의 능력을 스트림에 통합할 수 있습니다. IBM SPSS Modeler Text Analytics는 바로 그것을 수행하기 위한 여러 가지 텍스트 마이닝을 제공합니다. 이들 노드는 노드 팔레트의 IBM SPSS Modeler Text Analytics 탭에 저장됩니다.

다음 노드가 포함됩니다.

- **파일 목록 소스 노드**는 텍스트 마이닝 프로세스에 대한 입력으로 문서 이름의 목록을 생성합니다. 이것은 텍스트가 데이터베이스나 다른 구조화된 파일이 아니라 외부 문서에 상주할 때 유용합니다. 노드는 나열된 각 문서나 폴더에 대해 하나의 레코드를 갖는 단일 필드를 출력하는데, 이것을 후속 텍스트 마이닝 노드의 입력으로 선택할 수 있습니다. 자세한 정보는 11 페이지의 『파일 목록 노드』의 내용을 참조하십시오.
- **웹 피드 소스 노드**는 RSS 또는 HTML 형식의 블로그 또는 뉴스 피드 같은 웹 피드에서 텍스트를 읽고 이 데이터를 텍스트 마이닝 프로세스에서 사용할 수 있게 합니다. 노드는 피드에서 발견되는 각 레코드에 대한 하나 이상의 필드를 출력하는데, 이것은 후속 텍스트 마이닝 노드에서 출력으로 선택할 수 있습니다. 자세한 정보는 13 페이지의 『웹 피드 노드』의 내용을 참조하십시오.
- **텍스트 마이닝 노드**는 언어 방법을 사용하여 텍스트에서 핵심 개념을 추출하고, 이들 개념 및 기타 데이터로 범주를 작성할 수 있게 하고, 알려진 패턴을 바탕으로 개념 사이의 관계 및 연관을 식별하는 기능(텍스트 링크 분석이라고 부름)을 제공합니다. 이 노드를 사용하여 텍스트 데이터 내용을 탐색하거나 개념 모델 또는 범주 모델을 생성할 수 있습니다. 개념 및 범주를 인구 통계 같은 기존의 구조화된 데이터와 결합하고 모델링에 적용할 수 있습니다. 자세한 정보는 21 페이지의 『텍스트 마이닝 모델링 노드』의 내용을 참조하십시오.
- **텍스트 링크 분석 노드**는 개념을 추출하며 텍스트 내에서 알려진 패턴을 바탕으로 개념 사이의 관계를 식별합니다. 패턴 추출은 개념 사이의 관계뿐 아니라 이들 개념에 첨부된 모든 의견이나 규정자를 발견하는 데 사용할 수 있습니다. 텍스트 링크 분석 노드는 텍스트에서 패턴을 식별 및 추출한 후 스트림의 데이터 세트에 패턴 결과를 추가하는 보다 직접적인 방법을 제공합니다. 그러나 텍스트 마이닝 모델링 노드에서 대화형 워크벤치 세션을 사용하여 TLA를 수행할 수도 있습니다. 자세한 정보는 51 페이지의 『텍스트 링크 분석 노드』의 내용을 참조하십시오.
- **변환 노드**를 사용하면 아랍어, 중국어 및 페르시아어 같은 지원되는 언어에서 영어나 모델링 목적의 기타 언어로 텍스트를 변환할 수 있습니다. 이것은 그렇지 않은 경우 지원되지 않는 2바이트 언어로 된 문서를 마이닝할 수 있게 하며 분석자가 문제가 되는 언어를 말할 수 없는 경우에도 이들 문서에서 개념을 추출할 수 있게 합니다. 동일한 기능을 텍스트 모델링 노드 중 하나에서 호출할 수 있지만, 별도의 변환 노드 사용은 다중 노드에서 변환을 캐시하고 재사용할 수 있게 만듭니다. 자세한 정보는 59 페이지의 『변환 노드』의 내용을 참조하십시오.

- 외부 문서에서 텍스트를 마이닝할 때, **텍스트 마이닝 출력 노드**를 사용하여 개념이 추출된 문서에 대한 링크를 포함하는 HTML 페이지를 생성할 수 있습니다. 자세한 정보는 63 페이지의 『파일 뷰어 노드』의 내용을 참조하십시오.

애플리케이션

일반적으로, 일상적으로 큰 볼륨의 문서를 검토하여 추가 탐색을 위한 핵심 요소를 식별해야 하는 사람은 IBM SPSS Modeler Text Analytics를 활용할 수 있습니다.

일부 특정 애플리케이션은 다음을 포함합니다.

- **과학 및 의학 연구.** 특허 보고서, 저널 기사, 프로토콜 서적 같은 보조 연구 자료를 탐색하십시오. 이전에 알려진 연관(예: 특정 제품과 연관된 의사)을 식별하여 추가 탐색을 위한 길을 표시하십시오. 약 발견 프로세스에서 소비되는 시간을 최소화하십시오. 유전자 연구에서의 도움으로 사용하십시오.
- **투자 연구.** 일일 분석 보고서, 뉴스 기사 및 회사 보도 자료를 검토하여 핵심 전략 포인트 또는 시장 변동을 식별하십시오. 그런 정보의 추세 분석은 기간 동안 회사 또는 산업에 대한 새로운 이슈나 기회를 드러냅니다.
- **사기 발견.** 비정상을 발견하고 많은 양의 텍스트에서 위험 신호를 발견하려면 금융 및 건강 관리 사기에서 사용하십시오.
- **시장 조사.** 개방형 설문조사 응답에서 핵심 주제를 식별하기 위해 시장 조사 시도에서 사용하십시오.
- **블로그 및 웹 피드 분석.** 뉴스 피드, 블로그 등에서 발견된 핵심 아이디어를 사용하여 모델을 탐색 및 작성하십시오.
- **CRM.** 이메일, 트랜잭션, 설문조사 같은 모든 고객 접촉 지점의 데이터를 사용하여 모델을 작성하십시오.

제 2 장 소스 텍스트에서 읽기

텍스트 마이닝을 위한 데이터는 데이터베이스를 포함하여 IBM SPSS Modeler가 사용하는 표준 형식 중 하나 또는 데이터를 행과 열로 나타내는 다른 "직사각형" 형식 또는 이 구조를 따르지 않는 Microsoft Word, Adobe PDF, HTML 같은 문서 형식으로 상주할 수 있습니다.

- Microsoft Word, Microsoft Excel, Microsoft PowerPoint뿐 아니라 Adobe PDF, XML, HTML 등을 포함하여 표준 데이터 구조를 따르지 않는 문서에서 텍스트를 읽으려면 파일 목록 노드를 사용하여 텍스트 마이닝 프로세스에 대한 입력으로 문서 또는 폴더의 목록을 생성할 수 있습니다. 자세한 정보는 『파일 목록 노드』의 내용을 참조하십시오.
- 블로그 또는 RSS나 HTML 형식의 뉴스 피드 같은 웹 피드에서 텍스트를 읽기 위해 웹 피드 노드를 사용하여 웹 피드 데이터를 텍스트 마이닝 프로세스에 대한 입력으로 형식화할 수 있습니다. 자세한 정보는 13 페이지의 『웹 피드 노드』의 내용을 참조하십시오.
- 고객 의견을 위한 하나 이상의 텍스트 필드를 갖는 데이터베이스 같이 IBM SPSS Modeler가 사용하는 표준 데이터 형식 중 하나로부터 텍스트를 읽기 위해 IBM SPSS Modeler에 기본인 표준 소스 노드 중 하나를 사용할 수 있습니다. 자세한 정보는 IBM SPSS Modeler 노드 문서를 참조하십시오.

파일 목록 노드

Microsoft Word, Microsoft Excel, Microsoft PowerPoint뿐 아니라 Adobe PDF, XML, HTML 및 기타와 같은 형식으로 저장된 비정형 문서로부터 텍스트를 읽기 위해, 파일 목록 노드를 사용하여 텍스트 마이닝 프로세스에 대한 입력으로 문서 또는 폴더의 목록을 생성할 수 있습니다. 이것은 비정형 텍스트 문서는 IBM SPSS Modeler가 사용하는 다른 데이터와 동일한 방식으로 필드 및 레코드(행과 열)에 의해 표시될 수 없기 때문에 필요합니다. 이 노드는 텍스트 마이닝 팔레트에서 찾을 수 있습니다.

파일 목록 노드는 소스 노드로서 기능합니다. 그러나 소스 파일의 실제 데이터를 읽고 출력할 뿐 아니라, 이 노드를 사용하여 지정된 루트 아래에 있는 문서 또는 디렉토리의 이름을 읽고 이들을 목록으로 생성할 수 있습니다. 문서 또는 디렉토리 이름을 읽는 데 사용될 때, 출력은 나열되는 각 파일에 대한 하나의 레코드를 갖는 단일 필드이며 후속 텍스트 마이닝 또는 텍스트 링크 분석 노드에 대한 입력으로 선택할 수 있습니다.

IBM SPSS Modeler 창의 맨 아래에 있는 노드 팔레트의 IBM SPSS Modeler Text Analytics 탭에서 이 노드를 찾을 수 있습니다. 자세한 정보는 9 페이지의 『IBM SPSS Modeler Text Analytics 노드』의 내용을 참조하십시오.

중요사항: 머신 로컬 인코딩에 포함되지 않는 문자를 포함한 디렉토리 이름과 파일 이름은 지원되지 않습니다. 파일 목록 노드를 포함하는 스트림을 실행하려고 할 때, 이러한 문자를 포함하는 파일 또는 디렉토리 이름을 사용하면 스트림 실행에 실패합니다. 프랑스어 로케일의 일본어 파일 이름과 같이, 외부 언어 디렉토리 이름 또는 파일 이름을 사용하는 경우에 이러한 상황이 발생할 수 있습니다.

로컬 데이터 지원. 원격 IBM SPSS Modeler Text Analytics Server에 연결되어 있고 파일 목록 노드와의 스트림이 있는 경우, 데이터는 IBM SPSS Modeler Text Analytics Server와 동일한 머신에 상주해야 하거나 서버 머신이 파일 목록 노드의 소스 데이터가 저장되는 폴더에 액세스할 수 있어야 합니다.

참고: IBM SPSS Collaboration and Deployment Services - Scoring 구성 내에서 스코어링을 위해 파일 목록 노드를 사용할 수 없습니다.

파일 목록 노드: 설정 탭

이 탭에서 디렉토리, 파일 확장자 및 이 노드에서 바람직한 출력을 정의할 수 있습니다.

참고: 텍스트 마이닝 추출은 비Microsoft Windows 플랫폼에서 Microsoft Office 및 Adobe PDF 파일을 처리할 수 없습니다. 그러나, XML, HTML 또는 텍스트 파일은 항상 처리할 수 있습니다.

머신 로컬 인코딩에 포함되지 않는 문자를 포함한 디렉토리 이름과 파일 이름은 지원되지 않습니다. 파일 목록 노드를 포함하는 스트림을 실행하려고 할 때, 이러한 문자를 포함하는 파일 또는 디렉토리 이름을 사용하면 스트림 실행에 실패합니다. 프랑스어 로케일의 일본어 파일 이름과 같이, 외부 언어 디렉토리 이름 또는 파일 이름을 사용하는 경우에 이러한 상황이 발생할 수 있습니다.

디렉토리. 나열하려는 문서를 포함하는 루트 폴더를 지정합니다.

- **하위 디렉토리 포함.** 하위 디렉토리도 스캔해야 함을 지정합니다.

목록에 포함할 파일 유형: 사용하려는 파일 유형 및 확장자를 선택 또는 선택 취소할 수 있습니다. 파일 확장자를 선택 취소하면 해당 확장자를 갖는 파일은 무시됩니다. 다음 확장자로 필터링할 수 있습니다.

표 1. 파일 확장자별 파일 유형 필터.

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

참고: 자세한 정보는 11 페이지의 『파일 목록 노드』 주제를 참조하십시오.

확장자가 없거나 후미 도트 확장자를 갖는 파일이 있는 경우(예: `File01` 또는 `File01.`), **확장자 없음** 옵션을 사용하여 선택하십시오.

출력 필드 표시. 출력 필드의 형식을 선택하십시오. 선택 사항은 다음과 같습니다.

- **실제 텍스트.** 필드가 정확한 텍스트를 포함할 경우 이 옵션을 선택하십시오. 그러면 다음 목록에서 **입력 인코딩** 값을 선택할 수 있습니다.
  - Automatic(유럽)
  - Automatic(일본어)
  - UTF-8
  - UTF-16
  - ISO-8859-1

  - US ascii
  - CP850
  - Shift-JIS
- **문서의 경로명.** 출력 필드가 문서가 상주하는 위치에 대한 하나 이상의 경로명을 포함하는 경우 이 옵션을 선택하십시오.

중요! 버전 14 이후, '디렉토리 목록' 옵션은 더 이상 사용할 수 없으며 유일한 출력은 파일의 목록입니다.

파일 목록 노드: 기타 탭

유형 탭은 주석 탭과 마찬가지로 IBM SPSS Modeler 노드의 표준 탭입니다.

텍스트 마이닝에서 파일 목록 노드 사용

파일 목록 노드는 텍스트 데이터가 Microsoft Word, Microsoft Excel, Microsoft PowerPoint뿐 아니라 Adobe PDF, XML, HTML 등과 같은 형식으로 된 외부 비정형 문서에 상주할 때 사용됩니다. 실제 텍스트 출력 외에, 이 노드를 사용하여 텍스트 마이닝 프로세스(예: 후속 텍스트 마이닝 또는 텍스트 링크 분석 노드)에 대한 입력으로 문서 또는 폴더의 목록을 생성할 수 있습니다.

파일 목록 노드를 사용하여 실제 텍스트 대신 문서 목록을 생성하는 경우, 나중에 텍스트 마이닝 또는 텍스트 링크 분석 노드 중 하나를 사용할 때 마이닝할 실제 텍스트를 포함하는 대신 선택된 문서가 텍스트가 위치하는 문서에 대한 경로를 포함함을 표시하기 위해 **텍스트 필드**가 **문서의 경로명**을 나타내도록 지정하십시오.

예를 들어, 외부 문서에 상주하는 텍스트를 제공하기 위해 파일 목록 노드를 텍스트 마이닝 노드에 연결했다고 가정하십시오.

1. **파일 목록 노드(설정 탭).** 먼저, 이 노드를 스트림에 추가하여 텍스트 문서가 저장되는 장소를 지정했습니다. 텍스트 마이닝을 수행하려는 모든 문서를 포함하는 디렉토리를 선택했습니다.
2. **텍스트 마이닝 노드(필드 탭).** 다음으로, 파일 목록 노드에 텍스트 마이닝 노드를 추가하고 연결했습니다. 이 노드에서, 입력 형식, 자원 템플리트 및 출력 형식을 정의했습니다. 파일 목록 노드에서 생성된 필드 이름을 선택하고, 텍스트 필드가 다른 설정뿐만 아니라 **문서의 경로명**을 나타내는 옵션을 선택했습니다. 자세한 정보는 32 페이지의 『스트림에서 텍스트 마이닝 노드 사용』의 내용을 참조하십시오.

텍스트 마이닝 노드 사용에 대한 자세한 정보는 21 페이지의 『텍스트 마이닝 모델링 노드』의 내용을 참조하십시오.

웹 피드 노드

웹 피드 노드는 텍스트 마이닝 프로세스에 대해 웹 피드의 텍스트 데이터를 준비하기 위해 사용됩니다. 이 노드는 두 가지 형식의 웹 피드를 승인합니다.

- RSS 형식. RSS는 웹 내용에 대한 단순한 XML 기반 표준화된 형식입니다. 이 형식의 URL은 신디케이트된 뉴스 소스 및 블로그와 같은 링크된 기사 세트가 있는 페이지를 가리킵니다. RSS는 표준화된 형식이므

로, 링크된 각 기사는 결과 데이터 스트림에서 별도의 레코드로 식별되고 처리됩니다. 필터링 기술을 텍스트에 적용하지 않으면 피드에서 중요한 텍스트 데이터와 레코드를 식별할 수 있도록 추가 입력이 필요한 것은 아닙니다.

- HTML 형식. 입력 탭에서 HTML 페이지에 대한 하나 이상의 URL을 정의할 수 있습니다. 그런 다음 레코드 탭에서 레코드 시작 태그를 정의하고 대상 내용을 구분하는 태그를 식별한 후 선택하는 출력 필드(설명, 제목, 수정된 날짜 등)에 해당 태그를 지정하십시오. 자세한 정보는 15 페이지의 『웹 피드 노드: 레코드 탭』의 내용을 참조하십시오.

중요! 프록시 서버를 통해 웹에서 정보를 검색하려고 시도하는 경우에는 IBM SPSS Modeler Text Analytics 클라이언트와 서버 둘 모두에 대해 `net.properties` 파일에서 프록시 서버를 사용으로 설정해야 합니다. 이 파일 내부에 설명된 지시사항을 따르십시오. 이는 웹 피드 노드를 통해 웹에 액세스할 때나 SDL SaaS(Software as a Service) 사용권을 검색할 때 적용됩니다. 이러한 연결은 Java™을 통과하기 때문입니다. 이 파일은 기본적으로 *C:\Program Files\IBM\SPSS\Modeler\17.1\jre\lib\net.properties*에 있습니다.

이 노드의 출력은 레코드를 설명하기 위해 사용되는 필드 세트입니다. **설명** 필드는 대부분의 텍스트 내용을 포함하므로 가장 일반적으로 사용됩니다. 그러나 레코드의 간단한 설명(**간단한 설명** 필드)이나 레코드의 제목(**제목** 필드)과 같은 다른 필드에 관심이 있을 수도 있습니다. 출력 필드는 후속 텍스트 마이닝 노드의 입력을 선택할 수 있습니다.

참고: IBM SPSS Collaboration and Deployment Services - Scoring 구성 내에서 스코어링에 대해 웹 피드 노드를 사용할 수 없습니다.

IBM SPSS Modeler 창의 맨 아래에 있는 노드 팔레트의 IBM SPSS Modeler Text Analytics 탭에서 이 노드를 찾을 수 있습니다. 자세한 정보는 9 페이지의 『IBM SPSS Modeler Text Analytics 노드』의 내용을 참조하십시오.

웹 피드 노드: 입력 탭

입력 탭은 텍스트 데이터를 캡처하기 위해 하나 이상의 웹 주소나 URL을 지정하기 위해 사용됩니다. 텍스트 마이닝의 컨텍스트에서, 텍스트 데이터를 포함하는 피드에 대한 URL을 지정할 수 있습니다.

중요! RSS 이외 데이터에 대해 작업할 때, 내용 수집 후 다른 소스 노드를 사용하여 해당 도구에서 출력을 참조하는 것을 자동화하기 위해 WebQL®과 같은 웹 스크래핑 도구를 사용하는 것을 선호할 수 있습니다.

다음 매개변수를 설정할 수 있습니다.

URL 입력 또는 붙여넣기. 이 필드에서, 하나 이상의 URL을 입력하거나 붙여넣을 수 있습니다. 두 개 이상을 입력하는 경우, 행당 하나만 입력하며 **Enter/Return** 키를 사용하여 행을 구분하십시오. 파일의 전체 URL 경로를 입력하십시오. 이 URL은 두 형식 중 하나로 된 피드 URL일 수 있습니다.

- *RSS 형식.* RSS는 웹 내용에 대한 단순한 XML 기반 표준화된 형식입니다. 이 형식의 URL은 신디케이트된 뉴스 소스 및 블로그와 같은 링크된 기사 세트가 있는 페이지를 가리킵니다. RSS는 표준화된 형식이므

로, 링크된 각 기사는 결과 데이터 스트림에서 별도의 레코드로 식별되고 처리됩니다. 필터링 기술을 텍스트에 적용하지 않으면 피드에서 중요한 텍스트 데이터와 레코드를 식별할 수 있도록 추가 입력이 필요한 것은 아닙니다.

- *HTML 형식.* 입력 탭에서 HTML 페이지에 대한 하나 이상의 URL을 정의할 수 있습니다. 그런 다음 레코드 탭에서 레코드 시작 태그를 정의하고 대상 내용을 구분하는 태그를 식별한 후 선택하는 출력 필드(설명, 제목, 수정된 날짜 등)에 해당 태그를 지정하십시오. RSS 이외 데이터에 대해 작업할 때, 내용 수집 후 다른 소스 노드를 사용하여 해당 도구에서 출력을 참조하는 것을 자동화하기 위해 WebQL®과 같은 웹 스크래핑 도구를 사용하는 것을 선호할 수 있습니다. 자세한 정보는 『웹 피드 노드: 레코드 탭』의 내용을 참조하십시오.

URL마다 읽을 최근 항목 수. 이 필드는 피드에서 발견된 첫 번째 레코드에서 시작하여 필드에 나열된 각 URL에 대해 읽을 최대 레코드 수를 지정합니다. 텍스트 양은 텍스트 마이닝 노드 또는 텍스트 링크 분석 노드에서 추출 다운스트림 동안 처리 속도에 영향을 줍니다.

가능할 때 이전 웹 피드를 저장하고 재사용하십시오. 이 옵션을 사용하면, 웹 피드가 스캔되고 처리 결과가 캐시됩니다. 그런 다음 후속 스트림 실행 시, 지정된 피드의 내용이 변경되지 않았거나 피드에 액세스할 수 없는 경우(예: 인터넷 가동 중단), 캐시된 버전이 사용되어 처리 시간을 가속화합니다. 이 피드에서 발견되는 새 내용은 다음에 노드를 실행할 때도 캐시됩니다.

- **레이블. 가능할 때 이전 웹 피드 저장 및 재사용**을 선택하는 경우, 결과의 레이블 이름을 지정해야 합니다. 이 레이블은 서버에서 캐시된 피드를 설명하기 위해 사용됩니다. 지정된 레이블이 없거나 레이블이 인식되지 않는 경우, 재사용이 가능하지 않습니다. IBM SPSS Text Analytics Administration Console의 세션 테이블에서 웹 피드 캐시를 관리할 수 있습니다. 자세한 정보는 IBM SPSS Text Analytics Administration Console 사용자 안내서를 참조하십시오.

웹 피드 노드: 레코드 탭

레코드 탭은 각각의 새 레코드가 시작하는 위치와 각 레코드에 관한 다른 관련 정보를 식별하여 비RSS 피드의 텍스트 내용을 지정하기 위해 사용됩니다. 비RSS 피드(HTML)에 여러 레코드에 있는 텍스트가 포함되어 있는 것을 알면, 여기에서 레코드 시작 태그를 식별해야 합니다. 그렇지 않으면 텍스트는 하나의 레코드로 처리됩니다. RSS 피드가 표준화되어 이 탭에서 태그를 지정하지 않아도 되지만, 미리보기 탭에서 내용을 미리볼 수 있습니다.

중요! RSS 이외 데이터에 대해 작업할 때, 내용 수집 후 다른 소스 노드를 사용하여 해당 도구에서 출력을 참조하는 것을 자동화하기 위해 WebQL®과 같은 웹 스크래핑 도구를 사용하는 것을 선호할 수 있습니다.

URL. 이 드롭 다운 목록에는 입력 탭에 입력된 URL의 목록이 포함됩니다. HTML 및 RSS 형식화 피드 모두가 제시됩니다. URL 주소가 드롭 다운 목록에 대해 너무 길면, 잘린 텍스트를 바꾸기 위해 생략 기호를 사용하여 중간에서 자동으로 잘립니다(예: *http://www.ibm.com/example/start-of-address...rest-of-address/path.htm*).

- **HTML 형식화 피드**를 사용할 때 피드에 두 개 이상의 레코드(또는 항목)가 있는 경우, 테이블에 표시된 필드에 해당되는 데이터를 포함하는 HTML 태그를 정의할 수 있습니다. 예를 들어, 새 레코드가 시작되었음을 표시하는 시작 태그, 수정된 날짜 태그 또는 작성자 이름을 정의할 수 있습니다.

- **RSS 형식화 피드**를 사용하는 경우에는 RSS가 표준화 형식이므로 태그를 입력하도록 요청하는 프롬프트가 표시되지 않습니다. 그러나 원하면 미리보기 탭에서 표본 결과를 볼 수 있습니다. 인식되는 모든 RSS 피드 앞에는 RSS 로고 이미지가 붙습니다.

소스 탭. 이 탭에서, HTML 피드에 대한 소스 코드를 볼 수 있습니다. 이 코드는 편집할 수 없습니다. 찾기 필드를 사용하여 특정 태그나 정보를 이 페이지에서 찾을 수 있습니다. 그런 다음 아래 테이블로 복사하여 붙여넣을 수 있습니다. 찾기 필드에서는 대소문자가 구분되지 않으므로 부분 문자열을 매치합니다.

미리보기 탭. 이 탭에서는, 웹 피드 노드에서 레코드가 읽혀지는 방법을 미리볼 수 있습니다. 이는 HTML 피드의 경우 특히 유용합니다. 미리보기 탭 아래의 테이블에서 HTML 태그를 정의하여 레코드를 읽을 방법을 변경할 수 있기 때문입니다.

비RSS 레코드 시작 태그. 이 옵션은 비RSS 피드에만 적용됩니다. HTML 피드에 여러 레코드로 분리하려고 하는 다중 텍스트가 있는 경우, 레코드(예: 기사 또는 블로그 항목) 시작을 알리는 HTML 태그를 여기에 지정하십시오. 비RSS 피드에 대해 이 태그를 정의하지 않으면, 전체 페이지가 하나의 단일 레코드로 처리되고, 전체 내용이 **설명** 필드에 출력되며, 노드 실행 날짜가 **수정된 날짜** 및 **출판된 날짜** 둘 다로 사용됩니다.

필드 표. 이 옵션은 비RSS 피드에만 적용됩니다. 이 표에서, 사전정의된 출력 필드 중 하나에 대해 시작 태그를 입력하여 특정 출력 필드로 텍스트 내용을 분리할 수 있습니다. 시작 태그만 입력하십시오. HTML을 구문분석하고 표 내용을 HTML에서 발견된 속성과 태그 이름에 매치하여 모든 매치가 수행됩니다. 정의한 태그를 복사하고 다른 피드에 재사용하기 위해 맨 아래에 있는 단추를 사용할 수 있습니다.

표 2. 비RSS 피드에 가능한 출력 필드(HTML 형식)

출력 필드 이름	예상된 태그 내용
제목	레코드 제목을 구분하는 태그. (선택사항)
간단한 설명	간단한 설명 또는 레이블을 구분하는 태그. (선택사항)
설명	주 텍스트를 구분하는 태그. 공백으로 남겨둘 경우, 이 필드는 <body> 태그(단일 레코드가 있는 경우)의 다른 모든 내용이나 현재 레코드에서 발견된 내용(레코드 구분자가 지정된 경우)을 포함합니다.
작성자	텍스트 작성자를 구분하는 태그. (선택사항)
기여자	기여자의 이름을 구분하는 태그. (선택사항)
출판된 날짜	텍스트가 출판된 날짜를 구분하는 태그. 공백으로 남겨두면, 이 필드는 노드가 데이터를 읽을 때 날짜를 포함합니다.
수정된 날짜	텍스트가 수정된 날짜를 구분하는 태그. 공백으로 남겨두면, 이 필드는 노드가 데이터를 읽을 때 날짜를 포함합니다.

테이블에 태그를 입력할 때, 피드는 정확히 일치보다 매치시킬 최소 태그로 이 태그를 사용하여 스캔됩니다. 즉, 제목 필드에 대해 를 입력한 경우, 이는 지정된 속성(예:)을 가지고 있는 태그를 비롯하여 피드에서 태그를 매치하고(가 루트 태그( )와 같도록) 속성을 포함하는 파생어를 매치한 후 제목 출력 필드에 대해 해당 내용을 사용합니다. 루트 태그를 입력하면, 추가 속성도 포함됩니다.

표 3. 출력 필드에 대해 텍스트 식별에 사용되는 HTML 태그의 예

다음을 입력하는 경우:	다음을 매치함:	다음도 매치함:	다음은 매치하지 않음:
			기타 태그
<p class="auth">	<p class="auth">	<p color="black" class="auth" id="85643">	<p color="black">

웹 피드 노드: 내용 필터 탭

내용 필터 탭은 RSS 피드 내용에 필터 기술을 적용하기 위해 사용됩니다. 이 탭은 HTML 피드에 적용되지 않습니다. 필드에 헤더, 꼬리말, 메뉴, 광고 등의 양식으로 많은 텍스트를 포함하는 경우 필터를 원할 수 있습니다. 이 탭을 사용하여 원하지 않는 HTML 태그, JavaScript, 그리고 내용의 간단한 단어 또는 행을 완전히 제거할 수 있습니다.

내용 필터. 정리 기술을 적용하지 않으려면, **없음**을 선택하십시오. 그렇지 않으면, **RSS 내용 정리기**를 선택하십시오.

RSS 내용 정리기 옵션. **RSS 내용 정리기**를 선택하는 경우, 특정 기준을 기초로 행을 삭제할 것을 선택할 수 있습니다. 행은 <p> 및 와 같은 HTML 태그(, <b> 및 <font>와 같은 인라인 태그 제외)로 구분됩니다.
 태그는 행 바꿈으로 처리됩니다.

- **짧은 행 삭제**. 이 옵션은 여기에서 정의되는 **최소 단어 수**를 포함하지 않는 행을 무시합니다.
- **짧은 단어가 있는 행 삭제**. 이 옵션은 여기에서 정의되는 **최소 평균 단어 길이**보다 긴 행을 무시합니다.
- **많은 단일 문자 단어가 있는 행 삭제**. 이 옵션은 특정의 **단일 문자 단어 비율**보다 더 포함하는 행을 무시합니다.
- **특정 태그 포함 행 삭제**. 이 옵션은 필드에 지정된 태그를 포함하는 행에서 텍스트를 무시합니다.
- **특정 텍스트를 포함하는 행 삭제**. 이 옵션은 필드에 지정된 텍스트를 포함하는 행을 무시합니다.

텍스트 마이닝에서 웹 피드 노드 사용

웹 피드 노드는 텍스트 마이닝 프로세스에 대해 인터넷 웹 피드의 텍스트 데이터를 준비하기 위해 사용됩니다. 이 노드는 HTML 또는 RSS 형식의 웹 피드를 승인합니다. 이 피드는 텍스트 마이닝 프로세스의 입력 역할을 합니다(후속 텍스트 마이닝 또는 텍스트 링크 분석 노드).

웹 피드 노드를 사용하는 경우, 해당 피드가 각 기사 또는 블로그 항목에 직접 링크됨을 표시하기 위해 텍스트 필드가 텍스트 마이닝 또는 텍스트 링크 분석 노드에서 **실제 텍스트**를 나타냄을 지정하도록 해야 합니다.

중요! 프록시 서버를 통해 웹에서 정보를 검색하려고 시도하는 경우에는 IBM SPSS Modeler Text Analytics 클라이언트와 서버 둘 모두에 대해 `net.properties` 파일에서 프록시 서버를 사용으로 설정해야 합니다. 이 파일 내부에 설명된 지시사항을 따르십시오. 이는 웹 피드 노드를 통해 웹에 액세스할 때나 SDL SaaS(Software as a Service) 사용권을 검색할 때 적용됩니다. 이러한 연결은 Java을 통과하기 때문입니다. 이 파일은 기본적으로 *C:\Program Files\IBM\SPSS\Modeler\17.1\jre\lib\net.properties*에 있습니다.

예: 텍스트 마이닝 모델링 노드가 있는 웹 피드 노드(RSS 피드)

예로서, RSS 피드의 텍스트를 텍스트 마이닝 프로세스에 제공하기 위해 텍스트 마이닝 노드에 웹 피드 노드를 연결한다고 가정해 보십시오.

1. **웹 피드 노드(입력 탭).** 먼저, 피드 내용이 위치되는 곳을 지정하고 내용 구조를 확인하기 위해 스트림에 이 노드를 추가했습니다. 첫 번째 탭에서, URL을 RSS 피드에 제공했습니다. 이 예는 RSS 피드에 대한 것이므로, 형식화가 이미 정의되어 있어서 레코드 탭에서 변경할 필요는 없습니다. 그러나 적용되지 않은 경우에는 RSS 피드에 대해 선택적 내용 필터링 알고리즘을 사용할 수 있습니다.
2. **텍스트 마이닝 노드(필드 탭).** 다음으로, 웹 피드 노드에 텍스트 마이닝 노드를 추가하고 연결했습니다. 이 탭에서는 웹 피드 노드에 의해 출력된 텍스트 필드를 정의했습니다. 이 경우에, **설명** 필드를 사용하려고 했습니다. 또한 텍스트 필드가 **실제 텍스트**를 나타내는 옵션과 다른 설정을 선택했습니다.
3. **텍스트 마이닝 노드(모델 탭).** 다음으로, 모델 탭에서 작성 모드 및 자원을 선택했습니다. 이 예에서는, 기본 자원 템플리트를 사용하여 노드에서 직접 개념 모델을 작성할 것을 선택했습니다.

텍스트 마이닝 노드 사용에 대한 자세한 정보는 21 페이지의 『텍스트 마이닝 모델링 노드』의 내용을 참조하십시오.

제 3 장 개념 및 범주 마이닝

텍스트 마이닝 모델링 노드는 다음 두 개의 텍스트 마이닝 모델 너깃 중 하나를 생성하는 데 사용됩니다.

- *개념 모델 너깃*은 구조화 또는 비정형 텍스트 데이터에서 핵심적인 개념을 드러내서 추출합니다.
- *범주 모델 너깃*은 문서 및 레코드를 스코어링하고 범주에 지정합니다. 범주는 추출된 개념(및 패턴)으로 구성됩니다.

추출된 개념 및 패턴뿐 아니라 모델 너깃의 범주가 모든 인구 통계 같은 기존의 구조화된 데이터와 결합되고 IBM SPSS Modeler의 전체 도구 모음을 사용하여 적용되어 더 좋고 더욱 집중된 의사결정을 내릴 수 있습니다. 예를 들어, 고객이 온라인 계정 관리 태스크를 완료하기 위한 1차적인 장애로 로그인 문제를 자주 나열하는 경우, "로그인 문제"를 모델에 통합하기 원할 수 있습니다.

또한, 텍스트 마이닝 모델링 노드는 IBM SPSS Modeler와 완전히 통합되므로 PredictiveCallCenter 같은 애플리케이션에서 비정형 데이터의 실시간 스코어링을 위해 IBM SPSS Modeler Solution Publisher를 통해 텍스트 마이닝 스트림을 배치할 수 있습니다. 이들 스트림을 배치하는 기능은 성공적은 폐쇄 루프 텍스트 마이닝 구현을 보장합니다. 예를 들어, 사용자 조직이 이제 예측 모형을 적용하여 마케팅 메시지의 정확도를 실시간으로 늘려서 인바운드 또는 아웃바운드 호출자의 메모철을 분석할 수 있습니다. 스트림에서 텍스트 마이닝 모델 결과 사용은 예측 데이터 모델의 정확도를 개선하기 위해 표시되었습니다.

참고: IBM SPSS Modeler Solution Publisher와 함께 IBM SPSS Modeler Text Analytics를 실행하려면 `<install_directory>/ext/bin/spss.TMWBServer` 디렉토리를 `$LD_LIBRARY_PATH` 환경 변수에 추가하십시오.

IBM SPSS Modeler Text Analytics에서 종종 추출된 개념 및 범주를 참조합니다. 개념 및 범주는 예비 작업 및 모델 작성 중에 더 많은 정보가 제공된 결정을 내리는 데 도움이 될 수 있으므로 개념 및 범주의 의미를 이해하는 것이 중요합니다.

개념 및 개념 모델 너깃

추출 프로세스 중에 텍스트 데이터가 스캔되고 `election` 또는 `peace` 및 `presidential election`, `election of the president` 또는 `peace treaties` 같은 단어 구 같이 관심이 있거나 관련 단일 단어를 식별하기 위해 분석됩니다. 이러한 단어와 구문을 집합적으로 *용어*라고 부릅니다. 언어학적 자원을 사용하여 관련 용어가 추출되고, 비슷한 용어가 **개념**이라는 리드 용어 아래에 함께 그룹화됩니다.

이런 방식으로, 개념은 텍스트와 사용 중인 언어학적 자원 세트에 따라 많은 기초적인 용어를 표시할 수 있습니다. 예를 들어, 직원 만족도 설문 조사가 있고 개념 `salary`가 추출되었다고 가정합니다. 또한 `salary`와 연관된 레코드를 볼 때 `salary`가 항상 텍스트에 표시되지 않고 `wage`, `wages` 및 `salaries` 등과 유사한 것을 포함하는 특정 레코드에 표시된다고 해 봅시다. 이러한 용어는 `salary` 아래에 그룹화됩니다. 추출 엔진이 이

들을 유사한 것으로 간주하거나 처리 규칙이나 언어학적 자원을 기반으로 이들이 동의어라고 판별했기 때문입니다. 이 경우, 이러한 용어를 포함하는 모든 문서 또는 레코드는 이들이 단어 `salary`를 포함하는 것처럼 처리됩니다.

어떤 용어가 개념 아래에 그룹화되는지 보려는 경우, 대화형 워크벤치 내에서 개념을 탐색하거나 개념 모델에 표시되는 동의어를 조사할 수 있습니다. 자세한 정보는 35 페이지의 『개념 모델의 기본 용어』의 내용을 참조하십시오.

개념 모델 너깃은 개념을(그의 모든 동의어 또는 그룹화된 용어 포함) 포함하는 레코드 또는 문서를 식별하는 데 사용할 수 있는 개념 세트를 포함합니다. 개념 모델은 두 가지 방법으로 사용할 수 있습니다. 첫 번째는 원래 소스 텍스트에서 발견된 개념을 탐색 및 분석하거나 관심있는 문서를 빨리 식별하는 것입니다. 두 번째는 이 모델을 새 텍스트 레코드나 문서에 적용하여 콜센터의 메모철 데이터에 있는 핵심 개념의 실시간 발견 같이 새 문서/레코드에서 동일한 핵심 개념을 빨리 식별하는 것입니다.

자세한 정보는 33 페이지의 『텍스트 마이닝 너깃: 개념 모델』의 내용을 참조하십시오.

범주 및 범주 모델 너깃

본질적으로 상위 레벨 개념이나 주제를 나타내는 **범주**를 작성하여 텍스트에서 표현되는 주요 아이디어, 지식 및 태도를 캡처할 수 있습니다. 범주는 *개념*, *유형*, *규칙* 같은 디스크립터의 세트로 구성됩니다. 이들 디스크립터는 함께 사용되어 레코드나 문서가 주어진 범주에 속하는지 여부를 식별합니다. 문서나 레코드를 스캔하여 그의 텍스트 중 하나가 디스크립터와 매치하는지 확인할 수 있습니다. 매치가 발견되면 문서/레코드가 해당 범주에 지정됩니다. 이 프로세스를 **범주화**라고 부릅니다.

범주는 제품의 강력한 자동화 기법 세트를 사용하여 자동으로, 사용자가 데이터에 관하여 가질 수 있는 추가 직관력을 사용하여 수동으로 또는 둘의 조합으로 작성될 수 있습니다. 또한 이 노드의 모델 탭을 통해 텍스트 분석 패키지로부터 사전 작성된 범주 세트를 로드할 수 있습니다. 범주의 수동 작성이나 범주 세분화는 대화형 워크벤치를 통해서만 수행될 수 있습니다. 자세한 정보는 24 페이지의 『텍스트 마이닝 노드: 모델 탭』의 내용을 참조하십시오.

범주 모델 너깃은 범주 세트를 해당 디스크립터와 함께 포함하고 있습니다. 이 모델을 사용하면 각 문서/레코드에 있는 텍스트를 기반으로 문서 또는 레코드의 세트를 범주화할 수 있습니다. 모든 문서나 레코드를 읽고 디스크립터 매치가 발견된 각 범주에 지정합니다. 이 방법으로 문서나 레코드가 둘 이상의 범주에 지정될 수 있습니다. 범주 모델 너깃을 사용하여 개방형 설문조사 응답이나 예를 들어 블로그 항목 세트에서 본질적인 아이디어를 볼 수 있습니다.

자세한 정보는 42 페이지의 『텍스트 마이닝 너깃: 범주 모델』의 내용을 참조하십시오.

텍스트 마이닝 모델링 노드

텍스트 마이닝 노드는 언어 및 빈도 기법을 사용하여 텍스트에서 핵심 개념을 추출하고 이들 개념과 다른 데이터로 범주를 작성합니다. 이 노드를 사용하여 텍스트 데이터 컨텐츠를 탐색하거나 개념 모델 너깃이나 범주 모델 너깃을 생성할 수 있습니다. 이 모델링 노드를 실행할 때, 내부 언어학적 추출 엔진이 자연어 처리 방법을 사용하여 개념, 패턴 및/또는 범주를 추출하고 구성합니다.

텍스트 마이닝 노드를 실행하고 **직접 생성** 옵션을 사용하여 자동으로 개념 또는 범주 모델 너깃을 생성할 수 있습니다. 또는 개념을 추출하고 범주를 작성하고 언어학적 자원을 세분화할 뿐 아니라 텍스트 링크 분석을 수행하고 군집을 탐색할 수도 있는 **대화형 작성** 모드를 사용하여 더 실무적이고 예비적인 방식을 사용할 수 있습니다. 자세한 정보는 24 페이지의 『텍스트 마이닝 노드: 모델 탭』의 내용을 참조하십시오.

IBM SPSS Modeler 창의 맨 아래에 있는 노드 팔레트의 IBM SPSS Modeler Text Analytics 탭에서 이 노드를 찾을 수 있습니다. 자세한 정보는 9 페이지의 『IBM SPSS Modeler Text Analytics 노드』의 내용을 참조하십시오.

요구사항. 텍스트 마이닝 모델링 노드는 웹 피드 노드, 파일 목록 노드 또는 표준 소스 노드 중 하나의 텍스트 데이터를 수락합니다. 이 노드는 IBM SPSS Modeler Text Analytics와 함께 설치되며 IBM SPSS Modeler Text Analytics 팔레트에서 액세스할 수 있습니다.

참고: 이 노드는 모든 사용자에 대한 텍스트 추출 노드를 대체하며 일본어 사용자에 대한 이전 텍스트 마이닝 노드를 대체하는데, 후자는 Text Mining for Clementine의 이전 버전에서 제공되었습니다. 이들 노드나 모델 너깃을 사용하는 이전 스트림이 있는 경우, 새 텍스트 마이닝 노드를 사용하여 스트림을 다시 작성해야 합니다.

텍스트 마이닝 노드: 필드 탭

필드 탭은 개념을 추출 중인 데이터에 대한 필드 설정을 지정하는 데 사용됩니다. 처리 시간을 가속화하기 위해 더 큰 데이터 세트에 대해 작업할 때 이 노드에서 샘플 노드 업스트림 사용을 고려하십시오. 자세한 정보는 31 페이지의 『시간 절약을 위한 업스트림 표본추출』의 내용을 참조하십시오.

다음 매개변수를 설정할 수 있습니다.

텍스트 필드. 마인드할 텍스트를 포함하는 필드, 문서 경로 이름 또는 문서의 디렉토리 경로 이름을 선택하십시오. 이 필드는 데이터 소스에 따라 다릅니다.

텍스트 필드 제시. 이전 설정에 지정된 텍스트 필드에 포함된 것을 표시합니다. 선택 사항은 다음과 같습니다.

- **실제 텍스트.** 필드에 개념을 추출해야 하는 정확한 텍스트가 포함된 경우 이 옵션을 선택하십시오.
- **문서의 경로명.** 필드에 텍스트 문서가 상주하는 위치에 대한 하나 이상의 경로 이름이 포함된 경우 이 옵션을 선택하십시오.

문서 유형. 이 옵션은 텍스트 필드가 **문서의 경로명**을 나타냄을 지정한 경우에만 사용할 수 있습니다. 문서 유형은 텍스트의 구조를 지정합니다. 다음 유형 중 하나를 선택하십시오.

- **전체 텍스트.** 대부분의 문서 또는 텍스트 소스에 대해 사용합니다. 추출을 위해 전체 텍스트 세트가 스캔됩니다. 다른 옵션과 달리, 이 옵션의 추가 설정은 없습니다.
- **구조화된 텍스트.** 식별하고 분석할 수 있는 일반 구조를 포함하는 파일, 서지 목록 양식 및 특허에 사용됩니다. 이 문서 유형은 추출 프로세스의 일부 또는 전체를 건너뛰기 위해 사용됩니다. 이 문서 유형을 사용하여 항 구분 문자를 정의하고, 유형을 지정하며, 최소 빈도 값을 부과할 수 있습니다. 이 옵션을 선택하면, **설정** 단추를 클릭하고 문서 설정 대화 상자의 **구조화된 텍스트 형식** 영역에서 텍스트 구분 문자를 입력해야 합니다. 자세한 정보는 『필드 탭의 문서 설정』의 내용을 참조하십시오.
- **XML 텍스트.** 추출될 텍스트를 포함하는 XML 태그를 지정하기 위해 사용합니다. 다른 모든 태그는 무시됩니다. 이 옵션을 선택하면, **설정** 단추를 클릭하고 문서 설정 대화 상자의 **XML 텍스트 형식** 영역에서 추출 프로세스 동안 읽을 텍스트를 포함하는 XML 요소를 명시적으로 지정해야 합니다. 자세한 정보는 『필드 탭의 문서 설정』의 내용을 참조하십시오.

텍스트 통합. 이 옵션은 텍스트 필드가 **문서의 경로명**을 나타냄을 지정하고 문서 유형으로 **전체 텍스트**를 선택한 경우에만 사용할 수 있습니다. 다음에서 추출 모드를 선택하십시오.

- **문서 모드.** 간단하고 의미적으로 동일한 문서(예: 통신사의 기사)에 사용하십시오.
- **단락 모드.** 웹 페이지와 태그가 없는 문서에 사용하십시오. 추출 프로세스는 내부 태그 및 구문과 같은 특성을 이용하여 문서를 의미적으로 나눕니다. 이 모드가 선택되는 경우, 스코어링은 단락별로 적용됩니다. 따라서, 예를 들어 `apple` 및 `orange`가 동일한 단락에서 발견되는 경우에만 `apple & orange` 규칙은 true입니다.

  참고: 텍스트가 PDF 문서에서 추출되는 방식으로 인해, **단락 모드**는 이러한 문서에 대해 작동하지 않습니다. 이는 추출이 캐리지 리턴 표식을 억제하기 때문입니다.

단락 모드 설정. 이 옵션은 텍스트 필드가 **문서의 경로명**을 나타냄을 지정하고 텍스트 통합 옵션을 **단락 모드**로 설정한 경우에만 사용할 수 있습니다. 추출에서 사용할 문자 임계값을 지정하십시오. 실제 *크기*는 가장 가까운 마침표로 반올림 또는 반내림됩니다. 문서 콜렉션의 텍스트에서 생성되는 단어 연관이 대표적이 되도록 하려면 너무 작은 추출 *크기*를 지정하지 않도록 하십시오.

- **최소.** 추출에서 사용될 최소 문자 수를 지정하십시오.
- **최대값.** 추출에서 사용될 최대 문자 수를 지정하십시오.

입력 인코딩. 이 옵션은 텍스트 필드가 **문서의 경로명**을 나타냄을 표시한 경우에만 사용 가능합니다. 기본 텍스트 인코딩을 지정합니다. 일본어를 제외한 모든 언어의 경우, 변환은 지정되거나 인식되는 인코딩에서 `ISO-8859-1`로 수행됩니다. 따라서 다른 인코딩을 지정하는 경우에도, 추출 엔진은 이 인코딩을 처리 전에 `ISO-8859-1`로 변환합니다. `ISO-8859-1` 인코딩 정의에 맞지 않는 문자는 공백으로 변환됩니다. 일본어 텍스트의 경우, `SHIFT_JIS`, `EUC_JP`, `UTF-8` 또는 `ISO-2022-JP` 인코딩 옵션 중 하나를 선택할 수 있습니다.

파티션 모드. 유형 노드 설정을 바탕으로 파티션하거나 또 다른 파티션을 선택할지 여부를 선택하려면 파티션 모드를 사용하십시오. 파티션화는 데이터를 학습 및 테스트 샘플로 분리합니다.

필드 탭의 문서 설정

구조화된 텍스트 형식

데이터를 구조화하였거나 텍스트 처리 방법에 대해 규칙을 부과하기 위해 추출 프로세스의 일부 또는 전체를 건너뛰려는 경우, **구조화된 텍스트** 문서 유형 옵션을 사용하고 문서 설정 대화 상자의 **구조화된 텍스트 형식** 섹션에서 텍스트를 포함하는 태그 또는 필드를 선언하십시오. 추출된 용어는 선언된 필드 또는 태그(및 하위 태그) 내에 포함된 텍스트에서만 파생됩니다. 선언되지 않은 필드 또는 태그는 무시됩니다.

특정 컨텍스트에서, 언어 처리는 필요하지 않으므로 언어적 추출 엔진이 명시적 선언에 의해 대체될 수 있습니다. 키워드 필드가 세미콜론(;) 또는 콤마(,)와 같은 구분 문자로 분리되는 도서 목록 파일에서는, 두 개의 구분 문자 사이에서 문자열을 추출하는 것으로 충분합니다. 이러한 이유로, 전체 추출 프로세스를 일시중단시키고 대신 용어 구분 문자를 선언하거나, 유형을 추출된 텍스트에 지정하거나, 추출에 대해 최소 빈도 수를 부과하기 위해 특수 처리 규칙을 정의할 수 있습니다.

구조화된 텍스트 요소를 선언할 때 다음 규칙을 사용하십시오.

- 행마다 단 하나의 필드, 태그 또는 요소를 선언할 수 있습니다. 데이터에는 존재하지 않아도 됩니다.
- 선언에서는 대소문자가 구분됩니다.
- `<title id="1234">`와 같은 속성을 가지고 있는 태그를 선언하고 있고 모든 변동 또는 이 경우, 모든 ID를 포함하도록 하려면, 속성이나 닫는 꺾쇠괄호(>) 없이 태그를 추가하십시오(예: `<title`).
- 필드 또는 태그 이름 뒤에 콜론을 추가하여 구조화된 텍스트임을 표시하십시오. 필드 또는 태그 바로 뒤에, 그리고 구분 문자, 유형 또는 빈도 값 이전에 이 콜론을 추가하십시오(예: `author:` 또는 `<place>:`).
- 여러 용어가 필드 또는 태그에 포함되고 구분 문자가 개별 용어를 지정하기 위해 사용됨을 표시하려면 콜론 다음에 구분 문자를 선언하십시오(예: `author:,` 또는 `<section>:;`).
- 태그에서 발견된 내용에 유형을 지정하려면, 콜론 및 구분 문자 다음에 유형 이름을 선언하십시오(예: `author:,Person` 또는 `<place>:;Location`). 자원 편집기에 표시되는 대로 이름을 사용하여 유형을 선언하십시오.
- 필드 또는 태그에 대한 최소 빈도 수를 정의하려면, 행의 끝에서 수를 선언하십시오(예: `author:,Person1` 또는 `<place>:;Location5`). 여기서 n은 사용자가 정의한 빈도 수이고, 필드 또는 태그에서 발견되는 용어는 추출할 전체 문서 또는 레코드 세트에서 최소 n번 발생해야 합니다. 또한 구분 문자도 정의해야 합니다.
- 콜론을 포함하는 태그가 있으면, 선언이 무시되지 않도록 백슬래시 문자를 콜론 앞에 붙여야 합니다. 예를 들어, `<topic:source>` 필드가 있으면, `<topic\:source>`로 입력하십시오.

명령문을 설명하기 위해, 다음과 같은 반복되는 도서 목록 필드가 있다고 가정합니다.

```
author:Morel, Kawashima
abstract:This article describes how fields are declared.
publication:Text Mining Documentation
datepub:March 2010
```

이 예에 대해, 추출 프로세스가 작성자 및 요약에 초점을 맞추고 내용의 나머지는 무시하기를 원한 경우, 다음 필드만 선언합니다.

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

이 예에서, author:,Person1 필드 선언에서는 언어 처리가 필드 내용에서 일시중단되었음을 알려줍니다. 그 대신, 작성자 필드에 두 개 이상 이름이 포함되고(콤마 구분 문자에 의해 다음 이름과 구분되어서) 이 이름은 사람 유형에 지정되어야 한다고 알려주고, 이름이 전체 문서 또는 레코드 세트에서 최소 한 번 발생하는 경우 이를 추출하도록 알립니다. 필드 abstract:는 다른 선언 없이 나열되어 있으므로, 필드는 추출 및 표준 언어 처리 동안 스캔되고 유형 지정이 적용됩니다.

XML 텍스트 형식

특정 XML 태그 내의 텍스트로만 추출 프로세스를 제한하려면, **XML 텍스트** 문서 유형 옵션을 사용하고 문서 설정 대화 상자의 **XML 텍스트 형식** 섹션에서 텍스트를 포함하는 태그를 선언하십시오. 추출된 용어는 이 태그 또는 해당되는 하위 태그 내에 포함된 텍스트에서만 파생됩니다.

중요! 추출 프로세스를 건너뛰고 용어 구분 문자에 대해 규칙을 부과하거나, 추출된 텍스트에 유형을 지정하거나, 추출된 용어에 대해 빈도 수를 부과하려면, 다음에 설명되는 **구조화된 텍스트** 옵션을 사용하십시오.

XML 텍스트 형식에 대해 태그를 선언할 때 다음 규칙을 사용하십시오.

- 행마다 하나의 XML 태그만 선언할 수 있습니다.
- 태그 요소에서는 대소문자가 구분됩니다.
- 태그에 <title id="1234">와 같은 속성이 있고 모든 변동 또는 이 경우, 모든 ID를 포함하도록 하려면, 속성이나 닫는 꺾쇠괄호(>) 없이 태그를 추가하십시오(예: <title).

명령문을 설명하기 위해, 다음과 같은 XML 문서를 가지고 있다고 가정합니다.

```
<section>Rules of the Road
     <title id="01234">Traffic Signals</title>
     <p>Road signs are helpful.</p>
</section>
<p>Learning the rules is important.</p>
```

이 예의 경우 다음 태그를 선언합니다.

```
<section>
<title
```

이 예에서, 태그 <section>을 선언했기 때문에, 이 태그 및 해당되는 중첩 태그의 텍스트 Traffic Signals 및 Road signs are helpful은 추출 프로세스 동안 스캔됩니다. 그러나 태그 <p>가 명시적으로 선언되지 않았고 선언된 태그 내에 중첩된 태그도 아니므로 Learning the rules is important는 무시됩니다.

텍스트 마이닝 노드: 모델 탭

모델 탭은 작성 방법 및 노드 출력에 대한 일반 모델 설정을 지정하는 데 사용됩니다.

다음 매개변수를 설정할 수 있습니다.

모델 이름 목표나 ID 필드(또는 이러한 필드가 지정되지 않은 경우에는 모델 유형)를 기준으로 하여 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

파티션된 데이터 사용. 파티션 필드가 정의된 경우 이 옵션을 사용하면 훈련 파티션의 데이터만 사용하여 모델을 작성합니다.

작성 모드. 이 텍스트 마이닝 노드와의 스트림이 실행될 때 모델 너깃이 생성되는 방법을 지정합니다. 또는 개념을 추출하고 범주를 작성하고 언어학적 자원을 세분화할 뿐 아니라 텍스트 링크 분석을 수행하고 군집을 탐색할 수도 있는 **대화형 작성** 모드를 사용하여 더 실무적이고 예비적인 방식을 사용할 수 있습니다.

- **대화형 작성.** 스트림이 실행될 때 이 옵션은 개념 및 패턴을 추출하고 추출 결과를 탐색 및 미세 조정하고 범주를 작성 및 세분화하고 언어학적 자원(템플리트, 동의어, 유형, 라이브러리 등)을 세분화하고 범주 모델 너깃을 작성할 수 있는 대화형 인터페이스를 시작합니다. 자세한 정보는 『대화형 작성』의 내용을 참조하십시오.
- **직접 생성.** 이 옵션은 스트림이 실행될 때 모델이 자동으로 작성되고 모델 팔레트에 추가되어야 함을 표시합니다. 대화형 워크벤치와는 달리, 노드에서 정의된 설정 외에 실행 시에 사용자의 추가 조작이 필요 없습니다. 이 옵션을 선택하는 경우, 생성하려는 모델의 유형을 정의할 수 있는 모델 특정 옵션이 나타납니다. 자세한 정보는 27 페이지의 『직접 생성』의 내용을 참조하십시오.

자원 복사 출처. 텍스트를 마이닝할 때, 추출은 전문가 탭의 설정뿐 아니라 언어학적 자원도 기반으로 합니다. 이들 자원은 개념, 유형 및 가끔은 패턴을 얻기 위해 추출 중에 텍스트를 처리 및 취급하는 방법의 기초로 작용합니다. 자원 템플리트 또는 텍스트 분석 패키지로부터 이 노드로 자원을 복사할 수 있습니다. 하나를 선택한 후 **로드**를 클릭하여 자원이 복사될 패키지나 템플리트를 정의하십시오. 로드하는 순간에 자원의 사본이 노드에 저장됩니다. 그러므로 업데이트된 템플리트 또는 TAP를 사용하려는 경우 여기에서 또는 대화형 워크벤치 세션에서 다시 로드해야 합니다. 사용자 편의를 위해 자원이 복사 및 로드된 날짜 및 시간이 노드에 표시됩니다. 자세한 정보는 27 페이지의 『템플리트 및 TAP에서 자원 복사』의 내용을 참조하십시오.

텍스트 언어. 마이닝할 텍스트의 언어를 식별합니다. 노드에서 복사된 자원은 제시된 언어 옵션을 제어합니다. 자원이 조정된 언어를 선택하거나 **ALL** 옵션을 선택할 수 있습니다. 텍스트 데이터에 맞는 정확한 언어를 지정해야 하지만, 확실하지 않으면 **ALL** 옵션을 선택할 수 있습니다. **ALL**은 일본어 텍스트에 대해 사용할 수 없습니다. 이 **ALL** 옵션을 사용하면 실행 시간이 길어집니다. 먼저 텍스트 언어를 식별하기 위해 모든 문서와 레코드를 스캔하기 위해 자동 언어 인식이 사용되기 때문입니다. 이 옵션을 사용하는 경우, 지원되고 라이센스가 부여된 모든 레코드 또는 문서는 언어에 적절한 내부 사전을 사용하여 추출 엔진에 의해 준비됩니다. 자세한 정보는 221 페이지의 『언어 식별자』의 내용을 참조하십시오. 현재 액세스 권한이 없는 지원되는 언어에 대한 라이센스 구매에 관심이 있는 경우 영업 담당자에게 문의하십시오.

대화형 작성

텍스트 마이닝 모델링 노드의 모델 탭에서 모델 너깃에 대한 작성 모드를 선택할 수 있습니다. **대화형 작성**을 선택하는 경우, 스트림을 실행할 때 대화형 인터페이스가 열립니다. 이 대화형 워크벤치에서 다음을 수행할 수 있습니다.

- 개념을 포함하고 텍스트 데이터에서 핵심적인 아이디어를 발견하기 위해 노력하여 추출하고 추출 결과를 탐색합니다.
- 다양한 방법을 사용하여 개념, 유형, TLA 패턴 및 규칙에서 범주를 작성 및 확장함으로써 문서 및 레코드를 이들 범주로 스코어링할 수 있습니다.

- 언어학적 자원(자원 템플리트, 라이브러리, 사전, 동의어 등)을 세분화함으로써 개념이 추출, 검사 및 세분화되는 반복적 프로세스를 통해 결과를 개선할 수 있습니다.
- 텍스트 링크 분석(TLA)을 수행하고 발견된 TLA 패턴을 사용하여 더 좋은 범주 모델 너깃을 작성합니다. 텍스트 링크 분석 노드는 동일한 예비 옵션이나 모델링 기능을 제공하지 않습니다.
- 새 관계를 발견하기 위한 군집을 생성하고 시각화 분할창에서 개념, 유형, 패턴 및 범주 사이의 관계를 탐색합니다.
- 세분화된 범주 모델 너깃을 IBM SPSS Modeler의 모델 팔레트에 생성하고 이들을 다른 스트림에서 사용합니다.

참고: IBM SPSS Collaboration and Deployment Services 작업을 작성 중인 경우 대화형 모델을 작성할 수 없습니다.

최신 노드 업데이트로부터 세션 작업(범주, TLA, 자원 등)을 사용하십시오. 대화형 워크벤치 세션에서 작업할 때, 세션 데이터(추출 매개변수, 자원, 범주 정의 등)로 노드를 업데이트할 수 있습니다. **세션 작업 사용** 옵션으로 저장된 세션 데이터를 사용하여 대화형 워크벤치를 다시 시작할 수 있습니다. 이 옵션은 세션 데이터가 저장될 수 없었으므로 이 노드를 처음 사용할 때는 사용 불가능합니다. 이 옵션을 사용할 수 있도록 세션 데이터로 노드를 업데이트하는 방법을 알려면 88 페이지의 『모델링 노드 업데이트 및 저장』의 내용을 참조하십시오.

이 옵션을 *사용하여* 세션을 시작하는 경우, 대화형 워크벤치 세션으로부터 수행한 마지막 노드 업데이트의 추출 설정, 범주, 자원 및 다른 모든 작업이 다음에 세션을 시작할 때 사용 가능합니다. 저장된 세션 데이터가 이 옵션에서 사용되므로, 아래 템플리트로부터 복사된 자원 같은 특정 컨텐츠 및 기타 탭은 사용 불가능하고 무시됩니다. 그러나 이 옵션 *없이* 세션을 시작하는 경우 노드가 현재 정의되는 그대로의 노드 컨텐츠만 사용되며, 사용자가 워크벤치에서 수행한 모든 이전 작업이 사용 불가능함을 의미합니다.

참고: 추출 결과가 **세션 작업 사용...** 옵션으로 캐싱된 후 스트림에 대한 소스 노드를 변경하는 경우, 추출 결과가 업데이트되기 원하는 경우에 대화형 워크벤치 세션이 시작된 후 새 추출을 실행해야 합니다.

추출을 건너뛰고 캐시된 데이터 및 결과를 재사용하십시오. 대화형 워크벤치 세션에서 모든 캐시된 추출 결과 및 데이터를 재사용할 수 있습니다. 이 옵션은 특히 시간을 절약하고 세션이 시작될 때 완전히 새로운 추출이 수행되기를 기다리기 보다는 추출 결과를 재사용하기 원할 때 유용합니다. 이 옵션을 사용하려면 이전에 대화형 워크벤치 세션 안에서 이 노드를 업데이트했고 **세션 작업 보존 및 재사용을 위해 추출 결과와 함께 텍스트 데이터 캐시** 옵션을 선택했어야 합니다. 이 옵션을 사용할 수 있도록 세션 데이터로 노드를 업데이트하는 방법을 알려면 88 페이지의 『모델링 노드 업데이트 및 저장』의 내용을 참조하십시오.

세션 시작 방법. 대화형 워크벤치 세션을 시작할 때 처음 발생하기 원하는 보기 및 조치를 표시하는 옵션을 선택하십시오. 시작하는 보기와 상관없이, 세션에 있는 동안은 임의의 보기로 전환할 수 있습니다.

- **추출 결과를 사용한 범주 작성.** 이 옵션은 범주 및 개념 보기에서 대화형 워크벤치를 시작하고 적용 가능한 경우 추출을 수행합니다. 이 보기에서 범주를 작성하고 범주 모델을 생성할 수 있습니다. 또한 다른 보기로 전환할 수도 있습니다. 자세한 정보는 77 페이지의 제 8 장 『대화형 워크벤치 모드』의 내용을 참조하십시오.

- **텍스트 링크 분석(TLA) 결과 탐색.** 이 옵션은 의견이나 텍스트 링크 분석 보기의 다른 링크 같은 텍스트 내의 개념 사이의 관계를 추출 및 식별하여 실행하고 시작합니다. 이 옵션을 사용하고 결과를 얻기 위해서는 TLA 패턴 규칙을 포함하는 템플리트 또는 텍스트 분석 패키지를 선택해야 합니다. 더 큰 데이터 세트에 대해 작업 중인 경우 TLA 추출에 다소 시간이 걸릴 수 있습니다. 이 경우에는 샘플 노드 업스트림 사용을 고려하기 원할 수 있습니다. 자세한 정보는 159 페이지의 제 12 장 『텍스트 링크 분석 탐색』의 내용을 참조하십시오.
- **상관 단어 군집 분석.** 이 옵션은 군집 보기에서 시작하고 모든 오래된 추출 결과를 업데이트합니다. 이 보기에서 상관 단어 군집 분석을 수행할 수 있는데 이것은 군집 세트를 생성합니다. 상관 단어 군집화는 주어진 레코드나 문서에서 동시 발생을 기반으로 두 개념 사이의 링크 값의 강도를 평가하여 시작하고 강하게 링크된 개념을 군집으로 그룹화하여 종료하는 프로세스입니다. 자세한 정보는 77 페이지의 제 8 장 『대화형 워크벤치 모드』의 내용을 참조하십시오.

직접 생성

텍스트 마이닝 모델링 노드의 모델 탭에서 모델 너깃에 대한 작성 모드를 선택할 수 있습니다. **직접 생성**을 선택하는 경우 노드에서 옵션을 설정한 후 스트림을 바로 실행할 수 있습니다. 출력은 개념 모델 너깃으로, 모델 팔레트에 바로 배치되었습니다. 대화형 워크벤치와는 달리, 노드에서 이 옵션에 대해 정의되는 빈도 설정 외에는 실행 시에 추가 조작이 필요하지 않습니다.

모델에 포함시킬 개념의 최대 수. 자동으로(비대화형) 모델을 작성할 때만 적용되는 이 옵션은 개념 모델을 작성하기 원함을 표시합니다. 또한 이 모델이 지정된 개념 수보다 많지 않은 개념을 포함해야 함을 말합니다.

- **최고 빈도를 바탕으로 개념 선택. 최상위 개념 수.** 최고 빈도를 갖는 개념으로 시작할 때 이것은 검사할 개념의 수입니다. 여기에서 빈도는 문서/레코드의 전체 세트에서 개념(및 그의 모든 기본 용어)이 나타나는 횟수를 의미합니다. 한 개념이 한 레코드에서 여러 번 나타날 수 있으므로 이 숫자는 레코드 개수보다 더 높을 수 있습니다.
- **너무 많은 레코드에서 발생하는 개념 선택 취소. 레코드 백분율.** 사용자가 지정한 숫자보다 더 높은 레코드 수 백분율을 갖는 개념을 선택 취소합니다. 이 옵션은 텍스트나 모든 레코드에서 자주 발생하지만 분석에서 의미가 없는 개념을 제외하는 데 유용합니다.

스코어링 속도에 최적화. 기본적으로 선택되는 이 옵션은 작성되는 모델이 최소이며 높은 속도로 스코어링되도록 보장합니다. 이 옵션을 선택 취소하면 더 느리게 스코어링하는 훨씬 더 큰 모델이 작성됩니다. 그러나 더 큰 모델은 생성된 개념 모델에서 초기에 표시되는 스코어가 모델 너깃을 사용하여 동일한 텍스트를 스코어링할 때 얻는 스코어와 동일하도록 보장합니다.

템플리트 및 TAP에서 자원 복사

텍스트를 마이닝할 때, 추출은 전문가 탭의 설정뿐 아니라 언어학적 자원도 기반으로 합니다. 이들 자원은 개념, 유형 및 가끔은 패턴을 얻기 위해 추출 중에 텍스트를 처리 및 취급하는 방법의 기초로 작용합니다. *자원 템플리트*에서 이 노드로 자원을 복사할 수 있으며, 텍스트 마이닝 노드에 있는 경우 *텍스트 분석 패키지*(TAP)를 선택할 수도 있습니다.

기본적으로, 자원은 노드를 캔버스에 추가할 때 제품에 대한 라이센스가 있는 언어에 대한 기본 템플리트로부터 노드로 복사됩니다. 다중 언어에 대한 라이센스가 있는 경우 선택된 첫 번째 언어가 자동으로 로드할 템플리트를 판별하는 데 사용됩니다.

로드하는 순간에 선택된 자원의 사본이 노드에 저장됩니다. 템플리트 또는 TAP의 컨텐츠만 복사되는 반면 템플리트나 TAP 자체는 노드에 링크되지 않습니다. 이것은 이 템플리트 또는 TAP가 나중에 업데이트되는 경우 이들 업데이트가 노드에서 자동으로 사용 가능하지 않음을 의미합니다. 요약하면, 템플리트 또는 TAP의 사본으로 재로드하지 않는 한 또는 텍스트 마이닝 노드를 업데이트하고 **세션 작업 사용** 옵션을 선택하지 않는 한 노드로 로드된 자원이 항상 사용됩니다. **세션 작업 사용**에 대한 자세한 정보는 이 주제에서 추가로 확인하십시오.

템플리트 또는 TAP를 선택할 때 텍스트 데이터와 동일한 언어를 갖는 것을 선택하십시오. 라이센스가 있는 언어로 된 템플리트나 TAP만 사용할 수 있습니다. 텍스트 링크 분석을 수행하려는 경우 TLA 패턴을 포함하는 템플리트를 선택해야 합니다. 템플리트가 TLA 패턴을 포함하는 경우, 자원 템플리트 로드 대화 상자의 TLA 열에 아이콘이 나타납니다.

참고: TAP를 텍스트 링크 분석 노드로 로드할 수 없습니다.

자원 템플리트

자원 템플리트는 라이브러리 및 특정 도메인이나 사용법을 위해 미세 조정된 고급 언어 및 비언어학적 자원의 사전 정의된 세트입니다. 텍스트 마이닝 모델링 노드에서, 노드를 스트림에 추가할 때 기본 템플리트의 자원 사본이 이미 노드에 로드되었지만, **자원 템플리트** 또는 **텍스트 분석 패키지** 중 하나를 선택한 후 **로드**를 클릭하여 템플리트를 변경하거나 텍스트 분석 패키지를 로드할 수 있습니다. 템플리트의 경우 자원 템플리트 로드 대화 상자에서 템플리트를 선택할 수 있습니다.

참고: 목록에서 원하는 템플리트를 보지 않지만 사용자 머신에 내보내진 사본이 있는 경우 지금 가져올 수 있습니다. 또한 이 대화 상자에서 내보내어 다른 사용자와 공유할 수도 있습니다. 자세한 정보는 183 페이지의 『템플리트 가져오기 및 내보내기』의 내용을 참조하십시오.

텍스트 분석 패키지(TAP)

텍스트 분석 패키지(TAP)는 하나 이상의 사전 정의된 범주 세트와 함께 번들로 제공되는 라이브러리 및 고급 언어 및 비언어학적 자원의 사전 정의된 세트입니다. IBM SPSS Modeler Text Analytics는 영어 텍스트 및 일본어 텍스트에 대한 여러 개의 사전 작성된 TAP를 제공하는데, 각각은 특정 도메인에 대해 미세 조정됩니다. 이들 TAP를 편집할 수 없지만 이들을 사용하여 범주 모델 작성을 재개할 수 있습니다. 또한 대화형 세션에서 사용자 자신의 TAP를 작성할 수도 있습니다. 자세한 정보는 148 페이지의 『텍스트 분석 패키지 로드』의 내용을 참조하십시오.

참고: TAP를 텍스트 링크 분석 노드로 로드할 수 없습니다.

"세션 작업 사용" 옵션 사용(모델 탭)

자원이 모델 탭의 노드에 복사되는 동안, 나중에 대화형 세션에서 자원을 변경할 수도 있으며 이런 최종 변경으로 텍스트 마이닝 모델링 노드를 업데이트하기 원할 수도 있습니다. 이 경우 텍스트 마이닝 모델링 노드의 모델 탭에서 **세션 작업 사용** 옵션을 선택할 수 있습니다.

세션 작업 사용을 선택하는 경우, **로드** 단추가 노드에서 사용 안함으로 설정되어 대화형 워크벤치에서 온 자원이 이전에 여기에 로드된 자원 대신에 사용됨을 표시합니다.

세션 작업 사용 옵션을 선택한 후 자원을 변경하기 위해 자원 편집기 보기를 통해 자원을 대화형 워크벤치 세션 안에서 직접 편집 또는 전환할 수 있습니다. 자세한 정보는 181 페이지의 『로드 후 노드 자원 업데이트』의 내용을 참조하십시오.

텍스트 마이닝 노드: 전문가 탭

전문가 탭에는 텍스트가 추출 및 처리되는 방법에 영향을 주는 특정 고급 매개변수가 들어 있습니다. 이 대화상자의 매개변수는 추출 프로세스의 기본 작동뿐 아니라 몇 가지 고급 작동을 제어합니다. 그러나 이들은 사용 가능한 옵션의 일부만 표시합니다. 또한 추출 결과에 영향을 미치는 많은 언어학적 자원 및 옵션이 있는데, 이것은 모델 탭에서 선택하는 자원 템플리트에 의해 제어됩니다. 자세한 정보는 24 페이지의 『텍스트 마이닝 노드: 모델 탭』의 내용을 참조하십시오.

참고: 이 전체 탭은 모델 탭에서 저장된 대화형 워크벤치 정보를 사용하여 **대화형 작성** 모드를 선택한 경우 사용 불가능하며, 이 경우 추출 설정은 마지막 저장된 워크벤치 세션에서 가져옵니다.

네덜란드어, 영어, 프랑스어, 독일어, 이탈리아어, 포르투갈어 및 스페인어 텍스트의 경우

영어, 스페인어, 프랑스어, 독일어 등과 같은 일본어 이외의 언어에 대해 추출할 때마다 다음 매개변수를 설정할 수 있습니다.

참고: 일본어 텍스트에 대한 전문가 설정에 대한 정보는 이 주제를 추가로 확인하십시오.

최소한 [n] 전역 빈도를 사용하는 개념으로 추출 제한. 텍스트의 단어 또는 구문이 추출되기 위해 최소한 발생해야 하는 횟수를 지정합니다. 이 방식에서, 값 5는 전체 레코드 또는 문서 세트에서 최소 5번 발생하는 단어 또는 구문으로 추출을 제한합니다.

일부 경우에, 이 한계를 변경하여 추출 결과와 결국 범주에 큰 차이가 발생할 수 있습니다. 식당 데이터에 대해 작업할 때 이 옵션에 대해 1 이상으로 한계를 증가시키지 마십시오. 이러한 경우, 추출 결과에서 *피자(1), 씬 피자(2), 시금치 피자(2), 즐겨먹는 피자(2)*를 볼 수 있습니다. 그러나 추출을 전역 빈도 5 이상으로 제한하고 다시 추출하면, 더 이상 이 세 개의 개념을 얻을 수 없습니다. 대신 *피자(7)*를 얻게 됩니다. *피자*는 가장 단순한 양식이고 이 단어는 이미 가능한 후보로 존재하고 있기 때문입니다. 텍스트의 나머지에 따라서, 텍스트에 피자가 있는 다른 구문이 계속 있는지 여부에 따라 실제로 8 이상의 빈도를 가질 수 있습니다. 또한 *시금치 피자*가 이미 범주 디스크립터인 경우, 모든 레코드를 캡처하는 대신 *피자*를 디스크립터로 추가해야 할 수 있습니다. 이러한 이유로, 범주가 이미 작성된 경우에는 항상 주의하여 이 한계를 변경하십시오.

이는 추출 전용 기능입니다. 템플리트에 용어(보통 수행되는)가 있고 템플리트에 대한 용어가 텍스트에서 발견되는 경우, 용어는 해당 빈도에 관계없이 색인화됩니다.

예를 들어, 코어 라이브러리에서 <Location> 유형 아래에 "los angeles"를 포함하는 기본 자원 템플리트를 사용한다고 가정하십시오. 문서에 Los Angeles가 한 번만 포함되면, Los Angeles는 개념 목록의 일부가 됩니다. 이를 방지하기 위해서는 **최소한 [n] 전역 빈도를 사용하는 개념으로 추출 제한** 필드에 입력한 값과 동일한 횟수만큼 발생하는 개념을 표시하도록 필터를 설정해야 합니다.

구두점 오류를 조정하십시오. 이 옵션은 추출 중에 구두점 오류가 있는 텍스트(예를 들어, 부적절한 사용법)를 일시적으로 표준화하여 개념의 추출 가능성을 향상시킵니다. 이 옵션은 텍스트가 짧고 품질이 저조한 경우(예를 들어, 개방형 설문조사 반응, 이메일 및 CRM 데이터) 또는 텍스트에 많은 약어가 포함된 경우에 특히 유용합니다.

[n]의 최소 루트 문자 제한의 맞춤법 오류를 수용합니다. 이 옵션은 맞춤법이 자주 틀리는 단어나 맞춤법이 유사한 단어를 하나의 개념으로 그룹화하는 퍼지 그룹화 기술을 적용합니다. 퍼지 그룹화 알고리즘은 일시적으로 모든 모음(맨 처음 것은 제외)을 지우고 추출된 단어에서 이중/삼중 자음을 지운 다음 modeling과 modelling이 함께 그룹화될 수 있도록 이들이 동일한지 비교합니다. 그러나 각 용어가 <Unknown> 유형을 제외하고 서로 다른 유형에 지정된 경우에는 퍼지 그룹화 기술은 적용되지 않습니다.

퍼지 그룹화를 사용하기 전에 필요한 루트 문자의 최소 수를 정의할 수도 있습니다. 용어에서 루트 문자의 수는 모든 문자를 합한 후 굴절접사를 형성하는 문자와 복합어의 경우에는 한정사 및 전치사를 형성하는 문자를 빼서 계산합니다. 예를 들어, exercises 용어는 "exercise" 양식의 8개 루트 문자가 있는 것으로 간주됩니다. 단어 끝의 *s*자는 굴절(복수형)이기 때문입니다. 마찬가지로, apple sauce는 10개의 루트 문자로 간주되고("apple sauce") manufacturing of cars는 16개의 루트 문자("manufacturing car")로 간주됩니다. 이 계산 방법은 퍼지 그룹화를 적용해야 하는지 여부를 확인하는 데에만 사용되고 단어가 매치하는 방법에는 영향을 미치지 않습니다.

참고: 특정 단어가 나중에 잘못 그룹화되는 경우에는 이를 고급 자원 탭의 **퍼지 그룹화: 예외** 섹션에 명시적으로 선언하여 이 기술에서 단어 쌍을 제외할 수 있습니다. 자세한 정보는 214 페이지의 『퍼지 그룹화』의 내용을 참조하십시오.

단일어 추출. 이 옵션은 단어가 복합어의 일부가 아니거나 명사이거나 인식되지 않은 품사인 경우에만 단일어를 추출합니다.

비언어 엔티티 추출. 이 옵션은 전화 번호, 주민등록번호, 시간, 날짜, 통화, 숫자, 백분율, 이메일 주소 및 HTTP 주소 등과 같은 비언어 엔티티를 추출합니다. 고급 자원 탭의 **비언어 엔티티: 구성** 섹션에서 비언어 엔티티의 특정 유형을 포함하거나 제외할 수 있습니다. 불필요한 엔티티를 사용 안함으로 설정하면 추출 엔진은 처리 시간을 낭비하지 않습니다. 자세한 정보는 219 페이지의 『구성』의 내용을 참조하십시오.

대문자 알고리즘. 이 옵션은 용어의 첫 글자가 대문자인 한 내장된 사전에 없는 단순 및 복합어를 추출합니다. 이 옵션은 가장 적합한 명사를 추출하기 위한 좋은 방법을 제공합니다.

가능한 경우 부분 및 전체 사람 이름을 함께 그룹화. 이 옵션은 텍스트에 다르게 나타나는 이름을 그룹화합니다. 이름은 종종 시작부에는 전체 이름이 언급되고 나중에는 약어로만 표시되기 때문에 이 기능이 유용합니다. 이 옵션은 <Unknown> 유형의 단일어를 <Person>으로 입력되는 복합어와 매치시키려고 시도합니다. 예를 들

어, *doe*가 있고 처음에는 <Unknown>으로 입력되는 경우에는, 추출 엔진은 <Person> 유형에 있는 복합어가 *doe*를 마지막 단어로 포함하는지 여부를 확인합니다(예: *john doe*). 이 옵션은 이름에는 적용되지 않습니다. 이름의 대부분은 단일어로 추출되지 않기 때문입니다.

최대 비기능 단어 순열. 이 옵션은 순열 기술을 적용할 때 존재할 수 있는 비기능 단어 최대 수를 지정합니다. 이 순열 기술은 서로 굴절과는 관계 없이 포함된 비기능 단어(예: of 및 the)만 다른 유사한 구를 그룹화합니다. 예를 들어, 이 값을 최소 두 개의 단어로 설정하고 company officials 및 officials of the company 둘 모두가 추출되었다고 해 봅시다. 이 경우, 추출된 두 용어는 모두 마지막 개념 목록에 그룹화됩니다. 두 용어 모두 of the가 무시될 때 동일한 것으로 간주되기 때문입니다.

참고: 텍스트 링크 분석 결과의 추출이 가능하게 하려면 **텍스트 링크 분석 결과 탐색** 옵션으로 세션을 시작하고 TLA 정의를 포함하는 자원을 선택해야 합니다. 항상 추출 설정 대화 상자를 통해 대화형 워크벤치 세션 중에 나중에 TLA 결과를 추출할 수 있습니다. 자세한 정보는 92 페이지의 『데이터 추출』의 내용을 참조하십시오.

일본어 텍스트의 경우

추출 프로세스가 몇 가지 차이를 갖고 있으므로 대화 상자에는 일본어 텍스트를 위한 여러 가지 옵션이 있습니다. 일본어 텍스트에 대해 작업하려면 이 노드의 모델 탭에서 일본어에 대해 조정된 템플리트나 텍스트 분석 패키지도 선택해야 합니다. 자세한 정보는 27 페이지의 『템플리트 및 TAP에서 자원 복사』의 내용을 참조하십시오.

2차 분석. 추출이 실행될 때 기본 키워드 추출은 기본 유형 세트를 사용하여 수행됩니다. 그러나, 2차 분석기를 선택하면, 추출기가 이제는 불변화사 및 보조 동사를 개념의 일부로 포함하므로 더 많은 수 또는 더 풍부한 개념을 확보할 수 있습니다. 정서 분석의 경우 수많은 추가 유형 또한 포함됩니다. 게다가, 2차 분석기를 선택하면 텍스트 링크 분석 결과를 생성할 수도 있습니다.

참고: 2차 분석기가 호출되면 추출 프로세스를 완료하는 데 더 오래 걸립니다.

- **종속성 분석**. 이 옵션을 선택하면 기본 유형과 키워드 추출로부터 추출 개념의 확장된 불변화사가 생깁니다. 종속 항목 텍스트 링크 분석(TLA)으로부터 더 풍부한 패턴 결과를 얻을 수도 있습니다.
- **정서 분석**. 이 분석기를 선택하면 추가로 추출된 개념이 생기고, 적용 가능한 경우 TLA 패턴 결과 추출이 생깁니다. 기본 유형에 추가로, 80개가 넘는 정서 유형의 혜택을 얻을 수도 있습니다. 이러한 유형은 감정, 정서 및 의견의 표현을 통해 텍스트에서 개념과 패턴을 찾아내는 데 사용됩니다. 정서 분석의 초점을 지시하는 세 개의 옵션, **모든 정서**, **대표 정서만** 및 **결론만**이 있습니다.
- **2차 분석기 없음**. 이 옵션은 모든 2차 분석기를 끕니다. 2차 분석기가 TLA 결과를 얻기 위해 필요하므로 모델 탭에서 **텍스트 링크 분석(TLA) 결과 탐색** 옵션이 선택된 경우 이 옵션은 숨겨집니다. 이 옵션을 선택하지만 나중에 **텍스트 링크 분석(TLA) 결과 탐색** 옵션을 선택하는 경우 스트림 실행 중에 오류가 발생합니다.

시간 절약을 위한 업스트림 표본추출

많은 양의 데이터가 있을 때 처리 시간은 특히 대화형 워크벤치 세션을 사용할 때 몇 분에서 몇 시간까지 걸릴 수 있습니다. 데이터의 크기가 클수록, 추출 및 범주화 프로세스에 시간이 더 걸립니다. 효율적으로 작업하

기 위해 텍스트 마이닝 노드에서 IBM SPSS Modeler의 표본 노드 업스트림 중 하나를 추가할 수 있습니다. 이 표본 노드를 사용하여 문서 또는 레코드의 더 작은 서브세트를 사용하는 임의 표본을 취하여 처음 몇 번의 패스를 수행하십시오.

더 작은 표본이 종종 자원을 편집하고 모든 범주가 아닌 대부분을 작성하는 방법을 결정하기에 완벽하게 충분합니다. 그리고 더 작은 데이터 세트에 대해 실행하고 결과에 만족한 후에 동일한 기법을 전체 데이터 세트에 대한 범주를 작성하는 데 적용할 수 있습니다. 그런 다음 사용자가 작성한 범주에 맞지 않는 문서나 레코드를 찾고 필요에 따라 조정할 수 있습니다.

참고: 표본 노드는 표준 IBM SPSS Modeler 노드입니다.

스트림에서 텍스트 마이닝 노드 사용

텍스트 마이닝 모델링 노드는 데이터에 액세스하고 스트림에서 개념을 추출하는 데 사용됩니다. 데이터베이스 노드, 변수 파일 노드, 웹 피드 노드 또는 고정 파일 노드 같은 모든 소스 노드를 사용하여 데이터에 액세스할 수 있습니다. 외부 문서에 상주하는 텍스트의 경우 파일 목록 노드를 사용할 수 있습니다.

예 1: 개념 모델 너깃을 직접 작성하기 위한 파일 목록 노드 및 텍스트 마이닝 노드

다음 예는 파일 목록 노드를 텍스트 마이닝 모델링 노드와 함께 사용하여 개념 모델 너깃을 생성하는 방법을 보여줍니다. 파일 목록 노드 사용에 대한 자세한 정보는 11 페이지의 『파일 목록 노드』를 참조하십시오.

1. **파일 목록 노드(설정 탭).** 먼저, 이 노드를 스트림에 추가하여 텍스트 문서가 저장되는 장소를 지정했습니다. 텍스트 마이닝을 수행하려는 모든 문서를 포함하는 디렉토리를 선택했습니다.
2. **텍스트 마이닝 노드(필드 탭).** 다음으로, 파일 목록 노드에 텍스트 마이닝 노드를 추가하고 연결했습니다. 이 노드에서, 입력 형식, 자원 템플리트 및 출력 형식을 정의했습니다. 파일 목록 노드에서 생성된 필드 이름을 선택하고, 텍스트 필드가 다른 설정뿐만 아니라 **문서의 경로명**을 나타내는 옵션을 선택했습니다. 자세한 정보는 『스트림에서 텍스트 마이닝 노드 사용』의 내용을 참조하십시오.
3. **텍스트 마이닝 노드(모델 탭)**. 다음, 모델 탭에서 이 노드에서 직접 개념 모델 너깃을 생성하는 작성 모드를 선택했습니다. 다른 자원 템플리트를 선택하거나 기본 자원을 유지할 수 있습니다.

예 2: 범주 노드를 대화식으로 작성하기 위한 Excel 파일 및 텍스트 마이닝 노드

이 예는 텍스트 마이닝 노드가 대화형 워크벤치 세션을 시작할 수 있는 방법을 보여줍니다. 대화형 워크벤치에 대한 자세한 정보는 77 페이지의 제 8 장 『대화형 워크벤치 모드』를 참조하십시오.

1. **Excel 소스 노드(데이터 탭).** 먼저, 이 노드를 스트림에 추가하여 텍스트가 저장되는 위치를 지정했습니다.
2. **텍스트 마이닝 노드(필드 탭).** 다음, 텍스트 마이닝 노드를 추가하고 연결했습니다. 이 첫 번째 탭에서 입력 형식을 정의했습니다. 소스 노드에서 필드 이름을 선택하고 데이터가 Excel 소스 노드에서 직접 오므로 '텍스트 필드가 **실제 텍스트**를 표시' 옵션을 선택했습니다.
3. **텍스트 마이닝 노드(모델 탭).** 다음, 모델 탭에서 범주 모델 너깃을 대화식으로 작성하고 추출 결과를 사용하여 범주를 자동으로 작성할 것을 선택했습니다. 이 예에서 텍스트 분석 패키지로부터 자원 사본 및 범주 세트를 로드했습니다.

4. **대화형 워크벤치 세션**. 다음, 스트림을 실행했으며 대화형 워크벤치 인터페이스가 열렸습니다. 추출이 수행된 후 데이터 탐색 및 범주 개선을 시작했습니다.

텍스트 마이닝 너깃: 개념 모델

텍스트 마이닝 개념 모델 너깃은 모델 탭에서 **모델을 직접 생성** 옵션을 선택한 텍스트 마이닝 모델 노드를 성공적으로 실행할 때마다 작성됩니다. 텍스트 마이닝 개념 모델 너깃은 콜센터의 메모철 데이터 같은 다른 텍스트 데이터에서 핵심 개념의 실시간 발견에 사용됩니다.

개념 모델 너깃 자체는 개념의 목록으로 구성되는데, 이들은 유형에 지정되었습니다. 다른 데이터에 대한 스코어링을 위해 해당 모델에 있는 데이터의 일부 또는 전부를 선택할 수 있습니다. 텍스트 마이닝 모델 너깃을 포함하는 스트림을 실행할 때, 모델을 작성하기 전에 텍스트 마이닝 모델링 노드의 모델 탭에서 선택된 작성 모드에 따라서 새 필드가 데이터에 추가됩니다. 자세한 정보는 『개념 모델: 모델 탭』의 내용을 참조하십시오.

모델 너깃이 변환된 문서를 사용하여 생성된 경우, 스코어링은 변환된 언어로 수행됩니다. 마찬가지로, 모델 너깃이 언어로 영어를 사용하여 생성된 경우, 문서가 영어로 변환되므로 모델 너깃에서 변환 언어를 지정할 수 있습니다.

텍스트 마이닝 모델 너깃은 생성될 때 모델 너깃 팔레트(IBM SPSS Modeler 창의 상단 오른쪽에 있는 모델 탭에 있는)에 위치됩니다.

결과 보기

모델 너깃에 대한 정보를 보려면, 모델 너깃 팔레트를 마우스 오른쪽 단추로 클릭하고 컨텍스트 메뉴에서 **찾아보기**를(또는 스트림의 노드의 경우 **편집**을) 선택하십시오.

스트림에 모델 추가

모델 너깃을 스트림에 추가하려면, 모델 너깃 팔레트에서 아이콘을 클릭하고 노드를 위치시키려는 스트림 캔버스를 클릭하십시오. 또는 아이콘을 마우스 오른쪽 단추로 클릭하고 컨텍스트 메뉴로부터 **스트림에 추가**를 선택하십시오. 그리고 나서 노드에 스트림을 연결하십시오. 그러면 예측을 생성하기 위해 데이터를 전달할 준비가 됩니다.

주의: 스코어링 너깃을 사용하여 범주 모델 및 사용된 템플리트를 둘 다 포함하는 모델링 노드를 다시 생성하려는 경우, 스코어링 너깃을 생성하기 전에 TAP를 작성하고 이를 모델링 노드 대신 대화형 세션에서 사용할 것을 권장합니다.

개념 모델: 모델 탭

개념 모델에서 모델 탭은 추출된 개념의 세트를 표시합니다. 개념은 각 개념에 대해 하나의 행을 갖는 테이블 형식으로 제공됩니다. 이 탭의 목적은 스코어링에 사용될 개념을 선택하는 것입니다.

참고: 범주 모델 너깃을 대신 생성한 경우, 이 탭은 다른 정보를 표시합니다. 자세한 정보는 43 페이지의 『범주 모델 너깃: 모델 탭』의 내용을 참조하십시오.

가장 왼쪽 열의 선택란에 표시되는 것처럼 기본적으로 모든 개념이 스코어링을 위해 선택됩니다. 선택된 상자는 개념이 스코어링에 사용될 것임을 의미합니다. 선택되지 않은 상자는 개념이 스코어링에서 제외될 것임을 의미합니다. 복수 행을 선택하고 선택에 있는 선택란 중 하나를 클릭하여 다중 행을 선택할 수 있습니다.

각 개념에 대해 자세히 알기 위해 다음 열의 각각에서 제공되는 추가 정보를 찾을 수 있습니다.

개념. 이것은 추출된 리드 단어 또는 구입니다. 어떤 경우에는 이 개념이 개념 이름뿐 아니라 이 개념과 연관된 다른 어떤 기본 용어를 나타냅니다. 어떤 기본 용어가 개념의 일부인지 알려면, 이 탭 안에서 기본 용어 분할창을 표시하고 개념을 선택하여 대화 상자의 맨 아래에 있는 대응하는 용어를 보십시오. 자세한 정보는 35 페이지의 『개념 모델의 기본 용어』의 내용을 참조하십시오.

글로벌. 여기에서, 글로벌(빈도)은 개념(및 그의 모든 기본 용어)이 문서/레코드의 전체 세트에서 나타나는 횟수를 의미합니다.

- **막대형 차트.** 텍스트 데이터에서 이 개념의 글로벌 빈도가 막대형 차트로 제공됩니다. 막대는 유형을 시각적으로 구별하기 위해 개념이 지정되는 유형의 색상을 갖습니다.
- **%.** 텍스트 데이터에서 이 개념의 글로벌 빈도가 퍼센트로 제공됩니다.
- **N.** 텍스트 데이터에서 이 개념의 실제 발생 수입니다.

문서. 여기에서 문서는 문서 개수를 말하며, 개념(및 그의 모든 기본 용어)이 나타나는 문서 또는 레코드의 수를 의미합니다.

- **막대형 차트.** 이 개념의 문서 개수가 막대형 차트로 제공됩니다. 막대는 유형을 시각적으로 구별하기 위해 개념이 지정되는 유형의 색상을 갖습니다.
- **%.** 이 개념의 문서 개수가 퍼센트로 제공됩니다.
- **N.** 이 개념을 포함하는 문서 또는 레코드의 실제 수입니다.

유형. 개념이 지정된 유형입니다. 각 개념에 대해 글로벌 및 문서 열이 이 개념이 지정된 유형을 표시하기 위해 색상으로 나타납니다. **유형**은 개념의 시맨틱 그룹입니다. 자세한 정보는 197 페이지의 『유형 사전』의 내용을 참조하십시오.

개념에 대한 작업

테이블에서 셀을 마우스 오른쪽 단추로 클릭하여 다음을 수행할 수 있는 컨텍스트 메뉴를 표시할 수 있습니다.

- **모두 선택.** 테이블의 모든 행이 선택됩니다.
- **복사.** 선택된 개념이 클립보드에 복사됩니다.
- **필드 포함 복사** 선택된 개념이 열 머리말과 함께 클립보드에 복사됩니다.
- **선택 확인.** 테이블에서 선택된 행에 대한 모든 선택란을 선택하므로 해당 개념을 스코어링에 포함시킵니다.
- **선택 취소.** 테이블에서 선택된 행에 대한 선택란을 선택 취소합니다.
- **모두 선택.** 테이블의 모든 선택란을 선택합니다. 그러면 모든 개념이 최종 출력에서 사용됩니다.

- **모두 선택 취소.** 테이블의 모든 선택란을 선택 취소합니다. 개념을 선택 취소하는 것은 개념이 최종 출력에서 사용되지 않음을 의미합니다.
- **개념 포함.** 개념 포함 대화 상자를 표시합니다. 자세한 정보는 『스코어링에 개념 포함을 위한 옵션』의 내용을 참조하십시오.

스코어링에 개념 포함을 위한 옵션

스코어링에 사용될 개념을 빨리 선택 또는 선택 취소하려면 **개념 포함**에 대한 도구 모음 단추를 클릭하십시오.

그림 1. 개념 포함 도구 모음 단추

이 도구 모음 단추를 클릭하면 규칙을 기반으로 개념을 선택할 수 있는 개념 포함 대화 상자가 열립니다. 모델 탭에서 선택 표시를 갖는 모든 개념이 스코어링에 포함됩니다. 이 하위 대화 상자의 규칙을 적용하여 스코어링에 사용될 개념을 변경하십시오.

다음 옵션 중에서 선택할 수 있습니다.

최고 빈도를 바탕으로 개념 선택. 최상위 개념 수. 최고 글로벌 빈도를 갖는 개념으로 시작할 때 이것은 검사할 개념의 수입니다. 여기에서 빈도는 문서/레코드의 전체 세트에서 개념(및 그의 모든 기본 용어)이 나타나는 횟수를 의미합니다. 한 개념이 한 레코드에서 여러 번 나타날 수 있으므로 이 숫자는 레코드 개수보다 더 높을 수 있습니다.

문서 개수를 기반으로 개념 선택. 최소 빈도. 이것은 개념이 선택되기 위해 필요한 최저 문서 개수입니다. 여기에서 문서 개수는 개념(및 그의 모든 기본 용어)이 나타나는 문서/레코드의 수를 의미합니다.

유형에 지정된 개념 선택. 드롭 다운 목록에서 유형을 선택하여 이 유형에 지정된 모든 개념을 선택하십시오. 개념이 추출 프로세스 중에 자동으로 유형에 지정됩니다. **유형**은 개념의 시맨틱 그룹입니다. 유형은 상위 레벨 개념, 긍정적 및 부정적 단어와 규정자, 이름, 장소, 조직 등과 같은 것을 포함합니다. 자세한 정보는 197 페이지의 『유형 사전』의 내용을 참조하십시오.

너무 많은 레코드에서 발생하는 개념 선택 취소. 레코드 백분율. 사용자가 지정한 숫자보다 더 높은 레코드 수 백분율을 갖는 개념을 선택 취소합니다. 이 옵션은 텍스트나 모든 레코드에서 자주 발생하지만 분석에서 의미가 없는 개념을 제외하는 데 유용합니다.

유형에 지정된 개념 선택 취소. 드롭 다운 목록에서 선택하는 유형과 매치하는 개념을 선택 취소합니다.

개념 모델의 기본 용어

테이블에서 선택한 개념에 대해 정의되는 기본 용어를 볼 수 있습니다. 도구 모음의 기본 용어 전환 단추를 클릭하여 대화 상자의 맨 아래에 있는 분할된 분할창에서 기본 용어 테이블을 표시할 수 있습니다.

이런 기본 용어는 언어학적 자원에서 정의되는 동의어(텍스트에서 발견되었는지 여부는 상관없음) 및 모델 너깃, 재배치된 용어, 퍼지 그룹화로부터의 용어 등을 생성하는 데 사용된 텍스트에서 발견되는 모든 추출된 복

수형/단수형 양식을 포함합니다.

그림 2. 기본 용어 도구 모음 단추 표시

참고: 기본 용어의 목록을 편집할 수 없습니다. 이 목록은 대체, 동의어 정의(대체 사전에서), 퍼지 그룹화 등을 통해 생성되는데, 이들은 모두 언어학적 자원에서 정의됩니다. 용어가 개념 아래에 그룹화되는 방법이나 용어가 처리되는 방법을 변경하려면, 자원(대화형 워크벤치의 자원 편집기 또는 템플리트 편집기에서 편집 가능하며 노드에 재로드)에서 직접 변경한 후 스트림을 재실행하여 업데이트된 결과를 갖는 새 모델 너깃을 확보해야 합니다.

기본 용어나 개념을 포함하는 셀을 마우스 오른쪽 단추로 클릭하여 다음을 수행할 수 있는 컨텍스트 메뉴를 표시할 수 있습니다.

- **복사.** 선택된 셀이 클립보드에 복사됩니다.
- **필드 포함 복사.** 선택된 셀이 열 머리말과 함께 클립보드에 복사됩니다.
- **모두 선택.** 테이블의 모든 셀이 선택됩니다.

개념 모델: 설정 탭

설정 탭은 필요한 경우 새 입력 데이터에 대한 텍스트 필드 값을 정의하는 데 사용됩니다. 또한 출력을 위한 데이터 모델을 정의하는 장소입니다(스코어링 모드).

참고: 이 탭은 모델 너깃이 캔버스에 배치될 때만 나타납니다. 모델 팔레트에서 직접 이 대화 상자에 액세스 중일 때는 존재하지 않습니다.

스코어링 모드: 레코드로서의 개념

이 스코어링 모드를 사용하면 각 개념/문서 쌍에 대해 새 레코드가 작성됩니다. 일반적으로 출력에는 입력에 있는 것보다 더 많은 레코드가 있습니다.

입력 필드 외에, 다음의 새 필드가 데이터에 추가됩니다.

표 4. "레코드로서의 개념"에 대한 출력 필드.

필드	설명
Concept	텍스트 데이터 필드에서 발견되는 추출된 개념 이름을 포함합니다.
Type	*위치* 또는 *사용자* 같은 전체 유형 이름으로 개념의 유형을 저장합니다. **유형**은 개념의 시맨틱 그룹입니다. 자세한 정보는 197 페이지의 『유형 사전』의 내용을 참조하십시오.
Count	텍스트 본문(레코드/문서)에서 해당 개념(및 그의 기본 용어)에 대한 발생 수를 표시합니다.

이 옵션을 선택할 때, **구두점 오류 수용**을 제외한 다른 모든 옵션이 표시됩니다.

스코어링 모드: 필드로서의 개념

개념 모델에서 각 입력 레코드에 대해, 주어진 문서에서 발견되는 모든 개념에 대해 새 레코드가 작성됩니다. 그러므로 입력에 있는 경우 만큼의 출력 레코드가 있습니다. 그러나 각 레코드(행)는 이제 모델 탭에서(선택 표시를 사용하여) 선택된 각 개념에 대한 하나의 새 필드(열)를 포함합니다. 각 개념 필드에 대한 값은 이 탭에서 **플래그** 또는 **개수**를 필드 값으로 선택하는지 여부에 따라 다릅니다.

참고: 예를 들어 DB2 데이터베이스에서 매우 큰 데이터 세트를 사용 중인 경우, **필드로서의 개념**을 사용하면 데이터량으로 인해 처리 문제가 발생할 수 있습니다. 이 경우 **레코드로서의 개념**을 대신 사용할 것을 권장합니다.

필드 값. 각 개념에 대한 새 필드가 개수 또는 플래그 값을 포함할지 여부를 선택하십시오.

- **플래그.** 이 옵션은 *Yes*/*No*, *True*/*False*, *T*/*F* 또는 *1*과 *2* 같이 출력에서 두 개의 고유한 값을 갖는 플래그를 얻는 데 사용됩니다. 저장 유형이 선택된 값을 반영하도록 자동으로 설정됩니다. 예를 들어, 플래그에 대해 숫자 값을 입력하는 경우 자동으로 정수 값으로 처리됩니다. 플래그의 저장 유형은 문자열, 정수, 실수 또는 날짜/시간입니다. **True** 및 **False**에 대한 플래그 값을 입력하십시오.
- **개수.** 개념이 주어진 레코드에서 발생한 빈도를 얻는 데 사용됩니다.

필드 이름 확장. 필드 이름의 확장을 지정하십시오. 필드 이름은 개념 이름에 이 확장을 더해서 사용하여 생성됩니다.

- **추가 위치.** 확장이 필드 이름에 추가될 위치를 지정하십시오. 확장을 문자열의 시작에 추가하려면 **접두문자**를 선택하십시오. 확장을 문자열의 끝에 추가하려면 **접미문자**를 선택하십시오.

구두점 오류를 조정하십시오. 이 옵션은 추출 중에 구두점 오류가 있는 텍스트(예를 들어, 부적절한 사용법)를 일시적으로 표준화하여 개념의 추출 가능성을 향상시킵니다. 이 옵션은 텍스트가 짧고 품질이 저조한 경우(예를 들어, 개방형 설문조사 반응, 이메일 및 CRM 데이터) 또는 텍스트에 많은 약어가 포함된 경우에 특히 유용합니다.

참고: 구두점 오류 수용 옵션은 일본어 텍스트에 대해 작업할 때 적용되지 않습니다.

개념 모델: 필드 탭

필드 탭은 필요한 경우 새 입력 데이터에 대한 텍스트 필드 값을 정의하는 데 사용됩니다.

참고: 이 탭은 모델 너깃이 스트림에 배치될 때만 나타납니다. 모델 팔레트에서 직접 이 출력에 액세스 중일 때는 존재하지 않습니다.

텍스트 필드. 마인드할 텍스트를 포함하는 필드, 문서 경로 이름 또는 문서의 디렉토리 경로 이름을 선택하십시오. 이 필드는 데이터 소스에 따라 다릅니다.

텍스트 필드 제시. 이전 설정에 지정된 텍스트 필드에 포함된 것을 표시합니다. 선택 사항은 다음과 같습니다.

- **실제 텍스트.** 필드에 개념을 추출해야 하는 정확한 텍스트가 포함된 경우 이 옵션을 선택하십시오.

- **문서의 경로명.** 필드에 텍스트 문서가 상주하는 위치에 대한 하나 이상의 경로 이름이 포함된 경우 이 옵션을 선택하십시오.

문서 유형. 이 옵션은 텍스트 필드가 **문서의 경로명**을 나타냄을 지정한 경우에만 사용할 수 있습니다. 문서 유형은 텍스트의 구조를 지정합니다. 다음 유형 중 하나를 선택하십시오.

- **전체 텍스트.** 대부분의 문서 또는 텍스트 소스에 대해 사용합니다. 추출을 위해 전체 텍스트 세트가 스캔됩니다. 다른 옵션과 달리, 이 옵션의 추가 설정은 없습니다.
- **구조화된 텍스트.** 식별하고 분석할 수 있는 일반 구조를 포함하는 파일, 서지 목록 양식 및 특허에 사용됩니다. 이 문서 유형은 추출 프로세스의 일부 또는 전체를 건너뛰기 위해 사용됩니다. 이 문서 유형을 사용하여 항 구분 문자를 정의하고, 유형을 지정하며, 최소 빈도 값을 부과할 수 있습니다. 이 옵션을 선택하면, **설정** 단추를 클릭하고 문서 설정 대화 상자의 **구조화된 텍스트 형식** 영역에서 텍스트 구분 문자를 입력해야 합니다. 자세한 정보는 22 페이지의 『필드 탭의 문서 설정』의 내용을 참조하십시오.
- **XML 텍스트.** 추출될 텍스트를 포함하는 XML 태그를 지정하기 위해 사용합니다. 다른 모든 태그는 무시됩니다. 이 옵션을 선택하면, **설정** 단추를 클릭하고 문서 설정 대화 상자의 **XML 텍스트 형식** 영역에서 추출 프로세스 동안 읽을 텍스트를 포함하는 XML 요소를 명시적으로 지정해야 합니다. 자세한 정보는 22 페이지의 『필드 탭의 문서 설정』의 내용을 참조하십시오.

입력 인코딩. 이 옵션은 텍스트 필드가 **문서의 경로명**을 나타냄을 표시한 경우에만 사용 가능합니다. 기본 텍스트 인코딩을 지정합니다. 일본어를 제외한 모든 언어의 경우, 변환은 지정되거나 인식되는 인코딩에서 ISO-8859-1로 수행됩니다. 따라서 다른 인코딩을 지정하는 경우에도, 추출 엔진은 이 인코딩을 처리 전에 ISO-8859-1로 변환합니다. ISO-8859-1 인코딩 정의에 맞지 않는 문자는 공백으로 변환됩니다. 일본어 텍스트의 경우, SHIFT_JIS, EUC_JP, UTF-8 또는 ISO-2022-JP 인코딩 옵션 중 하나를 선택할 수 있습니다.

텍스트 언어. 마이닝될 텍스트의 언어를 식별합니다. 이것은 추출 중에 발견되는 기본 언어입니다. 현재 액세스 권한이 없는 지원되는 언어에 대한 라이센스 구매에 관심이 있는 경우 영업 담당자에게 문의하십시오.

개념 모델: 요약 탭

요약 탭은 모델 자체(*분석* 폴더), 모델에서 사용되는 필드(*필드* 폴더), 모델을 작성할 때 사용된 설정(*작성 설정* 폴더), 모델 학습(*학습 요약* 폴더)에 관한 정보를 제공합니다.

처음 모델링 노드를 찾아볼 때 요약 탭의 폴더는 접혀있습니다. 관심있는 결과를 보려면 폴더 왼쪽에 있는 펼치기 제어를 사용하여 결과를 표시하거나 **모두 확장** 단추를 클릭하여 모든 결과를 표시하십시오. 결과를 본 후에 결과를 숨기려면 펼치기 제어를 사용하여 숨기려는 특정 폴더를 접거나 **모두 접기** 단추를 클릭하여 모든 폴더를 접으십시오.

스트림에서 개념 모델 너깃 사용

텍스트 마이닝 모델링 노드를 사용할 때, 개념 모델 너깃 또는 범주 모델 너깃(대화형 워크벤치 세션을 통해) 중 하나를 생성할 수 있습니다. 다음 예는 단순 스트림에서 개념 모델을 사용하는 방법을 보여줍니다.

예: 개념 모델 너깃을 갖는 통계량 파일 노드

다음 예는 텍스트 마이닝 개념 모델 너깃 사용 방법을 보여줍니다.

그림 3. 예제 스트림: 텍스트 마이닝 개념 모델 너깃을 갖는 통계량 파일 노드

1. **통계량 파일 노드(데이터 탭).** 먼저, 이 노드를 스트림에 추가하여 텍스트 문서가 저장되는 장소를 지정했습니다.

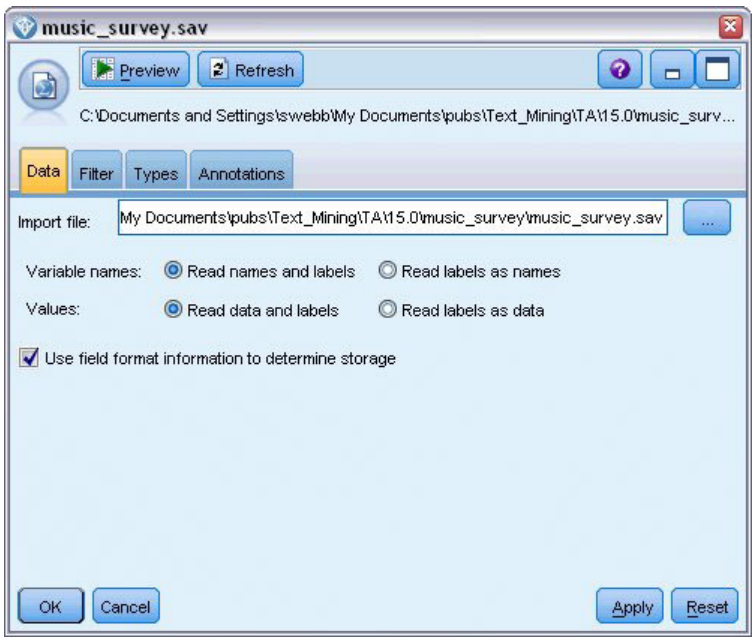


그림 4. 통계량 파일 노드 대화 상자: 데이터 탭

2. **텍스트 마이닝 개념 모델 너깃(모델 탭).** 다음, 통계량 파일 노드에 개념 모델 너깃을 추가하고 연결했습니다. 데이터를 스코어링하는 데 사용하려는 개념을 선택했습니다.

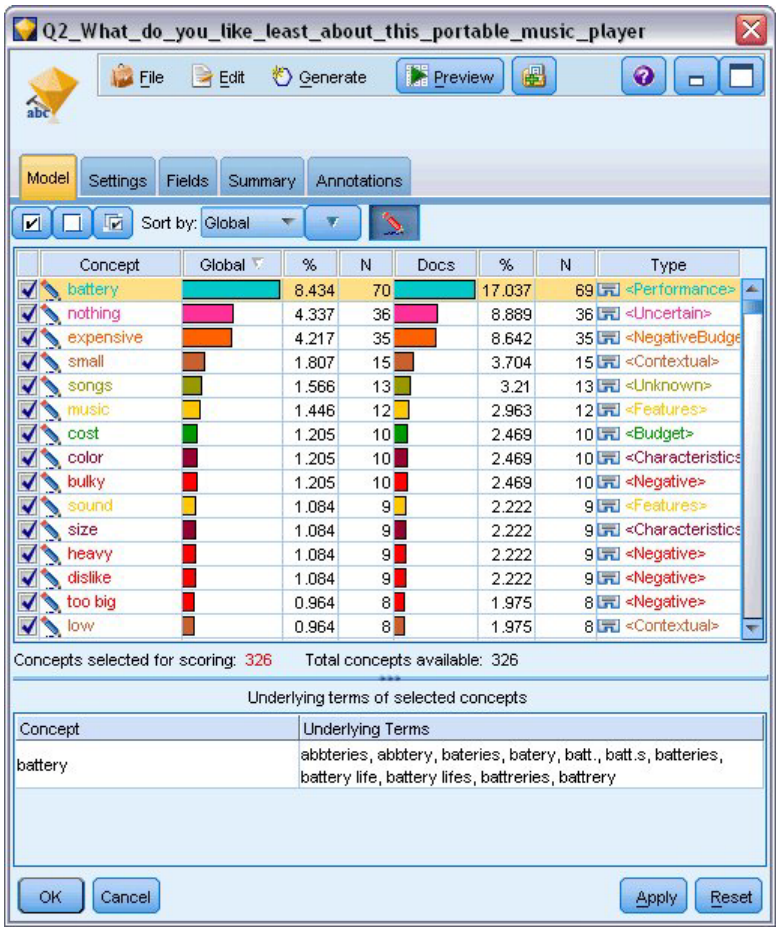


그림 5. 텍스트 마이닝 모델 너깃 대화 상자: 모델 탭

3. **텍스트 마이닝 개념 모델 너깃(설정 탭).** 다음, 출력 형식을 정의하고 *필드로서의 개념*을 선택했습니다. 모델 탭에서 선택되는 각 개념에 대해 하나의 새 필드가 출력에 작성됩니다. 각 필드 이름은 개념 이름과 접두문자 "*Concept_*"으로 구성됩니다.

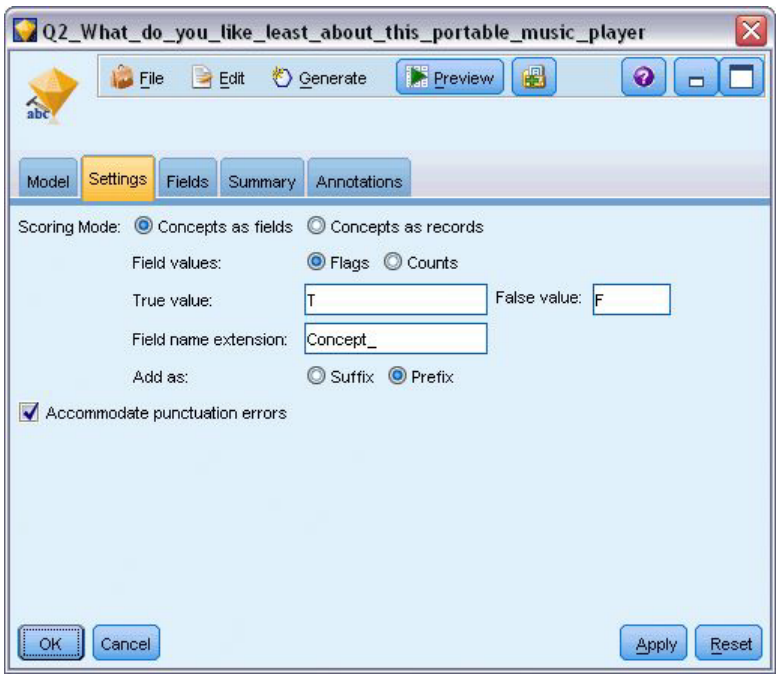


그림 6. 텍스트 마이닝 개념 모델 너깃 대화 상자: 설정 탭

4. **텍스트 마이닝 개념 모델 너깃(필드 탭).** 다음, 텍스트 필드 **Q2_What_do_you_like_least_about_this_portable_music_player**를 선택했는데, 이것은 통계량 파일 노드에서 온 필드 이름입니다. 또한 **텍스트 필드 표시: 실제 텍스트** 옵션을 선택했습니다.

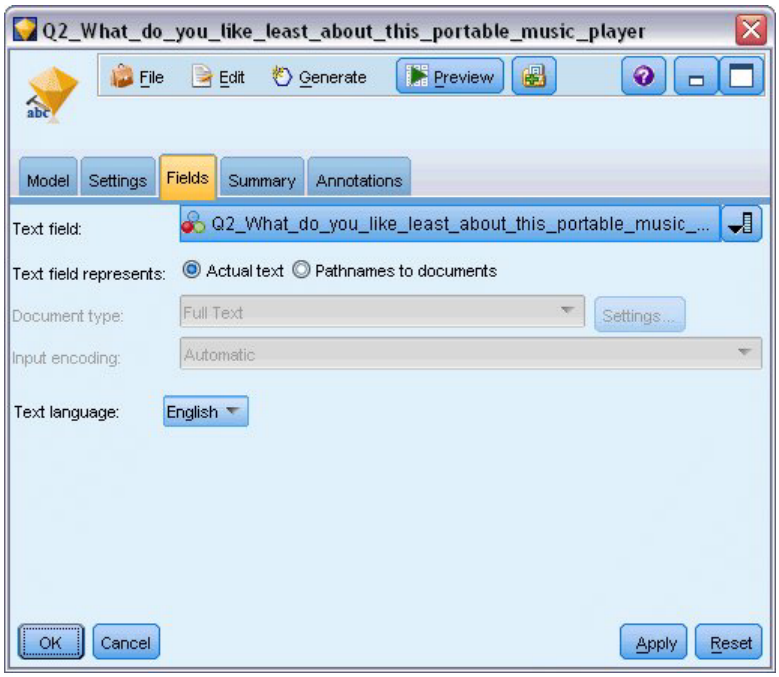
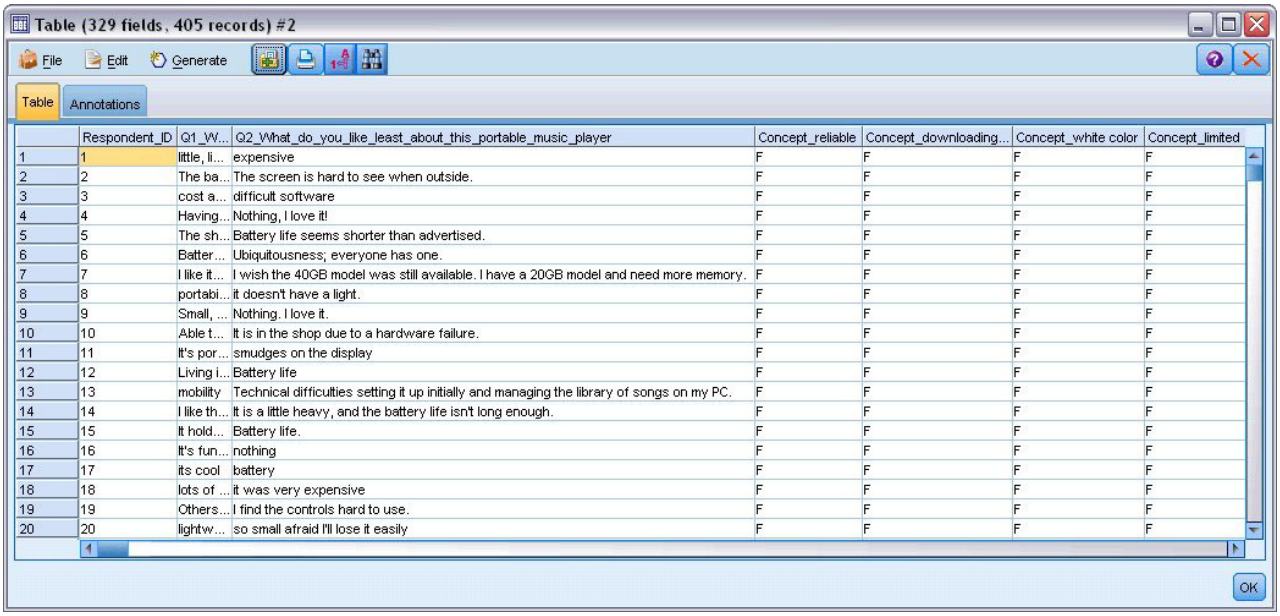


그림 7. 텍스트 마이닝 개념 모델 너깃 대화 상자: 필드 탭

5. **테이블 노드.** 다음, 테이블 노드를 첨부하여 결과를 보고 스트림을 실행했습니다. 테이블 출력이 화면에 열립니다.



	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

그림 8. 개념 플래그를 표시하기 위해 화면 이동된 테이블 출력

텍스트 마이닝 너깃: 범주 모델

텍스트 마이닝 범주 모델 너깃은 대화형 워크벤치 안에서 범주 모델을 생성할 때마다 작성됩니다. 이 모델링 너깃은 범주 세트를 포함하고 있는데, 그의 정의는 개념, 유형, TLA 패턴 및/또는 범주 규칙으로 구성됩니다. 너깃은 설문조사 반응, 블로그 항목, 기타 웹 피드, 다른 모든 텍스트 데이터를 범주화하는 데 사용됩니다.

모델링 노드에서 대화형 워크벤치 세션을 시작하는 경우, 범주 모델을 생성하기 전에 추출 결과를 탐색하고 자원을 세분화하고 범주를 미세 조정할 수 있습니다. 텍스트 마이닝 모델 너깃을 포함하는 스트림을 실행할 때, 모델을 작성하기 전에 텍스트 마이닝 모델링 노드의 모델 탭에서 선택된 작성 모드에 따라서 새 필드가 데이터에 추가됩니다. 자세한 정보는 43 페이지의 『범주 모델 너깃: 모델 탭』의 내용을 참조하십시오.

모델 너깃이 변환된 문서를 사용하여 생성된 경우, 스코어링은 변환된 언어로 수행됩니다. 마찬가지로, 모델 너깃이 언어로 영어를 사용하여 생성된 경우, 문서가 영어로 변환되므로 모델 너깃에서 변환 언어를 지정할 수 있습니다.

텍스트 마이닝 모델 너깃은 생성될 때 모델 너깃 팔레트(IBM SPSS Modeler 창의 상단 오른쪽에 있는 모델 탭에 있는)에 위치됩니다.

결과 보기

모델 너깃에 대한 정보를 보려면, 모델 너깃 팔레트를 마우스 오른쪽 단추로 클릭하고 컨텍스트 메뉴에서 **찾아보기**를(또는 스트림의 노드의 경우 **편집**을) 선택하십시오.

스트림에 모델 추가

모델 너깃을 스트림에 추가하려면, 모델 너깃 팔레트에서 아이콘을 클릭하고 노드를 위치시키려는 스트림 캔버스를 클릭하십시오. 또는 아이콘을 마우스 오른쪽 단추로 클릭하고 컨텍스트 메뉴로부터 **스트림에 추가**를 선택하십시오. 그리고 나서 노드에 스트림을 연결하십시오. 그러면 예측을 생성하기 위해 데이터를 전달할 준비가 됩니다.

주의: 스코어링 너깃을 사용하여 범주 모델 및 사용된 템플리트를 둘 다 포함하는 모델링 노드를 다시 생성하려는 경우, 스코어링 너깃을 생성하기 전에 TAP를 작성하고 이를 모델링 노드 대신 대화형 세션에서 사용할 것을 권장합니다.

범주 모델 너깃: 모델 탭

범주 모델의 경우, 모델 탭이 왼쪽에 범주 모델에 있는 범주의 목록을 표시하고 오른쪽에 선택된 범주에 대한 디스크립터를 표시합니다. 각 범주는 많은 디스크립터로 구성됩니다. 사용자가 선택하는 각 범주에 대해, 연관된 디스크립터가 테이블에 나타납니다. 이들 디스크립터는 개념, 범주 규칙, 유형 및 TLA 패턴을 포함할 수 있습니다. 각 디스크립터의 유형뿐 아니라 각 디스크립터가 나타내는 것의 예도 표시됩니다.

이 탭에서 목적은 스코어링에 사용하려는 범주를 선택하는 것입니다. 범주 모델의 경우 문서 및 레코드가 범주로 스코어링됩니다. 문서나 레코드가 텍스트나 임의의 기본 용어에 하나 이상의 디스크립터를 포함하는 경우, 해당 문서나 레코드는 디스크립터가 속하는 범주에 지정됩니다. 이런 기본 용어는 언어학적 자원에서 정의되는 동의어(텍스트에서 발견되었는지 여부는 상관없음) 및 모델 너깃, 재배치된 용어, 퍼지 그룹화로부터의 용어 등을 생성하는 데 사용된 텍스트에서 발견되는 모든 추출된 복수형/단수형 용어를 포함합니다.

참고: 개념 모델 너깃을 대신 생성한 경우 이 탭은 다른 결과를 포함합니다. 자세한 정보는 33 페이지의 『개념 모델: 모델 탭』의 내용을 참조하십시오.

범주 트리

각 범주에 대해 자세히 알려면 해당 범주를 선택하고 해당 범주에 있는 디스크립터에 대해 나타나는 정보를 검토하십시오. 각 디스크립터에 대해 다음 정보를 검토할 수 있습니다.

- **디스크립터** 이름. 이 필드에는 디스크립터의 종류가 무엇인지를 나타내는 아이콘뿐 아니라 디스크립터 이름이 들어 있습니다.

표 5. 디스크립터 아이콘

	개념		TLA 패턴
	유형		범주 규칙

- **유형.** 이 필드에는 디스크립터의 유형 이름이 들어 있습니다. 유형은 조직 이름, 제품 또는 긍정적 의견 같이 비슷한 개념의 콜렉션(시맨틱 그룹화)입니다. 규칙은 유형에 지정되지 않습니다.

- **세부사항.** 이 필드에는 해당 디스크립터에 포함되는 것의 목록이 들어 있습니다. 매치의 수에 따라서, 대화 상자에서의 크기 한계로 인해 각 디스크립터에 대한 전체 목록을 보지 못할 수 있습니다.

범주 선택 및 복사

왼쪽 분할창의 선택란에 표시되는 것처럼 기본적으로 모든 최상위 범주가 스코어링을 위해 선택됩니다. 선택된 상자는 범주가 스코어링에 사용될 것임을 의미합니다. 선택되지 않은 상자는 해당 범주가 스코어링에서 제외될 것임을 의미합니다. 복수 행을 선택하고 선택에 있는 선택란 중 하나를 클릭하여 다중 행을 선택할 수 있습니다. 또한, 범주 또는 하위 범주가 선택되지만 그의 하위 범주 중 하나가 선택되지 않은 경우, 선택란은 파란색 배경을 표시하여 선택된 범주의 하위에서 부분 선택만 있음을 표시합니다.

트리에서 범주를 마우스 오른쪽 단추로 클릭하면 다음을 수행할 수 있는 컨텍스트 메뉴를 표시할 수 있습니다.

- **선택 확인.** 테이블에서 선택된 행에 대한 모든 선택란을 선택합니다.
- **선택 취소.** 테이블에서 선택된 행에 대한 선택란을 선택 취소합니다.
- **모두 선택.** 테이블의 모든 선택란을 선택합니다. 그러면 모든 범주가 최종 출력에서 사용됩니다. 또한 도구 모음의 대응하는 선택란 아이콘을 사용할 수도 있습니다.
- **모두 선택 취소.** 테이블의 모든 선택란을 선택 취소합니다. 범주를 선택 취소하는 것은 범주가 최종 출력에서 사용되지 않음을 의미합니다. 도구 모음의 대응하는 빈 선택란 아이콘을 사용할 수도 있습니다.

디스크립터 테이블에서 셀을 마우스 오른쪽 단추로 클릭하여 다음을 수행할 수 있는 컨텍스트 메뉴를 표시할 수 있습니다.

- **복사.** 선택된 개념이 클립보드에 복사됩니다.
- **필드 포함 복사.** 선택된 디스크립터가 열 머리말과 함께 클립보드에 복사됩니다.
- **모두 선택.** 테이블의 모든 행이 선택됩니다.

범주 모델 너깃: 설정 탭

설정 탭은 필요한 경우 새 입력 데이터에 대한 텍스트 필드 값을 정의하는 데 사용됩니다. 또한 출력을 위한 데이터 모델을 정의하는 장소입니다(스코어링 모드).

참고: 이 탭은 모델 너깃이 캔버스나 스트림에 위치할 때만 노드 대화 상자에 나타납니다. 모델 팔레트에서 직접 이 너깃에 액세스 중일 때는 존재하지 않습니다.

스코어링 모드: 필드로서의 범주

이 옵션을 사용하면 입력에 있는 경우 만큼의 출력 레코드가 있습니다. 그러나 각 레코드는 이제 모델 탭에서 (선택 표시를 사용하여) 선택된 모든 범주에 대한 하나의 새 필드를 포함합니다. 각 필드에 대해 *Yes*/*No*, *True*/*False*, *T*/*F* 또는 *1* 및 *2* 같이 **True** 및 **False**에 대한 플래그 값을 입력하십시오. 저장 유형은 선택된 값을 반영하도록 자동으로 설정됩니다. 예를 들어, 플래그에 대해 숫자 값을 입력하는 경우 자동으로 정수 값으로 처리됩니다. 플래그의 저장 유형은 문자열, 정수, 실수 또는 날짜/시간입니다.

참고: 예를 들어 DB2 데이터베이스에서 매우 큰 데이터 세트를 사용 중인 경우, **필드로서의 범주**를 사용하면 데이터량으로 인해 처리 문제가 발생할 수 있습니다. 이 경우 **레코드로서의 범주**를 대신 사용할 것을 권장합니다.

필드 이름 확장. 필드 이름에 대해 확장 접두문자/접미문자를 지정하거나 범주 코드를 사용할 것을 선택할 수 있습니다. 필드 이름은 범주 이름에 이 확장을 더해서 사용하여 생성됩니다.

- **추가 위치.** 확장이 필드 이름에 추가될 위치를 지정하십시오. 확장을 문자열의 시작에 추가하려면 **접두문자**를 선택하십시오. 확장을 문자열의 끝에 추가하려면 **접미문자**를 선택하십시오.

하위 범주가 선택되지 않은 경우. 이 옵션으로 스코어링을 위해 선택되지 않은 하위 범주에 속하는 디스크립터가 처리되는 방법을 지정할 수 있습니다. 두 가지 옵션이 있습니다.

- **해당 디스크립터를 스코어링에서 완전히 제외** 옵션은 체크 표시가 없는(선택되지 않은) 하위 범주의 디스크립터가 무시되어 스코어링에서 사용되지 않도록 합니다.
- **디스크립터를 상위 범주에 있는 것과 통합** 옵션은 체크 표시가 없는(선택되지 않은) 하위 범주의 디스크립터가 상위 범주(해당 하위 범주 위의 범주)의 디스크립터로 사용되도록 합니다. 몇 개의 하위 범주 레벨이 선택되지 않은 경우 디스크립터는 첫 번째 사용 가능한 상위 범주 아래에서 롤업됩니다.

구두점 오류를 조정하십시오. 이 옵션은 추출 중에 구두점 오류가 있는 텍스트(예를 들어, 부적절한 사용법)를 일시적으로 표준화하여 개념의 추출 가능성을 향상시킵니다. 이 옵션은 텍스트가 짧고 품질이 저조한 경우(예를 들어, 개방형 설문조사 반응, 이메일 및 CRM 데이터) 또는 텍스트에 많은 약어가 포함된 경우에 특히 유용합니다.

참고: 구두점 오류 수용 옵션은 일본어 텍스트에 대해 작업할 때 적용되지 않습니다.

스코어링 모드: 레코드로서의 범주

이 옵션을 사용하면 각 범주, 문서 쌍에 대해 새 레코드가 작성됩니다. 일반적으로 출력에는 입력에 있는 것보다 더 많은 레코드가 있습니다. 입력 필드 외에, 모델의 종류에 따라서 데이터에 새 필드도 추가됩니다.

표 6. "레코드로서의 범주"에 대한 출력 필드.

새 출력 필드	설명
범주(Category)	텍스트 문서가 지정된 범주 이름이 들어 있습니다. 범주가 다른 범주의 하위 범주인 경우, 범주 이름에 대한 전체 경로는 이 대화 상자에서 선택한 값에 의해 제어됩니다.

계층 구조 범주의 값. 이 옵션은 하위 범주의 이름이 출력에 표시되는 방법을 제어합니다.

- **전체 범주 경로.** 이 옵션은 적용 가능한 경우 슬래시를 사용하여 범주 이름을 하위 범주 이름과 구분하여 범주의 이름 및 상위 범주의 전체 경로를 출력합니다.
- **짧은 범주 경로.** 이 옵션은 범주의 이름만 출력하지만 생략 기호를 사용하여 문제가 되는 범주에 대한 상위 범주의 수를 표시합니다.
- **최하위 레벨 범주.** 이 옵션은 전체 경로 또는 상위 범주가 표시되지 않으면서 범주의 이름만 출력합니다.

하위 범주가 선택되지 않은 경우. 이 옵션으로 스코어링을 위해 선택되지 않은 하위 범주에 속하는 디스크립터가 처리되는 방법을 지정할 수 있습니다. 두 가지 옵션이 있습니다.

- **해당 디스크립터를 스코어링에서 완전히 제외** 옵션은 체크 표시가 없는(선택되지 않은) 하위 범주의 디스크립터가 무시되어 스코어링에서 사용되지 않도록 합니다.
- **디스크립터를 상위 범주에 있는 것과 통합** 옵션은 체크 표시가 없는(선택되지 않은) 하위 범주의 디스크립터가 상위 범주(해당 하위 범주 위의 범주)의 디스크립터로 사용되도록 합니다. 몇 개의 하위 범주 레벨이 선택되지 않은 경우 디스크립터는 첫 번째 사용 가능한 상위 범주 아래에서 롤업됩니다.

구두점 오류를 조정하십시오. 이 옵션은 추출 중에 구두점 오류가 있는 텍스트(예를 들어, 부적절한 사용법)를 일시적으로 표준화하여 개념의 추출 가능성을 향상시킵니다. 이 옵션은 텍스트가 짧고 품질이 저조한 경우(예를 들어, 개방형 설문조사 반응, 이메일 및 CRM 데이터) 또는 텍스트에 많은 약어가 포함된 경우에 특히 유용합니다.

참고: 구두점 오류 수용 옵션은 일본어 텍스트에 대해 작업할 때 적용되지 않습니다.

범주 모델 너깃: 기타 탭

범주 모델 너깃에 대한 필드 탭과 설정 탭은 개념 모델 너깃의 경우와 동일합니다.

- 필드 탭. 자세한 정보는 37 페이지의 『개념 모델: 필드 탭』의 내용을 참조하십시오.
- 요약 탭. 자세한 정보는 38 페이지의 『개념 모델: 요약 탭』의 내용을 참조하십시오.

스트림에서 범주 모델 너깃 사용

텍스트 마이닝 범주 모델 너깃은 대화형 워크벤치 세션에서 생성됩니다. 스트림에서 이 모델 너깃을 사용할 수 있습니다.

예: 범주 모델 너깃을 갖는 통계량 파일 노드

다음 예는 텍스트 마이닝 모델 너깃 사용 방법을 보여줍니다.

그림 9. 예제 스트림: 텍스트 마이닝 범주 모델 너깃을 갖는 통계량 파일 노드

1. **통계량 파일 노드(데이터 탭).** 먼저, 이 노드를 스트림에 추가하여 텍스트 문서가 저장되는 장소를 지정했습니다.

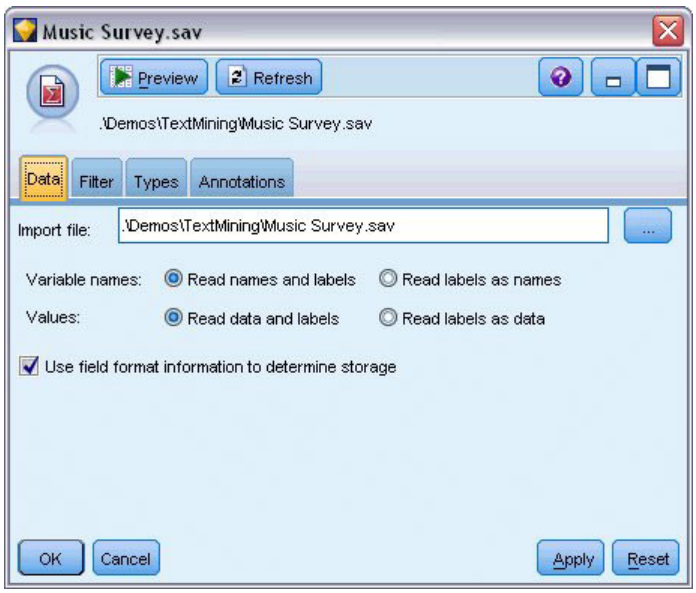


그림 10. 통계량 파일 노드 대화 상자: 데이터 탭

2. **텍스트 마이닝 범주 모델 너깃(모델 탭).** 다음, 통계량 파일 노드에 범주 모델 너깃을 추가하고 연결했습니다. 데이터를 스코어링하는 데 사용하려는 범주를 선택했습니다.

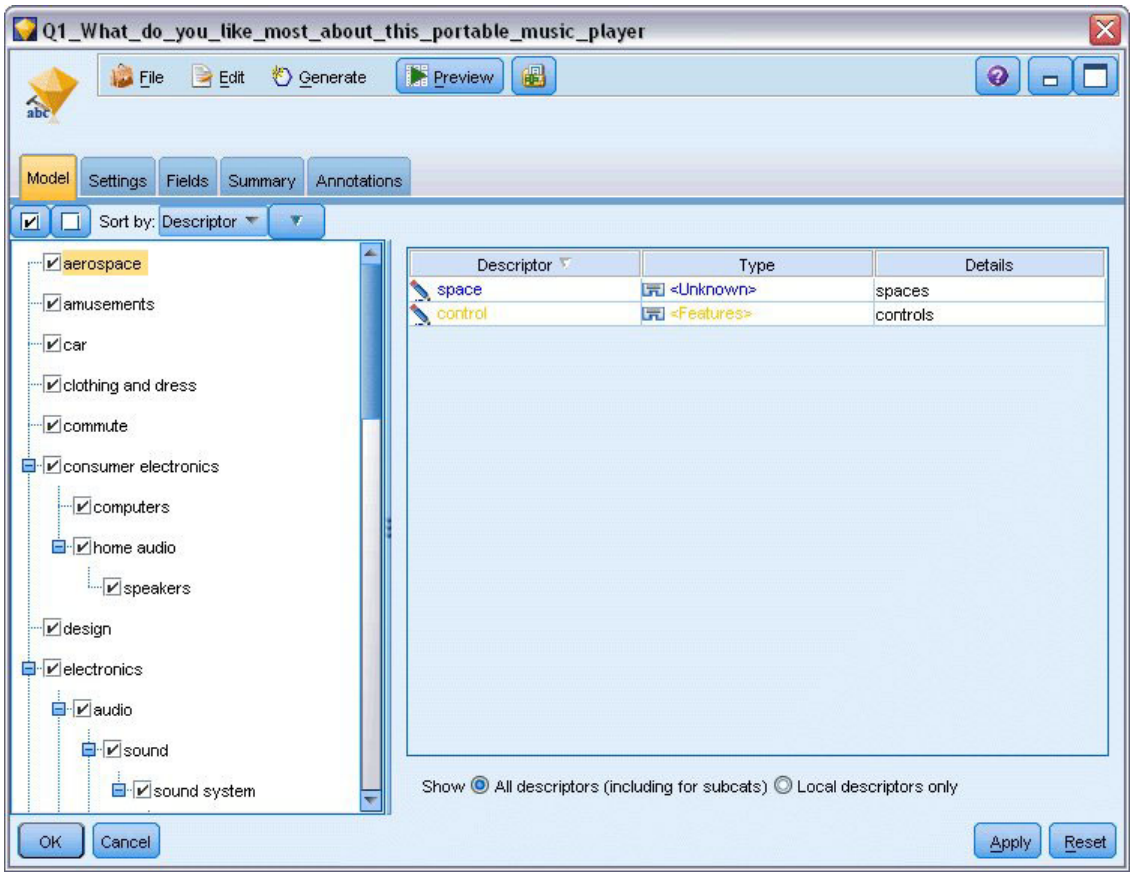


그림 11. 텍스트 마이닝 모델 너깃 대화 상자: 모델 탭

3. **텍스트 마이닝 모델 너깃(설정 탭).** 다음, 출력 형식 **필드로서의 범주**를 정의했습니다.

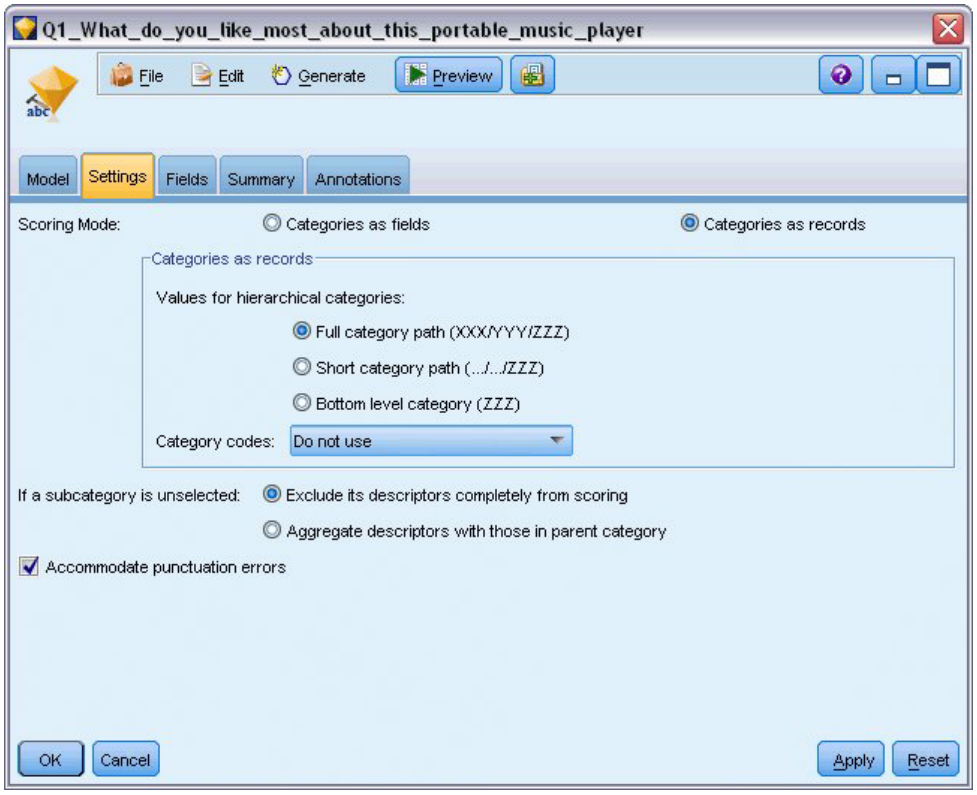


그림 12. 범주 모델 너깃 대화 상자: 설정 탭

4. **텍스트 마이닝 범주 모델 너깃(필드 탭).** 다음, 통계량 파일 노드로부터 오는 필드 이름인 텍스트 필드 변수를 선택했고 옵션 '텍스트 필드가 **실제 텍스트**를 나타냄' 및 기타 설정을 선택했습니다.

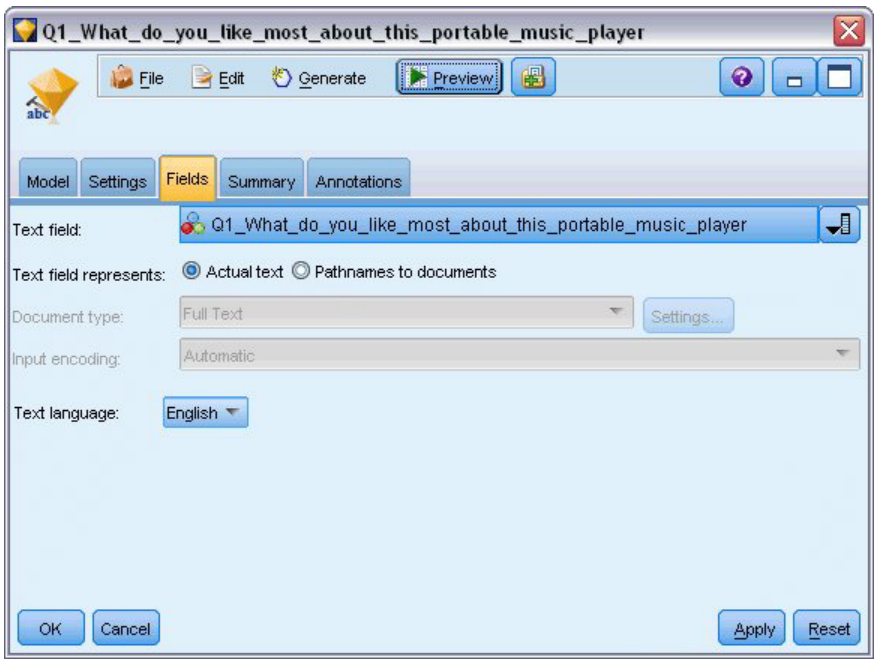
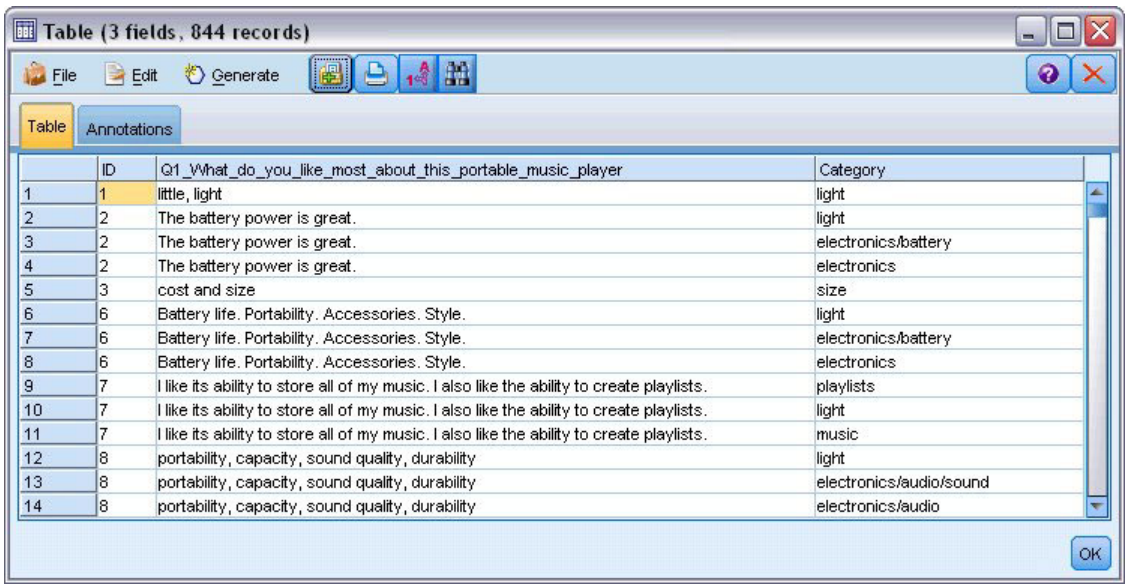


그림 13. 텍스트 마이닝 모델 너깃 대화 상자: 필드 탭

5. **테이블 노드.** 다음, 테이블 노드를 첨부하여 결과를 보고 스트림을 실행했습니다.



	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

그림 14. 테이블 출력

제 4 장 텍스트 링크 마이닝

텍스트 링크 분석 노드

텍스트 링크 분석(TLA) 노드는 알려진 패턴을 바탕으로 텍스트에 있는 개념 사이의 관계를 식별하기 위해 텍스트 마이닝의 개념 추출에 패턴 매치 기술을 추가합니다. 이들 관계는 고객이 제품에 대해 어떻게 느끼는지, 어떤 회사가 함께 비즈니스를 수행 중인지, 또는 유전자 또는 약품 사이의 관계를 설명할 수 있습니다.

예를 들어, 경쟁자의 제품 이름을 추출하는 것은 사용자에게 충분히 흥미롭지 않을 수 있습니다. 이 노드를 사용하면 해당 의견이 데이터에 존재하는 경우 사람들이 이 제품에 대해 어떻게 느끼는지를 알 수도 있습니다. 관계 및 연관은 알려진 패턴을 텍스트 데이터에 매치시켜서 식별 및 추출됩니다.

IBM SPSS Modeler Text Analytics와 함께 제공되는 특정 자원 템플리트 안에서 TLA 패턴 규칙을 사용하거나 사용자 자신의 규칙을 작성/편집할 수 있습니다. 패턴 규칙은 매크로, 단어 목록 및 단어 간격으로 구성되어 입력 텍스트에 대해 비교되는 부울 쿼리 또는 규칙을 형성합니다. TLA 패턴 규칙이 텍스트와 매치할 때마다, 이 텍스트를 TLA 결과로 추출하고 출력 데이터로 재구성할 수 있습니다. 자세한 정보는 223 페이지의 제 19 장 『텍스트 링크 규칙에 대한 정보』의 내용을 참조하십시오.

텍스트 링크 분석 노드는 텍스트에서 TLA 패턴 결과를 식별 및 추출한 후 스트림의 데이터 세트에 결과를 추가하는 보다 직접적인 방법을 제공합니다. 그러나 텍스트 링크 분석 노드가 텍스트 링크 분석을 수행할 수 있는 유일한 방법은 아닙니다. 또한 텍스트 마이닝 모델링 노드에서 대화형 워크벤치 세션을 사용할 수도 있습니다.

대화형 워크벤치에서, TLA 패턴 결과를 탐색하고 이들을 범주 디스크립터로 사용하거나 드릴다운 및 그래프를 사용하여 결과에 대해 자세히 배울 수 있습니다. 자세한 정보는 159 페이지의 제 12 장 『텍스트 링크 분석 탐색』의 내용을 참조하십시오. 사실, 텍스트 마이닝 노드를 사용하여 TLA 결과를 추출하는 것은 나중에 TLA 노드에서 직접 사용하기 위해 템플리트를 탐색하고 데이터에 대해 미세 조정하는 좋은 방법입니다.

출력은 최대 6개의 슬롯 또는 패턴으로 표현될 수 있습니다. 일본어 패턴은 1 - 2개의 슬롯으로만 출력됩니다. 자세한 정보는 55 페이지의 『TLA 노드 출력』의 내용을 참조하십시오.

IBM SPSS Modeler 창의 맨 아래에 있는 노드 팔레트의 IBM SPSS Modeler Text Analytics 탭에서 이 노드를 찾을 수 있습니다. 자세한 정보는 9 페이지의 『IBM SPSS Modeler Text Analytics 노드』의 내용을 참조하십시오.

요구사항. 텍스트 링크 분석 노드는 표준 소스 노드(데이터베이스 노드, 플랫 파일 노드 등) 중 하나를 사용하여 필드로 읽어들이거나 파일 목록 노드나 웹 피드 노드에 의해 생성된 외부 문서에 대한 경로를 나열하는 필드로 읽어들인 텍스트 데이터를 수락합니다.

강도. 텍스트 링크 분석 노드는 개념 *사이의* 관계뿐 아니라 데이터에서 드러날 수 있는 관련된 의견이나 규정자에 관한 정보를 제공하기 위해 기본 개념 추출을 초과합니다.

텍스트 링크 분석 노드: 필드 탭

필드 탭은 개념을 추출 중인 데이터에 대한 필드 설정을 지정하는 데 사용됩니다. 다음 매개변수를 설정할 수 있습니다.

ID 필드. 텍스트 레코드의 식별자를 포함하는 필드를 선택하십시오. 식별자는 정수여야 합니다. ID 필드는 개별 텍스트 레코드에 대한 색인 역할을 수행합니다. 텍스트 필드가 마이닝될 텍스트를 나타내는 경우 ID 필드를 사용하십시오. 텍스트 필드가 **문서의 경로명**을 나타내는 경우 ID 필드를 사용하지 마십시오.

텍스트 필드. 마인드할 텍스트를 포함하는 필드, 문서 경로 이름 또는 문서의 디렉토리 경로 이름을 선택하십시오. 이 필드는 데이터 소스에 따라 다릅니다.

텍스트 필드 제시. 이전 설정에 지정된 텍스트 필드에 포함된 것을 표시합니다. 선택 사항은 다음과 같습니다.

- **실제 텍스트.** 필드에 개념을 추출해야 하는 정확한 텍스트가 포함된 경우 이 옵션을 선택하십시오.
- **문서의 경로명.** 필드에 텍스트 문서가 상주하는 위치에 대한 하나 이상의 경로 이름이 포함된 경우 이 옵션을 선택하십시오.

문서 유형. 이 옵션은 텍스트 필드가 **문서의 경로명**을 나타냄을 지정한 경우에만 사용할 수 있습니다. 문서 유형은 텍스트의 구조를 지정합니다. 다음 유형 중 하나를 선택하십시오.

- **전체 텍스트.** 대부분의 문서 또는 텍스트 소스에 대해 사용합니다. 추출을 위해 전체 텍스트 세트가 스캔됩니다. 다른 옵션과 달리, 이 옵션의 추가 설정은 없습니다.
- **구조화된 텍스트.** 식별하고 분석할 수 있는 일반 구조를 포함하는 파일, 서지 목록 양식 및 특허에 사용됩니다. 이 문서 유형은 추출 프로세스의 일부 또는 전체를 건너뛰기 위해 사용됩니다. 이 문서 유형을 사용하여 항 구분 문자를 정의하고, 유형을 지정하며, 최소 빈도 값을 부과할 수 있습니다. 이 옵션을 선택하면, **설정** 단추를 클릭하고 문서 설정 대화 상자의 **구조화된 텍스트 형식** 영역에서 텍스트 구분 문자를 입력해야 합니다. 자세한 정보는 22 페이지의 『필드 탭의 문서 설정』의 내용을 참조하십시오.
- **XML 텍스트.** 추출될 텍스트를 포함하는 XML 태그를 지정하기 위해 사용합니다. 다른 모든 태그는 무시됩니다. 이 옵션을 선택하면, **설정** 단추를 클릭하고 문서 설정 대화 상자의 **XML 텍스트 형식** 영역에서 추출 프로세스 동안 읽을 텍스트를 포함하는 XML 요소를 명시적으로 지정해야 합니다. 자세한 정보는 22 페이지의 『필드 탭의 문서 설정』의 내용을 참조하십시오.

텍스트 통합. 이 옵션은 텍스트 필드가 **문서의 경로명**을 나타냄을 지정하고 문서 유형으로 **전체 텍스트**를 선택한 경우에만 사용할 수 있습니다. 다음에서 추출 모드를 선택하십시오.

- **문서 모드.** 간단하고 의미적으로 동일한 문서(예: 통신사의 기사)에 사용하십시오.
- **단락 모드.** 웹 페이지와 태그가 없는 문서에 사용하십시오. 추출 프로세스는 내부 태그 및 구문과 같은 특성을 이용하여 문서를 의미적으로 나눕니다. 이 모드가 선택되는 경우, 스코어링은 단락별로 적용됩니다. 따라서, 예를 들어 `apple` 및 `orange`가 동일한 단락에서 발견되는 경우에만 `apple & orange` 규칙은 true입니다.

  참고: 텍스트가 PDF 문서에서 추출되는 방식으로 인해, **단락 모드**는 이러한 문서에 대해 작동하지 않습니다. 이는 추출이 캐리지 리턴 표식을 억제하기 때문입니다.

단락 모드 설정. 이 옵션은 텍스트 필드가 **문서의 경로명**을 나타냄을 지정하고 텍스트 통합 옵션을 **단락 모드**로 설정한 경우에만 사용할 수 있습니다. 추출에서 사용할 문자 임계값을 지정하십시오. 실제 *크기*는 가장 가까운 마침표로 반올림 또는 반내림됩니다. 문서 콜렉션의 텍스트에서 생성되는 단어 연관이 대표적이 되도록 하려면 너무 작은 추출 *크기*를 지정하지 않도록 하십시오.

- **최소.** 추출에서 사용될 최소 문자 수를 지정하십시오.
- **최대값.** 추출에서 사용될 최대 문자 수를 지정하십시오.

입력 인코딩. 이 옵션은 텍스트 필드가 **문서의 경로명**을 나타냄을 표시한 경우에만 사용 가능합니다. 기본 텍스트 인코딩을 지정합니다. 일본어를 제외한 모든 언어의 경우, 변환은 지정되거나 인식되는 인코딩에서 ISO-8859-1로 수행됩니다. 따라서 다른 인코딩을 지정하는 경우에도, 추출 엔진은 이 인코딩을 처리 전에 ISO-8859-1로 변환합니다. ISO-8859-1 인코딩 정의에 맞지 않는 문자는 공백으로 변환됩니다. 일본어 텍스트의 경우, SHIFT_JIS, EUC_JP, UTF-8 또는 ISO-2022-JP 인코딩 옵션 중 하나를 선택할 수 있습니다.

자원 복사 출처. 텍스트를 마이닝할 때, 추출은 전문가 탭의 설정뿐 아니라 언어학적 자원도 기반으로 합니다. 이들 자원은 개념, 유형 및 TLA 패턴을 얻기 위해 추출 중에 텍스트를 처리 및 취급하는 방법의 기초로 작용합니다. 자원 템플리트로부터 이 노드로 자원을 복사할 수 있습니다.

자원 템플리트는 라이브러리 및 특정 도메인이나 사용법을 위해 미세 조정된 고급 언어 및 비언어학적 자원의 사전 정의된 세트입니다. 이들 자원은 추출 중에 데이터를 취급 및 처리하는 방법에 대한 기초로 작용합니다. **로드**를 클릭하고 자원을 복사할 템플리트를 선택하십시오.

템플리트는 스트림이 실행될 때가 아니라 사용자가 템플리트를 선택할 때 로드됩니다. 로드하는 순간에 자원의 사본이 노드에 저장됩니다. 그러므로 업데이트된 템플리트를 사용하기 원한 경우 여기에서 재로드할 필요가 있습니다. 자세한 정보는 27 페이지의 『템플리트 및 TAP에서 자원 복사』의 내용을 참조하십시오.

텍스트 언어. 마이닝할 텍스트의 언어를 식별합니다. 노드에서 복사된 자원은 제시된 언어 옵션을 제어합니다. 자원이 조정된 언어를 선택하거나 **ALL** 옵션을 선택할 수 있습니다. 텍스트 데이터에 맞는 정확한 언어를 지정해야 하지만, 확실하지 않으면 **ALL** 옵션을 선택할 수 있습니다. **ALL**은 일본어 텍스트에 대해 사용할 수 없습니다. 이 **ALL** 옵션을 사용하면 실행 시간이 길어집니다. 먼저 텍스트 언어를 식별하기 위해 모든 문서와 레코드를 스캔하기 위해 자동 언어 인식이 사용되기 때문입니다. 이 옵션을 사용하는 경우, 지원되고 라이센스가 부여된 모든 레코드 또는 문서는 언어에 적절한 내부 사전을 사용하여 추출 엔진에 의해 준비됩니다. 자세한 정보는 221 페이지의 『언어 식별자』의 내용을 참조하십시오. 현재 액세스 권한이 없는 지원되는 언어에 대한 라이센스 구매에 관심이 있는 경우 영업 담당자에게 문의하십시오.

텍스트 링크 분석 노드: 모델 탭

모델 탭에는 추출 프로세스의 속도 및 정확도에 영향을 주는 단일 옵션이 들어 있습니다.

스코어링 속도에 최적화. 기본적으로 선택되는 이 옵션은 작성되는 모델이 최소이며 높은 속도로 스코어링되도록 보장합니다. 이 옵션을 선택 취소하면 더 느리게 스코어링하지만 완전한 개념-유형 일관성을 보장하는 모델이 작성됩니다. 즉, 주어진 개념이 절대 둘 이상의 유형에 지정되지 않게 합니다.

텍스트 링크 분석 노드: 전문가 탭

이 노드에서 텍스트 링크 분석(TLA) 패턴 결과의 추출이 자동으로 사용 가능합니다. 전문가 탭에는 텍스트가 추출되고 처리되는 방법에 영향을 주는 특정한 추가 매개변수가 들어 있습니다. 이 대화 상자의 매개변수는 추출 프로세스의 기본 작동뿐 아니라 몇 가지 고급 작동을 제어합니다. 또한 추출 결과에 영향을 미치는 많은 언어학적 자원 및 옵션이 있는데, 이것은 사용자가 선택하는 자원 템플리트에 의해 제어됩니다.

네덜란드어, 영어, 프랑스어, 독일어, 이탈리아어, 포르투갈어 및 스페인어 텍스트의 경우

구두점 오류를 조정하십시오. 이 옵션은 추출 중에 구두점 오류가 있는 텍스트(예를 들어, 부적절한 사용법)를 일시적으로 표준화하여 개념의 추출 가능성을 향상시킵니다. 이 옵션은 텍스트가 짧고 품질이 저조한 경우(예를 들어, 개방형 설문조사 반응, 이메일 및 CRM 데이터) 또는 텍스트에 많은 약어가 포함된 경우에 특히 유용합니다.

[n]의 최소 루트 문자 제한의 맞춤법 오류를 수용합니다. 이 옵션은 맞춤법이 자주 틀리는 단어나 맞춤법이 유사한 단어를 하나의 개념으로 그룹화하는 퍼지 그룹화 기술을 적용합니다. 퍼지 그룹화 알고리즘은 일시적으로 모든 모음(맨 처음 것은 제외)을 지우고 추출된 단어에서 이중/삼중 자음을 지운 다음 `modeling`과 `modelling`이 함께 그룹화될 수 있도록 이들이 동일한지 비교합니다. 그러나 각 용어가 `<Unknown>` 유형을 제외하고 서로 다른 유형에 지정된 경우에는 퍼지 그룹화 기술은 적용되지 않습니다.

퍼지 그룹화를 사용하기 전에 필요한 루트 문자의 최소 수를 정의할 수도 있습니다. 용어에서 루트 문자의 수는 모든 문자를 합한 후 굴절접사를 형성하는 문자와 복합어의 경우에는 한정사 및 전치사를 형성하는 문자를 빼서 계산합니다. 예를 들어, `exercises` 용어는 "exercise" 양식의 8개 루트 문자가 있는 것으로 간주됩니다. 단어 끝의 *s*자는 굴절(복수형)이기 때문입니다. 마찬가지로, `apple sauce`는 10개의 루트 문자로 간주되고("apple sauce") `manufacturing of cars`는 16개의 루트 문자("manufacturing car")로 간주됩니다. 이 계산 방법은 퍼지 그룹화를 적용해야 하는지 여부를 확인하는 데에만 사용되고 단어가 매치하는 방법에는 영향을 미치지 않습니다.

참고: 특정 단어가 나중에 잘못 그룹화되는 경우에는 이를 고급 자원 탭의 **퍼지 그룹화: 예외** 섹션에 명시적으로 선언하여 이 기술에서 단어 쌍을 제외할 수 있습니다. 자세한 정보는 214 페이지의 『퍼지 그룹화』의 내용을 참조하십시오.

단일어 추출. 이 옵션은 단어가 복합어의 일부가 아니거나 명사이거나 인식되지 않은 품사인 경우에만 단일어를 추출합니다.

비언어 엔티티 추출. 이 옵션은 전화 번호, 주민등록번호, 시간, 날짜, 통화, 숫자, 백분율, 이메일 주소 및 HTTP 주소 등과 같은 비언어 엔티티를 추출합니다. 고급 자원 탭의 **비언어 엔티티: 구성** 섹션에서 비언어 엔티티의 특정 유형을 포함하거나 제외할 수 있습니다. 불필요한 엔티티를 사용 안함으로 설정하면 추출 엔진은 처리 시간을 낭비하지 않습니다. 자세한 정보는 219 페이지의 『구성』의 내용을 참조하십시오.

대문자 알고리즘. 이 옵션은 용어의 첫 글자가 대문자인 한 내장된 사전에 없는 단순 및 복합어를 추출합니다. 이 옵션은 가장 적합한 명사를 추출하기 위한 좋은 방법을 제공합니다.

가능한 경우 부분 및 전체 사람 이름을 함께 그룹화. 이 옵션은 텍스트에 다르게 나타나는 이름을 그룹화합니다. 이름은 종종 시작부에는 전체 이름이 언급되고 나중에는 약어로만 표시되기 때문에 이 기능이 유용합니다. 이 옵션은 <Unknown> 유형의 단일어를 <Person>으로 입력되는 복합어와 매치시키려고 시도합니다. 예를 들어, *doe*가 있고 처음에는 <Unknown>으로 입력되는 경우에는, 추출 엔진은 <Person> 유형에 있는 복합어가 *doe*를 마지막 단어로 포함하는지 여부를 확인합니다(예: *john doe*). 이 옵션은 이름에는 적용되지 않습니다. 이름의 대부분은 단일어로 추출되지 않기 때문입니다.

최대 비기능 단어 순열. 이 옵션은 순열 기술을 적용할 때 존재할 수 있는 비기능 단어 최대 수를 지정합니다. 이 순열 기술은 서로 굴절과는 관계 없이 포함된 비기능 단어(예: of 및 the)만 다른 유사한 구를 그룹화합니다. 예를 들어, 이 값을 최소 두 개의 단어로 설정하고 company officials 및 officials of the company 둘 모두가 추출되었다고 해 봅시다. 이 경우, 추출된 두 용어는 모두 마지막 개념 목록에 그룹화됩니다. 두 용어 모두 of the가 무시될 때 동일한 것으로 간주되기 때문입니다.

일본어 텍스트의 경우

일본어 텍스트에서는 적용할 2차 분석기를 선택할 수 있습니다.

2차 분석. 추출이 실행될 때 기본 키워드 추출은 기본 유형 세트를 사용하여 수행됩니다. 그러나, 2차 분석기를 선택하면, 추출기가 이제는 불변화사 및 보조 동사를 개념의 일부로 포함하므로 더 많은 수 또는 더 풍부한 개념을 확보할 수 있습니다. 정서 분석의 경우 수많은 추가 유형 또한 포함됩니다. 게다가, 2차 분석기를 선택하면 텍스트 링크 분석 결과를 생성할 수도 있습니다.

참고: 2차 분석기가 호출되면 추출 프로세스를 완료하는 데 더 오래 걸립니다.

- **종속성 분석**. 이 옵션을 선택하면 기본 유형과 키워드 추출로부터 추출 개념의 확장된 불변화사가 생깁니다. 종속 항목 텍스트 링크 분석(TLA)으로부터 더 풍부한 패턴 결과를 얻을 수도 있습니다.
- **정서 분석**. 이 분석기를 선택하면 추가로 추출된 개념이 생기고, 적용 가능한 경우 TLA 패턴 결과 추출이 생깁니다. 기본 유형에 추가로, 80개가 넘는 정서 유형의 혜택을 얻을 수도 있습니다. 이러한 유형은 감정, 정서 및 의견의 표현을 통해 텍스트에서 개념과 패턴을 찾아내는 데 사용됩니다. 정서 분석의 초점을 지시하는 세 개의 옵션, **모든 정서**, **대표 정서만** 및 **결론만**이 있습니다.

TLA 노드 출력

텍스트 링크 분석 노드를 실행한 후 데이터가 구조변환됩니다. 텍스트 마이닝이 데이터를 구조변환하는 방법을 이해하는 것이 중요합니다. 데이터 마이닝을 위해 다른 구조를 원하는 경우, 필드 작업 팔레트의 노드를 사용하여 이를 수행할 수 있습니다. 예를 들어, 각 행이 텍스트 레코드를 나타낸 데이터에 대해 작업 중인 경우, 소스 텍스트 데이터에서 발견되는 각 패턴에 대해 한 행이 작성됩니다. 출력의 각 행에 대해 15개의 필드가 있습니다.

- 6개 필드(**Concept1**, **Concept2**, ..., **Concept6** 같은 **Concept#**)는 패턴 매치에서 발견되는 모든 개념을 나타냅니다.
- 6개 필드(**Type1**, **Type2**, ..., **Type6** 같은 **Type#**)는 각 개념에 대한 유형을 나타냅니다.
- **규칙 이름**은 텍스트와 매치하고 출력을 생성하는 데 사용되는 텍스트 링크 규칙의 이름을 나타냅니다.

- 노드에서 사용자가 지정한 ID 필드의 이름을 사용하고 레코드 또는 문서 ID를 입력 데이터에 있었던 대로 나타내는 필드
- **매치된 텍스트**는 TLA 패턴에 매치된 원래 레코드나 문서에 있는 텍스트 데이터의 부분을 나타냅니다.

참고: 일본어 텍스트에 대한 텍스트 링크 분석 패턴 규칙은 1 - 2개의 슬롯 패턴 결과만 생성합니다.

참고: 5.0 이전 릴리스의 텍스트 링크 분석 노드를 포함하는 미리 존재하는 모든 스트림은 사용자가 노드를 업데이트할 때까지 완전히 실행 가능하지 않을 수 있습니다. IBM SPSS Modeler의 최신 버전에서의 특정 개선은 이전 노드가 최신 버전으로 대체되어야 하며, 이것은 배치 가능하면서 더욱 강력합니다.

또한 특정 언어의 자동 변환을 수행할 수도 있습니다. 이 기능은 사용자가 말하거나 읽을 수 없는 언어로 된 문서를 마이닝할 수 있게 합니다. 변환 기능을 사용하려는 경우, SDL SaaS(Software as a Service)에 대한 액세스 권한이 있어야 합니다. 자세한 정보는 61 페이지의 『변환 설정』의 내용을 참조하십시오.

TLA 결과 캐싱

캐시하는 경우 텍스트 링크 분석 결과는 스트림에 있습니다. 스트림이 실행될 때마다 텍스트 링크 분석 결과의 추출 반복을 피하려면 텍스트 링크 분석 노드를 선택하고 메뉴에서 **편집 > 노드 > 캐시 > 사용**을 선택하십시오. 다음에 스트림이 실행될 때 출력이 노드에 캐시됩니다. 노드 아이콘은 캐시가 채워질 때 흰색에서 녹색으로 변하는 작은 "문서" 그래픽을 표시합니다. 캐시는 세션의 기간 동안 유지됩니다. 다른 날(스트림이 닫히고 다시 열린 후)을 위해 캐시를 유지하려면 노드를 선택한 후 메뉴에서 **편집 > 노드 > 캐시 > 캐시 저장**을 선택하십시오. 다음에 스트림을 열 때, 변환을 다시 실행하지 않고 저장된 캐시를 다시 로드할 수 있습니다.

또는 노드를 마우스 오른쪽 단추로 클릭하고 컨텍스트 메뉴에서 **캐시**를 선택하여 노드 캐시를 저장 또는 사용으로 설정할 수 있습니다.

스트림에서 텍스트 링크 분석 노드 사용

텍스트 링크 분석 노드는 데이터에 액세스하고 스트림에서 개념을 추출하는 데 사용됩니다. 임의의 소스 노드를 사용하여 데이터에 액세스할 수 있습니다.

예: 텍스트 링크 분석 노드를 갖는 통계량 파일 노드

다음 예는 텍스트 링크 분석 노드 사용 방법을 보여줍니다.

그림 15. 예: 텍스트 링크 분석 노드를 갖는 통계량 파일 노드

1. **통계량 파일 노드(데이터 탭).** 먼저, 이 노드를 스트림에 추가하여 텍스트가 저장되는 위치를 지정했습니다.

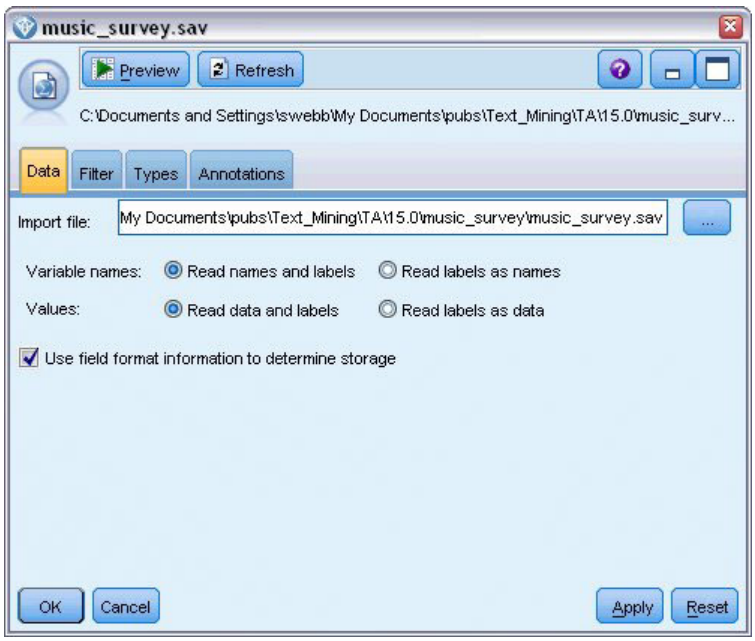


그림 16. 통계량 파일 노드 대화 상자: 데이터 탭

2. **텍스트 링크 분석 노드(필드 탭).** 다음, 다운스트림 모델링 또는 보기를 위한 개념을 추출하기 위해 스트림에 이 노드를 첨부했습니다. ID 필드 및 데이터를 포함하는 텍스트 필드 이름뿐 아니라 다른 설정도 지정했습니다.

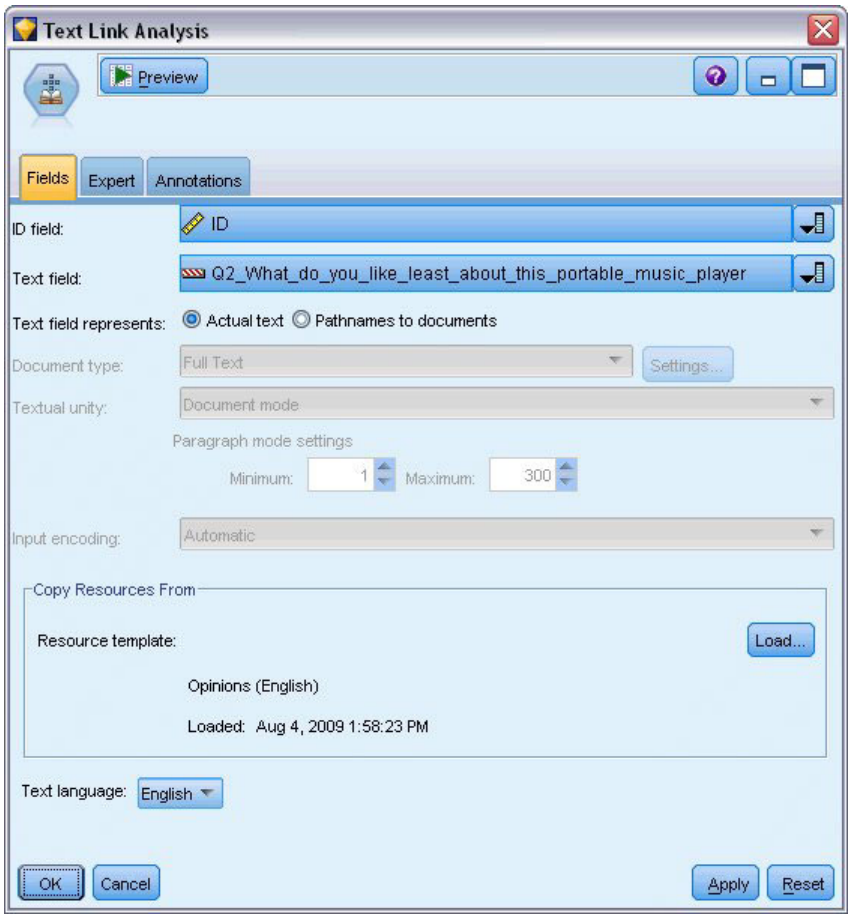
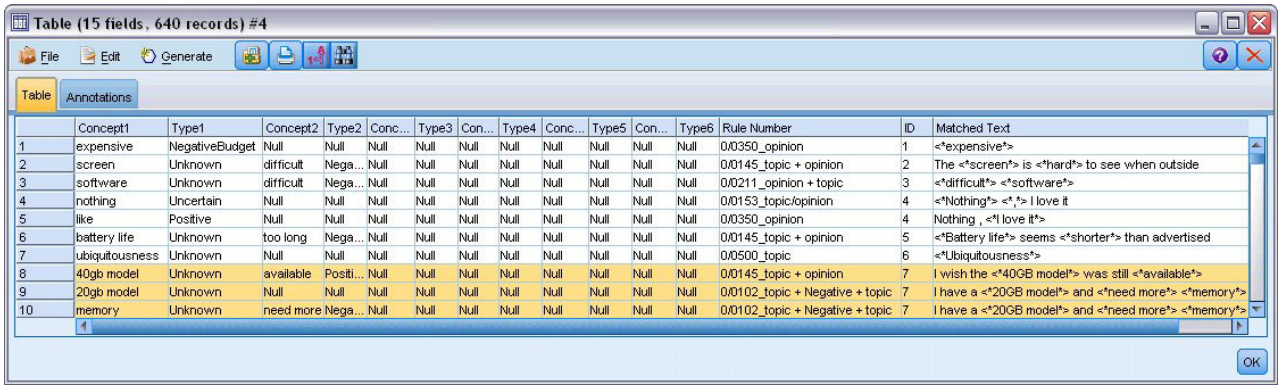


그림 17. 텍스트 링크 분석 노드 대화 상자: 필드 탭

3. **테이블 노드.** 마지막으로, 텍스트 문서에서 추출된 개념을 보기 위해 테이블 노드를 첨부했습니다. 표시된 테이블 출력에서, 이 스트림이 텍스트 링크 분석 노드를 사용하여 실행된 후 데이터에서 발견된 TLA 패턴 결과를 볼 수 있습니다. 일부 결과는 단 하나의 개념/유형이 매치되었음을 표시합니다. 다른 경우에는 결과가 더 복잡하고 여러 가지 유형 및 개념을 포함합니다. 또한, 텍스트 링크 분석 노드를 통해 데이터를 실행하고 개념을 추출한 결과로 데이터의 여러 측면이 변경됩니다. 본 예제의 원 데이터는 8개 필드와 405개의 레코드를 포함했습니다. 텍스트 링크 분석 노드를 실행한 후 이제는 15개 필드와 640개 레코드가 있습니다. 이제 발견된 각 TLA 패턴 결과에 대한 한 행이 있습니다. 예를 들어, 세 개의 TLA 패턴 결과가 추출되었기 때문에 ID 7은 원본에서 세 행이 되었습니다. 이 출력 데이터를 다시 원 데이터에 병합하려는 경우 병합 노드를 사용할 수 있습니다.



	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*>

그림 18. 테이블 출력 노드

제 5 장 추출을 위한 텍스트 변환

변환 노드

번역 노드를 사용하면 IBM SPSS Modeler Text Analytics를 사용한 분석을 위해 아랍어, 중국어 및 페르시아어 같은 지원되는 언어에서 영어로 텍스트를 번역할 수 있습니다. 이것은 그렇지 않은 경우 지원되지 않는 2바이트 언어로 된 문서를 마이닝할 수 있게 하며 분석자가 문제가 되는 언어를 이해할 수 없는 경우에도 외국어 문서에서 개념을 추출할 수 있게 합니다. 번역 노드를 사용할 수 있으려면 SDL의 SaaS(Software as a Service)에 연결할 수 있어야 합니다.

이들 언어 중 하나로 된 텍스트를 마이닝할 때, 단순히 사용자 스트림에서 텍스트 마이닝 모델링 노드 앞에 번역 노드를 추가하십시오. 또한 번역 노드에서 캐싱을 사용 가능하게 하여 스트림이 실행될 때마다 번역을 반복하는 것을 피할 수 있습니다.

IBM SPSS Modeler 창의 맨 아래에 있는 노드 팔레트의 IBM SPSS Modeler Text Analytics 탭에서 이 노드를 찾을 수 있습니다. 자세한 정보는 9 페이지의 『IBM SPSS Modeler Text Analytics 노드』의 내용을 참조하십시오.

번역 캐싱. 번역을 캐시하는 경우, 번역된 텍스트는 외부 파일이 아니라 스트림에 저장됩니다. 스트림이 실행될 때마다 번역 반복을 피하려면 번역 노드를 선택하고 메뉴에서 **편집 > 노드 > 캐시 > 사용**을 선택하십시오. 다음에 스트림이 실행될 때 번역의 결과가 노드에 캐시됩니다. 노드 아이콘은 캐시가 채워질 때 흰색에서 녹색으로 변하는 작은 "문서" 그래픽을 표시합니다. 캐시는 세션의 기간 동안 유지됩니다. 다른 날(스트림이 닫히고 다시 열린 후)을 위해 캐시를 유지하려면 노드를 선택한 후 메뉴에서 **편집 > 노드 > 캐시 > 캐시 저장**을 선택하십시오. 다음에 스트림을 열 때, 번역을 다시 실행하지 않고 저장된 캐시를 다시 로드할 수 있습니다.

또는 노드를 마우스 오른쪽 단추로 클릭하고 컨텍스트 메뉴에서 **캐시**를 선택하여 노드 캐시를 저장 또는 사용으로 설정할 수 있습니다.

중요! 프록시 서버를 통해 웹에서 정보를 검색하려고 시도하는 경우에는 IBM SPSS Modeler Text Analytics 클라이언트와 서버 둘 모두에 대해 net.properties 파일에서 프록시 서버를 사용으로 설정해야 합니다. 이 파일 내부에 설명된 지시사항을 따르십시오. 이는 웹 피드 노드를 통해 웹에 액세스할 때나 SDL SaaS(Software as a Service) 사용권을 검색할 때 적용됩니다. 이러한 연결은 Java을 통과하기 때문입니다. 이 파일은 기본적으로 *C:\Program Files\IBM\SPSS\Modeler\17.1\jre\lib\net.properties*에 있습니다.

참고: IBM SPSS Collaboration and Deployment Services - Scoring 구성 안에서 스코어링을 위해 번역 노드를 사용할 수 없습니다.

변환 노드: 변환 탭

텍스트 필드 마이닝될 텍스트, 문서 경로명 또는 문서에 대한 디렉토리 경로명을 포함하는 필드를 선택하십시오. 이 필드는 데이터 소스에 따라 다릅니다. Direction=None 또는 Type=Typeless를 갖는 경우에도 모든 문자열 필드를 지정할 수 있습니다.

텍스트 필드 표시. 이전 설정에 지정된 텍스트 필드에 포함된 것을 표시합니다. 선택 사항은 다음과 같습니다.

- **실제 텍스트** 필드가 개념이 추출되어야 하는 정확한 텍스트를 포함하는 경우 이 옵션을 선택하십시오.
- **문서의 경로명** 필드가 추출할 텍스트를 포함하는 외부 문서가 상주하는 하나 이상의 경로명을 포함하는 경우 이 옵션을 선택하십시오. 예를 들어 파일 목록 노드가 문서 목록에서 읽는 데 사용되는 경우 이 옵션을 선택해야 합니다. 자세한 정보는 11 페이지의 『파일 목록 노드』의 내용을 참조하십시오.

입력 인코딩 소스 텍스트의 인코딩을 선택하십시오. **자동** 옵션을 선택하여 시작할 수 있지만 일부 파일이 적절하게 처리되지 않고 있음을 아는 경우 여기에 있는 목록에서 실제 인코딩을 선택할 것을 권장합니다. 자동 옵션은 짧은 데이터베이스 레코드 같은 짧은 텍스트를 다룰 때 인코딩을 올바르지 않게 식별할 수 있습니다. 이 노드의 텍스트 출력은 UTF-8로 인코딩됩니다.

설정 스트림에 대한 변환 설정을 지정하십시오.

- **언어 쌍 연결**. 사용하려는 언어 쌍을 선택하십시오. **변환 설정** 대화 상자에서 SDL 서비스에 대한 링크를 설정한 후 사용 가능한 언어 쌍이 자동으로 이 목록에 표시됩니다. 자세한 정보는 61 페이지의 『변환 설정』의 내용을 참조하십시오.
- **터치 포인트**. 이전에 *SDL TouchPoints*를 작성한 경우, 변환과의 연결에서 사용될 터치 포인트를 선택하십시오.
- **가능할 때 이전에 변환된 텍스트 저장 및 재사용** 변환 결과가 저장되어야 하며 다음에 스트림이 실행될 때 동일한 수의 레코드/문서가 존재하는 경우 컨텐츠는 동일하다고 간주되고 처리 시간을 절약하기 위해 변환 결과가 재사용되도록 지정합니다. 이 옵션이 런타임 시에 선택되고 레코드 수가 마지막으로 저장된 것과 매치하지 않는 경우, 텍스트가 완전히 변환된 후 다음 실행을 위한 레이블 이름 아래에 저장됩니다. 이 옵션은 SDL 변환 언어를 선택한 경우에만 사용할 수 있습니다.

  참고: 텍스트가 스트림에 저장되는 경우 변환 노드에서도 캐싱을 사용할 수 있습니다. 이 경우, 변환 결과가 재사용될 뿐 아니라 캐시가 사용 가능할 때마다 모든 업스트림도 무시됩니다.

- **레이블 가능할 때 이전에 변환된 텍스트 저장 및 재사용**을 선택하는 경우, 결과에 대한 레이블 이름을 지정해야 합니다. 이 레이블이 이전에 변환된 텍스트를 식별하는 데 사용됩니다. 레이블이 지정되지 않는 경우, 스트림을 실행하고 재사용이 불가능할 때 경고가 스트림 특성에 추가됩니다.

변환 설정

이 대화 상자에서, 변환할 때 언제든지 재사용할 수 있는 SDL SaaS(Software as a Service) 변환 연결을 정의 및 관리할 수 있습니다. 여기에서 연결을 정의한 후, 모든 연결 설정을 다시 입력할 필요없이 변환 시에 언어 쌍 연결을 빨리 선택할 수 있습니다.

언어 쌍 연결은 소스 및 변환 언어뿐 아니라 서버에 대한 URL 연결 세부사항을 식별합니다. 예를 들어, *중국어 - 영어*는 소스 텍스트는 중국어이고 결과 변환은 영어임을 의미합니다. SDL 온라인 서비스를 통해 액세스하는 각 연결을 수동으로 정의해야 합니다.

중요! 프록시 서버를 통해 웹에서 정보를 검색하려고 시도하는 경우에는 IBM SPSS Modeler Text Analytics 클라이언트와 서버 둘 모두에 대해 `net.properties` 파일에서 프록시 서버를 사용으로 설정해야 합니다. 이 파일 내부에 설명된 지시사항을 따르십시오. 이는 웹 피드 노드를 통해 웹에 액세스할 때나 SDL SaaS(Software as a Service) 사용권을 검색할 때 적용됩니다. 이러한 연결은 Java을 통과하기 때문입니다. 이 파일은 기본적으로 *C:\Program Files\IBM\SPSS\Modeler\17.1\jre\lib\net.properties*에 있습니다.

연결 URL SDL Software as a Service 연결에 대한 URL을 입력하십시오.

API 키 SDL이 사용자에게 제공하는 키를 입력하십시오.

계정 ID SDL이 사용자에게 제공하는 고유 ID를 입력하십시오.

사용자 ID SDL이 사용자에게 제공하는 고유 ID를 입력하십시오.

테스트 연결이 적절하게 구성되었는지 검증하고 해당 연결에서 발견되는 언어 쌍을 보려면 **테스트**를 클릭하십시오.

변환 노드 사용

아랍어, 중국어 또는 페르시아어 같은 지원되는 변환 언어에서 개념을 추출하려면 사용자 스트림에서 임의의 텍스트 마이닝 노드 앞에 변환 노드를 추가하십시오.

변환될 텍스트가 하나 이상의 외부 파일에 포함된 경우 파일 목록 노드를 사용하여 이름의 목록에서 읽을 수 있습니다. 이 경우에 변환 노드는 파일 목록 노드와 모든 후속 텍스트 마이닝 노드 사이에 추가되며, 출력은 변환된 텍스트가 상주하는 위치입니다.

제 6 장 외부 소스 텍스트 찾아보기

파일 뷰어 노드

문서 콜렉션을 마이닝하고 있을 때, 파일의 전체 경로 이름을 텍스트 마이닝 모델링 및 변환 노드에 직접 지정할 수 있습니다. 그러나, 테이블 노드로 출력할 때 그 안에 있는 텍스트가 아니라 문서의 전체 경로 이름만 표시됩니다. 파일 뷰어 노드는 테이블 노드와 유사하게 사용될 수 있으며 이 노드를 사용하여 모두를 단일 파일로 병합할 필요 없이 각 문서 내의 실제 텍스트에 액세스할 수 있습니다.

파일 뷰어 노드는 개념이 추출된 소스 또는 변환되지 않은 텍스트에 대한 액세스 권한을 제공함으로써 텍스트 추출의 결과를 더 잘 이해하는 데 도움을 줄 수 있습니다. 그렇지 않으면 스트림에서 액세스할 수 없기 때문입니다. 이 노드는 모든 파일에 대한 링크의 목록을 얻기 위해 파일 목록 노드 뒤에서 스트림에 추가됩니다.

이 노드의 결과는 개념을 추출하기 위해 읽고 사용한 모든 문서 요소를 표시하는 창입니다. 이 창에서 도구 모음 아이콘을 클릭하여 문서 이름을 하이퍼링크로 나열하는 외부 브라우저에서 보고서를 실행할 수 있습니다. 링크를 클릭하여 콜렉션의 대응하는 문서를 열 수 있습니다. 자세한 정보는 『파일 뷰어 노드 사용』의 내용을 참조하십시오.

IBM SPSS Modeler 창의 맨 아래에 있는 노드 팔레트의 IBM SPSS Modeler Text Analytics 탭에서 이 노드를 찾을 수 있습니다. 자세한 정보는 9 페이지의 『IBM SPSS Modeler Text Analytics 노드』의 내용을 참조하십시오.

참고: 클라이언트-서버 모드에서 작업 중이고 파일 뷰어 노드가 스트림의 일부일 때, 문서 콜렉션이 서버의 웹 서버 디렉토리에 저장되어야 합니다. 텍스트 마이닝 출력 노드가 웹 서버 디렉토리에 저장되는 문서의 목록을 생성하므로, 웹 서버의 보안 설정이 이들 문서에 대한 권한을 관리합니다.

파일 뷰어 노드 설정

파일 뷰어 노드에 대한 다음 설정을 지정할 수 있습니다.

문서 필드. 표시될 문서의 전체 이름 및 경로를 포함하는 데이터에서 필드를 선택하십시오.

생성된 HTML 페이지에 대한 제목. 문서의 목록을 포함하는 페이지의 맨 위에 나타날 제목을 작성하십시오.

파일 뷰어 노드 사용

다음 예는 파일 뷰어 노드 사용법을 보여줍니다.

예: 파일 목록 노드 및 파일 뷰어 노드

그림 19. 파일 뷰어 노드의 사용을 설명하는 스트림

1. **파일 목록 노드(설정 탭).** 먼저, 문서가 위치하는 장소를 지정하기 위해 이 노드를 추가했습니다.

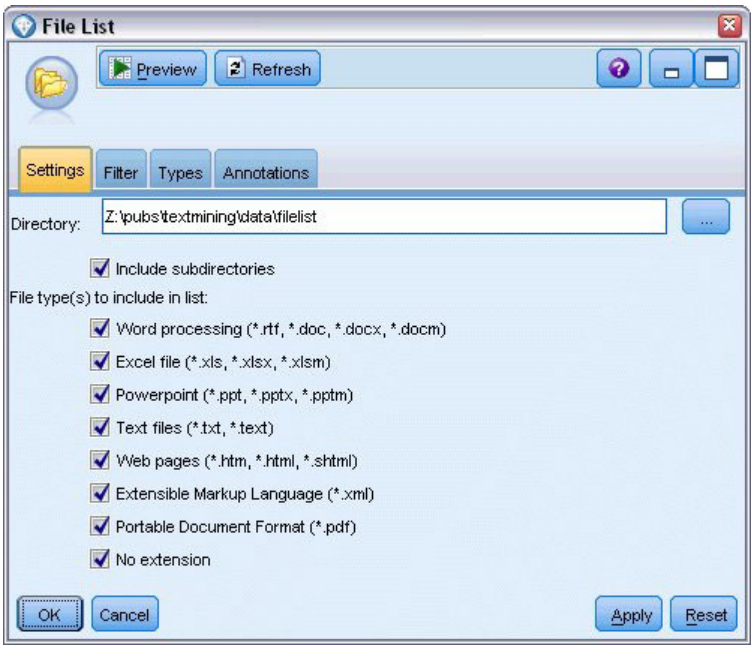


그림 20. 파일 목록 노드 대화 상자: 설정 탭

2. **파일 뷰어 노드(설정 탭).** 다음, 파일 뷰어 노드를 첨부하여 문서의 HTML 목록을 생성했습니다.

그림 21. 파일 뷰어 노드 대화 상자: 설정 탭

3. **파일 뷰어 출력 대화 상자.** 다음, 문서 목록을 새 창에서 출력하는 스트림을 실행했습니다.

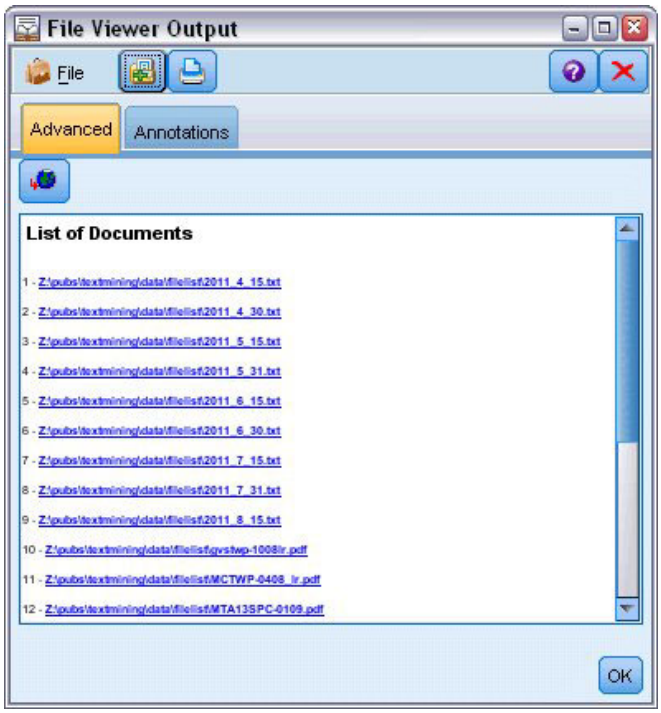


그림 22. 파일 뷰어 출력

4. 문서를 보기 위해 빨간색 화살표를 갖는 지구본을 표시하는 도구 모음 단추를 클릭했습니다. 이것은 브라우저에 문서 하이퍼링크의 목록을 열었습니다.

제 7 장 스크립팅을 위한 노드 특성

IBM SPSS Modeler는 명령행에서 스트림을 실행할 수 있게 하는 스크립팅 언어를 갖고 있습니다. 여기에서 IBM SPSS Modeler Text Analytics와 함께 제공되는 각 노드에 특정한 노드 특성에 대해 배울 수 있습니다. IBM SPSS Modeler와 함께 제공되는 표준 노드 세트에 대한 자세한 정보는 Scripting and Automation Guide를 참조하십시오.

파일 목록 노드: filelistnode

스크립팅에 대해 다음 표의 특성을 사용할 수 있습니다. 노드 자체는 `filelistnode`라고 부릅니다.

표 7. 파일 목록 노드 스크립팅 특성

스크립팅 속성	데이터 유형
`path`	*string*
`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*

참고: '목록 작성' 매개변수는 더 이상 사용할 수 없으며 해당 옵션을 포함하는 모든 스크립트는 자동으로 '파일' 출력으로 변환됩니다.

웹 피드 노드: webfeednode

스크립팅에 대해 다음 표의 특성을 사용할 수 있습니다. 노드 자체를 `webfeednode`라고 합니다.

표 8. 웹 피드 노드 스크립팅 특성

스크립팅 속성	데이터 유형	특성 설명
`urls`	*string1 string2 ...stringn*	각 URL은 목록 구조로 지정됩니다. URL 목록은 “\n”으로 구분됩니다.
`recent_entries`	*flag*	
`limit_entries`	*integer*	URL마다 읽을 최근 항목 수.
`use_previous`	*flag*	웹 피드 캐시를 저장하고 재사용합니다.
`use_previous_label`	*string*	저장된 웹 캐시의 이름.
`start_record`	*string*	비RSS 시작 태그.

표 8. 웹 피드 노드 스크립팅 특성 (계속)

스크립팅 속성	데이터 유형	특성 설명
url *n* .title	*string*	목록에서 각 URL에 대해, 여기에서도 정의해야 합니다. 첫 번째는 url1.title입니다. 숫자는 URL 목록에서 해당 위치와 매치됩니다. 이는 내용의 제목을 포함하는 시작 태그입니다.
url *n* .short_description	*string*	url *n* .title의 경우와 같습니다.
url *n* .description	*string*	url *n* .title의 경우와 같습니다.
url *n* .authors	*string*	url *n* .title의 경우와 같습니다.
url *n* .contributors	*string*	url *n* .title의 경우와 같습니다.
url *n* .published_date	*string*	url *n* .title의 경우와 같습니다.
url *n* .modified_date	*string*	url *n* .title의 경우와 같습니다.
html_alg	None HTMLCleaner	내용 필터링 방법.
discard_lines	*flag*	짧은 행을 삭제합니다. min_words와 함께 사용됩니다.
min_words	*integer*	최소 단어 수.
discard_words	*flag*	짧은 행을 삭제합니다. min_avg_len과 함께 사용됩니다.
min_avg_len	*integer*	
discard_scw	*flag*	많은 단일 문자 단어가 있는 행을 삭제합니다. max_scw와 함께 사용됩니다.
max_scw	*integer*	행에서 단일 문자 단어의 최대 비율 0-100 퍼센트
discard_tags	*flag*	특정 태그를 포함하는 행을 삭제합니다.
태그	*string*	특수 문자는 백슬래시 문자 \로 이스케이프해야 합니다.
discard_spec_words	*flag*	특정 문자열을 포함하는 행을 삭제합니다.
words	*string*	특수 문자는 백슬래시 문자 \로 이스케이프해야 합니다.

텍스트 마이닝 노드: TextMiningWorkbench

다음 매개변수를 사용하여 스크립팅을 통해 노드를 정의하거나 업데이트할 수 있습니다. 노드 자체는 TextMiningWorkbench라고 부릅니다.

중요! 스크립팅을 통해 다른 자원 템플리트를 지정하는 것은 불가능합니다. 템플리트가 필요하다고 생각하는 경우 노드 대화 상자에서 선택해야 합니다.

표 9. 텍스트 마이닝 모델링 노드 스크립팅 특성

스크립팅 속성	데이터 유형	특성 설명
text	*field*	
method	ReadText ReadPath	
docType	*integer*	가능한 값은 (0,1,2)이며, 0 = Full Text, 1 = Structured Text, 및 2 = XML

표 9. 텍스트 마이닝 모델링 노드 스크립팅 특성 (계속)

스크립팅 속성	데이터 유형	특성 설명
encoding	Automatic "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	"UTF-8"과 같은 특수 문자가 있는 값은 수학적 연산자와 혼동을 피하기 위해 따옴표로 묶어야 합니다.
unity	*integer*	가능한 값은 (0,1)이며, 0 = Paragraph 및 1 = Document
para_min	*integer*	
para_max	*integer*	
mtag	*string*	모든 mtag 설정(XML 파일의 경우 설정 대화 상자의) 포함
mclef	*string*	모든 mclef 설정(구조화된 텍스트 파일의 경우 설정 대화 상자의) 포함
partition	*field*	
custom_field	*flag*	파티션 필드가 지정될지 여부를 표시합니다.
use_model_name	*flag*	
model_name	*string*	
use_partitioned_data	*flag*	파티션 필드가 정의되는 경우, 훈련 데이터만 모델 작성에 사용합니다.
model_output_type	Interactive Model	대화형의 결과는 범주 모델입니다. 모델의 결과는 개념 모델입니다.
use_interactive_info	*flag*	워크벤치 세션에서만 대화식으로 작성하기 위한 것입니다.
reuse_extraction_results	*flag*	워크벤치 세션에서만 대화식으로 작성하기 위한 것입니다.
interactive_view	Categories TLA Clusters	워크벤치 세션에서만 대화식으로 작성하기 위한 것입니다.
extract_top	*integer*	이 매개변수는 model_type = Concept일 때 사용됩니다.
use_check_top	*flag*	
check_top	*integer*	
use_uncheck_top	*flag*	
uncheck_top	*integer*	

표 9. 텍스트 마이닝 모델링 노드 스크립팅 특성 (계속)

스크립팅 속성	데이터 유형	특성 설명
언어	de en es fr it ja nl pt	
frequency_limit	*integer*	14.0에서 더 이상 사용되지 않습니다.
concept_count_limit	*integer*	최소한 이 값의 글로벌 빈도를 사용하는 개념으로 추출을 제한합니다. 일본어 텍스트의 경우 사용할 수 없습니다.
fix_punctuation	*flag*	일본어 텍스트의 경우 사용할 수 없습니다.
fix_spelling	*flag*	일본어 텍스트의 경우 사용할 수 없습니다.
spelling_limit	*integer*	일본어 텍스트의 경우 사용할 수 없습니다.
extract_uniterm	*flag*	일본어 텍스트의 경우 사용할 수 없습니다.
extract_nonlinguistic	*flag*	일본어 텍스트의 경우 사용할 수 없습니다.
upper_case	*flag*	일본어 텍스트의 경우 사용할 수 없습니다.
group_names	*flag*	일본어 텍스트의 경우 사용할 수 없습니다.
permutation	*integer*	최대 비기능 단어 치환(기본값: 3). 일본어 텍스트의 경우 사용할 수 없습니다.
jp_algorithmset conclusions only Representative only All Sentiments	0 1 2	일본어 텍스트 추출 전용. 0 = 정서 2차 추출 1 = 종속 항목 추출 2 = 설정된 2차 분석기 없음
jp_algorithm_sense_mode	0 1 2	일본어 텍스트 추출 전용. 0 = 결론만 2 = 대표적인 것만 3 = 모든 정서.

텍스트 마이닝 모델 너깃: TMWBModelApplier

스크립팅에 대해 다음 표의 특성을 사용할 수 있습니다. 너깃 자체는 TMWBModelApplier라고 부릅니다.

표 10. 텍스트 마이닝 모델 너깃 특성

스크립팅 속성	데이터 유형	특성 설명
scoring_mode	Fields Records	
field_values	Flags Counts	이 옵션은 범주 모델 너깃에서 사용할 수 없습니다. Flags의 경우 TRUE 또는 FALSE로 설정하십시오.
true_value	*string*	Flags를 사용하여, true에 대한 값을 정의하십시오.
false_value	*string*	Flags를 사용하여, false에 대한 값을 정의하십시오.

표 10. 텍스트 마이닝 모델 너깃 특성 (계속)

스크립팅 속성	데이터 유형	특성 설명
extension_concept	*string*	필드 이름의 확장을 지정하십시오. 필드 이름은 개념 이름에 이 확장을 더해서 사용하여 생성됩니다. add_as 값을 사용하여 이 확장을 넣을 위치를 지정하십시오.
extension_category	*string*	필드 이름 확장입니다. 필드 이름에 대해 확장 접두문자/접미문자를 지정하거나 범주 코드를 사용할 것을 선택할 수 있습니다. 필드 이름은 범주 이름에 이 확장을 더해서 사용하여 생성됩니다. add_as 값을 사용하여 이 확장을 넣을 위치를 지정하십시오.
add_as	Suffix Prefix	
fix_punctuation	*flag*	
excluded_subcategories_descriptors	RollUpToParent Ignore	범주 모델의 경우에만 사용됩니다. 하위 범주가 선택 취소되는 경우 이 옵션으로 스코어링을 위해 선택되지 않은 하위 범주에 속하는 디스크립터가 처리되는 방법을 지정할 수 있습니다. 두 가지 옵션이 있습니다. • Ignore. '스코어링에서 디스크립터를 완전히 제외' 옵션은 선택 표시가 없는(선택 취소된) 하위 범주의 디스크립터가 스코어링 중에 무시되고 사용되지 않게 합니다. • RollUpToParent. '상위 범주에 있는 것과 디스크립터 통합' 옵션은 선택 표시가 없는(선택 취소된) 하위 범주의 디스크립터가 상위 범주(이 하위 범주 위에 있는 범주)에 대한 디스크립터로 사용되도록 합니다. 여러 레벨의 하위 범주가 있고 선택되지 않은 경우, 디스크립터는 첫 번째 사용 가능한 상위 범주 아래에 롤업됩니다.
check_model	*flag*	버전 14에서 더 이상 사용되지 않음
text	*field*	
method	ReadText ReadPath	
docType	*integer*	가능한 값은 (0,1,2)이며, 0 = Full Text, 1 = Structured Text, 및 2 = XML
encoding	Automatic "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	"UTF-8"과 같은 특수 문자가 있는 값은 수학적 연산자와 혼동을 피하기 위해 따옴표로 묶어야 합니다.

표 10. 텍스트 마이닝 모델 너깃 특성 (계속)

스크립팅 속성	데이터 유형	특성 설명
언어	de en es fr it ja nl pt	

텍스트 링크 분석 노드: textlinkanalysis

다음 테이블의 매개변수를 사용하여 스크립팅을 통해 노드를 정의 또는 업데이트할 수 있습니다. 노드 자체는 textlinkanalysis라고 부릅니다.

중요! 스크립팅을 통해 자원 템플리트를 지정하는 것은 불가능합니다. 템플리트를 선택하려면 노드 대화 상자 안에서 선택해야 합니다.

표 11. 텍스트 링크 분석(TLA) 노드 스크립팅 특성

스크립팅 속성	데이터 유형	특성 설명
id_field	*field*	
text	*field*	
method	ReadText ReadPath	
docType	*integer*	가능한 값은 (0,1,2)이며, 0 = Full Text, 1 = Structured Text, 및 2 = XML
encoding	Automatic "UTF-8" "UTF-16" "ISO-8859-1" "US-ASCII" "CP850" "EUC-JP" "SHIFT-JIS" "ISO2022-JP"	"UTF-8"과 같은 특수 문자가 있는 값은 수학적 연산자와 혼동을 피하기 위해 따옴표로 묶어야 합니다.
unity	*integer*	가능한 값은 (0,1)이며, 0 = Paragraph 및 1 = Document
para_min	*integer*	
para_max	*integer*	
mtag	*string*	모든 mtag 설정(XML 파일의 경우 설정 대화 상자의) 포함
mclef	*string*	모든 mclef 설정(구조화된 텍스트 파일의 경우 설정 대화 상자의) 포함

표 11. 텍스트 링크 분석(TLA) 노드 스크립팅 특성 (계속)

스크립팅 속성	데이터 유형	특성 설명
언어	de en es fr it ja nl pt	
concept_count_limit	*integer*	최소한 이 값의 글로벌 빈도를 사용하는 개념으로 추출을 제한합니다. 일본어 텍스트의 경우 사용할 수 없습니다.
fix_punctuation	*flag*	일본어 텍스트의 경우 사용할 수 없습니다.
fix_spelling	*flag*	일본어 텍스트의 경우 사용할 수 없습니다.
spelling_limit	*integer*	일본어 텍스트의 경우 사용할 수 없습니다.
extract_uniterm	*flag*	일본어 텍스트의 경우 사용할 수 없습니다.
extract_nonlinguistic	*flag*	일본어 텍스트의 경우 사용할 수 없습니다.
upper_case	*flag*	일본어 텍스트의 경우 사용할 수 없습니다.
group_names	*flag*	일본어 텍스트의 경우 사용할 수 없습니다.
permutation	*integer*	최대 비기능 단어 치환(기본값: 3). 일본어 텍스트의 경우 사용할 수 없습니다.
jp_algorithmset conclusions only Representative only All Sentiments	0 1 2	일본어 텍스트 추출 전용. 0 = 정서 2차 추출 1 = 종속 항목 추출 2 = 설정된 2차 분석기 없음
jp_algorithm_sense_mode	0 1 2	일본어 텍스트 추출 전용. 0 = 결론만 2 = 대표적인 것만 3 = 모든 정서.

변환 노드: translatenode

스크립팅에 대해 다음 표의 특성을 사용할 수 있습니다. 노드 자체는 translatenode라고 부릅니다.

표 12. 변환 노드 특성

스크립팅 속성	데이터 유형	특성 설명
text	*field*	
method	ReadText ReadPath	

표 12. 변환 노드 특성 (계속)

스크립팅 속성	데이터 유형	특성 설명
encoding	Automatic "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"	"UTF-8" 같은 특수 문자를 갖는 값은 수학 연산자와의 혼동을 피하기 위해 따옴표로 묶어야 함을 주의하십시오.
lw_server_type	LOC WAN HTTP	
lw_hostname	*string*	
lw_port	*integer*	
url	*string*	변환 서버의 URL
apiKey	*string*	
user_id	*string*	
lpid	*integer*	*language_from* 또는 *language_from_id*가 설정된 경우 사용되지 않습니다.

표 12. 변환 노드 특성 (계속)

스크립팅 속성	데이터 유형	특성 설명
translate_from	Arabic, Chinese, Traditional Chinese, Czech, Danish,Dutch, English,French, German, Greek, Hindi, Hungarian, Italian, Japanese, Korean, Persian, Polish, Portuguese, Romanian, Russian, Spanish, Somali, Swedish	
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	English	
translate_to_id	eng	
translation_accuracy	*integer*	변환 프로세스에 대해 원하는 정확도 수준을 지정합니다. 1에서 3의 값을 선택하십시오.
use_previous_translation	*flag*	변환 결과가 이미 이전 실행으로부터 존재하며 재사용될 수 있도록 지정합니다.
translation_label	*string*	재사용하기 위해 변환 결과를 식별할 레이블을 입력하십시오.

제 8 장 대화형 워크벤치 모드

텍스트 마이닝 모델링 노드에서 스트림 실행 중에 대화형 워크벤치 세션을 실행하도록 선택할 수 있습니다. 이 워크벤치에서, 텍스트 데이터에서 주요 개념을 추출하고 범주를 작성하며 텍스트 링크 분석 패턴 및 군집을 탐색하고 범주 모델을 생성할 수 있습니다. 이 장 에서는 다음을 포함하여 작업할 주요 요소와 함께 상위 레벨 관점에서 워크벤치 인터페이스에 대해 설명합니다.

- **추출 결과.** 추출이 수행된 후 이는 데이터 텍스트에서 식별 및 추출된 주요 단어 및 문구입니다(*개념*이라고도 함). 이러한 개념은 *유형*으로 그룹화됩니다. 이러한 개념과 유형을 사용하면 범주를 작성하는 것은 물론 데이터도 탐색할 수 있습니다. **범주 및 개념** 보기에서 관리됩니다.
- **범주.** 디스크립터(예: 추출 결과, 패턴, 규칙)를 정의로 사용하면 범주 정의의 일부가 포함되었는지 여부에 관계없이 문서 및 레코드가 지정되는 범주 세트를 수동으로 또는 자동으로 작성할 수 있습니다. **범주 및 개념** 보기에서 관리됩니다.
- **군집.** *군집*은 개념 간의 관계를 표시하는 링크가 발견된 개념 집단입니다. 개념은 다른 요인 중에서 두 개념이 함께 나타나는 빈도를 개별적으로 나타나는 빈도와 비교하는 복합 알고리즘을 사용하여 그룹화됩니다. **군집** 보기에서 관리됩니다. 또한 군집을 구성하는 개념을 범주에 추가할 수 있습니다.
- **텍스트 링크 분석 패턴.** 언어학적 자원에 텍스트 링크 분석(TLA) 패턴 규칙이 있거나 일부 TLA 규칙이 이미 있는 자원 템플리트를 사용 중인 경우, 텍스트 데이터에서 패턴을 추출할 수 있습니다. 이러한 패턴을 사용하여 데이터에 있는 개념 간의 흥미로운 관계를 쉽게 알아낼 수 있습니다. 또한 이러한 패턴을 범주에서 디스크립터로 사용할 수 있습니다. **텍스트 링크 분석** 보기에서 관리됩니다. 일본어 텍스트의 경우, 2차 분석기를 선택하고 TLA 추출을 설정해야 합니다.
- **언어학적 자원.** 추출 프로세스는 매개변수 및 언어 정의 세트에 의존하여 텍스트 추출 및 처리 방법을 제어합니다. **자원 편집기** 보기에서 템플리트 및 라이브러리 양식으로 관리됩니다.

범주 및 개념 보기

애플리케이션 인터페이스는 몇 개의 보기로 구성됩니다. 범주 및 개념 보기는 범주를 작성하고 탐색하는 것은 물론 추출 결과를 탐색하고 조정할 수 있는 창입니다. **범주**는 스코어링 프로세스를 통해 문서와 레코드가 지정된 밀접하게 관련된 아이디어 및 패턴 그룹입니다. 반면에 **개념**은 범주에 디스크립터라는 구성 요소로 사용할 사용 가능한 가장 기본적인 수준의 추출 결과를 나타냅니다.

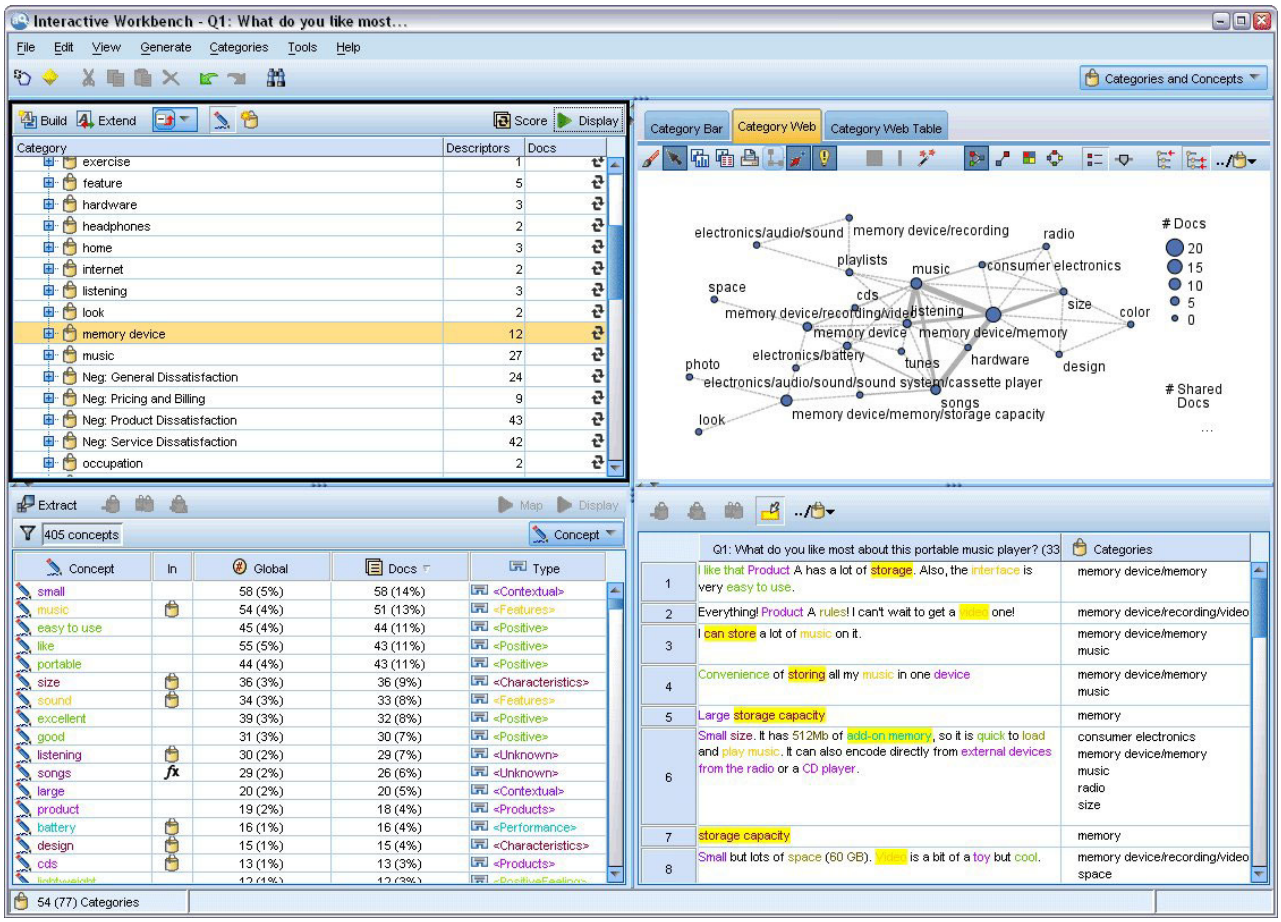


그림 23. 범주 및 개념 보기

범주 및 개념 보기는 네 개의 분할창으로 구성되며, 보기 메뉴에서 해당 이름을 선택하여 각각 숨기거나 표시할 수 있습니다. 자세한 정보는 105 페이지의 제 10 장 『텍스트 데이터 범주화』의 내용을 참조하십시오.

범주 분할창

왼쪽 상단 모서리에 위치한 이 영역은 작성하는 범주를 관리할 수 있는 테이블을 제공합니다. 텍스트 데이터에서 개념과 유형을 추출한 후에는 시맨틱 네트워크 및 개념 포함과 같은 기술을 사용하거나 수동으로 작성하여 범주 작성을 시작할 수 있습니다. 범주 이름을 두 번 클릭하면 범주 정의 대화 상자가 열려 정의를 구성하는 모든 디스크립터(예: 개념, 유형, 규칙)를 표시합니다. 자세한 정보는 105 페이지의 제 10 장 『텍스트 데이터 범주화』의 내용을 참조하십시오. 모든 언어에 자동 기술을 모두 사용할 수 있는 것은 아닙니다.

분할창에서 행을 선택하면 데이터 및 시각화 분할창에 해당 문서/레코드 또는 디스크립터에 대한 정보를 표시할 수 있습니다.

추출 결과 분할창

왼쪽 상단 모서리에 위치한 이 영역은 추출 결과를 제공합니다. 추출을 실행하면 추출 엔진이 텍스트 데이터를 읽고 관련 개념을 식별하며, 각각에 유형을 지정합니다. **개념**은 텍스트 데이터에서 추출된 단어 또는 문구입니

다. **유형**은 유형 사전 양식으로 저장된 개념에 대한 시맨틱 집단입니다. 추출이 완료되면 개념과 유형이 추출 결과 분할창에 색상 코딩으로 나타납니다. 자세한 정보는 91 페이지의 『추출 결과: 개념 및 유형』의 내용을 참조하십시오.

마우스를 개념 이름 위에 올리면 개념의 기본 용어 세트를 볼 수 있습니다. 그렇게 하면 개념 이름을 표시하는 도구팁과 해당 개념 아래에 그룹화되는 몇몇 용어 라인이 표시됩니다. 이러한 기본 용어에는 언어학적 자원(텍스트에서 발견되는지 여부와 관계 없이)에 정의되는 동의어뿐만 아니라 추출된 복수/단수 용어, 순열된 용어, 퍼지 그룹화의 용어 등이 포함됩니다. 이러한 용어를 복사하거나 개념 이름을 마우스 오른쪽 단추로 클릭하고 컨텍스트 메뉴 옵션을 선택하여 전체 기본 용어 세트를 볼 수 있습니다.

텍스트 마이닝은 텍스트 데이터 컨텍스트에 따라 추출 결과를 검토하고 미세 조정하여 새 결과를 생성한 후 재평가하는 대화형 프로세스입니다. 언어학적 자원을 수정하여 추출 결과를 세분화할 수 있습니다. 이 미세 조정을 추출 결과 또는 데이터 분할창에서 직접 부분적으로 수행할 수 있지만 자원 편집기 보기에서도 직접 수행할 수 있습니다. 자세한 정보는 84 페이지의 『자원 편집기 보기』의 내용을 참조하십시오.

시각화 분할창

오른쪽 상단 모서리에 위치한 이 영역은 문서/레코드 범주화의 공통성에 대한 여러 퍼스펙티브를 제공합니다. 각 그래프 또는 차트는 유사한 정보를 제공하지만 다른 방식으로 또는 다른 세부사항 수준으로 정보를 제공합니다. 이러한 차트와 그래프를 사용하여 범주화 결과를 분석하고 쉽게 범주를 미세 조정하거나 보고서를 작성할 수 있습니다. 예를 들어, 그래프에서 너무 유사하거나(예: 레코드의 75% 이상을 공유) 너무 다른 범주를 발견할 수 있습니다. 그래프 또는 차트의 컨텐츠는 다른 분할창의 선택사항에 해당합니다. 자세한 정보는 165 페이지의 『범주 그래프 및 도표』의 내용을 참조하십시오.

데이터 분할창

데이터 분할창은 오른쪽 하단 모서리에 있습니다. 이 분할창은 보기의 다른 영역에서의 선택사항에 해당하는 문서 또는 레코드가 포함된 테이블을 제공합니다. 선택사항에 따라 해당 텍스트만 데이터 분할창에 표시됩니다. 선택했으면 **표시** 단추를 클릭하여 데이터 분할창을 해당 텍스트로 채우십시오.

다른 분할창에 선택사항이 있으면 텍스트에서 쉽게 식별할 수 있도록 해당 문서 또는 레코드가 색상으로 강조표시된 개념을 표시합니다. 색상 코드화된 항목 위에 마우스를 올려 놓아 추출된 개념과 지정된 유형을 표시하는 도구팁도 표시할 수 있습니다. 자세한 정보는 114 페이지의 『데이터 분할창』의 내용을 참조하십시오.

범주 및 개념 보기에서 검색 및 찾기

일부 경우 특정 섹션에서 빨리 정보를 찾아야 합니다. 찾기 도구 모음을 사용하면 검색할 문자열을 입력하고 다른 검색 기준(예: 대소문자 구분 또는 검색 방향)을 정의할 수 있습니다. 그리고 나서 검색할 분할창을 선택할 수 있습니다.

찾기 기능 사용 방법

1. 범주 및 개념 보기의 메뉴에서 **편집 > 찾기**를 선택하십시오. 찾기 도구 모음이 범주 분할창 및 시각화 분할창 위에 나타납니다.

2. 텍스트 상자에 검색할 단어 문자열을 입력하십시오. 도구 모음 단추를 사용하여 대소문자 구분, 부분 매치 및 검색 방향을 제어할 수 있습니다.
3. 도구 모음에서 검색할 분할창 이름을 클릭하십시오. 매치가 발견되면 창에서 텍스트가 강조표시됩니다.
4. 다음 매치를 찾으려면 분할창 이름을 다시 클릭하십시오.

군집 보기

군집 보기에서 텍스트 데이터에서 찾은 군집 결과를 작성하고 탐색할 수 있습니다. **군집**은 개념이 발생하는 빈도와 함께 나타나는 빈도를 기반으로 한 군집화 알고리즘을 통해 생성된 개념 집단입니다. 범주의 목적은 범주에 포함된 텍스트가 각 범주에 대해 디스크립터(개념, 규칙, 패턴)를 매치하는 방법을 기반으로 문서 또는 레코드를 그룹화하는 것인 반면 군집의 목적은 함께 동시 발생하는 개념을 그룹화하는 것입니다.

군집 내에서 개념이 함께 연결되어 발생하는 빈도가 높을수록 다른 개념과 함께 발생하는 빈도가 낮으며, 군집은 흥미로운 개념 관계를 더 잘 식별합니다. 두 개의 개념이 모두 동일한 문서 또는 레코드에 나타나는 경우(또는 동의어나 용어 중 하나가 나타나는 경우) 두 개념은 동시 발생합니다. 자세한 정보는 153 페이지의 제 11 장 『군집 분석』의 내용을 참조하십시오.

군집을 작성하고, 그렇지 않으면 찾는 데 시간이 너무 많이 걸리는 개념 간의 관계를 파악하는 데 도움이 되는 차트 및 그래프 세트에서 탐색할 수 있습니다. 전체 군집을 범주에 추가할 수는 없지만 군집 정의 대화 상자를 통해 군집의 개념을 범주에 추가할 수 있습니다. 자세한 정보는 157 페이지의 『군집 정의』의 내용을 참조하십시오.

군집화 설정을 변경하여 결과에 영향을 줄 수 있습니다. 자세한 정보는 155 페이지의 『군집 작성』의 내용을 참조하십시오.

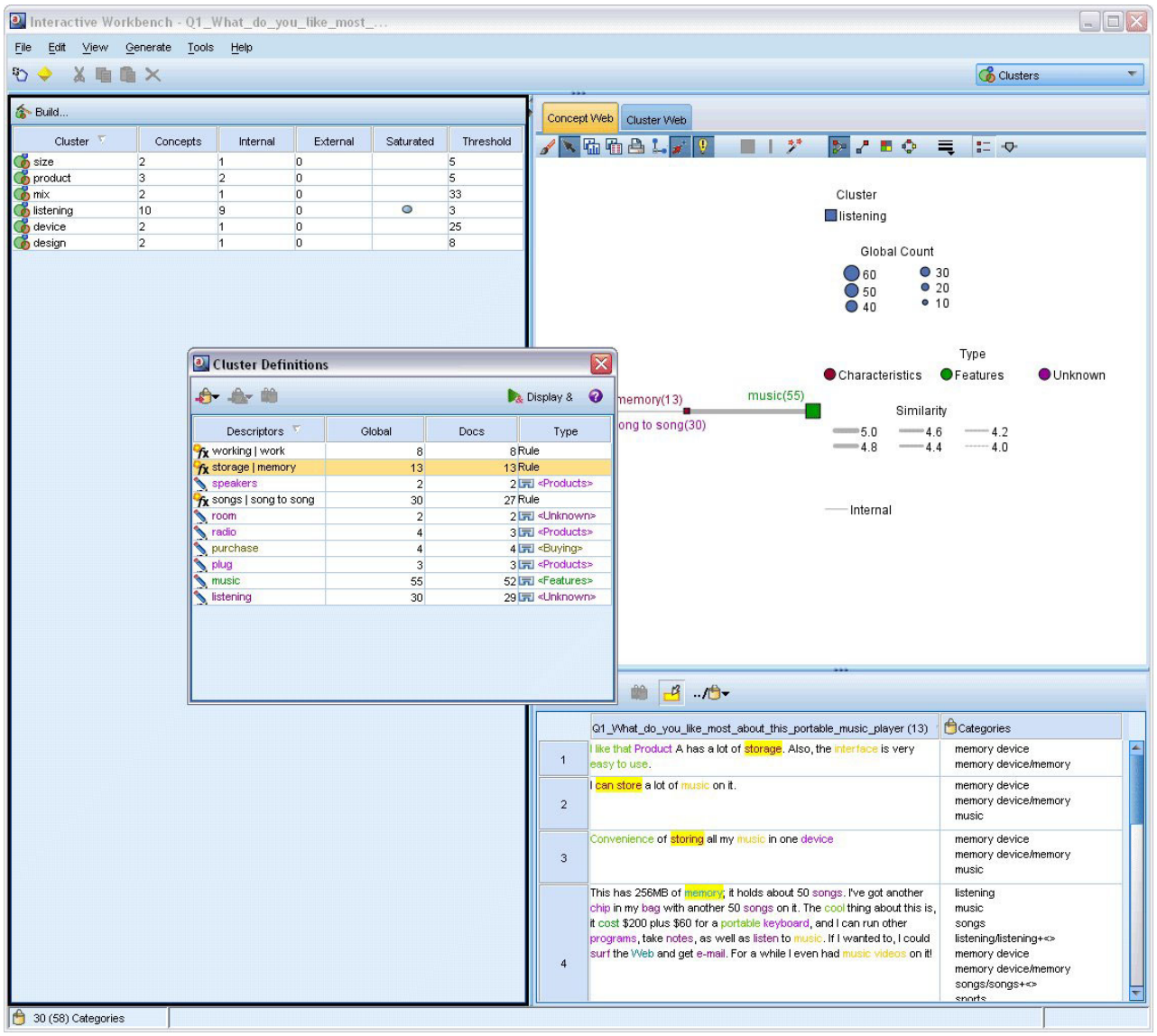


그림 24. 군집 보기

군집 보기는 세 개의 분할창으로 구성되며, 보기 메뉴에서 해당 이름을 선택하여 각각 숨기거나 표시할 수 있습니다. 일반적으로 군집 분할창과 시각화 분할창만 표시됩니다.

군집 분할창

왼쪽에 위치한 이 분할창은 텍스트 데이터에서 발견된 군집을 제공합니다. **작성** 단추를 클릭하여 군집화 결과를 작성할 수 있습니다. 군집은 함께 자주 발생하는 개념을 식별하려고 시도하는 군집화 알고리즘을 통해 형성됩니다.

새 추출이 발생될 때마다 군집 결과가 지워지며, 최신 결과를 얻으려면 군집을 다시 작성해야 합니다. 군집을 작성할 때 일부 설정(예: 작성할 최대 군집 수, 군집에 포함될 수 있는 최대 개념 수 또는 군집이 가질 수 있는 외부 개념과의 최대 링크 수)을 변경할 수 있습니다. 자세한 정보는 157 페이지의 『군집 탐색』의 내용을 참조하십시오.

시각화 분할창

오른쪽 상단 모서리에 위치한 이 분할창은 군집화에 대한 두 개의 퍼스펙티브 즉, 개념 웹 그래프와 군집 웹 그래프를 제공합니다. 표시되지 않으면 보기 메뉴에서 이 분할창에 액세스할 수 있습니다(**보기 > 시각화**). 군집 분할창에서의 선택사항에 따라 군집 간의 또는 군집 내 해당 상호작용을 볼 수 있습니다. 결과는 다음과 같이 다중 형식으로 제공됩니다.

- **개념 웹.** 선택된 군집 내의 모든 개념은 물론 군집 외부의 링크된 개념도 표시하는 웹 그래프입니다.
- **군집 웹.** 선택된 군집에서 다른 군집으로의 링크는 물론 해당 다른 군집 간의 링크도 보여주는 웹 그래프입니다.

참고: 군집 웹 그래프를 표시하려면 외부 링크가 있는 군집을 이미 작성했어야 합니다. 외부 링크는 별도 군집의 개념 쌍(한 군집 내 개념과 다른 군집에서 외부의 개념) 간의 링크입니다. 자세한 정보는 167 페이지의 『군집 그래프』의 내용을 참조하십시오.

데이터 분할창

데이터 분할창은 오른쪽 하단 모서리에 있으며 기본적으로 숨겨집니다. 이러한 군집은 여러 문서/레코드에 걸쳐 있기 때문에 군집 분할창에서 데이터 분할창 결과를 표시할 수 없어 데이터 결과가 흥미롭지 못하게 됩니다. 그러나 군집 정의 대화 상자의 선택사항에 해당하는 데이터를 볼 수 있습니다. 해당 대화 상자에서의 선택사항에 따라 해당 텍스트만 데이터 분할창에 표시됩니다. 선택했으면 **표시 &** 단추를 클릭하여 모든 개념을 함께 포함하는 문서 또는 레코드로 데이터 분할창을 채우십시오.

해당 문서 또는 레코드는 텍스트에서 쉽게 식별할 수 있도록 색상으로 강조표시된 개념을 표시합니다. 색상 코드화된 항목 위에 마우스를 올려 놓아 추출된 개념과 지정된 유형도 표시할 수 있습니다. 데이터 분할창에는 여러 개의 열이 포함될 수 있지만 텍스트 필드 열은 항상 표시됩니다. 추출 중에 사용된 텍스트 필드 이름 또는 여러 파일에 텍스트 데이터가 있는 경우에는 문서 이름을 수반합니다. 기타 열은 사용 가능합니다. 자세한 정보는 114 페이지의 『데이터 분할창』의 내용을 참조하십시오.

텍스트 링크 분석 보기

텍스트 링크 분석(TLA) 보기에서 텍스트 데이터에서 찾은 텍스트 링크 분석 패턴을 작성하고 탐색할 수 있습니다. 텍스트 링크 분석(TLA)은 TLA 규칙을 정의하고 이를 텍스트에서 찾은 실제로 추출된 개념 및 관계와 비교할 수 있게 하는 패턴 매치 기술입니다.

패턴은 특정 주제에 대한 의견 또는 개념 간의 관계를 발견하려고 시도할 때 가장 유용합니다. 몇 가지 예로 설문조사 데이터에서 제품에 대한 의견, 의학 연구 논문에서 유전자 관계 또는 지능형 데이터에서 개체 간 또는 위치 간의 관계를 추출하려고 하는 경우를 들 수 있습니다.

일부 TLA 패턴을 추출했으면 데이터 또는 시각화 분할창에서 탐색하고 범주 및 개념 보기에서 범주에 추가할 수 있습니다. TLA 결과를 추출하려면 사용 중인 자원 템플리트 또는 라이브러리에 일부 TLA 규칙이 정의되어 있어야 합니다. 자세한 정보는 223 페이지의 제 19 장 『텍스트 링크 규칙에 대한 정보』의 내용을 참조하십시오.

TLA 패턴 결과를 추출하기로 선택했으면 결과가 이 보기에 제공됩니다. 이렇게 하도록 선택하지 않았으면 추출 단추를 사용하고 패턴 추출 사용 옵션을 선택해야 합니다.

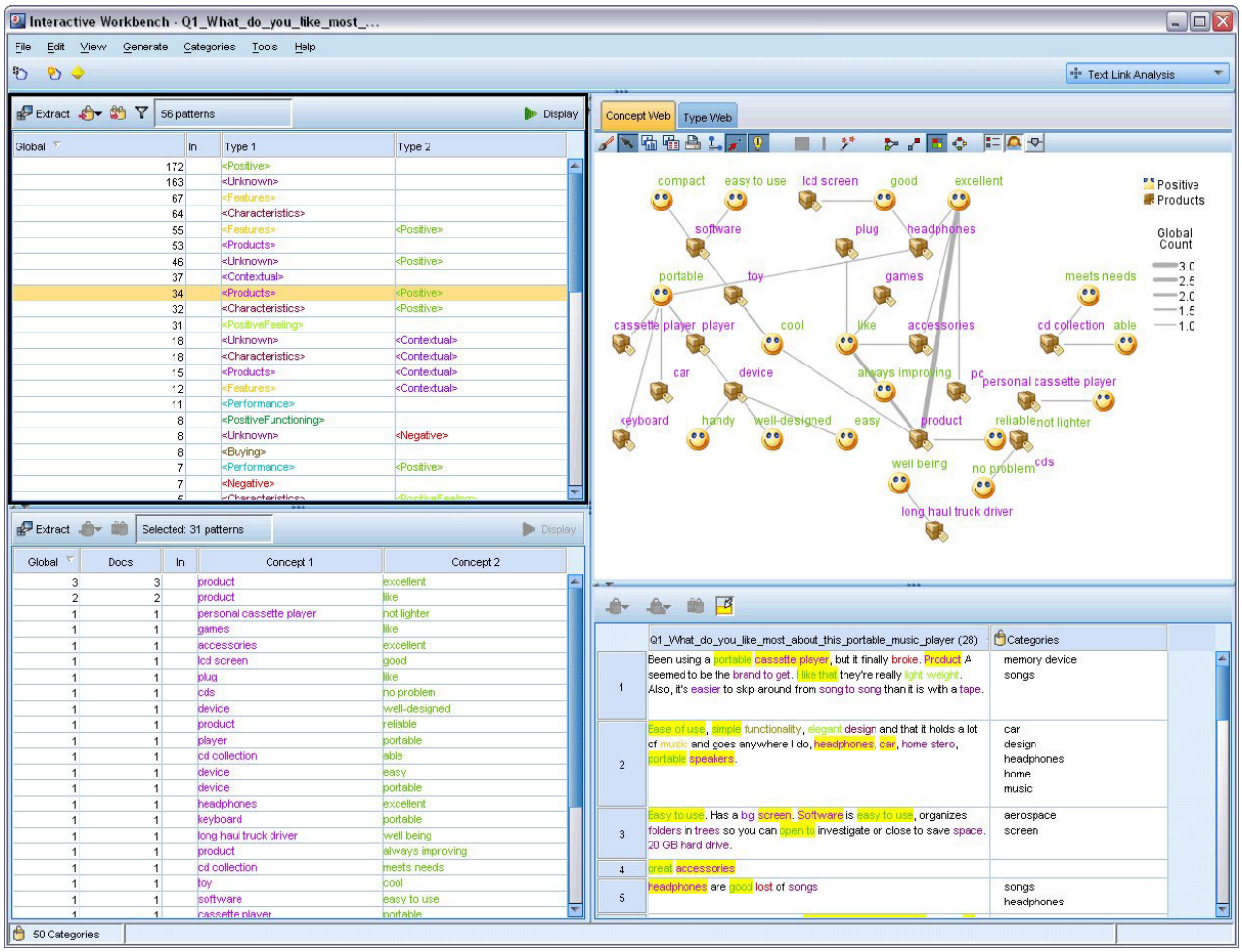


그림 25. 텍스트 링크 분석 보기

텍스트 링크 분석 보기는 네 개의 분할창으로 구성되며, 보기 메뉴에서 해당 이름을 선택하여 각각 숨기거나 표시할 수 있습니다. 자세한 정보는 159 페이지의 제 12 장『텍스트 링크 분석 탐색』의 내용을 참조하십시오.

유형 및 개념 패턴 분할창

왼쪽에 위치한 유형 및 개념 패턴 분할창은 TLA 패턴 결과를 탐색하고 선택할 수 있는 두 개의 상호 연결된 분할창입니다. 패턴은 최대 6개의 일련의 유형 또는 6개의 개념으로 구성됩니다. 일본어 텍스트의 경우 패턴은 최대 1 - 2개의 일련의 유형 또는 개념일 뿐입니다. 언어학적 자원에서 정의된 대로 TLA 패턴 규칙은 패턴 결과의 복잡도를 지시합니다. 자세한 정보는 223 페이지의 제 19 장『텍스트 링크 규칙에 대한 정보』의 내용을 참조하십시오.

패턴 결과는 먼저 유형 수준에서 그룹화된 후 개념 패턴으로 나뉩니다. 이러한 이유로 두 개의 다른 결과 분할창 즉, 유형 패턴(왼쪽 상단)과 개념 패턴(왼쪽 하단)이 있습니다.

- **유형 패턴.** 유형 패턴 분할창은 TLA 패턴 규칙과 매치하는 두 개 이상의 관련 유형으로 구성되는 추출된 패턴을 제공합니다. 유형 패턴은 특정 위치의 조직에 대한 긍정적 피드백을 제공할 수 있는 <Organization> + <Location> + <Positive>로 표시됩니다.
- **개념 패턴.** 개념 패턴 분할창은 그 위의 유형 패턴 분할창에서 현재 선택된 모든 유형 패턴에 대한 추출된 패턴을 개념 수준에서 제공합니다. 개념 패턴은 hotel + paris + wonderful과 같은 구조를 따릅니다.

범주 및 개념 보기에서의 추출 결과의 경우와 마찬가지로 여기서 결과를 검토할 수 있습니다. 이러한 패턴을 구성하는 유형 및 개념에 대해 수행할 세분화가 있으면 범주 및 개념 보기의 추출 결과 분할창에서 또는 자원 편집기에서 직접 세분화를 수행하고 패턴을 재추출하십시오.

시각화 분할창

텍스트 링크 분석 보기의 오른쪽 상단 모서리에 위치한 이 분할창은 유형 패턴 또는 개념 패턴으로 선택된 패턴의 웹 그래프를 제공합니다. 표시되지 않으면 보기 메뉴에서 이 분할창에 액세스할 수 있습니다(**보기 > 시각화**). 다른 분할창에서의 선택사항에 따라 문서/레코드 및 패턴 간의 해당 상호작용을 볼 수 있습니다.

결과는 다음과 같이 다중 형식으로 제공됩니다.

- **개념 그래프.** 이 그래프는 선택된 패턴의 모든 개념을 제공합니다. 개념 그래프에서 선 *굵기*와 노드 *크기*(유형 아이콘이 표시되지 않은 경우)는 선택된 테이블에서 글로벌 발생 수를 표시합니다.
- **유형 그래프.** 이 그래프는 선택된 패턴의 모든 유형을 제공합니다. 그래프에서 선 *굵기*와 노드 *크기*(유형 아이콘이 표시되지 않은 경우)는 선택된 테이블에서 글로벌 발생 수를 표시합니다. 노드는 유형 색상 또는 아이콘으로 표시됩니다.

자세한 정보는 168 페이지의 『텍스트 링크 분석 그래프』의 내용을 참조하십시오.

데이터 분할창

데이터 분할창은 오른쪽 하단 모서리에 있습니다. 이 분할창은 보기의 다른 영역에서의 선택사항에 해당하는 문서 또는 레코드가 포함된 테이블을 제공합니다. 선택사항에 따라 해당 텍스트만 데이터 분할창에 표시됩니다. 선택했으면 **표시** 단추를 클릭하여 데이터 분할창을 해당 텍스트로 채우십시오.

다른 분할창에 선택사항이 있으면 텍스트에서 쉽게 식별할 수 있도록 해당 문서 또는 레코드가 색상으로 강조표시된 개념을 표시합니다. 색상 코드화된 항목 위에 마우스를 올려 놓아 추출된 개념과 지정된 유형을 표시하는 도구팁도 표시할 수 있습니다. 자세한 정보는 114 페이지의 『데이터 분할창』의 내용을 참조하십시오.

자원 편집기 보기

IBM SPSS Modeler Text Analytics 은 강력한 추출 엔진을 사용하여 텍스트 데이터로부터 주요 개념을 빠르고 정확하게 캡처합니다. 이 엔진은 얼마나 많은 양의 구조화되지 않은 텍스트 데이터가 분석되고 해석되어야 하는지를 지시하기 위해 언어학적 자원에 *크게* 의존합니다.

자원 편집기 보기에서는 개념을 추출하는 데 사용된 언어학적 자원을 보고 세부 조정하고, 이들을 유형별로 그룹화하고, 텍스트 데이터에서 패턴을 찾아내는 등의 작업을 할 수 있습니다. IBM SPSS Modeler Text Analytics

에서는 몇몇 사전에 구성된 자원 템플리트를 제공합니다. 또한 몇몇 언어에서는 텍스트 분석 패키지에서 자원을 사용할 수도 있습니다. 자세한 정보는 147 페이지의 『텍스트 분석 패키지 사용』의 내용을 참조하십시오.

이러한 자원이 항상 데이터 컨텍스트에 완벽하게 적합하지는 않을 수 있으므로, 자원 편집기에서 특정 컨텍스트 또는 도메인에서 사용자 고유의 자원을 작성, 편집 및 관리할 수 있습니다. 자세한 정보는 187 페이지의 제 16 장 『라이브러리에 대한 작업』 주제를 참조하십시오.

언어학적 자원을 세부 조정하는 프로세스를 단순화하기 위해 추출 결과 및 데이터 분할창의 컨텍스트 메뉴를 통해 범주 및 개념 보기에서 직접 공통 사전 작업을 수행할 수 있습니다. 자세한 정보는 99 페이지의 『추출 결과 세분화』 주제를 참조하십시오.

참고: 일본어 텍스트에 맞춰진 자원 인터페이스가 약간 다릅니다.

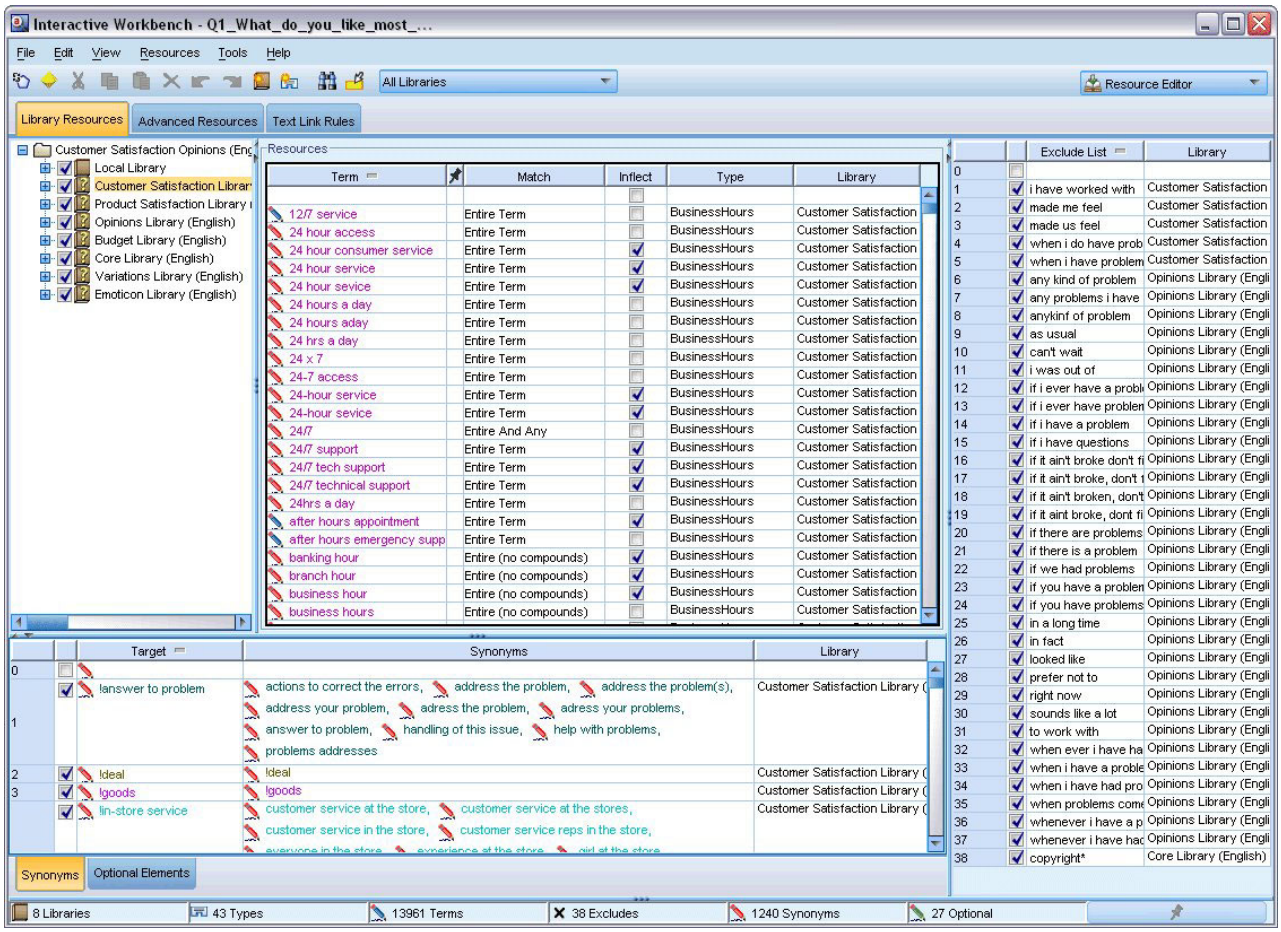


그림 26. 자원 편집기 보기

자원 편집기 보기에서 수행하는 작업은 언어학적 자원의 관리 및 세부 조정을 중심으로 합니다. 이러한 자원은 템플리트 및 라이브러리 양식으로 저장됩니다. 자원 편집기 보기는 라이브러리 트리 분할창, 유형 사전 분할창, 대체 사전 분할창, 제외 사전 분할창의 네 개의 파트로 구성됩니다.

참고: 자세한 정보는 176 페이지의 『편집기 인터페이스』 주제를 참조하십시오.

옵션 설정

옵션 대화 상자에서 IBM SPSS Modeler Text Analytics의 일반 옵션을 설정할 수 있습니다. 이 대화 상자에는 다음 탭이 있습니다.

- **세션.** 이 탭은 일반 옵션과 구분자를 포함합니다.
- **표시.** 이 탭은 인터페이스에서 사용되는 색상에 대한 옵션을 포함합니다.
- **사운드.** 이 탭은 사운드 큐 옵션을 포함합니다.

옵션 편집 방법

1. 메뉴에서 **도구 > 옵션**을 선택하십시오. 옵션 대화 상자가 열립니다.
2. 변경할 정보가 포함된 탭을 선택하십시오.
3. 옵션을 변경하십시오.
4. **확인**을 클릭하여 변경사항을 저장하십시오.

옵션: 세션 탭

이 탭에서, 일부 기본 설정을 정의할 수 있습니다.

데이터 분할창 및 범주 그래프 표시. 이 옵션은 범주 및 개념 보기의 시각화 분할창과 데이터 분할창에서 데이터가 표시되는 방법에 영향을 줍니다.

- **데이터 분할창 및 범주 웹에 대한 표시 한계.** 이 옵션은 범주 및 개념 보기에서 데이터 분할창이나 그래프 및 도표를 채우기 위해 사용하거나 표시할 최대 문서 수를 설정합니다.
- **표시할 때 문서/레코드에 대한 범주 표시.** 선택하면, 문서나 레코드가 속하는 범주가 데이터 분할창의 범주 열과 범주 그래프에 표시될 수 있도록 표시를 클릭할 때마다 문서 또는 레코드가 스코어링됩니다. 일부 경우에는(특히 큰 데이터 세트가 있는 경우) 데이터와 그래프가 더 빨리 표시되도록 이 옵션을 끌 수도 있습니다.

데이터 분할창에서 범주에 추가. 이 옵션은 문서와 레코드가 데이터 분할창에서 추가될 때 범주에 추가될 사항에 영향을 줍니다.

- **범주 및 개념 보기에서, 복사.** 이 보기에서 데이터 분할창으로부터 문서 또는 레코드를 추가하면 **개념만** 또는 **개념 및 패턴** 둘 다 복사됩니다.
- **텍스트 링크 분석 보기에서, 복사.** 이 보기에서 데이터 분할창으로부터 문서 또는 레코드를 추가하면 **패턴만** 또는 **개념 및 패턴** 둘 다 복사됩니다.

자원 편집기 구분자. 자원 편집기 보기에서 개념, 동의어 및 선택적 요소와 같은 요소를 입력할 때 구분자로서 사용될 문자를 선택하십시오.

옵션: 표시 탭

이 탭에서 애플리케이션의 전반적인 모양과 느낌에 영향을 주는 옵션과 요소를 구별하는 데 사용되는 색상을 편집할 수 있습니다.

참고: 제품의 모양과 느낌을 클래식 모양과 느낌 또는 이전 릴리스의 모양과 느낌으로 전환하려면 기본 IBM SPSS Modeler 창의 도구 메뉴에서 사용자 옵션을 여십시오.

사용자 정의 색상. 화면에 나타나는 요소의 색상을 편집하십시오. 테이블의 각 요소에 대해 색상을 변경할 수 있습니다. 사용자 정의 색상을 지정하려면 변경할 요소 오른쪽의 색상 영역을 클릭하고 드롭 다운 목록에서 색상을 선택하십시오.

- **추출되지 않은 텍스트.** 아직 추출되지 않았지만 데이터 분할창에 표시 가능한 텍스트 데이터입니다.
- **강조 배경.** 분할창에서 요소를 선택하거나 데이터 분할창에서 텍스트를 선택할 때 텍스트 선택 배경 색상입니다.
- **추출 필요 배경.** 라이브러리에 변경이 수행되었고 추출이 필요함을 나타내는 추출 결과, 패턴 및 군집 분할창의 배경 색상입니다.
- **범주 피드백 배경.** 작업 후 나타나는 범주 배경 색상입니다.
- **기본 유형.** 데이터 분할창과 추출 결과 분할창에 나타나는 유형 및 개념의 기본 색상입니다. 이 색상은 자원 편집기에서 작성하는 사용자 정의 유형에 적용됩니다. 자원 편집기에서 이러한 유형 사전의 특성을 편집하여 사용자 정의 유형 사전의 이 기본 색상을 대체할 수 있습니다. 자세한 정보는 199 페이지의 『유형 작성』의 내용을 참조하십시오.
- **스트라이프 테이블 1.** 각 선 세트를 구별하기 위해 강제 실행된 개념 편집 대화 상자의 테이블에서 대체 방식으로 사용되는 두 개 색상 중 첫 번째입니다.
- **스트라이프 테이블 2.** 각 선 세트를 구별하기 위해 강제 실행된 개념 편집 대화 상자의 테이블에서 대체 방식으로 사용되는 두 개 색상 중 두 번째입니다.

참고: **기본값으로 재설정** 단추를 클릭하면 이 대화 상자의 모든 옵션이 이 제품을 처음 설치했을 때의 값으로 재설정됩니다.

옵션: 사운드 탭

이 탭에서 사운드에 영향을 주는 옵션을 편집할 수 있습니다. 사운드 이벤트에서, 이벤트가 발생하면 알리는 데 사용되는 사운드를 지정할 수 있습니다. 다수의 사운드를 사용할 수 있습니다. 생략 기호 단추(**...**)를 사용하여 사운드를 찾아 선택하십시오. IBM SPSS Modeler Text Analytics용 사운드를 작성하는 데 사용되는 *.wav* 파일은 설치 디렉토리의 *media* 서브디렉토리에 저장됩니다. 사운드를 재생하지 않으려면 **모든 사운드 음소거**를 선택하십시오. 사운드는 기본적으로 음소거됩니다.

참고: **기본값으로 재설정** 단추를 클릭하면 이 대화 상자의 모든 옵션이 이 제품을 처음 설치했을 때의 값으로 재설정됩니다.

도움말에 대한 Microsoft Internet Explorer 설정

Microsoft Internet Explorer 설정

이 애플리케이션에서 대부분의 도움말 기능은 Microsoft Internet Explorer에 기반한 기술을 사용합니다. Internet Explorer 일부 버전(Microsoft Windows XP, 서비스팩 2와 함께 제공되는 버전 포함)은 로컬 컴퓨터의 Internet Explorer 창에서 "액티브 컨텐츠"로 간주하는 내용을 기본적으로 차단합니다. 이 기본 설정으로 인해 도움말 기능에서 일부 컨텐츠가 차단될 수 있습니다. 도움말 컨텐츠를 모두 보려면 Internet Explorer의 기본 작동을 변경할 수 있습니다.

1. Internet Explorer 메뉴에서 다음을 선택하십시오.

   도구 > 인터넷 옵션...
2. **고급** 탭을 클릭하십시오.
3. **보안** 섹션으로 아래로 스크롤하십시오.
4. **내 컴퓨터에 있는 파일에서 액티브 컨텐츠가 실행되는 것을 허용**을 선택(체크)하십시오.

모델 너깃 및 모델링 노드 생성

대화형 세션에 있을 때 완료한 작업을 사용하여 다음 둘 중 하나를 생성할 수 있습니다.

- **텍스트 마이닝 모델링 노드.** 대화형 워크벤치 세션에서 생성된 모델링 노드는 설정과 옵션이 열린 대화형 세션에 저장된 설정과 옵션을 반영하는 텍스트 마이닝 노드입니다. 이는 원래 텍스트 마이닝 노드가 더 이상 없거나 새 버전을 작성하려고 할 때 유용합니다. 자세한 정보는 19 페이지의 제 3 장 『개념 및 범주 마이닝』의 내용을 참조하십시오.
- **범주 모델 너깃.** 대화형 워크벤치 세션에서 생성된 모델 너깃은 범주 모델 너깃입니다. 범주 모델 너깃을 생성하려면 범주 및 개념 보기에 하나 이상의 범주가 있어야 합니다. 자세한 정보는 42 페이지의 『텍스트 마이닝 너깃: 범주 모델』의 내용을 참조하십시오.

텍스트 마이닝 모델링 노드 생성 방법

1. 메뉴에서 **생성 > 모델링 노드 생성**을 선택하십시오. 텍스트 마이닝 모델링 노드가 현재 워크벤치 세션의 모든 설정을 사용하여 작업 중인 캔버스에 추가됩니다. 노드는 텍스트 필드의 이름을 따서 이름 지정됩니다.

범주 모델 너깃 생성 방법

1. 메뉴에서 **생성 > 모델 생성**을 선택하십시오. 모델 너깃은 기본 이름을 가진 모델 팔레트에 직접 생성됩니다.

모델링 노드 업데이트 및 저장

대화형 세션에서 작업하는 동안 이따금씩 모델링 노드를 업데이트하여 변경사항을 저장하는 것이 좋습니다. 또한 대화형 워크벤치 세션에서 작업을 완료하고 작업을 저장하려고 할 때마다 모델링 노드를 업데이트해야 합니다. 모델링 노드를 업데이트할 때 워크벤치 세션 컨텐츠는 대화형 워크벤치 세션을 시작한 텍스트 마이닝 노드에 다시 저장됩니다. 출력 창은 닫히지 않습니다.

중요! 이 업데이트는 스트림을 저장하지 않습니다. 스트림을 저장하려면, 모델링 노드를 업데이트한 후 기본 IBM SPSS Modeler 창에서 저장을 수행하십시오.

모델링 노드 업데이트 방법

1. 메뉴에서 **파일 > 모델링 노드 업데이트**를 선택하십시오. 보유한 옵션 및 범주와 더불어 작성 및 추출 설정으로 모델링 노드가 업데이트됩니다.

세션 닫기 및 종료

세션에서 작업을 완료하면 세 가지의 다른 방법으로 세션에서 나갈 수 있습니다.

- **저장.** 이 옵션을 사용하면 먼저 다른 세션에서 재사용할 수 있도록 라이브러리를 출판할 뿐만 아니라, 나중 세션을 위해 원래의 모델링 노드에 작업을 다시 저장할 수 있습니다. 자세한 정보는 192 페이지의 『라이브러리 공유』의 내용을 참조하십시오. 저장하고 나면, 세션 창이 닫히고 세션은 IBM SPSS Modeler 창의 출력 관리자에서 삭제됩니다.
- **종료.** 이 옵션은 저장되지 않은 작업을 삭제하고, 세션 창을 닫은 후, IBM SPSS Modeler 창의 출력 관리자에서 세션을 삭제합니다. 메모리를 사용 가능하도록 하려면, 중요한 작업을 저장하고 세션을 종료하는 것이 좋습니다.
- **닫기.** 이 옵션은 어떤 작업도 저장하거나 삭제하지 않습니다. 이 옵션은 세션 창을 닫지만 세션이 계속 실행됩니다. IBM SPSS Modeler 창의 출력 관리자에서 이 세션을 선택하여 다시 세션 창을 열 수 있습니다.

워크벤치 세션을 닫으려면 다음을 수행하십시오.

1. 메뉴에서 **파일 > 닫기**를 선택하십시오.

내게 필요한 옵션의 키보드 기능

대화식 워크벤치 인터페이스는 제품의 기능성을 한층 액세스 가능하도록 만들기 위해 키보드 단축키를 제공합니다. 가장 기본적인 레벨에서 Alt + 해당 키를 눌러 창 메뉴를 활성화(예: Alt+F를 눌러 파일 메뉴에 액세스)하거나 Tab 키를 눌러 대화 상자 제어를 스크롤할 수 있습니다. 이 섹션에서는 대체 탐색에 대한 키보드 단축키에 대해 다룹니다. IBM SPSS Modeler 인터페이스의 경우 다른 키보드 단축키가 있습니다.

표 13. 일반 키보드 단축키

단축키	기능
Ctrl+1	탭이 있는 분할창에서 첫 번째 탭을 표시합니다.
Ctrl+2	탭이 있는 분할창에서 두 번째 탭을 표시합니다.
Ctrl+A	초점이 있는 분할창에 대한 모든 요소를 선택합니다.
Ctrl+C	선택한 텍스트를 클립보드로 복사합니다.
Ctrl+E	범주 및 개념 보기와 텍스트 링크 분석 보기에서 추출을 시작합니다.
Ctrl+F	자원 편집기/템플리트 편집기에서 찾기 도구 모음을 표시하고(아직 볼 수 없는 경우) 초점을 그 도구 모음에 둡니다.
Ctrl+I	범주 및 개념 보기에서, 선택된 범주에 대한 범주 정의 대화 상자를 시작합니다. 군집 보기에서, 선택된 군집에 대한 군집 정의 대화 상자를 시작합니다.
Ctrl+R	자원 편집기/템플리트 편집기에서 용어 추가 대화 상자를 엽니다.
Ctrl+T	자원 편집기/템플리트 편집기에서 새 유형을 작성하기 위해 유형 특성 대화 상자를 엽니다.
Ctrl+V	클립보드 내용을 붙여넣습니다.

표 13. 일반 키보드 단축키 (계속)

단축키	기능
Ctrl+X	자원 편집기/템플리트 편집기에서 선택된 항목을 잘라냅니다.
Ctrl+Y	보기에서 마지막 조치를 다시 실행합니다.
Ctrl+Z	보기에서 마지막 조치를 실행 취소합니다.
F1	도움말을 표시하거나, 대화 상자에 있는 경우 항목에 대한 컨텍스트 도움말을 표시합니다.
F2	테이블 셀에서 편집 모드 안팎으로 토글합니다.
F6	활성 보기에서 기본 분할창 사이에 초점을 이동합니다.
F8	크기를 조정하기 위해 분할창 분할기 막대로 초점을 이동합니다.
F10	기본 파일 메뉴를 펼칩니다.
위쪽 화살표, 아래로 화살표	분할기 막대가 선택될 때 수직으로 분할창 크기를 조정합니다.
왼쪽 화살표, 우측 화살표	분할기 막대가 선택될 때 수평으로 분할창 크기를 조정합니다.
Home, End	분할기 막대가 선택될 때 최소 또는 최대 크기로 분할창 크기를 조정합니다.
Tab	창, 분할창 또는 대화 상자에서 항목 사이에 앞으로 이동합니다.
Shift+F10	항목의 컨텍스트 메뉴를 표시합니다.
Shift+Tab	창 또는 대화 상자에서 항목 사이에 뒤로 이동합니다.
Shift+화살표	편집 모드(F2)에 있을 때 편집 필드에서 문자를 선택합니다.
Ctrl+Tab	창에서 다음 주 영역으로 초점을 앞으로 이동합니다.
Shift+Ctrl+Tab	창에서 이전 주 영역으로 초점을 뒤로 이동합니다.

대화 상자의 단축키

대화 상자에 대해 작업할 때 몇 개의 단축키 및 스크린 리더 키가 도움이 됩니다. 대화 상자에 입력할 때 첫 번째 제어에 초점을 맞추고 스크린 리더를 초기화하기 위해 Tab 키를 눌러야 할 수도 있습니다. 특수 키보드 및 스크린 리더 단축키의 전체 목록은 다음 테이블에 제공됩니다.

표 14. 대화 상자 단축키

단축키	함수
Tab	창 또는 대화 상자에서 항목 사이에 앞으로 이동합니다.
Ctrl+Tab	텍스트 상자에서 다음 항목으로, 앞으로 이동합니다.
Shift+Tab	창 또는 대화 상자에서 항목 사이에 뒤로 이동합니다.
Shift+Ctrl+Tab	텍스트 상자에서 이전 항목으로, 뒤로 이동합니다.
스페이스바	초점이 있는 제어 또는 단추를 선택합니다.
Esc	변경사항을 취소시키고 대화 상자를 닫습니다.
Enter	변경사항의 유효성을 검증하고 대화 상자를 닫습니다(확인 단추와 같음). 텍스트 상자에 있는 경우, 먼저 Ctrl+Tab을 눌러서 텍스트 상자에서 나가야 합니다.

제 9 장 개념 및 유형 추출

대화형 워크벤치를 실행하는 스트림을 실행할 때마다 스트림의 텍스트 데이터에서 추출이 자동으로 수행됩니다. 이 추출의 결과는 개념, 유형 및 TLA 패턴이 언어학적 자원에 존재하는 경우 패턴 세트입니다. 추출 결과 분할창에서 개념 및 유형을 보고 이에 대한 작업을 할 수 있습니다. 자세한 정보는 5 페이지의 『추출 작동 방법』의 내용을 참조하십시오.

추출 결과를 미세 조정하려면 언어학적 자원을 수정하고 재추출할 수 있습니다. 자세한 정보는 99 페이지의 『추출 결과 세분화』의 내용을 참조하십시오. 추출 프로세스는 자원 및 추출 대화 상자의 매개변수에 의존하여 결과 추출 및 구성 방법을 지시합니다. 추출 결과를 사용하여 모두는 아니더라도 범주 정의의 더 나은 파트를 정의할 수 있습니다.

추출 결과: 개념 및 유형

추출 프로세스 동안에 모든 텍스트 데이터가 스캔되고 관련 개념이 식별되고, 추출되고 유형에 지정됩니다. 추출이 완료되면 결과는 범주와 개념 보기의 왼쪽 하단 구석에 있는 추출 결과 분할창에 나타납니다. 세션을 처음 실행하면 노드에서 선택한 언어학적 자원 템플리트가 이러한 개념과 유형을 추출하고 구성하는 데 사용됩니다.

추출되는 개념, 유형 및 TLA 패턴은 집합적으로 **추출 결과**라고 불리고, 이들은 범주의 디스크립터 또는 구성요소의 역할을 합니다. 범주 규칙에서 개념, 유형 및 패턴을 사용할 수도 있습니다. 또한, 자동 기법은 개념과 유형을 사용하여 범주를 작성합니다.

텍스트 마이닝 은 추출 결과가 텍스트 데이터의 컨텍스트에 따라 검토되고 새로운 결과를 생성하기 위해 세부 조정한 다음 다시 평가되는 반복적인 프로세스입니다. 추출 후에는 결과를 검토하고 언어학적 자원을 수정하여 필요에 따라 변경해야 합니다. 자원을 부분적으로 추출 결과 분할창, 데이터 분할창, 범주 정의 대화 상자 또는 군집 정의 대화 상자에서 직접 세부 조정할 수 있습니다. 자세한 정보는 99 페이지의 『추출 결과 세분화』 주제를 참조하십시오. 자원 편집기 보기에서 직접 수행할 수도 있습니다. 자세한 정보는 84 페이지의 『자원 편집기 보기』 주제를 참조하십시오.

세부 조정 후에는 새로운 결과를 보기 위해서 다시 추출할 수 있습니다. 추출 결과를 처음부터 세부 조정하여 다시 추출할 때마다 데이터의 컨텍스트에 완벽하게 적응된, 범주 정의에서 동일한 결과를 얻도록 할 수 있습니다. 이런 방식으로 문서/레코드는 보다 정확하고 반복 가능한 방식으로 범주 정의에 지정됩니다.

개념

추출 프로세스 동안에 텍스트에서 텍스트 데이터가 스캔되고 관심 또는 관련된 단일어(예: `election` 또는 `peace`) 및 단어 구(예: `presidential election`, `election of the president` 또는 `peace treaties`)를 식별하기 위해 분석됩니다. 이러한 단어와 구문을 집합적으로 *용어*라고 부릅니다. 언어학적 자원을 사용하여 관련 용어가 추출된 다음 유사한 용어는 **개념**이라는 리드 용어 하에 그룹화됩니다.

마우스를 개념 이름 위에 올리면 개념의 기본 용어 세트를 볼 수 있습니다. 그렇게 하면 개념 이름을 표시하는 도구팁과 해당 개념 아래에 그룹화되는 몇몇 용어 라인이 표시됩니다. 이러한 기본 용어에는 언어학적 자원(텍스트에서 발견되는지 여부와 관계 없이)에 정의되는 동의어뿐만 아니라 추출된 복수/단수 용어, 순열된 용어, 퍼지 그룹화의 용어 등이 포함됩니다. 이러한 용어를 복사하거나 개념 이름을 마우스 오른쪽 단추로 클릭하고 컨텍스트 메뉴 옵션을 선택하여 전체 기본 용어 세트를 볼 수 있습니다.

기본적으로, 개념은 소문자로 표시되고 문서 개수(문서 열) 의 내림차순으로 정렬되어 있습니다. 개념이 추출되면 유사한 개념을 그룹화하기 위해 유형이 지정됩니다. 이들은 이 유형에 따라 색상 코딩됩니다. 색상은 자원 편집기 내에서 유형 특성에 정의됩니다. 자세한 정보는 197 페이지의 『유형 사전』의 내용을 참조하십시오.

개념, 유형 또는 패턴이 범주 정의에 사용될 때마다 아이콘이 정렬 가능한 **위치** 열에 나타납니다.

유형

유형은 개념의 시맨틱 그룹입니다. 개념이 추출되면 유사한 개념을 그룹화하기 위해 유형이 지정됩니다. 몇몇 내장된 유형은 IBM SPSS Modeler Text Analytics (예: <Location>, <Organization>, <Person>, <Positive>, <Negative> 등)과 함께 제공됩니다. 예를 들어, <위치> 유형은 지리적 키워드와 장소를 그룹화합니다. 이 유형은 chicago, paris 및 tokyo와 같은 개념에 지정됩니다. 유형 사전에는 없고 텍스트에서는 추출된 대부분의 언어의 경우, 개념은 자동으로 <Unknown>으로 유형이 지정됩니다.자세한 정보는 198 페이지의 『내장 유형』 주제를 참조하십시오.

유형 보기를 선택할 때 추출된 유형은 기본적으로 글로벌 빈도순으로 내림차순으로 나타납니다. 또한 유형은 식별하기 쉽도록 색상 코딩되어 있음을 볼 수 있습니다. 색상은 유형 특성의 일부입니다. 자세한 정보는 199 페이지의 『유형 작성』의 내용을 참조하십시오. 사용자만의 유형을 작성할 수도 있습니다.

패턴

패턴은 또한 텍스트 데이터에서 추출할 수도 있습니다. 그러나 자원 편집기에 일부 텍스트 링크 분석(TLA) 패턴을 포함하는 라이브러리가 있어야 합니다. 또한 IBM SPSS Modeler Text Analytics 노드 설정에서 또는 추출 대화 상자에서 **텍스트 링크 분석 패턴 추출 사용** 옵션을 사용하여 이러한 패턴을 추출하도록 선택해야 합니다. 자세한 정보는 159 페이지의 제 12 장 『텍스트 링크 분석 탐색』의 내용을 참조하십시오.

데이터 추출

추출이 필요할 때마다 추출 결과 분할창은 노란색이 되고 **개념을 추출하려면 추출 단추를 누르십시오** 메시지가 이 분할창에서 도구 모음 아래에 나타납니다.

추출 결과가 아직 없거나, 언어학적 자원을 변경했거나, 추출 결과를 업데이트해야 하거나, 추출 결과를 저장하지 않은 세션을 다시 연 경우에(**도구 > 옵션**) 추출해야 할 수도 있습니다.

참고: 추출 결과가 **세션 작업 사용...** 옵션을 사용하여 캐싱된 후에 스트림의 소스 노드를 변경한 경우에는 추출 결과를 업데이트하려면 대화식 워크벤치 세션이 실행된 후에 새 추출을 실행해야 합니다.

추출을 실행하면 진행 표시기가 나타나서 추출 상태에 대한 피드백을 제공합니다. 이번에는 추출 엔진은 모든 텍스트 데이터를 읽고 관련 용어와 패턴을 식별하고 이를 추출하고 이를 유형에 지정합니다. 그런 다음 엔진은 동의어를 개념이라고 불리는 하나의 리드 용어 아래에 그룹화하려고 시도합니다. 프로세스가 완료되면 결과로 나오는 개념, 유형 및 패턴이 추출 결과 분할창에 나타납니다.

추출 프로세스는 개념 및 유형 세트뿐만 아니라 텍스트 링크 분석(TLA) 패턴(사용 가능한 경우)을 결과로 생성합니다. 이러한 개념과 유형을 범주 및 개념 보기의 추출 결과 분할창에서 보고 작업할 수 있습니다. TLA 패턴을 추출한 경우에는 이를 텍스트 링크 분석 보기에서 볼 수 있습니다.

참고: 데이터 세트의 크기와 추출 프로세스를 완료하는 데 걸리는 시간 간의 관계가 있습니다. 언제든지 표본 노드 업스트림을 삽입하거나 시스템의 구성 최적화를 고려할 수 있습니다.

데이터 추출 방법

1. 메뉴에서 **도구 > 추출**을 선택하십시오. 또는 **추출** 도구 모음 단추를 클릭하십시오.
2. 추출 설정 대화 상자를 항상 표시하도록 선택하면 이는 사용자가 변경할 수 있도록 나타납니다. 각 설정의 디스크립터에 대해서는 이 주제를 추가로 참조하십시오.
3. **추출**을 클릭하여 추출 프로세스를 시작하십시오. 추출이 시작되면 진행 대화 상자가 열립니다. 추출 후에는 결과가 추출 결과 분할창에 나타납니다. 기본적으로, 개념은 소문자로 표시되고 문서 개수(문서 열) 의 내림차순으로 정렬되어 있습니다.

결과를 다르게 정렬하고, 결과를 필터링하거나 다른 보기(개념 또는 유형)로 전환하려면 도구 모음 옵션을 사용하여 결과를 검토할 수 있습니다. 언어학적 자원에 대해 작업하여 추출 결과를 세분화할 수도 있습니다. 자세한 정보는 99 페이지의 『추출 결과 세분화』 주제를 참조하십시오.

네덜란드어, 영어, 프랑스어, 독일어, 이탈리아어, 포르투갈어 및 스페인어 텍스트의 경우

추출 설정 대화 상자에는 몇몇 기본 추출 옵션이 포함됩니다.

텍스트 링크 분석 패턴 추출을 사용으로 설정하십시오.텍스트 데이터에서 TLA 패턴을 추출하려 함을 지정합니다. 또한 자원 편집기에서 사용자의 라이브러리 중 하나에 TLA 패턴 규칙이 있다고 가정합니다. 이 옵션은 추출 시간을 현저하게 늘릴 수 있습니다. 자세한 정보는 159 페이지의 제 12 장 『텍스트 링크 분석 탐색』의 내용을 참조하십시오.

구두점 오류를 조정하십시오. 이 옵션은 추출 중에 구두점 오류가 있는 텍스트(예를 들어, 부적절한 사용법)를 일시적으로 표준화하여 개념의 추출 가능성을 향상시킵니다. 이 옵션은 텍스트가 짧고 품질이 저조한 경우(예를 들어, 개방형 설문조사 반응, 이메일 및 CRM 데이터) 또는 텍스트에 많은 약어가 포함된 경우에 특히 유용합니다.

[n]의 최소 루트 문자 제한의 맞춤법 오류를 수용합니다. 이 옵션은 맞춤법이 자주 틀리는 단어나 맞춤법이 유사한 단어를 하나의 개념으로 그룹화하는 퍼지 그룹화 기술을 적용합니다. 퍼지 그룹화 알고리즘은 일시적으로 모든 모음(맨 처음 것은 제외)을 지우고 추출된 단어에서 이중/삼중 자음을 지운 다음 `modeling`과 `modelling`이 함께 그룹화될 수 있도록 이들이 동일한지 비교합니다. 그러나 각 용어가 `<Unknown>` 유형을 제외하고 서로 다른 유형에 지정된 경우에는 퍼지 그룹화 기술은 적용되지 않습니다.

퍼지 그룹화를 사용하기 전에 필요한 루트 문자의 최소 수를 정의할 수도 있습니다. 용어에서 루트 문자의 수는 모든 문자를 합한 후 굴절접사를 형성하는 문자와 복합어의 경우에는 한정사 및 전치사를 형성하는 문자를 빼서 계산합니다. 예를 들어, exercises 용어는 "exercise" 양식의 8개 루트 문자가 있는 것으로 간주됩니다. 단어 끝의 *s*자는 굴절(복수형)이기 때문입니다. 마찬가지로, apple sauce는 10개의 루트 문자로 간주되고("apple sauce") manufacturing of cars는 16개의 루트 문자("manufacturing car")로 간주됩니다. 이 계산 방법은 퍼지 그룹화를 적용해야 하는지 여부를 확인하는 데에만 사용되고 단어가 매치하는 방법에는 영향을 미치지 않습니다.

참고: 특정 단어가 나중에 잘못 그룹화되는 경우에는 이를 고급 자원 탭의 **퍼지 그룹화: 예외** 섹션에 명시적으로 선언하여 이 기술에서 단어 쌍을 제외할 수 있습니다. 자세한 정보는 214 페이지의 『퍼지 그룹화』의 내용을 참조하십시오.

단일어 추출. 이 옵션은 단어가 복합어의 일부가 아니거나 명사이거나 인식되지 않은 품사인 경우에만 단일어를 추출합니다.

비언어 엔티티 추출. 이 옵션은 전화 번호, 주민등록번호, 시간, 날짜, 통화, 숫자, 백분율, 이메일 주소 및 HTTP 주소 등과 같은 비언어 엔티티를 추출합니다. 고급 자원 탭의 **비언어 엔티티: 구성** 섹션에서 비언어 엔티티의 특정 유형을 포함하거나 제외할 수 있습니다. 불필요한 엔티티를 사용 안함으로 설정하면 추출 엔진은 처리 시간을 낭비하지 않습니다. 자세한 정보는 219 페이지의 『구성』의 내용을 참조하십시오.

대문자 알고리즘. 이 옵션은 용어의 첫 글자가 대문자인 한 내장된 사전에 없는 단순 및 복합어를 추출합니다. 이 옵션은 가장 적합한 명사를 추출하기 위한 좋은 방법을 제공합니다.

가능한 경우 부분 및 전체 사람 이름을 함께 그룹화. 이 옵션은 텍스트에 다르게 나타나는 이름을 그룹화합니다. 이름은 종종 시작부에는 전체 이름이 언급되고 나중에는 약어로만 표시되기 때문에 이 기능이 유용합니다. 이 옵션은 <Unknown> 유형의 단일어를 <Person>으로 입력되는 복합어와 매치시키려고 시도합니다. 예를 들어, *doe*가 있고 처음에는 <Unknown>으로 입력되는 경우에는, 추출 엔진은 <Person> 유형에 있는 복합어가 *doe*를 마지막 단어로 포함하는지 여부를 확인합니다(예: *john doe*). 이 옵션은 이름에는 적용되지 않습니다. 이름의 대부분은 단일어로 추출되지 않기 때문입니다.

최대 비기능 단어 순열. 이 옵션은 순열 기술을 적용할 때 존재할 수 있는 비기능 단어 최대 수를 지정합니다. 이 순열 기술은 서로 굴절과는 관계 없이 포함된 비기능 단어(예: of 및 the)만 다른 유사한 구를 그룹화합니다. 예를 들어, 이 값을 최소 두 개의 단어로 설정하고 company officials 및 officials of the company 둘 모두가 추출되었다고 해 봅시다. 이 경우, 추출된 두 용어는 모두 마지막 개념 목록에 그룹화됩니다. 두 용어 모두 of the가 무시될 때 동일한 것으로 간주되기 때문입니다.

개념 맵의 색인 옵션 개념 맵을 나중에 빠르게 그릴 수 있도록 추출 시에 맵 색인 작성을 지정합니다. 색인 설정을 편집하려면 **설정**을 클릭하십시오. 자세한 정보는 99 페이지의 『개념 맵 지수 작성』의 내용을 참조하십시오.

추출을 시작하기 전에 항상 이 대화 상자를 표시.추출할 때마다 추출 설정 대화 상자를 표시하려는지 여부를 지정합니다. 도구 메뉴로 돌아가지 않는 한 이를 표시하지 않거나, 추출할 때마다 추출 설정을 편집하려는지 요청할지 여부를 지정합니다.

일본어 텍스트의 경우

추출 설정 대화 상자에는 일본어 텍스트 언어를 위한 몇몇 기본 추출 옵션이 있습니다. 기본적으로 대화 상자에 선택된 설정은 텍스트 마이닝 모델링 노드의 전문가 탭에서 선택된 것과 동일합니다. 일본어 텍스트에 대해 작업하려면 텍스트를 입력으로 사용할 뿐만 아니라 텍스트 마이닝 노드의 모델 탭에서 일본어 템플리트 또는 텍스트 분석 패키지를 선택해야 합니다. 자세한 정보는 27 페이지의 『템플리트 및 TAP에서 자원 복사』의 내용을 참조하십시오.

2차 분석. 추출이 실행될 때 기본 키워드 추출은 기본 유형 세트를 사용하여 수행됩니다. 그러나, 2차 분석기를 선택하면, 추출기가 이제는 불변화사 및 보조 동사를 개념의 일부로 포함하므로 더 많은 수 또는 더 풍부한 개념을 확보할 수 있습니다. 정서 분석의 경우 수많은 추가 유형 또한 포함됩니다. 게다가, 2차 분석기를 선택하면 텍스트 링크 분석 결과를 생성할 수도 있습니다.

참고: 2차 분석기가 호출되면 추출 프로세스를 완료하는 데 더 오래 걸립니다.

- **종속성 분석**. 이 옵션을 선택하면 기본 유형과 키워드 추출로부터 추출 개념의 확장된 불변화사가 생깁니다. 종속 항목 텍스트 링크 분석(TLA)으로부터 더 풍부한 패턴 결과를 얻을 수도 있습니다.
- **정서 분석**. 이 분석기를 선택하면 추가로 추출된 개념이 생기고, 적용 가능한 경우 TLA 패턴 결과 추출이 생깁니다. 기본 유형에 추가로, 80개가 넘는 정서 유형의 혜택을 얻을 수도 있습니다. 이러한 유형은 감정, 정서 및 의견의 표현을 통해 텍스트에서 개념과 패턴을 찾아내는 데 사용됩니다. 정서 분석의 초점을 지시하는 세 개의 옵션, **모든 정서**, **대표 정서만** 및 **결론만**이 있습니다.
- **2차 분석기 없음**. 이 옵션은 모든 2차 분석기를 끕니다. 이 옵션은 TLA 결과를 구하기 위해서 2차 분석기가 필요하므로 **텍스트 링크 분석 패턴 추출 사용** 옵션이 선택된 경우에는 선택할 수 없습니다.

텍스트 링크 분석 패턴 추출을 사용으로 설정하십시오.텍스트 데이터에서 TLA 패턴을 추출하려 함을 지정합니다. 또한 자원 편집기에서 사용자의 라이브러리 중 하나에 TLA 패턴 규칙이 있다고 가정합니다. 이 옵션은 추출 시간을 현저하게 늘릴 수 있습니다. 또한 TLA 패턴 결과를 추출하기 위해서는 2차 분석기가 선택되어야 합니다. 자세한 정보는 159 페이지의 제 12 장 『텍스트 링크 분석 탐색』의 내용을 참조하십시오.

추출 결과 필터링

매우 큰 데이터 세트에 대한 작업 시 추출 프로세스는 수백만 개의 결과를 생성할 수 있습니다. 많은 사용자가 이 양으로 인해 결과를 효과적으로 검토하기가 더 어렵습니다. 따라서 가장 흥미로운 해당 결과에 주목하려면 추출 결과 분할창에서 사용 가능한 필터 대화 상자를 통해 이러한 결과를 필터링할 수 있습니다.

이 필터 대화 상자의 모든 설정을 함께 사용하여 범주에 사용 가능한 추출 결과를 필터링함을 명심하십시오.

빈도별 필터. 일정 글로벌 또는 문서 빈도 값을 가진 해당 결과만 표시하도록 필터링할 수 있습니다.

- **글로벌 빈도**는 전체 문서 또는 레코드 세트에 개념이 나타나는 총 횟수이며 **글로벌** 열에 표시됩니다.
- **문서 빈도**는 개념이 나타나는 총 문서 또는 레코드 수이며 **문서** 열에 표시됩니다.

예를 들어, 개념 nato가 500개 레코드에 800번 나타났으면 이 개념의 글로벌 빈도는 800이고 문서 빈도는 500입니다.

유형별. 일정 유형에 속한 해당 결과만 표시하도록 필터링할 수 있습니다. 모든 유형 또는 특정 유형만 선택할 수 있습니다.

매치 텍스트별. 여기에 정의하는 규칙과 매치하는 해당 결과만 표시하도록 필터링할 수도 있습니다. **매치 텍스트** 필드에 매치될 문자 세트를 입력한 후 매치를 적용할 조건을 선택하십시오.

표 15. 매치 텍스트 조건

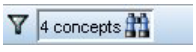
조건	설명
포함	문자열이 어딘가에 발생하면 텍스트가 매치됩니다(기본 선택사항).
시작 문자	개념 또는 유형이 지정된 텍스트로 시작하는 경우에만 텍스트가 매치됩니다.
끝 문자	개념 또는 유형이 지정된 텍스트로 끝나는 경우에만 텍스트가 매치됩니다.
정확히 일치	전체 문자열이 개념 또는 유형 이름과 매치해야 합니다.

순위별. 또한 글로벌 빈도(**글로벌**) 또는 문서 빈도(**문서**)에 따라 오름차순이나 내림차순으로 순위가 높은 개념만 표시하도록 필터링할 수 있습니다.

추출 결과 분할창에 표시된 결과

필터를 기반으로 추출 결과 분할창 도구 모음에 영어로 결과가 표시되는 방법의 몇 가지 예제는 다음과 같습니다.

표 16. 필터 피드백 예제

필터 피드백	설명
	도구 모음은 결과 수를 표시합니다. 텍스트 매치 필터가 없고 최대값을 충족하지 않았기 때문에 추가 아이콘이 표시되지 않습니다.
	도구 모음은 필터에 지정된 최대값(이 경우 300)으로 결과가 제한되었음을 표시합니다. 보라색 아이콘이 있는 경우 이는 최대 개념 수가 충족되었음을 의미합니다. 자세한 정보를 보려면 아이콘 위에 마우스를 올려 놓으십시오.
	도구 모음은 매치 텍스트 필터를 사용하여 결과가 제한되었음을 표시합니다. 이는 돋보기 아이콘으로 표시됩니다.

결과 필터링 방법

1. 메뉴에서 **도구 > 필터**를 선택하십시오. 필터 대화 상자가 열립니다.
2. 사용할 필터를 선택하고 세분화하십시오.
3. **확인**을 클릭하여 필터를 적용하고 추출 결과 분할창에서 새 결과를 확인하십시오.

개념 맵 탐색

개념 맵을 작성하여 개념이 상호 관련되는 방법을 탐색할 수 있습니다. 단일 개념을 선택하고 **맵**을 클릭하면 선택된 개념과 관련된 개념 세트를 탐색할 수 있도록 개념 맵 창이 열립니다. 포함시킬 유형, 검색할 관계 종류 등과 같은 설정을 편집하여 표시되는 개념을 필터링할 수 있습니다.

중요! 맵을 작성하려면 먼저 지수를 생성해야 합니다. 이를 수행하는 데 몇 분이 걸릴 수 있습니다. 그러나 일단 지수를 생성했으면 재추출할 때까지 다시 재생성하지 않아도 됩니다. 추출할 때마다 자동으로 지수를 생성

하려면 추출 설정에서 해당 옵션을 선택하십시오. 자세한 정보는 92 페이지의 『데이터 추출』의 내용을 참조하십시오.

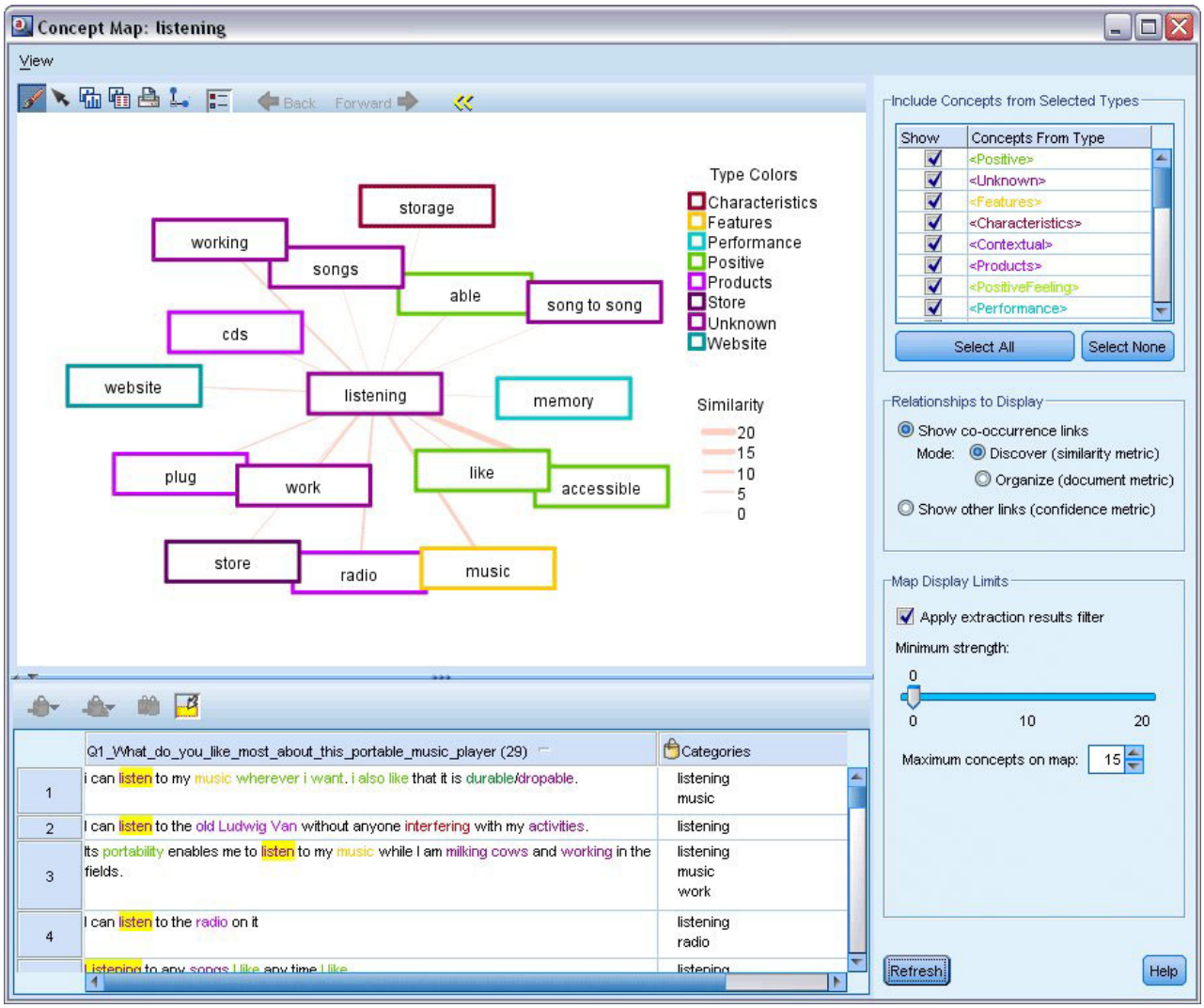


그림 27. 선택된 개념에 대한 개념 맵

개념 맵을 보는 방법

1. 추출 결과 분할창에서 단일 개념을 선택하십시오.
2. 이 분할창의 도구 모음에서 **맵** 단추를 클릭하십시오. 맵 지수가 이미 생성된 경우 개념 맵은 별도의 대화 상자에서 열립니다. 맵 지수가 생성되지 않았거나 오래된 경우에는 지수를 다시 작성해야 합니다. 이 프로세스는 몇 분이 걸릴 수 있습니다.
3. 탐색할 맵을 클릭하십시오. 링크된 개념을 두 번 클릭하면 맵이 저절로 다시 그려져 방금 두 번 클릭한 개념에 대해 링크된 개념을 보여줍니다.

4. 맨 위 도구 모음은 이전 맵으로 다시 이동, 관계 강도에 따라 링크 필터링, 표시할 관계 종류는 물론 나타나는 개념 유형을 제어하기 위한 필터 대화 상자 열기와 같은 몇 가지 기본 맵 도구를 제공합니다. 두 번째 도구 모음 줄은 그래프 편집 도구를 포함합니다. 자세한 정보는 169 페이지의 『그래프 도구 모음 및 팔레트 사용』의 내용을 참조하십시오.
5. 찾고 있는 링크 종류에 만족하지 않으면 맵 오른쪽에 표시된 이 맵 설정을 검토하십시오.

맵 설정: 선택된 유형의 개념 포함

테이블에서 선택된 유형에 속한 해당 개념만 맵에 표시됩니다. 일정 유형의 개념을 숨기려면 테이블에서 해당 유형을 선택 취소하십시오.

맵 설정: 표시할 관계

동시 발생 링크 표시. 동시 발생 링크를 표시하려면 모드를 선택하십시오. 모드는 링크 강도 계산 방법에 영향을 줍니다.

- *발견(유사성 메트릭)*. 이 메트릭을 사용하면 두 개념이 함께 나타나는 빈도는 물론 따로 나타나는 빈도도 고려하는 보다 복잡한 계산을 사용하여 링크 강도를 계산합니다. 강도 값이 높으면 개념 쌍이 따로 나타나는 것보다 더 자주 함께 나타나는 경향이 있음을 의미합니다. 다음 수식을 사용하면 부동 소수점 값이 정수로 변환됩니다.

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

그림 28. 유사성 계수 수식

  이 수식에서 C_I는 개념 I가 발생하는 문서 또는 레코드 수입니다.

  C_J는 개념 J가 발생하는 문서 또는 레코드 수입니다.

  C_{IJ}는 개념 쌍 I와 J가 문서 세트에서 동시 발생하는 문서 또는 레코드 수입니다.
- *구성(문서 메트릭)*. 이 메트릭을 사용하면 동시 발생 원래 수를 통해 링크 강도를 판별합니다. 일반적으로 더 빈번한 두 개의 개념이 때때로 함께 발생할 가능성이 높습니다. 강도 값이 높으면 개념 쌍이 자주 함께 나타남을 의미합니다.

다른 링크 표시(신뢰 메트릭). 표시할 다른 링크를 선택할 수 있습니다. 이는 시맨틱, 파생(형태론) 또는 포함(구문)이며 링크된 개념에서 제거된 단계 수와 관련됩니다. 이를 통해 자원 특히, 동의성을 조정하거나 모호성을 해소할 수 있습니다. 이러한 집단 기술 각각에 대한 간단한 설명은 118 페이지의 『고급 언어학적 설정』의 내용을 참조하십시오.

참고: 지수가 작성될 때 선택되지 않았거나 관계를 찾을 수 없으면 아무 것도 표시되지 않음을 명심하십시오. 자세한 정보는 99 페이지의 『개념 맵 지수 작성』의 내용을 참조하십시오.

맵 설정: 맵 표시 한계

추출 결과 필터 적용. 모든 개념을 사용하지는 않으려면 추출 결과 분할창에서 필터를 선택하여 표시되는 항목을 제한할 수 있습니다. 그리고 나서 이 옵션을 선택하면 IBM SPSS Modeler Text Analytics가 이 필터링된 세트를 사용하여 관련 개념을 찾습니다. 자세한 정보는 95 페이지의 『추출 결과 필터링』의 내용을 참조하십시오.

최소 강도. 여기서 최소 링크 강도를 설정하십시오. 이 한계보다 관계 강도가 낮은 관련된 개념은 맵에서 숨겨집니다.

맵의 최대 개념 수. 맵에 표시할 최대 관계 수를 지정하십시오.

개념 맵 지수 작성

맵을 작성하려면 먼저 개념 관계 지수를 생성해야 합니다. 개념 맵을 작성할 때마다 IBM SPSS Modeler Text Analytics는 이 지수를 참조합니다. 이 대화 상자에서 기술을 선택하여 지수화할 관계를 선택할 수 있습니다.

집단 기술. 하나 이상의 기술을 선택하십시오. 이러한 기술 각각에 대한 간단한 설명은 121 페이지의 『언어학적 기술 정보』의 내용을 참조하십시오. 모든 텍스트 언어에 모든 기술을 사용할 수 있는 것은 아닙니다.

특정 개념 쌍 방지. 프로세스가 출력에 두 개의 개념을 함께 그룹화하거나 쌍으로 만드는 프로세스를 중지하려면 이 선택란을 선택하십시오. 개념 쌍을 작성하거나 관리하려면 **쌍 관리**를 클릭하십시오. 자세한 정보는 121 페이지의 『링크 예외 쌍 관리』의 내용을 참조하십시오.

지수를 작성하는 데 몇 분이 걸릴 수 있습니다. 그러나 일단 지수를 생성했으면 재추출할 때까지 또는 더 많은 관계를 포함시키도록 설정을 변경하지 않는 한 다시 재생성하지 않아도 됩니다. 추출할 때마다 지수를 생성하려면 추출 설정에서 해당 옵션을 선택할 수 있습니다. 자세한 정보는 92 페이지의 『데이터 추출』의 내용을 참조하십시오.

추출 결과 세분화

추출은 결과를 추출 및 검토하고 추출을 변경한 후 다시 추출하여 결과를 업데이트하는 반복적 프로세스입니다. 정확도와 연속성이 성공적인 텍스트 마이닝 및 범주화에 필수적이므로, 시작부터 추출 결과를 미세 조정하는 것이 재추출할 때마다 범주 정의에서 정확하게 동일한 결과를 얻도록 보장합니다. 이 방법으로 레코드 및 문서가 더 정확하고 반복 가능한 방식으로 범주에 지정됩니다.

추출 결과는 범주에 대한 구성 요소의 역할을 합니다. 이들 추출 결과를 사용하여 범주를 작성할 때, 레코드 및 문서가 하나 이상의 범주 디스크립터와 매치하는 텍스트를 포함하는 경우 자동으로 범주에 지정됩니다. 언어학적 자원에 대한 세분화를 수행하기 전에 범주화를 시작할 수 있지만, 시작하기 전에 최소한 한 번은 추출 결과를 검토하는 것이 유용합니다.

결과를 검토할 때 추출 엔진이 상이하게 처리하기 원하는 요소를 발견할 수 있습니다. 다음 예제를 고려하십시오.

- **인식되지 않는 동의어.** `smart`, `intelligent`, `bright`, `knowledgeable` 같이 동의어인 것으로 간주하는 여러 개의 개념을 발견하고, 이들이 모두 추출 결과에 개별 개념으로 나타난다고 가정하십시오. `intelligent`,

bright, knowledgeable이 대상 개념 smart 아래에 모두 그룹화되는 동의어 정의를 작성할 수 있습니다. 그렇게 하면 이들 모두가 smart와 그룹화되고, 글로벌 빈도 수는 더 높아집니다. 자세한 정보는 101 페이지의 『동의어 추가』의 내용을 참조하십시오.

- **맞춤법이 틀린 개념.** 추출 결과의 개념들이 하나의 유형에 나타나며 이들 개념이 또 다른 유형에 지정되기 원한다고 가정하십시오. 또 다른 예제에서, 추출 결과에서 15개의 채소 개념을 발견하고 이들 모두가 <야채>라는 새 유형에 추가되길 원한다고 상상하십시오. 유형 사전에는 없고 텍스트에서는 추출된 대부분의 언어의 경우, 개념은 자동으로 <Unknown>으로 유형이 지정됩니다. 개념을 유형에 추가할 수 있습니다. 자세한 정보는 102 페이지의 『유형에 개념 추가』의 내용을 참조하십시오.
- **무의미한 개념.** 추출되었고 매우 높은 빈도 수를 갖는 개념을 발견한다고 가정하십시오. 즉 많은 레코드 또는 문서에서 발견됩니다. 그러나 이 개념이 사용자 분석에는 중요하지 않다고 간주합니다. 이 개념을 추출에서 제외시킬 수 있습니다. 자세한 정보는 103 페이지의 『추출에서 개념 제외』의 내용을 참조하십시오.
- **올바르지 않은 매치.** 특정 개념을 포함하는 레코드 또는 문서를 검토할 때 faculty와 facility와 같이 두 개의 단어가 올바르지 않게 그룹화되었음을 발견한다고 가정하십시오. 이 매치는 공통된 철자법 오류를 그룹화하기 위해 이중 또는 삼중 자음과 모음을 일시적으로 무시하는 퍼지 그룹화하는 내부 알고리즘이 원인일 수 있습니다. 이들 단어를 그룹화되지 않아야 하는 단어 쌍의 목록에 추가할 수 있습니다. 자세한 정보는 214 페이지의 『퍼지 그룹화』 주제를 참조하십시오. 퍼지 그룹화는 일본어 텍스트에 사용할 수 없습니다.
- **추출되지 않는 개념.** 추출된 특정 개념을 찾을 것으로 예상하지만 레코드 또는 문서 텍스트를 검토할 때 몇 개의 단어나 구가 추출되지 않았음을 알 것으로 가정하십시오. 종종 이들 단어는 사용자가 관심을 갖지 않는 동사나 형용사입니다. 그러나 가끔은 범주 정의의 일부로서 추출되지 않은 단어나 구를 사용하기 원합니다. 개념을 추출하기 위해 용어를 유형 사전에 강제 실행할 수 있습니다. 자세한 정보는 104 페이지의 『단어 강제 추출』의 내용을 참조하십시오.

이들 변경의 많은 수가 하나 이상의 요소를 선택하고 마우스 오른쪽 단추를 클릭하여 컨텍스트 메뉴에 액세스하여 추출 결과 분할창 , 데이터 분할창, 범주 정의 대화 상자 또는 클러스터 정의 대화 상자에서 직접 수행할 수 있습니다.

변경을 수행한 후, 분할창 배경 색상이 변하여 변경사항을 보려면 재추출해야 함을 표시합니다. 자세한 정보는 92 페이지의 『데이터 추출』의 내용을 참조하십시오. 더 큰 데이터 세트에 대해 작업 중인 경우, 각 변경 후가 아니라 여러 개의 변경을 수행한 후 다시 추출하는 것이 더 효율적일 수 있습니다.

참고: 자원 편집기 보기(보기 > 자원 편집기)에서 추출 결과를 생성하는 데 사용된 편집 가능한 언어학적 자원의 전체 세트를 볼 수 있습니다. 이들 자원은 이 보기에서 라이브러리 및 사전의 양식으로 나타납니다. 라이브러리 및 사전 안에서 개념과 유형을 직접 사용자 정의할 수 있습니다. 자세한 정보는 187 페이지의 제 16 장 『라이브러리에 대한 작업』의 내용을 참조하십시오.

동의어 추가

*동의어*는 동일한 의미를 가지고 있는 두 개 이상의 단어를 연관시킵니다. 동의어는 또한 용어를 약어와 그룹화하거나 공통적으로 맞춤법이 틀린 단어를 올바른 맞춤법과 그룹화하는 데 사용됩니다. 동의어를 사용하면 대상 개념의 빈도가 더 큰데, 이것은 텍스트 데이터에서 여러 가지 방법으로 제시되는 유사한 정보를 발견하기가 훨씬 더 쉽게 만듭니다.

제품과 함께 제공되는 언어학적 자원 템플리트와 라이브러리는 많은 사전 정의된 동의어를 포함하고 있습니다. 그러나 인식되지 않는 동의어를 발견하는 경우, 다음에 추출할 때는 인식되도록 동의어를 정의할 수 있습니다.

첫 번째 단계는 대상 또는 리드, 개념이 무엇인지 결정하는 것입니다. *대상 개념*은 최종 결과에서 모든 동의어 용어를 그룹화하려는 단어나 구입니다. 추출 중에 동의어는 이 대상 개념 아래에 그룹화됩니다. 두 번째 단계는 이 개념에 대한 모든 동의어를 식별하는 것입니다. 대상 개념이 최종 추출에서 모든 동의어에 대해 대체됩니다. 용어가 동의어가 되도록 추출되어야 합니다. 그러나 대체가 발생하기 위해 대상 개념이 추출될 필요는 없습니다. 예를 들어, `intelligent`가 `smart`로 대체되기 원하는 경우, `intelligent`는 동의어이고 `smart`는 대상 개념입니다.

새 동의어 정의를 작성하는 경우 새 대상 개념이 사전에 추가됩니다. 그런 다음 동의어를 해당 대상 개념에 추가해야 합니다. 동의어를 작성 또는 편집할 때마다, 이들 변경이 자원 편집기의 동의어 사전에 기록됩니다. 이들 동의어 사전의 전체 내용을 보거나 상당한 수의 변경을 작성하려는 경우 자원 편집기에서 직접 작업하는 것이 더 좋을 수 있습니다. 자세한 정보는 205 페이지의 『대체/동의어 사전』의 내용을 참조하십시오.

모든 새 동의어는 자동으로 자원 편집기 보기에 있는 라이브러리 트리에 나열되는 첫 번째 라이브러리에 저장되는데, 기본적으로 이것은 *로컬 라이브러리*입니다.

참고: 동의어 정의를 찾고 컨텍스트 메뉴를 통해서 또는 자원 편집기에서 직접 찾을 수 없는 경우, 내부 퍼지 그룹화 기법으로부터 매치가 발생했을 수 있습니다. 자세한 정보는 214 페이지의 『퍼지 그룹화』의 내용을 참조하십시오.

새 동의어를 작성하려면 다음을 수행하십시오.

1. 추출 결과 분할창 , 데이터 분할창, 범주 정의 대화 상자 또는 클러스터 정의 대화 상자 중 하나에서, 새 동의어를 작성하려는 개념을 선택하십시오.
2. 메뉴에서 **편집 > 동의어에 추가 > 새로 만들기**를 선택하십시오. 동의어 작성 대화 상자가 열립니다.
3. 대상 텍스트 상자에 대상 개념을 입력하십시오. 이것은 모든 동의어가 그 아래에 그룹화되는 개념입니다.
4. 더 많은 동의어를 추가하려면 동의어 목록 상자에서 입력하십시오. 각 동의어 용어를 구분하려면 글로벌 구분 문자를 사용하십시오. 자세한 정보는 86 페이지의 『옵션: 세션 탭』 주제를 참조하십시오.
5. 일본어 텍스트에 대해 작업 중인 경우, **유형의 동의어** 필드에서 유형 이름을 선택하여 이들 동의어에 대한 유형을 지정하십시오. 그러나 대상은 추출 중에 지정되는 유형을 갖습니다. 하지만 대상이 개념으로 추출되지 않은 경우, 이 열에 나열되는 유형이 추출 결과의 대상에 지정됩니다.
6. **확인**을 클릭하여 변경사항을 적용하십시오. 대화 상자가 닫히고 추출 결과 분할창 배경 색상이 변경되어 변경사항을 보려면 다시 추출해야 함을 표시합니다. 변경이 여러 개인 경우 재추출 전에 변경하십시오.

동의어를 추가하려면 다음을 수행하십시오.

1. 추출 결과 분할창 , 데이터 분할창, 범주 정의 대화 상자 또는 클러스터 정의 대화 상자 중 하나에서, 기존 동의어 정의에 추가하려는 개념을 선택하십시오.
2. 메뉴에서 **편집 > 동의어에 추가**를 선택하십시오. 메뉴는 목록의 맨 위에 가장 최근에 작성된 동의어의 세트를 표시합니다. 선택된 개념을 추가하려는 동의어의 이름을 선택하십시오. 찾고 있는 동의어가 보이면 해당 동의어를 선택하십시오. 개념이 해당 동의어 정의에 추가됩니다. 보이지 않는 경우 **기타**를 선택하여 모든 동의어 대화 상자를 표시하십시오.
3. 모든 동의어 대화 상자에서 목록을 자연적 정렬순(작성 순서)으로 또는 오름차순이나 내림차순으로 정렬할 수 있습니다. 선택된 개념을 추가하려는 동의어의 이름을 선택하고 **확인**을 클릭하십시오. 대화 상자가 닫히고, 개념이 동의어 정의에 추가됩니다.

유형에 개념 추가

추출이 실행될 때마다, 추출된 개념이 공통적인 어떤 것을 갖는 용어를 그룹화하기 위해 유형에 지정됩니다. IBM SPSS Modeler Text Analytics는 많은 내장된 유형과 함께 제공됩니다. 자세한 정보는 198 페이지의 『내장 유형』의 내용을 참조하십시오. 유형 사전에는 없고 텍스트에서는 추출된 대부분의 언어의 경우, 개념은 자동으로 <Unknown>으로 유형이 지정됩니다.

결과를 검토할 때 한 유형에 나타나는 일부 개념이 다른 유형에 지정되기 원하거나 단어 그룹이 그 자체가 실제로는 새로운 유형에 속함을 발견할 수 있습니다. 이런 경우, 개념을 다른 유형에 다시 지정하거나 새 유형을 함께 작성하기 원할 것입니다. 일본어 텍스트의 경우에는 새 유형을 작성할 수 없습니다.

예를 들어, 자동차 관련 설문조사 데이터에 대해 작업 중이며 차량의 여러 가지 영역에 집중하여 범주화하는 데 관심이 있다고 가정하십시오. <대시보드>라는 유형을 작성하여 차량의 대시보드에 있는 계기 및 손잡이와 관련된 모든 개념을 그룹화할 수 있습니다. 그런 다음 해당하는 새 유형에 연료 계기, 히터, 라디오, 주행 기록계 같은 개념을 지정할 수 있습니다.

또 다른 예에서, 대학 및 전문대학과 관련된 설문조사 데이터에서 Johns Hopkins(대학)를 <조직> 유형이 아니라 <사람> 유형으로 갖는 추출에 대해 작업 중이라고 가정하십시오. 이 경우 이 개념을 <조직> 유형에 추가할 수 있습니다.

유형을 작성하거나 유형의 용어 목록에 개념을 추가할 때마다, 이들 변경사항이 자원 편집기에 있는 언어학적 자원 라이브러리의 유형 사전에 기록됩니다. 이들 사전의 내용을 보거나 상당한 수의 변경을 작성하려는 경우 자원 편집기에서 직접 작업하는 것이 더 좋을 수 있습니다. 자세한 정보는 200 페이지의 『용어 추가』의 내용을 참조하십시오.

유형에 개념을 추가하려면 다음을 수행하십시오.

1. 추출 결과 분할창 , 데이터 분할창, 범주 정의 대화 상자 또는 클러스터 정의 대화 상자 중 하나에서, 기존 유형에 추가하려는 개념을 선택하십시오.
2. 마우스 오른쪽 단추를 클릭하여 컨텍스트 메뉴를 여십시오.

3. 메뉴에서 **편집 > 유형에 추가**를 선택하십시오. 메뉴는 목록의 맨 위에 가장 최근에 작성된 유형의 세트를 표시합니다. 선택된 개념을 추가하려는 유형 이름을 선택하십시오. 찾고 있는 유형 이름이 보이면 해당 유형을 선택하십시오. 개념이 해당 유형에 추가됩니다. 보이지 않는 경우 **기타**를 선택하여 모든 유형 대화 상자를 표시하십시오.
4. 모든 유형 대화 상자에서 목록을 자연적 정렬(작성 순서)로 또는 오름차순이나 내림차순으로 정렬할 수 있습니다. 선택된 개념을 추가하려는 유형의 이름을 선택하고 **확인**을 클릭하십시오. 대화 상자가 닫히고, 개념이 유형에 용어로서 추가됩니다.

참고: 일본어 텍스트의 경우, 용어의 유형을 변경해도 궁극적으로 최종 추출 목록에서 지정될 유형을 변경하지 않는 몇 가지 사례가 있습니다. 이것은 일부 기본 용어의 경우 추출 중에 우선권을 갖는 내부 사전 때문입니다.

새 유형을 작성하려면 다음을 수행하십시오.

1. 추출 결과 분할창 , 데이터 분할창, 범주 정의 대화 상자 또는 클러스터 정의 대화 상자 중 하나에서, 새 유형을 작성하려는 개념을 선택하십시오.
2. 메뉴에서 **편집 > 유형에 추가 > 새로 만들기**를 선택하십시오. 유형 특성 대화 상자가 열립니다.
3. 이름 텍스트 상자에 이 유형에 대한 새 이름을 입력하고 기타 필드를 필요에 따라 변경하십시오. 자세한 정보는 199 페이지의 『유형 작성』의 내용을 참조하십시오.
4. **확인**을 클릭하여 변경사항을 적용하십시오. 대화 상자가 닫히고 추출 결과 분할창 배경 색상이 변경되어 변경사항을 보려면 다시 추출해야 함을 표시합니다. 변경이 여러 개인 경우 재추출 전에 변경하십시오.

추출에서 개념 제외

결과를 검토할 때 가끔 임의의 자동화된 범주 작성 기법에 의해 추출 또는 사용되길 원하지 않은 개념을 찾을 수 있습니다. 어떤 경우에는 이들 개념이 아주 높은 빈도수를 갖고 있으며 사용자 분석에 전혀 의미가 없습니다. 이 경우에는 최종 추출에서 제외되도록 개념을 표시할 수 있습니다. 일반적으로 이 목록에 추가하는 개념은 연속성을 위해 텍스트에서 사용되지만 어떤 중요한 것을 추가하지 않으며 추출 결과를 어수선하게 만들 수 있는 채우기 단어나 구입니다. 개념을 제외 사전에 추가하면 해당 개념이 추출되지 않도록 보장할 수 있습니다.

개념을 제외시키면 제외된 개념의 모든 변종이 다음에 추출하는 추출 결과에서 사라집니다. 이 개념이 이미 범주에 디스크립터로 나타나는 경우 재추출 후 0의 개수를 갖고 범주에 남아 있습니다.

제외시킬 때 이들 변경은 자원 편집기의 제외 사전에 기록됩니다. 모든 제외 정의를 보고 직접 편집하려는 경우, 자원 편집기에서 직접 작업할 것을 선호할 수 있습니다. 자세한 정보는 208 페이지의 『제외 사전』의 내용을 참조하십시오.

참고: 일본어 텍스트를 사용할 때 용어나 유형을 제외하면 용어를 제외하는 결과가 아닌 일부 인스턴스가 있습니다. 이것은 일본어 자원에 대한 일부 기본 용어에 대해 추출 중에 우선권을 갖는 내부 사전 때문입니다.

개념을 제외하려면 다음을 수행하십시오.

1. 추출 결과 분할창 , 데이터 분할창, 범주 정의 대화 상자 또는 클러스터 정의 대화 상자 중 하나에서, 추출에서 제외하려는 개념을 선택하십시오.
2. 마우스 오른쪽 단추를 클릭하여 컨텍스트 메뉴를 여십시오.
3. **추출에서 제외**를 선택하십시오. 개념이 자원 편집기의 제외 사전에 추가되며 추출 결과 분할창 배경 색상이 변경되어 변경사항을 보려면 재추출해야 함을 표시합니다. 변경이 여러 개인 경우 재추출 전에 변경하십시오.

참고: 제외하는 모든 단어가 자동으로 자원 편집기 기본적으로 이것은 *로컬 라이브러리*입니다.

단어 강제 추출

추출 후 데이터 분할창에서 텍스트 데이터를 검토할 때, 일부 단어나 문구가 추출되지 않았음을 발견할 수 있습니다. 보통 이들 단어는 사용자가 관심이 없는 동사나 형용사입니다. 그러나 가끔은 범주 정의의 일부로서 추출되지 않은 단어나 구를 사용하기 원합니다.

이들 단어와 구가 추출되기 원하는 경우 용어를 강제로 유형 라이브러리로 넣을 수 있습니다. 자세한 정보는 203 페이지의 『용어 강제 실행』의 내용을 참조하십시오.

중요! 사전에서 용어를 강제 실행으로 표시하는 것은 간단하지 않습니다. 이것은 용어를 사전에 명시적으로 추가했음에도 불구하고 재추출한 후 추출 결과 분할창에 존재하지 않거나 나타나지만 사용자가 선언한 것처럼 정확하게 나타나지 않을 수 있음을 의미합니다. 이런 현상이 드물긴 하지만, 단어나 구가 이미 더 긴 구의 일부로 추출된 경우에 발생할 수 있습니다. 이를 방지하기 위해서 유형 사전에서 이 용어에 **전체(복합 아님)** 매치 옵션을 적용하십시오. 자세한 정보는 200 페이지의 『용어 추가』의 내용을 참조하십시오.

제 10 장 텍스트 데이터 범주화

범주 및 개념 보기 에서, 텍스트에서 표현되는 핵심 아이디어, 지식 및 태도를 캡처하는 본질적으로 상위 레벨 개념 또는 주제를 나타내는 **범주**를 작성할 수 있습니다.

IBM SPSS Modeler Text Analytics 14 릴리스 현재, 범주는 계층 구조를 가질 수 있는데, 하위 범주를 포함할 수 있고 하위 범주도 그 자신의 하위 범주를 가질 수 있음을 의미합니다. 이전에는 코드 프레임이라고 불렀고 계층 구조 범주를 갖는 사전 정의된 범주 구조를 가져오고 이들 계층 구조 범주를 제품 안에서 작성할 수 있습니다.

사실상, 계층 구조 범주는 사용자가 하나 이상의 하위 범주를 갖는 트리 구조를 작성하여 여러 가지 개념이나 주제 영역 같은 항목을 더 정확하게 그룹화할 수 있게 합니다. 단순한 예를 레저 활동과 관련시킬 수 있습니다. *시간이 더 있다면 어떤 활동을 하시겠습니까?* 같은 질문에 응답함으로써 *스포츠*, *예술 및 공예*, *낚시* 등과 같은 최상위 범주를 갖고, *스포츠* 아래에 *구기 종목*, *물 관련* 등인지를 보기 위한 하위 범주를 가질 수 있습니다.

범주는 *개념*, *유형*, *패턴*, *범주 규칙* 같은 디스크립터 세트로 구성됩니다. 이들 디스크립터는 함께 사용되어 문서 또는 레코드가 주어진 범주에 속하는지 여부를 식별합니다. 문서 또는 레코드 내의 텍스트를 스캔하여 임의의 텍스트가 디스크립터와 매치하는지 확인할 수 있습니다. 매치가 발견되면 문서/ 레코드가 해당 범주에 지정됩니다. 이 프로세스를 **범주화**라고 부릅니다.

범주 및 개념 보기의 4개의 분할창에 제공되는 데이터를 사용하여 범주를 작업, 작성 및 시각적으로 탐색할 수 있는데, 각 분할창은 보기 메뉴에서 이름을 선택하여 숨기거나 표시할 수 있습니다.

- **범주 분할창.** 이 분할창에서 범주를 작성 및 관리합니다. 자세한 정보는 107 페이지의 『범주 분할창』의 내용을 참조하십시오.
- **추출 결과 분할창.** 이 분할창에서 추출된 개념 및 유형을 탐색하고 그에 대해 작업합니다. 자세한 정보는 91 페이지의 『추출 결과: 개념 및 유형』 주제를 참조하십시오.
- **시각화 분할창.** 이 분할창에서 범주 및 범주가 상호작용하는 방법을 시각적으로 탐색합니다. 자세한 정보는 165 페이지의 『범주 그래프 및 도표』 주제를 참조하십시오.
- **데이터 분할창.** 이 분할창에서 선택에 대응하는 문서 및 레코드 안에 있는 텍스트를 탐색하고 검토합니다. 자세한 정보는 114 페이지의 『데이터 분할창』 주제를 참조하십시오.

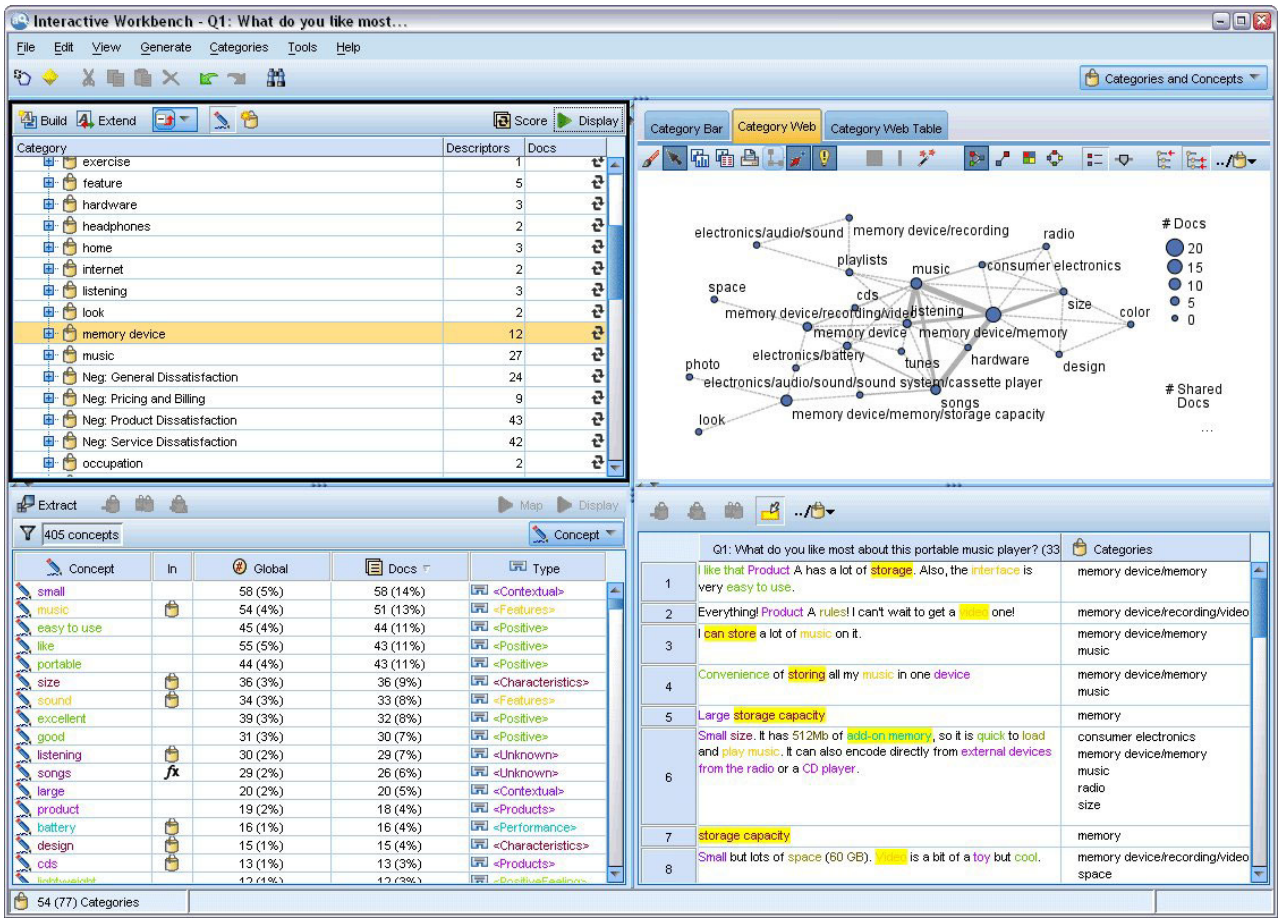


그림 29. 범주 및 개념 보기

텍스트 분석 패키지(TAP)의 범주 세트로 시작하거나 사전 정의된 범주 파일을 가져올 수 있지만, 사용자 스스로 범주를 작성해야 할 수도 있습니다. 범주는 제품의 강력한 자동화 기법 세트를 사용하여 자동으로 작성될 수 있는데, 이것은 추출 결과(개념, 유형 및 패턴)를 사용하여 범주 및 해당 디스크립터를 생성합니다. 범주는 또한 사용자가 데이터에 관하여 가질 수 있는 추가 직관을 사용하여 수동으로 작성할 수도 있습니다. 그러나 대화형 워크벤치를 통해서만 범주를 수동으로 작성하거나 세분화할 수 있습니다. 자세한 정보는 24 페이지의 『텍스트 마이닝 노드: 모델 탭』 주제를 참조하십시오. 추출 결과를 범주로 끌어다 놓아서 수동으로 범주 정의를 작성할 수 있습니다. 범주 규칙을 범주에 추가하거나 사용자 자신의 사전 정의된 범주를 사용하거나, 조합하여 이들 범주 또는 빈 범주를 강화할 수 있습니다.

각 기법과 방법은 특정 유형의 데이터 및 상황에 잘 맞지만, 보통 동일한 분석에서 기법을 조합하여 문서 또는 레코드의 전체 범위를 캡처하는 것이 도움이 됩니다. 또한 범주화 과정에서 언어학적 자원에 수행되는 기타 변경을 볼 수 있습니다.

범주 분할창

범주 분할창은 범주를 작성하고 관리할 수 있는 영역입니다. 이 분할창은 범주 및 개념 보기의 왼쪽 상단 구석에 위치합니다. 텍스트 데이터에서 개념 및 유형을 추출할 후, 개념 포함, 동시 발생 등과 같은 기법을 사용하여 자동으로 또는 수동으로 범주 작성을 시작할 수 있습니다. 자세한 정보는 116 페이지의 『범주 작성』의 내용을 참조하십시오.

범주가 작성 또는 업데이트될 때마다, 문서 또는 레코드는 **스코어** 단추를 클릭하여 스코어링하여 임의의 텍스트가 주어진 범주의 디스크립터와 매치하는지 여부를 확인할 수 있습니다. 매치가 발견되면 문서 또는 레코드가 해당 범주에 지정됩니다. 최종 결과는 전부는 아니더라도 대부분의 문서 또는 레코드가 범주의 디스크립터를 바탕으로 범주에 지정되는 것입니다.

범주 트리 테이블

이 분할창의 트리 테이블은 범주, 하위 범주 및 디스크립터의 세트를 나타냅니다. 트리는 또한 각 트리 항목의 정보를 제공하는 여러 개의 열을 갖고 있습니다. 표시할 수 있는 열은 다음과 같습니다.

- **코드.** 각 범주의 코드 값을 나열합니다. 이 열은 기본적으로 숨겨져 있습니다. **보기 > 범주 분할창** 메뉴를 사용하여 이 열을 표시할 수 있습니다.
- **범주.** 범주 및 하위 범주의 이름을 표시하는 포함 트리를 포함합니다. 또한 디스크립터 도구 모음이 클릭되는 경우 디스크립터 세트도 표시됩니다.
- **디스크립터.** 정의를 구성하는 디스크립터의 수를 제공합니다. 이 숫자는 하위 범주에 있는 디스크립터 수는 포함하지 않습니다. 디스크립터 이름이 **범주** 열에 표시되면 개수는 제공되지 않습니다. **보기 > 범주 분할창 > 모든 디스크립터** 메뉴를 통해 트리에서 디스크립터 자체를 표시하거나 숨길 수 있습니다.
- **문서.** 스코어링 후, 이 열은 범주 및 해당 범주의 모든 하위 범주로 범주화되는 문서 또는 레코드의 수를 제공합니다. 따라서 5개 레코드가 디스크립터를 바탕으로 최상위 범주와 매치하고 7개의 다른 레코드가 디스크립터를 바탕으로 하위 범주에 매치하는 경우, 최상위 범주에 대한 총 문서 수는 둘의 합이며, 이 경우에는 12입니다. 그러나 동일한 레코드가 최상위 범주 및 그의 하위 범주와 매치한 경우 개수는 11입니다.

범주가 없을 때 테이블은 여전히 두 개의 행을 포함합니다. **모든 문서** 라고 부르는 최상위 행은 문서 또는 레코드의 총 수입니다. **범주화 안됨**이라는 두 번째 열은 아직 범주화되지 않은 문서/레코드의 수를 표시합니다.

분할창의 각 범주에 대해 작은 노란색 버킷 아이콘이 범주 이름 앞에 표시됩니다. 범주를 두 번 클릭하거나 , 메뉴에서 **보기 > 범주 정의**를 클릭하는 경우, 범주 정의 대화 상자가 열리고 개념, 유형, 패턴 및 범주 규칙 같이 정의를 구성하는 **디스크립터**라는 모든 요소를 표시합니다. 자세한 정보는 113 페이지의 『범주 정보』의 내용을 참조하십시오. 기본적으로 범주 트리 테이블은 범주의 디스크립터를 표시하지 않습니다. 범주 정의 대화 상자에서가 아니라 트리에서 직접 디스크립터를 보려는 경우, 도구 모음의 연필 아이콘을 갖는 전환 단추를 클릭하십시오. 이 전환 단추가 선택되면 트리를 펼쳐서 디스크립터도 볼 수 있습니다.

범주 스코어링

범주 트리 테이블의 **문서** 열은 해당 특정 범주로 범주화되는 문서 또는 레코드의 수를 표시합니다. 숫자가 오래 되었거나 계산되지 않은 경우 해당 열에 아이콘이 나타납니다. 분할창 도구 모음의 **스코어**를 클릭하여 문서 수를 다시 계산할 수 있습니다. 더 큰 데이터 세트에 대해 작업 중일 때는 스코어링 프로세스가 다소 시간이 걸릴 수 있음을 기억하십시오.

트리에서 범주 선택

트리에서 선택할 때, 동위 범주만 선택할 수 있습니다. 즉, 최상위 범주를 선택하는 경우 하위 범주도 선택할 수는 없습니다. 또는 주어진 범주의 2 하위 범주를 선택하는 경우, 또 다른 범주의 하위 범주를 동시에 선택할 수 없습니다. 불연속적인 범주를 선택하면 이전 선택이 유실됩니다.

데이터 및 시각화 분할창에 표시

테이블에서 행을 선택할 때, **표시** 단추를 클릭하여 사용자 선택에 대응하는 정보로 시각화 및 데이터 분할창을 새로 고칠 수 있습니다. 분할창이 표시될 수 없는 경우 **표시**를 클릭하면 해당 분할창이 나타납니다.

범주 세분화

범주화가 첫 번째 시도에서 사용자 데이터에 대한 완벽한 결과를 생성하지 않을 수 있으며, 삭제하거나 다른 범주와 결합하기 원하는 범주도 있을 수 있습니다. 또한 추출 결과의 검토를 통해서 유용하다고 생각하는 몇 가지 범주가 작성되지 않았음을 발견할 수도 있습니다. 그런 경우, 결과를 수동으로 변경하여 특정 컨텍스트에 맞게 세분화할 수 있습니다. 자세한 정보는 150 페이지의 『범주 편집 및 세분화』 주제를 참조하십시오.

범주 작성을 위한 방법 및 전략

아직 추출하지 않았거나 추출 결과가 오래된 경우에는 범주 작성 또는 확장 기술 중 하나를 사용하면 추출하라는 메시지가 자동으로 프롬프트됩니다. 기술을 적용한 후에도 범주로 그룹화된 개념과 유형은 여전히 다른 기술로 범주 작성 시 사용할 수 있습니다. 즉, 이를 재사용하지 않도록 선택하지 않는 한 여러 범주에서 개념을 볼 수 있음을 의미합니다.

최상의 범주를 작성하기 위해서는 다음을 검토하십시오.

- **범주 작성 방법**
- **범주 작성 전략**
- **범주 작성 팁**

범주 작성 방법

모든 데이터 세트가 고유하므로 범주 작성 방법의 수와 이를 적용하는 순서는 시간에 따라 변경될 수 있습니다. 또한 텍스트 마이닝 목적은 한 데이터 세트와 다음 데이터 세트마다 다르므로, 여러 가지 방법을 시도하여 어떤 방법이 지정된 텍스트 데이터별로 최상의 결과를 낼 수 있는지를 살펴볼 필요가 있습니다. 자동 기술은 데이터를 완벽하게 범주화합니다. 그러므로 사용자의 데이터에 가장 적합한 하나 이상의 자동 기술을 찾아서 적용하는 것이 좋습니다.

범주 세트가 미리 작성된 텍스트 분석 패키지(TAP, *.*tap*)를 사용하는 것 외에도 다음 방법의 조합을 사용하여 반응을 범주화할 수도 있습니다.

- **자동 작성 기술.** 범주를 자동으로 작성하기 위해 몇몇 언어학적 기반 및 빈도 기반 범주 옵션을 사용할 수 있습니다. 자세한 정보는 116 페이지의 『범주 작성』의 내용을 참조하십시오.
- **자동 확장 기술.** 더 많은 레코드를 캡처할 수 있도록 디스크립터를 추가하거나 개선하여 기존 범주를 확장하는 데 여러 언어학적 기술을 사용할 수 있습니다. 자세한 정보는 127 페이지의 『범주 확장』의 내용을 참조하십시오.
- **수동 기술.** 끌어다 놓기 등과 같은 여러 수동 방법이 있습니다. 자세한 정보는 130 페이지의 『수동으로 범주 작성』의 내용을 참조하십시오.

범주 작성 전략

다음 전략 목록은 결코 완전하지는 않지만 범주 작성 방법에 대한 몇 가지 아이디어를 제공할 수 있습니다.

- 텍스트 마이닝 모드를 정의할 때, 몇몇 사전 정의된 범주의 분석을 시작할 수 있도록 텍스트 분석 패키지(TAP)에서 범주 세트를 선택하십시오. 이러한 범주는 텍스트를 처음부터 충분히 범주화할 수 있습니다. 그러나 더 많은 범주를 추가하려는 경우에는 범주 작성 설정(**범주 > 작성 설정**)을 편집할 수 있습니다. **고급 설정: 언어학** 대화 상자를 열고 범주 입력 옵션 **사용되지 않은 추출 결과**를 선택하고 추가 범주를 작성하십시오.
- 노드를 정의할 때, 대화식 워크벤치의 범주 및 개념 보기에 있는 TAP에서 범주 세트를 선택하십시오. 그런 다음 사용되지 않은 개념 또는 패턴을 적합하다고 생각되는 범주에 끌어다 놓으십시오. 그런 다음 방금 편집한 기존 범주(**범주 > 확장된 범주**)를 확장하여 기존 범주 디스크립터와 관련된 더 많은 디스크립터를 획득하십시오.
- 고급 언어학적 설정(**범주 > 범주 작성**)을 사용하여 자동으로 범주를 작성하십시오. 그런 다음 디스크립터를 삭제하고, 범주를 삭제하거나 결과로 나온 범주에 만족할 때까지 유사한 범주를 병합하여 범주를 수동으로 세분화하십시오. 또한 원래 **가능한 경우 와일드카드 일반화** 옵션을 사용하지 **않고** 범주를 작성한 경우에는 **일반화** 옵션을 사용하여 범주 확장을 사용하여 범주를 자동으로 단순화하려고 시도할 수도 있습니다.
- 설명적인 범주 이름 및/또는 주석을 사용하여 사전에 정의된 범주 파일을 가져오십시오. 또한 원래 가져올 옵션을 선택하거나 범주 이름에서 디스크립터를 생성하지 **않고** 가져온 경우에는 나중에 범주 확장 대화 상자를 사용하고 **범주 이름에서 생성된 디스크립터를 사용하여 빈 범주 확장** 옵션을 선택할 수 있습니다. 그런 다음 이러한 범주를 두 번째로 확장하지만 이번에는 그룹화 기술을 사용하십시오.
- 개념 또는 개념 패턴을 빈도 기준으로 정렬한 다음 가장 흥미로운 범주를 범주 분할창에 끌어다 놓는 방법으로 첫 번째 범주 세트를 수동으로 작성하십시오. 초기 범주 세트가 생긴 후에는 확장 기능(**범주 > 범주 확장**)을 사용하여 다른 관련된 디스크립터를 포함하고 더 많은 레코드와 매치할 수 있도록 선택한 모든 범주를 확장하고 세분화하십시오.

이러한 기술을 적용한 후에는 결과로 나온 범주를 검토하고 수동 기술을 사용하여 약간의 조정을 수행하고, 잘못된 분류를 제거하거나 누락된 레코드나 단어를 추가하는 것이 좋습니다. 또는 서로 다른 기술을 사용하면 중복된 범주가 생길 수 있으므로 필요에 따라 범주를 병합하거나 삭제할 수도 있습니다. 자세한 정보는 150 페이지의 『범주 편집 및 세분화』 주제를 참조하십시오.

범주 작성을 위한 팁

더 좋은 범주를 작성하기 위해서 접근 방식에서 결정하도록 도와줄 수 있는 몇 가지 팁을 검토할 수 있습니다.

범주 대 문서 비율에 대한 팁

문서 및 레코드가 지정되는 범주는 보통 다음 두 가지 이상의 이유로 정성적 텍스트 분석에서 상호 배타적이지 않습니다.

- 첫 번째, 일반적인 방법은 텍스트 문서 또는 레코드가 길수록, 표현되는 아이디어와 의견이 더 명확하다는 것입니다. 따라서 문서 또는 레코드가 다중 범주에 지정될 수 있는 기회가 크게 늘어납니다.
- 두 번째, 종종 논리적으로 구별되지 않는 텍스트 문서 또는 레코드를 그룹화하고 해석하는 다양한 방법이 있습니다. 반응자의 정치적 신념에 관한 개방형 질문을 갖는 설문조사의 경우, *진보*와 *보수* 또는 *공화당원*과 *민주당원* 같은 범주뿐 아니라 *사회적 진보*, *재정적 보수* 등과 같은 보다 특정한 범주를 작성할 수 있습니다. 이들 범주는 상호 배타적이고 철저할 필요가 없습니다.

작성할 범주의 수에 대한 팁

범주 작성은 데이터에서 직접 진행되어야 합니다. 데이터에 관해서 관심있는 어떤 것을 볼 때 해당 정보를 나타내기 위한 범주를 작성할 수 있습니다. 일반적으로 작성하는 범주 수에 대한 권장 상한은 없습니다. 그러나 너무 많은 범주를 작성하면 확실히 관리하기가 어려울 수 있습니다. 다음 두 가지 원리가 적용됩니다.

- **범주 빈도.** 범주가 유용하려면 최소 숫자의 문서 또는 레코드를 포함해야 합니다. 한두 개의 문서 가 아주 흥미로운 어떤 것을 포함할 수 있지만, 1,000개의 문서 중 한두 개가 있는 경우 거기에 포함된 정보는 실질적으로 유용하기 위해 인구에서 충분히 빈번하지 않을 수 있습니다.
- **복잡도.** 더 많은 범주를 작성할수록 분석을 완료한 후 더 많은 정보를 검토하고 요약해야 합니다. 그러나 너무 많은 범주는 복잡도를 추가하면서도 유용한 세부사항을 추가하지 않을 수 있습니다.

불행하게도, 얼마나 많은 범주가 너무 많은지 판별하거나 범주당 최소 레코드 수를 판별하기 위한 규칙은 없습니다. 사용자의 특정 상황의 수요를 바탕으로 그런 판단을 내려야 합니다.

하지만 시작할 위치에 대해 조언할 수 있습니다. 범주 수가 과도하지 않아야 하지만, 분석의 초기 단계에서는 너무 적은 범주를 갖기 보다는 너무 많은 범주를 갖는 것이 더 좋습니다. 사례를 새로운 범주로 나누기 보다는 상대적으로 비슷한 범주를 그룹화하는 것이 더 쉬우므로, 많은 범주에서 더 적은 범주로 작업하는 전략이 대개 최상의 방법입니다. 텍스트 마이닝의 반복적 본질과 이 소프트웨어 프로그램으로 수행할 수 있는 용이성이 주어지면, 더 많은 범주를 작성하는 것이 시작 시에 유용한 방법입니다.

최상의 디스크립터 선택

다음 정보에는 범주에 대한 최상의 디스크립터(개념, 유형, TLA 패턴, 범주 규칙) 선택 또는 작성을 위한 몇 가지 지침이 들어 있습니다. 디스크립터는 범주의 구성 요소입니다. 문서 또는 레코드에 있는 텍스트의 일부 또는 전부가 디스크립터와 매치할 때, 문서 또는 레코드가 범주와 매치합니다.

디스크립터가 추출된 개념이나 패턴을 포함하거나 대응하지 않는 한, 어떤 문서 또는 레코드에도 매치하지 않습니다. 그러므로 다음 단락에서 설명하는 대로 개념, 유형, 패턴 및 범주 규칙을 사용하십시오.

개념은 그 자체뿐 아니라 복수형/단수형부터 동의어, 철자법 변형까지의 범위를 가질 수 있는 기본 용어 세트를 나타내므로, 개념 자체만 디스크립터 또는 디스크립터의 일부로 사용되어야 합니다. 임의의 주어진 개념에 대한 기본 용어에 대해 자세히 알려면 범주 및 개념 보기의 추출 결과 분할창에서 개념 이름을 클릭하십시오. 개념 이름 위에 마우스를 움직일 때 도구팁이 나타나고 마지막 추출 중에 텍스트에서 발견된 기본 용어 중 하나를 표시합니다. 모든 개념이 기본 용어를 갖지는 않습니다. 예를 들어, 자동차와 차량은 동의어이지만 자동차는 차량을 기본 용어로 갖는 개념으로 추출된 경우, 차량을 갖는 문서 또는 레코드와 자동으로 매치하므로 디스크립터에서 자동차만 사용하기 원합니다.

디스크립터로서의 개념 및 유형

해당 개념(또는 그의 기본 용어 중 하나)을 포함하는 모든 문서 또는 레코드를 찾기 원할 때 개념을 디스크립터로 사용하십시오. 이 경우에 정확한 개념 이름이 충분하므로 더 복잡한 범주 규칙은 필요 없습니다. 의견을 추출하는 자원을 사용할 때 가끔 문장의 더 진실한 의미를 캡처하기 위해 TLA 패턴 추출 중에 개념이 변할 수 있음을 기억하십시오(TLA에 대한 다음 절의 예를 참조).

예를 들어, "*사과와 파인애플이 최고*" 같이 각 개인의 좋아하는 과일을 표시하는 설문조사 응답은 사과 및 파인애플의 추출을 가져옵니다. 사과 개념을 디스크립터로서 범주에 추가하여 사과(또는 그의 모든 기본 용어) 개념을 포함하는 모든 응답이 해당 범주에 매치됩니다.

그러나, 단순히 어떤 방법으로든지 *사과*를 언급하는 응답을 아는 것에 관심을 갖는 경우, * 사과 * 같은 범주 규칙을 작성할 수 있으며 사과, 사과 주스 또는 프랑스 사과 타르트 같은 개념을 포함하는 응답을 캡처합니다.

또한 <과일> 같이 유형을 디스크립터로서 직접 사용하여 동일한 방법으로 입력된 개념을 포함하는 모든 문서 또는 레코드를 캡처할 수도 있습니다. 유형에서는 *를 사용할 수 없음을 주의하십시오.

자세한 정보는 91 페이지의 『추출 결과: 개념 및 유형』 주제를 참조하십시오.

디스크립터로서의 텍스트 링크 분석(TLA) 패턴

더 미묘한 뉘앙스의 아이디어를 캡처하기 원할 때는 TLA 패턴 결과를 디스크립터로 사용하십시오. 텍스트가 TLA 추출 중에 분석될 때 텍스트는 전체 텍스트(문서 또는 레코드)를 보기 보다는 한 번에 하나의 문구나 절이 처리됩니다. 단일 문구의 모든 부분을 함께 고려함으로써, TLA는 의견, 두 요소 사이의 관계 또는 반대를 식별하고 예를 들어 더 진실한 의미를 이해할 수 있습니다. 개념 패턴이나 유형 패턴을 디스크립터로 사용할 수 있습니다. 자세한 정보는 161 페이지의 『유형 및 개념 패턴』 주제를 참조하십시오.

예를 들어, "*방이 그렇게 정리되지 않았습니다*"라는 텍스트가 있는 경우 방 및 정리 개념이 추출될 수 있습니다. 그러나 TLA 추출이 추출 설정에서 사용으로 설정된 경우, TLA는 정리가 부정적인 방식으로 사용되었고 실제로는 더러움의 동의어인 정리되지 않음에 해당함을 발견할 수 있습니다. 여기에서 그 자신에서 정리 개념을 디스크립터로 사용하는 것은 이 텍스트와 매치하지만 청결을 언급하는 다른 문서 또는 레코드도 캡처함을 알 수 있습니다. 그러므로 더러움을 갖는 TLA 개념 패턴을 출력 개념으로 사용하는 것이 더 좋을 수 있습니다. 이 개념은 이 텍스트와 매치하고 더 적합한 디스크립터일 수 있습니다.

디스크립터로서의 범주 비즈니스 규칙

범주 규칙은 문서 또는 레코드를 추출된 개념, 유형 및 패턴뿐만 아니라 부울 연산자를 사용하여 논리적 표현식을 기반으로 범주에 자동으로 분류하는 명령문입니다. 예를 들어, *추출된 개념* embassy*을 포함하지만* argentina*는 포함하지 않는 모든 레코드를 이 범주에 포함*을 의미하는 표현식을 작성할 수 있습니다.

범주에서 범주 규칙을 디스크립터로서 작성하고 사용하여 &, |, !() 부울을 사용하여 여러 가지 아이디어를 표현할 수 있습니다. 이들 규칙의 구문 및 규칙을 작성 및 편집하는 방법에 대한 상세한 정보는 132 페이지의 『범주 규칙 사용』을 참조하십시오.

- 2개 이상의 개념이 발생하는 문서 또는 레코드를 찾는 데 도움을 얻으려면 &(AND) 부울 연산자를 갖는 범주 규칙을 사용하십시오. & 연산자에 의해 연결되는 둘 이상의 개념이 동일한 문구나 구에서 발생할 필요는 없으며 범주 매치로 간주될 동일한 문서 또는 레코드의 어디에서나 발생할 수 있습니다. 예를 들어, 범주 규칙 food & cheap를 디스크립터로서 작성하는 경우, "*the food was pretty expensive, but the rooms were cheap*" 텍스트를 포함하는 레코드와 매치합니다. food가 cheap가 꾸미는 명사가 아님에도 불구하고 텍스트가 food와 cheap를 둘 다 포함하기 때문입니다.
- 일부가 발생하지만 다른 것은 발생하지 않는 문서 또는 레코드를 찾는 데 도움을 받으려면 !()(NOT) 부울 연산자를 갖는 범주 규칙을 디스크립터로 사용하십시오. 이것은 단어를 바탕으로 하면 관련된 것처럼 보이지만 컨텍스트를 바탕으로 하면 관련되지 않을 수 있는 정보 그룹화를 피하는 데 도움이 될 수 있습니다. 예를 들어, 범주 규칙 <Organization> & !(ibm)을 디스크립터로 작성하는 경우 *SPSS Inc. was a company founded in 1967* 텍스트와 매치하고 *the software company was acquired by IBM.* 텍스트와는 매치하지 않습니다.
- 여러 가지 개념이나 유형 중 하나를 포함하는 문서 또는 레코드 중 하나를 찾으려면 |(OR) 부울 연산자를 디스크립터로 갖는 범주 규칙을 사용하십시오. 예를 들어 범주 규칙 (personnel|staff|team|coworkers) & bad를 디스크립터로 작성하는 경우, bad 개념을 갖는 명사 중 하나가 발견되는 모든 문서 또는 레코드와 매치합니다.
- 규칙을 더 일반적이고 가능하면 더 배치 가능하게 만들려면 범주 규칙에서 유형을 사용하십시오. 예를 들어, 호텔 데이터에 대해 작업 중인 경우 고객이 호텔 직원에 대해 생각하는 바를 배우는 데 매우 관심이 있을 수 있습니다. 관련 용어는 접수 담당자, 웨이터, 웨이트리스, 접수 데스크, 프론트 데스크 등의 단어를 포함할 수 있습니다. 이 경우에 <HotelStaff>이라는 새 유형을 작성하고 해당 유형에 앞의 모든 용어를 추가할 수 있습니다. [* waitress * & nice], [* desk * & friendly], [* receptionist * & accommodating] 같은 모든 종류의 직원에 대한 하나의 범주 규칙을 작성할 수 있지만, <HotelStaff> 유형을 사용하여 더 일반적인 하나의 범주 규칙을 작성하여 [<HotelStaff> & <Positive>]의 양식으로 호텔 직원의 호의적인 의견을 갖는 모든 응답을 캡처할 수 있습니다.

참고: 규칙에 TLA 패턴을 포함할 때 범주 규칙에서 + 및 &를 둘 다 사용할 수 있습니다. 자세한 정보는 134 페이지의 『범주 규칙에서 TLA 패턴 사용』의 내용을 참조하십시오.

디스크립터로서의 개념, TLA 또는 범주 규칙이 상이하게 매치하는 방법의 예

다음 예는 개념을 디스크립터로서, 범주 규칙을 디스크립터로서 또는 TLA 패턴을 디스크립터로서 사용하는 것이 문서 또는 레코드가 범주화되는 방법에 어떤 영향을 주는지를 보여줍니다. 다음 5개 레코드가 있다고 가정합니다.

- A: *"굉장한 식당 직원, 탁월한 음식 및 편안하고 깨끗한 객실."*
- B: *"식당 직원은 끔찍했지만 객실은 깨끗했음."*
- C: *"안락하고 깨끗한 객실."*
- D: *"내 방은 그렇게 깨끗하지 않았음."*
- E: *"깨끗함."*

레코드가 *깨끗*이라는 단어를 포함하고 이 정보를 캡처하기 원하므로, 다음 테이블에 표시된 디스크립터 중 하나를 작성할 수 있습니다. 캡처하려는 본질을 바탕으로, 다른 디스크립터에 비해 한 종류의 디스크립터 사용이 상이한 결과를 생성할 수 있는 방법을 볼 수 있습니다.

표 17. 예제 레코드가 디스크립터와 매치한 방법.

디스크립터	A	B	C	D	E	설명
깨끗	*매치*	*매치*	*매치*	*매치*	*매치*	디스크립터가 추출된 개념입니다. 모든 레코드가 깨끗 개념을 포함했으며, TLA가 없으면 자동으로 "*깨끗하지 않음*"이 TLA 규칙에 의해 더러움을 의미한다고 알려지지 않았으므로 레코드 D도 포함했습니다.
깨끗 + .	-	-	-	-	*매치*	디스크립터는 그 자체가 깨끗을 나타내는 TLA 패턴입니다. TLA 추출 중에 연관된 개념 없이 깨끗이 추출된 레코드와만 매치합니다.
[깨끗]	*매치*	*매치*	*매치*	-	*매치*	디스크립터는 그 자체에서 또는 다른 어떤 것에서 깨끗을 포함하는 TLA 규칙을 찾는 범주 규칙입니다. 깨끗이 객실 같은 다른 개념에 링크되고 임의의 슬롯 위치에 있는지 여부와 상관없이 깨끗을 포함하는 TLA 출력이 발견된 모든 레코드와 매치했습니다.

범주 정보

범주는 서로 밀접하게 관련된 개념, 의견 또는 속성의 그룹을 가리킵니다. 범주는 중요 의미를 캡처하는 짧은 구문이나 레이블로 쉽게 설명되어야 유용하게 사용할 수 있습니다.

예를 들어, 새 세탁 세제에 대한 고객의 설문 반응을 분석 중이라면 제품의 향을 설명하는 모든 반응을 포함하는 *냄새*라는 레이블이 있는 범주를 작성할 수 있습니다. 그러나 이러한 범주는 향을 좋아하는 소비자와 이를 싫어하는 소비자를 분간하지는 못합니다. IBM SPSS Modeler Text Analytics 는 적합한 자원을 사용하면 의견을 추출할 수 있으므로 *냄새를 좋아한* 반응자와 *냄새를 싫어한* 반응자를 식별하기 위한 두 개의 다른 범주를 작성할 수 있습니다.

범주 및 개념 보기 창의 왼쪽 상단 분할창에 있는 범주 분할창에서 범주를 작성하고 이에 대해 작업할 수 있습니다. 각 범주는 하나 이상의 디스크립터로 정의됩니다. **디스크립터**는 개념, 유형 및 패턴뿐만 아니라 범주를 정의하는 데 사용된 범주 규칙입니다.

지정된 범주를 구성하는 디스크립터를 보고 싶은 경우에는 범주 분할창 도구 모음에서 연필 아이콘을 클릭한 다음 트리를 확장하여 디스크립터를 볼 수 있습니다. 또는 범주를 선택하고 범주 정의 대화 상자를 여십시오 (**보기 > 범주 정의**).

개념 포함 등과 같은 범주 작성 기술을 사용하여 범주를 자동으로 작성하면 기술은 개념 및 유형을 디스크립터로 사용하여 범주를 작성합니다. TLA 패턴을 추출하면 패턴 또는 이러한 패턴의 일부를 범주 디스크립터로 추가할 수 있습니다. 자세한 정보는 159 페이지의 제 12 장 『텍스트 링크 분석 탐색』의 내용을 참조하십시오. 군집을 작성하는 경우에는 군집에서 신규 또는 기존 범주에 개념을 추가할 수 있습니다. 마지막으로 범주에서 디스크립터로 사용하기 위한 범주 규칙을 수동으로 작성할 수 있습니다. 자세한 정보는 132 페이지의 『범주 규칙 사용』의 내용을 참조하십시오.

범주 특성

디스크립터에 추가로, 범주에는 또한 범주의 이름을 변경하고, 레이블을 추가하거나 주석을 추가하기 위해 편집할 수 있는 특성이 있습니다.

다음 특성이 있습니다.

- **이름.** 이 이름은 기본적으로 트리에 나타납니다. 자동화된 기술을 사용하여 범주가 작성되면 이는 자동으로 이름이 제공됩니다.
- **레이블.** 레이블 사용은 다른 제품에서나 다른 테이블 또는 그래프에서 사용하기 위해 보다 의미있는 범주 설명을 작성할 때 유용합니다. 레이블을 표시하기 위한 옵션을 선택하는 경우에는 레이블이 범주를 식별하기 위해 인터페이스에 사용됩니다.
- **코드.** 코드 번호는 이 범주의 코드 값에 해당합니다. .
- **주석.** 이 필드에서 각 범주의 짧은 설명을 추가할 수 있습니다. 범주가 범주 작성 대화 상자를 사용하여 생성된 경우에는 이 주석에 노트가 자동으로 추가됩니다. 텍스트를 선택하고 메뉴에서 **범주 > 주석에 추가**를 선택하여 데이터 분할창에서 직접 주석에 표본 텍스트를 추가할 수도 있습니다.

데이터 분할창

범주를 작성한 후 작업 중인 일부 텍스트 데이터를 검토하려는 경우가 있습니다. 예를 들어, 640개 문서가 범주화된 범주를 작성하는 경우 해당 문서 중 일부 또는 모두를 살펴 실제로 기록된 텍스트를 확인할 수 있습니다. 오른쪽 하단에 있는 데이터 분할창에서 레코드 또는 문서를 검토할 수 있습니다. 기본적으로 표시되지 않으면 메뉴에서 **보기 > 분할창 > 데이터**를 선택하십시오.

데이터 분할창은 일정 표시 한계까지 범주 분할창, 추출 결과 분할창 또는 범주 정의 대화 상자의 선택사항에 해당하는 문서 또는 레코드당 1행을 제공합니다. 기본적으로 데이터 분할창에 표시된 문서 또는 레코드 수는 데이터를 보다 빨리 볼 수 있도록 제한됩니다. 그러나 옵션 대화 상자에서 이를 조정할 수 있습니다. 자세한 정보는 86 페이지의 『옵션: 세션 탭』의 내용을 참조하십시오.

데이터 분할창 표시 및 새로 고침

큰 데이터 세트의 경우 자동 데이터 새로 고침을 완료하려면 약간의 시간이 걸리기 때문에 데이터 분할창은 자동으로 표시를 새로 고치지 않습니다. 따라서 이 보기의 다른 분할창 또는 범주 정의 대화 상자에서 선택할 때마다 **표시**를 클릭하여 데이터 분할창의 컨텐츠를 새로 고치십시오.

텍스트 문서 또는 레코드

텍스트 데이터가 레코드 양식으로 되어 있고 텍스트의 길이가 비교적 짧으면, 데이터 분할창의 텍스트 필드는 텍스트 데이터를 전부 표시합니다. 그러나 레코드와 큰 데이터 세트에 대한 작업을 할 때 텍스트 필드 열은 텍스트의 짧은 조각을 표시하고 테이블에서 선택한 레코드의 텍스트를 모두 또는 더 많이 표시할 수 있도록 오른쪽에 텍스트 미리보기 분할창을 엽니다. 텍스트 데이터가 개별 문서 양식으로 되어 있으면 데이터 분할창이 문서의 파일 이름을 표시합니다. 문서를 선택하면 선택된 문서의 텍스트와 함께 텍스트 미리보기 분할창이 열립니다.

색상 및 강조표시

데이터를 표시할 때마다 텍스트에서 쉽게 식별할 수 있도록 해당 문서 또는 레코드에서 찾은 개념 및 디스크립터가 색상으로 강조표시됩니다. 색상 코딩은 개념이 속한 유형에 해당합니다. 색상 코드화된 항목 위에 마우스를 올려 놓아 추출된 개념과 지정된 유형도 표시할 수 있습니다. 추출되지 않은 텍스트는 검은색으로 나타납니다. 일반적으로 추출되지 않은 이러한 단어는 접속사(*and* 또는 *with*), 대명사(*me* 또는 *they*), 동사(*is*, *have* 또는 *take*)인 경우가 많습니다.

데이터 분할창 열

텍스트 필드 열이 항상 표시되는 동안에는 다른 열도 표시할 수 있습니다. 다른 열을 표시하려면 메뉴에서 **보기 > 데이터 분할창**을 선택한 후 데이터 분할창에 표시할 열을 선택하십시오. 표시할 수 있는 열은 다음과 같습니다.

- **"텍스트 필드 이름" (#)/문서.** 개념과 유형이 추출된 텍스트 데이터에 열을 추가합니다. 데이터가 문서에 있는 경우, 열을 문서라고 하며 문서 파일 이름 또는 전체 경로만 표시됩니다. 해당 문서에 대한 텍스트를 보려면 텍스트 미리보기 분할창에서 보아야 합니다. 데이터 분할창의 행 수는 이 열 이름 다음에 괄호로 표시됩니다. 옵션 대화 상자에서 로드 속도를 늘리는 데 사용되는 한계 때문에 모든 문서 또는 레코드가 표시되지는 않는 경우가 있습니다. 최대값에 도달하면 숫자 뒤에 **- Max**가 옵니다. 자세한 정보는 86 페이지의 『옵션: 세션 탭』의 내용을 참조하십시오.
- **범주.** 레코드가 속한 범주를 각각 나열합니다. 이 열이 표시될 때마다 데이터 분할창을 새로 고치면 최신 정보를 표시하기 위해 시간이 약간 오래 걸립니다.
- **관련성 순위.** 단일 범주의 각 레코드에 대한 순위를 제공합니다. 이 순위는 해당 범주의 다른 레코드와 비교하여 레코드가 범주에 얼마나 잘 맞는지 보여줍니다. 순위를 보려면 범주 분할창(왼쪽 상단 분할창)에서 범주를 선택하십시오. 자세한 정보는 116 페이지의 『범주 관련성』의 내용을 참조하십시오.
- **범주 수.** 레코드가 속한 범주 수를 나열합니다.

범주 관련성

더 좋은 범주를 작성하기 위해 각 범주에 있는 문서 또는 레코드의 관련성뿐 아니라 문서 또는 레코드가 속하는 모든 범주의 관련성을 검토할 수 있습니다.

레코드에 대한 범주의 관련성

문서 또는 레코드가 데이터 분할창에 나타날 때마다, 그것이 속하는 모든 범주가 범주 열에 나열됩니다. 문서 또는 레코드가 다중 범주에 속하면 이 열의 범주는 관련성이 가장 큰 것부터 가장 작은 것의 순서로 나타납니다. 처음 나열되는 범주는 이 문서 또는 레코드에 최상으로 대응하는 것으로 생각됩니다. 자세한 정보는 114 페이지의 『데이터 분할창』 주제를 참조하십시오.

범주에 대한 레코드의 관련성

범주를 선택할 때 데이터 분할창의 관련성 순위 열에서 범주의 각 레코드의 관련성을 검토할 수 있습니다. 이 관련성 순위는 문서 또는 레코드가 해당 범주에 있는 다른 레코드와 비교하여 선택된 범주에 얼마나 잘 맞는지를 표시합니다. 단일 범주에 대한 레코드의 순위를 보려면 범주 분할창(왼쪽 상단 분할창)에서 이 범주를 선택하십시오. 문서 또는 레코드의 순위가 열에 나타납니다. 이 열은 기본적으로 표시되지 않지만 표시할 것을 선택할 수 있습니다. 자세한 정보는 114 페이지의 『데이터 분할창』 주제를 참조하십시오.

레코드 순위에 대한 숫자가 낮을수록, 이 레코드는 선택된 범주에 대해 더 잘 맞거나 더 큰 관련성을 가지며 1이 가장 잘 맞는 것입니다. 둘 이상의 레코드가 동일한 관련성을 갖는 경우, 각각은 동일한 순위를 갖고 나타나며 동일한 관련성을 갖고 있음을 표시하기 위해 등호(=)가 뒤따릅니다. 예를 들어 1=, 1=, 3, 4 등의 순위를 가질 수 있는데, 이것은 이 범주에 대해 최상의 매치로 동일하게 간주될 수 있는 두 개의 레코드가 있음을 의미합니다.

팁: 범주 주석(Annotation)에 가장 관련성이 큰 레코드의 텍스트를 추가하여 범주의 더 나은 설명을 제공할 수 있습니다. 텍스트를 선택하고 메뉴에서 **범주 > 주석에 추가**를 선택하여 데이터 분할창에서 직접 텍스트를 추가하십시오.

범주 작성

텍스트 분석 패키지에 범주가 있을 수 있지만 언어학적 기술이나 빈도 기술을 사용하여 자동으로 범주를 작성할 수도 있습니다. 범주 작성 설정 대화 상자를 통해 개념 또는 개념 패턴으로부터 범주를 생성하기 위해 자동화된 언어학적 및 빈도 기술을 적용할 수 있습니다.

일반적으로, 범주는 여러 유형의 디스크립터(유형, 개념, TLA 패턴, 범주 규칙)로 구성될 수 있습니다. 자동화된 범주 작성 기술을 사용하여 범주를 작성할 때 결과로 나오는 범주는 개념이나 개념 패턴(선택하는 입력에 따라 다름)을 따라 이름이 지정되고 각각에는 디스크립터 세트가 포함됩니다. 이러한 디스크립터는 범주 규칙이나 개념의 양식일 수 있으며 기술이 발견한 모든 관련 개념이 포함됩니다.

범주를 작성한 후에는 이를 범주 분할창에서 검토하고 그래프와 도표를 통해 탐색하여 범주에 대해 많은 것을 배울 수 있습니다. 그런 다음에는 수동 기술을 사용하여 경미한 조정을 하거나 잘못된 분류를 제거하거나 누락되었을 수 있는 레코드나 단어를 추가할 수 있습니다. 기술을 적용한 후에는 범주로 그룹화된 개념, 유형 및

패턴은 여전히 다른 기술에 사용 가능합니다. 또한, 다른 기술을 사용하면 중복되거나 부적합한 범주를 생성할 수도 있으므로 범주를 병합하거나 삭제할 수도 있습니다. 자세한 정보는 150 페이지의『범주 편집 및 세분화』주제를 참조하십시오.

중요! 이전 릴리스에서는 동시 발생과 동의어 규칙은 꺾쇠 괄호로 둘러싸였습니다. 이 릴리스에서는 꺾쇠 괄호는 이제는 텍스트 링크 분석 패턴 결과를 나타냅니다. 대신, 동시 발생 및 동의어 규칙은 (speaker systems|speakers)와 같은 소괄호로 캡슐화됩니다.

범주 작성

1. 메뉴에서 **범주 > 범주 작성**을 선택하십시오. 프롬프트하지 않기로 선택하지 않은 한 메시지 상자가 표시됩니다.
2. 지금 작성할지 또는 설정을 먼저 편집할지를 선택하십시오.

- 현재 설정을 사용하여 범주 작성을 시작하려면 **지금 작성**을 클릭하십시오. 기본적으로 선택된 설정은 종종 범주화 프로세스를 시작하기에 충분합니다. 범주 작성 프로세스가 시작되고 진행률 대화 상자가 나타납니다.
- 작성 설정을 검토하고 수정하려면 **편집**을 클릭하십시오.

참고: 표시할 수 있는 최대 범주 수는 10,000개입니다. 이 숫자에 도달했거나 초과되면 경고가 표시됩니다. 이 경우에는 범주 작성 또는 확장 옵션을 변경하여 작성된 범주 수를 줄여야 합니다.

입력

범주는 유형 패턴 또는 유형에서 파생된 디스크립터에서 작성됩니다. 테이블에서 범주 작성 프로세스를 포함하기 위한 개별 유형 또는 패턴을 선택할 수 있습니다.

유형 패턴. 유형 패턴을 선택하면 범주는 유형과 개념 대신 패턴으로부터 작성됩니다. 이런 방식으로 선택된 유형 패턴에 속하는 개념 패턴을 포함하는 모든 레코드 또는 문서가 범주화됩니다. 따라서 테이블에서 <Budget> 및 <Positive> 유형 패턴을 선택하면 cost & <Positive> 또는 rates & excellent 등과 같은 범주가 생성될 수 있습니다.

유형 패턴을 자동화된 범주 작성의 입력으로서 사용할 때는 기술이 범주 구조를 형성하기 위한 다양한 방식을 식별하는 때가 있습니다. 기술적으로 범주를 생성하는 한 가지의 옳은 방법이란 없습니다. 그러나 어떤 구조가 다른 구조보다는 사용자의 분석에 더 적합한지를 알아낼 수는 있습니다. 이 경우 출력을 사용자 정의하기 위해서는 유형을 선호 초점으로 지정할 수 있습니다. 생성된 모든 최상위 수준 범주는 다른 유형이 아니라 여기에서 선택한 유형의 개념에서 나옵니다. 모든 하위 범주에는 이 유형의 텍스트 링크 패턴이 포함됩니다. **패턴 유형별 구조 범주:** 필드에서 이 유형을 선택하면 테이블이 선택된 유형을 포함하는 적용 가능한 패턴만을 표시하기 위해 업데이트됩니다. 종종 <Unknown>이 미리 선택되어 있습니다. 그러면 <Unknown> (비일본어 텍스트의 경우) 유형을 포함하는 모든 패턴이 선택됩니다. . 테이블은 가장 많은 레코드 또는 문서(**문서** 개수)부터 시작하여 내림차순으로 유형을 표시합니다.

유형. 유형을 선택하면 범주는 선택된 유형에 속하는 개념으로부터 작성됩니다. 따라서 테이블에서 <Budget> 유형을 선택하는 경우 cost 또는 price 등과 같은 범주가 생성됩니다. cost 및 price는 <Budget> 유형에 지정되는 개념이기 때문입니다.

기본적으로 대부분의 레코드 또는 문서를 캡처하는 유형만이 선택됩니다. 이 사전 선택을 통해 가장 관심있는 유형에 빠르게 집중하고 관심없는 범주의 작성을 피할 수 있습니다. 테이블은 가장 많은 레코드 또는 문서(**문서** 개수)부터 시작하여 내림차순으로 유형을 표시합니다. Opinions 라이브러리의 유형은 기본적으로 유형 테이블에서 선택 취소되어 있습니다.

어떤 입력을 선택하는지가 어떤 범주를 얻게 되는지에 영향을 미칩니다. 유형을 입력으로 선택하면 명확하게 관련된 개념을 보다 쉽게 볼 수 있습니다. 예를 들어, 유형을 입력으로 사용하여 범주를 작성하는 경우에는 apple, pear, citrus fruits, orange 등과 같은 개념이 있는 Fruit 범주를 얻을 수 있습니다. 대신 유형 패턴을 입력으로 선택하고 <Unknown> + <Positive> 패턴을 예를 들어 선택하는 경우에는, fruit + tasty 및 apple + good 등과 같은 하나 또는 두 종류의 과일이 있는 fruit + <Positive> 범주를 얻을 수 있습니다. 이 두 번째 결과는 과일의 다른 발생이 반드시 절대적으로 자격이 있는 것은 아니므로 2개의 개념 패턴만을 보여줍니다. 이는 현재 텍스트 데이터에는 충분하지만 다른 문서 세트를 사용하는 장기적인 조사에서는 citrus fruit + positive 등과 같은 다른 디스크립터에서 수동으로 추가하거나 유형을 사용하려고 할 수도 있습니다. 유형만을 입력으로 단독 사용하면 가능한 모든 과일을 발견하는 데 도움이 됩니다.

기술

모든 데이터 세트가 고유하므로 방법의 수와 이를 적용하는 순서는 시간에 따라 변경될 수 있습니다. 텍스트 마이닝 목적은 한 데이터 세트와 다음 데이터 세트마다 다르므로, 여러 가지 방법을 시도하여 어떤 기술이 지정된 텍스트 데이터별로 최상의 결과를 낼 수 있는지를 살펴볼 필요가 있습니다.

이를 사용하기 위해서 이러한 설정의 전문가일 필요는 없습니다. 기본적으로 가장 공통된 평균 설정은 이미 선택되어 있습니다. 그러므로 고급 설정 대화 상자를 무시하고 범주 작성으로 바로 이동할 수 있습니다. 마찬가지로 여기에서 변경하면 마지막 설정은 항상 보존되므로 매번 설정 대화 상자로 돌아갈 필요가 없습니다.

언어학적 또는 빈도 기술을 선택하고 고급 설정 단추를 클릭하여 선택된 기술의 설정을 표시하십시오. 자동 기술은 데이터를 완벽하게 범주화합니다. 그러므로 사용자의 데이터에 가장 적합한 하나 이상의 자동 기술을 찾아서 적용하는 것이 좋습니다. 언어학적 및 빈도 기술을 동시에 사용하여 작성할 수는 없습니다.

- **고급 언어학적 기술**. 자세한 정보는 『고급 언어학적 설정』의 내용을 참조하십시오.
- **고급 빈도 기술**. 자세한 정보는 126 페이지의 『고급 빈도 설정』의 내용을 참조하십시오.

고급 언어학적 설정

범주를 작성할 때 *개념 루트 파생* (일본어에는 사용 불가능), *개념 포함*, *시맨틱 네트워크*(영어만) 및 *동시 발생 규칙*을 포함하여 여러 고급 언어학적 범주 작성 기술에서 선택할 수 있습니다. 이러한 기술은 범주를 작성하기 위해 개별적으로 또는 서로 결합하여 사용할 수 있습니다.

모든 데이터 세트가 고유하기 때문에 방법의 수와 이를 적용하는 순서는 시간에 따라 변경될 수 있음을 유의하십시오. 텍스트 마이닝 목적은 한 데이터 세트와 다음 데이터 세트마다 다르므로, 여러 가지 방법을 시도하여 어떤 기술이 지정된 텍스트 데이터별로 최상의 결과를 낼 수 있는지를 살펴볼 필요가 있습니다. 자동 기술은 데이터를 완벽하게 범주화합니다. 그러므로 사용자의 데이터에 가장 적합한 하나 이상의 자동 기술을 찾아서 적용하는 것이 좋습니다.

다음 영역과 필드는 고급 설정: 언어학적 대화 상자에서 사용 가능합니다.

입력 및 출력

범주 입력. 범주가 작성될 시작 위치를 선택하십시오.

- **미사용 추출 결과**. 이 옵션은 기존 범주에 사용되지 않는 추출 결과로부터 범주를 작성할 수 있게 해줍니다. 이는 레코드가 여러 범주에 매치하는 경향을 최소화하고 생성된 범주의 수를 제한합니다.
- **모든 추출 결과**. 이 옵션을 사용하면 추출 결과를 사용하여 범주를 작성할 수 있습니다. 이는 범주가 없거나 몇몇 범주만이 이미 존재하는 경우에 매우 유용합니다.

범주 출력. 범주가 작성될 일반 구조를 선택하십시오.

- **하위 범주가 있는 계층 구조**. 이 옵션을 사용하면 하위 범주와 하위 하위 범주를 작성할 수 있습니다. 작성할 수 있는 최대 수준 수(**작성된 최대 수준 수** 필드)를 선택하여 범주의 깊이를 설정할 수 있습니다. 3을 선택하면 범주에는 하위 범주가 포함되고 이러한 하위 범주에는 또 하위 범주가 있을 수 있습니다.
- **평면 범주(단일 수준만)**. 이 옵션을 사용하면 한 수준의 범주만이 작성됩니다. 즉 하위 범주가 생성되지 않습니다.

그룹화 기술

사용 가능한 각 기술은 특정 데이터 유형과 상황에 잘 맞지만, 종종 문서 또는 레코드의 전체 범위를 캡처하기 위해 동일한 분석으로 기술을 결합하는 것이 유용합니다. 다중 범주에서 개념을 확인하거나 중복 범주를 찾을 수 있습니다.

개념 루트 파생. 이 기술은 개념을 취하고 개념 구성요소가 형태소 분석으로 관련되거나 루트를 공유하는지 여부를 분석하여 관련되는 다른 개념을 찾아서 범주를 작성합니다. 이 기술은 동의 복합어 개념 식별에 아주 유용합니다. 생성된 각 범주의 개념은 동의어이거나 의미에서 거의 관련되기 때문입니다. 이는 다양한 길이의 데이터에 대해 작동하여 더 적은 수의 최소 범주를 생성합니다. 예를 들어, `opportunities to advance` 개념은 `opportunity for advancement` 및 `advancement opportunity` 개념을 사용하여 그룹화됩니다. 자세한 정보는 122 페이지의 『개념 루트 파생』의 내용을 참조하십시오. 이 옵션은 일본어 텍스트에는 사용할 수 없습니다.

시맨틱 네트워크. 이 기술은 광범위한 단어 색인 관계에서 각 개념의 가능한 의미를 식별하여 시작한 다음 관련된 개념을 그룹화하여 범주를 작성합니다. 이 기술은 개념이 시맨틱 네트워크에 알려져 있고 너무 애매하지 않을 경우에 가장 좋습니다. 텍스트에 네트워크에 알려지지 않은 용어나 특수화된 전문용어가 포함된 경우에는 덜 유용합니다. 하나의 예에서, 개념 `granny smith apple`은 `gala apple` 및 `winesap apple`과 그룹화될 수 있습니다. 이들은 granny smith의 형제어이기 때문입니다. 다른 예에서, 개념 `animal`은 `cat` 및 `kangaroo`와 그룹화될 수 있습니다. 이들은 `animal`의 하위어이기 때문입니다. 이 기술은 이 릴리스에서 영어 텍스트에만 사용할 수 있습니다. 자세한 정보는 124 페이지의 『시맨틱 네트워크』의 내용을 참조하십시오.

개념 포함. 이 기술은 다른 개념에서 단어의 서브세트 또는 수퍼세트인 단어를 포함하는지 여부를 기초로 다항어 개념(복합어)을 그룹화하여 범주를 작성합니다. 예를 들어, 개념 `seat`는 `safety seat`, `seat belt` 및 `seat belt buckle`과 함께 그룹화됩니다. 자세한 정보는 123 페이지의 『개념 포함』의 내용을 참조하십시오.

동시 발생. 이 기술은 텍스트에서 발견된 동시 발생에서 범주를 작성합니다. 개념 또는 개념 패턴이 종종 함께 문서 및 레코드에서 발견될 때, 동시 발생은 사용자 범주 정의의 값일 수 있는 기본적인 관계를 반영합니다. 단어가 현저하게 동시 발생하는 경우, 동시 발생 규칙이 작성되고 새 하위 범주에 대한 범주 디스크립터로 사용할 수 있습니다. 예를 들어, 많은 레코드에 단어 price 및 availability가 포함되어 있는 경우(그러나 몇 개의 레코드는 다른 하나 없이 하나만 포함함), 이 개념은 동시 발생 규칙으로 그룹화될 수 있고(price & available), 예를 들어 범주 price의 하위 범주에 지정됩니다. 자세한 정보는 125 페이지의 『동시 발생 규칙』의 내용을 참조하십시오.

최소 문서 수. 동시 발생 흥미 정도를 판별하기 위해, 범주에서 디스크립터로 사용되도록 지정된 동시 발생을 포함해야 하는 최소 문서 또는 레코드 수를 정의하십시오.

최대 검색 거리. 범주를 생성하기 전에 기술이 얼마나 멀리까지 검색하기를 원하는지를 선택하십시오. 값이 낮을수록 더 적은 수의 결과를 얻게 되지만, 이러한 결과는 잡음이 적고 서로 간에 의미 있게 링크되어 있거나 연관되어 있을 수 있습니다. 값이 높을수록 더 많은 수의 결과를 얻게 되지만 이러한 결과는 믿을 만하지 않거나 적절하지 않을 수 있습니다. 이러한 옵션은 모든 기술에 글로벌하게 적용되지만 동시 발생과 시맨틱 네트워크에 미치는 영향은 상당합니다.

특정 개념 쌍 방지. 프로세스가 출력에 두 개의 개념을 함께 그룹화하거나 쌍으로 만드는 프로세스를 중지하려면 이 선택란을 선택하십시오. 개념 쌍을 작성하거나 관리하려면 **쌍 관리**를 클릭하십시오. 자세한 정보는 121 페이지의 『링크 예외 쌍 관리』 주제를 참조하십시오.

가능한 곳에서 와일드카드 일반화. 별표 와일드카드를 사용하여 제품이 범주에서 일반 규칙을 생성하게 하려면 이 옵션을 선택하십시오. 예를 들어, [apple tart + .] 및 [apple sauce + .] 등과 같은 여러 디스크립터를 생성하는 대신에 와일드카드를 사용하면 [apple * + .]을 생성할 수 있습니다. 와일드카드를 사용하여 일반화하면 종종 앞서 한 것과 똑같은 수의 레코드 또는 문서 수를 얻게 됩니다. 그러나 이 옵션은 숫자를 줄이고 범주 디스크립터를 단순화하는 장점이 있습니다. 또한 이 옵션은 새 텍스트 데이터(예를 들어, 세로/파동 연구에서)에서 이러한 범주를 사용하여 더 많은 레코드 또는 문서를 범주화하는 기능을 증가시킵니다.

범주 작성을 위한 다른 옵션

적용할 그룹 기술을 선택하는 것 외에도 다음과 같은 몇몇 기타 작성 옵션을 편집할 수 있습니다.

작성된 최상위 수준 **범주의 최대 수.** 이 옵션을 사용하여 다음 번에 범주 작성 단추를 클릭할 때 생성할 수 있는 범주 수를 제한할 수 있습니다. 어떤 경우에는 이 값을 높게 설정한 다음에 관심없는 범주의 일부를 삭제하면 더 나은 결과를 얻을 수도 있습니다.

범주별 최대 디스크립터 및/또는 하위 범주의 수. 이 옵션을 사용하면 범주를 작성하기 위해 포함해야 하는 최소 디스크립터와 하위 범주 수를 정의할 수 있습니다. 이 옵션은 상당 수의 레코드 또는 문서를 캡처하지 않는 범주 작성을 제한하는 데 도움이 됩니다.

디스크립터가 둘 이상의 범주에 나타나게 허용. 이 옵션이 선택되면 디스크립터가 다음 번에 작성될 둘 이상의 범주에서 사용할 수 있게 됩니다. 이 옵션은 일반적으로 선택됩니다. 항목은 일반적으로 또는 "자연적으로" 둘 이상의 범주에 해당하고 이를 허용하면 더 높은 품질의 범주로 이어질 수 있기 때문입니다. 이 옵션을 선택하

지 않으면 여러 범주에서 레코드의 겹침을 줄이게 되는데, 가지고 있는 데이터의 유형에 따라서 이는 바람직하지 않을 수 있습니다. 그러나 대부분의 데이터 유형에서는 일반적으로 디스크립터를 단일 범주로 제한하면 품질이나 범주 범위가 손실됩니다. 예를 들어, car seat manufacturer 개념이 있다고 해봅시다. 이 옵션을 사용하면 이 개념은 car seat 텍스트를 기반으로 하나의 범주에 나타나거나 manufacturer를 기반으로 또 다른 범주에 나타날 수 있습니다. 그러나 이 옵션이 선택되지 않은 경우에는 두 범주를 모두 얻을 수는 있지만, car seat manufacturer 개념은 car seat 및 manufacturer가 각각 발생하는 레코드 수를 포함하여 여러 요소를 기반으로 가장 매치하는 범주에 디스크립터로만 나타납니다.

중복된 범주 이름 해결 기준 이름이 기존 범주와 같은 새 범주 또는 하위 범주를 처리하는 방법을 선택하십시오. 이름이 같은 기존 범주와 새 범주(및 해당 디스크립터)를 병합할 수 있습니다. 또는 기존 범주에서 중복 이름이 발견된 경우 범주의 작성을 건너뛰도록 선택할 수도 있습니다.

링크 예외 쌍 관리

범주 작성, 군집 및 개념 맵핑 동안 내부 알고리즘은 단어를 알려진 연관을 기준으로 그룹화합니다. 두 개의 개념이 쌍을 이루거나 서로 링크되는 것을 막으려면 **범주 작성 고급 설정** 대화 상자, **군집 작성** 대화 상자 및 **개념 맵 색인 설정** 대화 상자에서 이 기능을 켜고 **쌍 관리** 단추를 클릭하십시오.

결과로 나오는 **링크 예외 관리** 대화 상자에서 개념 쌍을 추가, 편집 또는 삭제할 수 있습니다. 행당 한 쌍을 입력하십시오. 여기에 쌍을 입력하면 범주, 군집 및 개념 맵핑을 작성하거나 확장할 때 쌍이 발생되는 것을 막습니다. 단어를 원하는 그대로 입력하십시오. 예를 들어, 단어의 액센트 버전은 단어의 액센트가 없는 버전과 같지 않습니다.

예를 들어, hot dog와 dog가 그룹화되지 않도록 하려면 쌍을 테이블에 별도의 행으로서 추가할 수 있습니다.

언어학적 기술 정보

범주를 작성하거나 확장할 때 *개념 루트 파생* (일본어에는 사용 불가능), *개념 포함*, *시맨틱 네트워크*(영어만) 및 *동시 발생 규칙*을 포함하여 여러 고급 언어학적 범주 작성 기술에서 선택할 수 있습니다. 이러한 기술은 범주를 작성하기 위해 개별적으로 또는 서로 결합하여 사용할 수 있습니다.

이를 사용하기 위해서 이러한 설정의 전문가일 필요는 없습니다. 기본적으로 가장 공통된 평균 설정은 이미 선택되어 있습니다. 원하는 경우 이 고급 설정 대화 상자를 무시하고 범주 작성이나 확장으로 바로 이동할 수 있습니다. 마찬가지로 여기에서 변경하는 경우에는 마지막으로 사용된 설정이 기억되므로 매번 설정 대화 상자로 돌아갈 필요가 없습니다.

그러나 모든 데이터 세트가 고유하기 때문에 방법의 수와 이를 적용하는 순서는 시간에 따라 변경될 수 있음을 유의하십시오. 텍스트 마이닝 목적은 한 데이터 세트와 다음 데이터 세트마다 다르므로, 여러 가지 방법을 시도하여 어떤 기술이 지정된 텍스트 데이터별로 최상의 결과를 낼 수 있는지를 살펴볼 필요가 있습니다. 자동 기술은 데이터를 완벽하게 범주화합니다. 그러므로 사용자의 데이터에 가장 적합한 하나 이상의 자동 기술을 찾아서 적용하는 것이 좋습니다.

범주 작성을 위한 기본 자동화된 언어학적 기술은 다음과 같습니다.

- **개념 루트 파생.** 이 기술은 개념을 사용하고 개념 구성요소가 형태학상으로 관련되어 있는지 여부를 분석하여 이와 관련된 다른 개념을 찾는 방법으로 범주를 작성합니다. 자세한 정보는 『개념 루트 파생』의 내용을 참조하십시오. 이 옵션은 일본어 텍스트에는 사용할 수 없습니다.
- **개념 포함.** 이 기술은 개념을 사용하여 이를 포함하는 다른 개념을 찾는 방법으로 범주를 작성합니다. 자세한 정보는 123 페이지의 『개념 포함』의 내용을 참조하십시오.
- **시맨틱 네트워크.** 이 기술은 광범위한 단어 색인 관계에서 각 개념의 가능한 의미를 식별하여 시작한 다음 관련된 개념을 그룹화하여 범주를 작성합니다. 자세한 정보는 124 페이지의 『시맨틱 네트워크』의 내용을 참조하십시오. 이 옵션은 영어 텍스트에만 사용 가능합니다.
- **동시 발생.** 이 기술은 새 범주를 작성하고, 범주를 확장하거나 다른 범주 기술에 대한 입력으로 사용될 수 있는 동시 발생 규칙을 작성합니다. 자세한 정보는 125 페이지의 『동시 발생 규칙』의 내용을 참조하십시오.

개념 루트 파생

참고: 이 기술은 일본어 텍스트에는 사용할 수 없습니다.

개념 루트 파생 기술은 개념을 사용하고 개념 구성요소가 형태학상으로 관련되어 있는지 여부를 분석하여 이와 관련된 다른 개념을 찾는 방법으로 범주를 작성합니다. 구성요소는 한 단어입니다. 기술은 개념에서 각 구성요소의 끝부분(접미문자)을 보고 여기에서 파생할 수 있는 다른 개념을 찾아서 개념을 그룹화하려고 시도합니다. 이 아이디어는 단어가 서로 파생되면 의미를 공유하거나 가까울 수 있다는 것입니다. 끝부분을 식별하기 위해 내부 언어 특정 규칙이 사용됩니다. 예를 들어, `opportunities to advance` 개념은 `opportunity for advancement` 및 `advancement opportunity` 개념을 사용하여 그룹화됩니다.

모든 종류의 텍스트에서 개념 루트 파생을 사용할 수 있습니다. 이는 스스로 소수의 범주를 생성하고 각 범주에는 소수의 개념이 포함되는 경향이 있습니다. 각 범주에서의 개념은 동의어이거나 상황적으로 관련이 있을 수 있습니다. 범주를 수동으로 작성하더라도 이 알고리즘을 사용하는 것이 유용할 수 있습니다. 여기에서 발견하는 동의어는 특히 관심이 있는 개념의 동의어일 수 있습니다.

참고: 개념을 명시적으로 지정하여 개념이 서로 그룹화되지 않도록 막을 수 있습니다. 자세한 정보는 121 페이지의 『링크 예외 쌍 관리』의 내용을 참조하십시오.

용어 컴포넌트화 및 굴절 해제

개념 루트 파생 또는 개념 포함 기술이 적용되면 용어는 먼저 구성요소(단어)로 구분된 다음 구성요소는 굴절이 해제됩니다. 기술이 적용되면 개념 및 해당 연관된 용어가 로드되고 공백, 하이픈 및 어포스트로피 등과 같은 구분 문자를 기반으로 구성요소로 나뉘어집니다. 예를 들어, `system administrator` 용어는 `{administrator, system}` 등과 같은 구성요소로 나뉘어집니다.

그러나 원래 용어의 일부분은 사용되지 않거나 검색 엔진에서 제외되는 단어로서 언급될 수도 있습니다. 영어에서는 이러한 무시 가능한 구성요소 중 일부는 `a`, `and`, `as`, `by`, `for`, `from`, `in`, `of`, `on`, `or`, `the`, `to` 및 `with`가 포함될 수 있습니다.

예를 들어, `examination of the data` 용어에는 `{data, examination}` 구성요소 세트가 있고, `of` 및 `the` 둘 모두가 무시 가능한 것으로 간주됩니다. 또한 구성요소 순서는 구성요소 세트에 없습니다. 이런 방식으로

cough relief for child, child relief from a cough 및 relief of child cough의 세 개의 용어는 동등할 수 있습니다. 이들 모두는 동일한 구성요소 세트 {child, cough, relief}를 가지고 있기 때문입니다. 용어 쌍이 동등한 것으로 식별될 때마다 해당하는 개념은 모든 용어를 참조하는 새 개념을 형성하기 위해 병합됩니다.

또는 용어의 구성요소가 굴절될 수 있으므로 복수 형태와 같은 굴절 변화와 관계 없이 동등한 용어를 식별하기 위해 언어 특정 규칙이 내부적으로 적용됩니다. 이런 방식으로 level of support 및 support levels는 동등한 것으로 식별될 수 있습니다. 굴절이 해제된 단수 양식은 level이기 때문입니다.

개념 루트 파생 작동 방법

용어가 컴포넌트화되고 굴절이 해제된 후(이전 섹션 참조) 개념 루트 파생 알고리즘은 구성요소 엔진 또는 접미문자를 분석하여 구성요소 루트를 찾은 다음 개념을 동일하거나 유사한 루트가 있는 다른 개념과 그룹화합니다. 끝부분은 텍스트 언어 특정 언어학적 파생 규칙 세트를 사용하여 식별됩니다. 예를 들어, 접미문자 ical이 있는 동일한 개념 구성요소 끝부분은 동일한 루트 어간과 접미문자 ic가 있는 끝부분에서 파생될 수 있음을 설명하는 영어 언어 텍스트의 파생 규칙이 있습니다. 이 규칙(및 굴절 해제)을 사용하면 알고리즘은 개념 epidemiologic study 및 epidemiological studies를 그룹화할 수 있습니다.

용어가 이미 컴포넌트화되었고 무시 가능한 구성요소(예: in 및 of)가 식별되었으므로, 개념 루트 파생 알고리즘은 개념 studies in epidemiology를 epidemiological studies와 그룹화할 수도 있습니다.

이 알고리즘으로 그룹화된 대부분의 개념이 동의어일 수 있도록 구성요소 파생 규칙 세트가 선택되었습니다. epidemiologic studies, epidemiological studies, studies in epidemiology 개념은 모두 동등한 용어입니다. 완전성을 늘리기 위해서 알고리즘이 상황적으로 관련된 개념을 그룹화할 수 있도록 해주는 몇몇 파생 규칙이 있습니다. 예를 들어, 알고리즘은 empire builder 및 empire building과 같은 개념을 그룹화할 수 있습니다.

개념 포함

개념 포함 기술은 개념을 사용하여 범주를 작성하고, 어휘 계열 알고리즘을 사용하여 다른 개념에 포함된 개념을 식별합니다. 이 아이디어는 개념에 있는 단어가 다른 개념의 서브세트이면 기본적인 시맨틱 관계를 반영한다는 것입니다. 포함은 모든 유형의 텍스트와 함께 사용할 수 있는 강력한 기술입니다.

이 기술은 시맨틱 네트워크와 결합하여 잘 작동하지만 별도로 사용될 수 있습니다. 개념 포함은 문서 또는 레코드에 많은 도메인 특정 용어 또는 전문어를 포함한 경우에 더 나은 결과를 제공할 수도 있습니다. 이는 특히 특수 용어가 추출되고 적절하게 그룹화될 수 있도록(동의어와) 사전에 사전을 조정한 경우에 특히 그렇습니다.

개념 포함 작동 방법

개념 포함 알고리즘을 적용하기 전에 용어가 컴포넌트화되고 굴절이 해제됩니다. 자세한 정보는 122 페이지의 『개념 루트 파생』의 내용을 참조하십시오. 그런 다음 개념 포함 알고리즘은 구성요소 세트를 분석합니다. 각 구성요소 세트마다 알고리즘은 첫 번째 구성요소 세트의 서브세트인 또 다른 구성요소 세트를 찾습니다.

예를 들어, 구성요소 세트 {breakfast, continental}이 있는 continental breakfast 개념이 있고 {breakfast} 구성요소 세트가 있는 breakfast 개념이 있는 경우에는 알고리즘은 continental breakfast가 breakfast의 종류라고 결론짓고 이들을 그룹화합니다.

더 큰 예에서 추출 결과 분할창에 seat 개념이 있고 이 알고리즘을 적용하는 경우에는 safety seat, leather seat, seat belt, seat belt buckle, infant seat carrier 및 car seat laws 등과 같은 개념 또한 해당 범주에서 그룹화됩니다.

용어가 이미 컴포넌트화되었고 무시 가능한 구성요소(예: in 및 of)가 식별된 경우에는 개념 포함 알고리즘은 advanced spanish course 개념이 course in spanish 개념을 포함한다고 인식합니다.

참고: 개념을 명시적으로 지정하여 개념이 서로 그룹화되지 않도록 막을 수 있습니다. 자세한 정보는 121 페이지의 『링크 예외 쌍 관리』의 내용을 참조하십시오.

시맨틱 네트워크

이 릴리스에서 시맨틱 네트워크 기술은 영어 텍스트에만 사용할 수 있습니다.

이 기법은 단어 관계의 내장된 네트워크를 사용하여 범주를 작성합니다. 이 때문에 이 기법은 용어가 구체적이고 너무 애매모호하지 않을 때 매우 좋은 결과를 생성할 수 있습니다. 그러나 이 기법이 매우 기술적/전문적 개념 사이의 많은 링크를 찾을 것으로 기대해서는 안 됩니다. 그런 개념을 다룰 때는 개념 포함 및 개념 루트 파생 기법이 더 유용함을 발견할 수 있습니다.

시맨틱 네트워크가 작업하는 방법

시맨틱 네트워크 기술 뒤에 있는 아이디어는 알려진 단어 관계를 활용하여 동의어 또는 하의어의 범주를 작성하는 것입니다. **하의어**는 하나의 개념이 일종의 두 번째 개념이어서 ISA 관계라고도 알려진 계층 구조 관계가 있을 때입니다. 예를 들어 동물이 하나의 개념일 때 고양이 및 캥거루는 동물의 일종이므로 동물의 하의어입니다.

동의어 및 하의어 관계 외에, 시맨틱 네트워크 기술은 <Location> 유형의 모든 개념 사이의 부분 및 전체 링크를 조사합니다. 예를 들어, 이 기법은 노르망디와 프로방스가 프랑스의 일부이기 때문에 노르망디, 프로방스 및 프랑스를 하나의 범주로 그룹화합니다.

시맨틱 네트워크는 시맨틱 네트워크에 있는 각 개념의 가능한 의미를 식별하여 시작합니다. 개념이 동의어 또는 하의어로서 식별될 때 하나의 범주로 그룹화됩니다. 예를 들어, 기법은 시맨틱 네트워크에 1) dessert apple은 eating apple의 동의어이고, 2) granny smith는 eating apple의 한 종류(eating apple의 하의어임을 의미)라는 정보가 들어 있으므로 eating apple, dessert apple, granny smith를 포함하는 단일 범주를 작성합니다.

개별적으로 취할 때, 많은 개념, 특히 단일어는 애매모호합니다. 예를 들어 뷔페란 개념은 식사의 한 종류 또는 가구의 일부를 나타낼 수 있습니다. 개념 세트가 식사, 가구 및 뷔페를 포함하면, 알고리즘은 뷔페를 식사 또는 가구와 그룹화 사이에서 선택하도록 강제 실행됩니다. 어떤 경우에는 알고리즘에 의해 이루어지는 선택이 레코드 또는 문서의 특정 세트의 컨텍스트에서 적합하지 않을 수 있음을 기억하십시오.

시맨틱 네트워크 기술은 특정 유형의 데이터를 갖는 개념 포함을 능가할 수 있습니다. 시맨틱 네트워크와 개념 포함이 둘 다 애플 파이가 파이의 한 종류임을 인식하지만, 시맨틱 네트워크만 타르트도 파이의 한 종류임을 인식합니다.

시맨틱 네트워크는 다른 기법과 결합하여 작동합니다. 예를 들어, 시맨틱 네트워크와 포함 기법을 둘 다 선택했고 시맨틱 네트워크가 선생님 개념을 튜터 개념과 그룹화(튜터는 선생님의 한 종류이므로)했다고 가정하십시오. 포함 알고리즘은 대졸 튜터를 튜터와 그룹화할 수 있어서 결국 두 알고리즘은 협력하여 튜터, 대졸 튜터, 선생님의 세 개념을 모두 포함하는 출력 범주를 생성합니다.

시맨틱 네트워크의 옵션

이 기법에서 관심을 가질 수 있는 많은 추가 설정이 있습니다.

- **최대 검색 거리**를 변경하십시오. 범주를 생성하기 전에 기술이 얼마나 멀리까지 검색하기를 원하는지를 선택하십시오. 값이 낮을수록 더 적은 수의 결과가 생성되지만, 이러한 결과는 잡음이 적고 서로 간에 의미 있게 링크되어 있거나 연관되어 있을 수 있습니다. 값이 높을수록 더 많은 수의 결과를 얻게 되지만 이러한 결과는 믿을 만하지 않거나 적절하지 않을 수 있습니다.

예를 들어, 거리에 따라서 이 알고리즘은 대니시 패스트리부터 커피롤(그의 상위)까지를 검색한 후, 번(조부모) 및 빵까지 검색합니다.

검색 거리를 줄임으로써 이 기법은 생성되는 범주가 너무 크거나 너무 많은 것을 그룹화한다고 느끼는 경우 작업하기에 더 쉬울 수 있는 더 작은 범주를 생성합니다.

중요! 또한, 일부 잘못된 그룹화가 결과에 부정적 영향을 크게 미칠 수 있으므로 이 기법을 사용할 때 퍼지 그룹화에 대해 **최소 루트 문자 한계에 대해 철자법 오류 수용**(노드의 전문가 탭이나 추출 대화 상자에서 정의됨) 옵션을 적용하지 않을 것을 권장합니다.

동시 발생 규칙

동시 발생 규칙을 사용하면 문서 또는 레코드 세트 내에서 강하게 관련되어 있는 개념을 발견하고 그룹화할 수 있습니다. 이 아이디어는 개념이 종종 문서와 레코드에서 발견될 때, 동시 발생은 범주 정의에 있는 값일 가능성이 있는 기본 관계를 반영한다는 것입니다. 이 기술은 새 범주를 작성하고, 범주를 확장하거나 다른 범주 기술에 대한 입력으로 사용될 수 있는 동시 발생 규칙을 작성합니다. 두 개의 개념이 레코드 세트에서 함께 자주 나타나고 다른 레코드에서 드물게 개별적으로 나타나는 경우에는 이들은 강력하게 동시 발생합니다. 이 기술은 최소 수백 개의 문서 또는 레코드가 있는 큰 데이터 세트에서 좋은 결과를 생성할 수 있습니다.

예를 들어, 많은 레코드에 `price` 및 `availability` 단어가 포함되는 경우에는 이러한 개념은 동시 발생 규칙, `(price & available)`로 그룹화될 수 있습니다. 다른 예에서 `peanut butter`, `jelly`, `sandwich` 개념이 따로 떨어지기보다는 자주 함께 나타나는 경우에는 이들은 개념 동시 발생 규칙 `(peanut butter & jelly & sandwich)`에 그룹화됩니다.

중요! 이전 릴리스에서는 동시 발생과 동의어 규칙은 꺾쇠 괄호로 둘러싸였습니다. 이 릴리스에서는 꺾쇠 괄호는 이제는 텍스트 링크 분석 패턴 결과를 나타냅니다. 대신, 동시 발생 및 동의어 규칙은 `(speaker systems|speakers)`와 같은 소괄호로 캡슐화됩니다.

동시 발생 규칙 작동 방법

이 기술은 함께 나타나는 경향이 있는 둘 이상의 개념을 찾기 위해 문서 또는 레코드를 스캔합니다. 둘 이상의 개념은 문서 또는 레코드 세트에 빈번하게 함께 나타나는 경우와 다른 문서 또는 레코드에서는 거의 개별적으로 나타나지 않는 경우에 강력하게 동시 발생합니다.

동시 발생하는 개념이 발견되면 범주 규칙이 형성됩니다. 이러한 규칙은 & 부울 연산자를 사용하여 연결된 둘 이상의 개념으로 구성되어 있습니다. 이러한 규칙은 규칙의 개념 세트가 모두 해당 문서 또는 레코드에서 동시 발생하는 경우 문서 또는 레코드를 범주로 자동으로 분류하는 논리문입니다.

동시 발생 규칙의 옵션

동시 발생 규칙 기술을 사용 중인 경우에는 결과로 나오는 규칙에 영향을 미치는 몇몇 설정을 세부 조정할 수 있습니다.

- **최대 검색 거리**를 변경하십시오. 기술이 얼마나 멀리까지 동시 발생을 검색하는지를 선택하십시오. 검색 거리를 늘릴 때 각 동시 발생에 필요한 최소 유사성 값은 낮아집니다. 따라서 많은 동시 발생 규칙이 생성될 수 있지만 유사성 값이 낮으면 중요성이 낮습니다. 검색 거리를 줄이면 필요한 최소 유사성 값이 늘어납니다. 그 결과로 생성되는 동시 발생 규칙의 수가 줄어들지만 보다 중요해지는(강력해지는) 경향이 있습니다.
- **문서.** 동시 발생으로 간주되기 위해서 지정된 개념 쌍을 포함해야 하는 최소 레코드 또는 문서 수입니다. 이 옵션을 낮게 설정할수록 동시 발생을 쉽게 찾을 수 있습니다. 값을 늘리면 동시 발생의 수는 줄어들지만 중요도는 높아집니다. 예를 들어, "사과"와 "배" 개념이 함께 2개의 레코드에서 발견되었고 두 개의 개념이 다른 레코드에서 발생하지 않는다고 가정합니다. **문서.** 2(기본값)로 설정된 동시 발생 기술은 범주 규칙(사과와 배)을 작성합니다. 값이 3으로 높아지면 규칙은 더 이상 작성되지 않습니다.

참고: 작은 데이터 세트(< 1000개 반응)의 경우 기본 설정이 있는 동시 발생을 찾을 수 없을 수도 있습니다. 그런 경우 검색 거리 값을 늘려 보십시오.

참고: 개념을 명시적으로 지정하여 개념이 서로 그룹화되지 않도록 막을 수 있습니다. 자세한 정보는 121 페이지의 『링크 예외 쌍 관리』의 내용을 참조하십시오.

고급 빈도 설정

간단하고 기계적인 빈도 기술을 기반으로 범주를 작성할 수 있습니다. 이 기술을 사용하면 주어진 레코드 또는 문서 개수를 넘어서 발견된 각 항목(유형, 개념 또는 패턴)마다 하나의 범주를 작성할 수 있습니다. 또한 덜 자주 발생하는 모든 항목에 대해 하나의 범주를 작성할 수 있습니다. 개수는 전체 텍스트에서 총 발생 수와 대조적으로 문제의 추출된 개념(및 해당 모든 동의어), 유형 또는 패턴을 포함하는 레코드 또는 문서 수를 가리킵니다.

자주 발생하는 항목 그룹화는 공통되거나 중요한 반응을 나타낼 수 있으므로 흥미로운 결과를 낼 수 있습니다. 이 기술은 다른 기술이 적용된 후에 사용되지 않은 추출 결과에서 매우 유용합니다. 또 다른 방법은 다른 범주가 하나도 없을 때 추출 직후에 이 기술을 실행하고, 결과를 편집하여 관심 없는 범주를 삭제한 다음 이러한 범주가 더 많은 레코드 또는 문서와 매치할 수 있도록 확장하는 것입니다. 자세한 정보는 127 페이지의 『범주 확장』의 내용을 참조하십시오.

이 기술을 사용하는 대신에 개념 또는 개념 패턴을 추출 결과 분할창에서 레코드 또는 문서 수의 내림차순으로 정렬한 다음 맨 위의 것을 범주 분할창으로 끌어다 놓는 방식으로 해당하는 범주를 작성할 수 있습니다.

다음 필드는 고급 설정: 빈도 대화 상자에서 사용 가능합니다.

범주 디스크립터 생성 위치. 디스크립터의 입력 종류를 선택하십시오. 자세한 정보는 116 페이지의 『범주 작성』의 내용을 참조하십시오.

- **개념 수준**. 이 옵션을 선택하면 개념 또는 개념 패턴 빈도가 사용됩니다. 유형이 범주 작성의 입력으로서 선택된 경우에는 개념이 사용되고, 패턴이 선택된 경우에는 개념 패턴이 사용됩니다. 일반적으로 이 기술을 개념 수준에 적용하면 보다 특정적인 결과가 생성됩니다. 개념과 개념 패턴은 더 낮은 측정 수준을 나타내기 때문입니다.
- **유형 수준**. 이 옵션을 선택하면 유형 또는 유형 패턴 빈도가 사용됩니다. 유형이 범주 작성의 입력으로서 선택된 경우에는 유형이 사용되고, 패턴이 선택된 경우에는 유형 패턴이 사용됩니다. 이 기술을 유형 수준에 적용하면 제공된 정보 유형과 관련한 빠른 보기를 볼 수 있습니다.

자체 범주를 가진 항목의 최소 문서 수. 이 옵션을 사용하면 자주 발생하는 항목으로부터 범주를 작성할 수 있습니다. 이 옵션은 출력을 최소 X개의 레코드 또는 문서에서 발생한 디스크립터를 포함하는 범주만으로 제한합니다. 여기서 X는 이 옵션에 입력할 값입니다.

남은 모든 항목을 호출된 범주로 그룹화. 이 옵션을 사용하면 덜 빈번하게 발생하는 모든 개념 또는 유형을 원하는 이름으로 된 하나의 '잡동사니' 범주로 그룹화할 수 있습니다. 기본적으로 이 범주의 이름은 *기타*입니다.

범주 입력. 기술을 적용할 그룹을 선택하십시오.

- **미사용 추출 결과**. 이 옵션은 기존 범주에 사용되지 않는 추출 결과로부터 범주를 작성할 수 있게 해줍니다. 이는 레코드가 여러 범주에 매치하는 경향을 최소화하고 생성된 범주의 수를 제한합니다.
- **모든 추출 결과**. 이 옵션을 사용하면 추출 결과를 사용하여 범주를 작성할 수 있습니다. 이는 범주가 없거나 몇몇 범주만이 이미 존재하는 경우에 매우 유용합니다.

중복된 범주 이름 해결 기준 이름이 기존 범주와 같은 새 범주 또는 하위 범주를 처리하는 방법을 선택하십시오. 이름이 같은 기존 범주와 새 범주(및 해당 디스크립터)를 병합할 수 있습니다. 또는 기존 범주에서 중복 이름이 발견된 경우 범주의 작성을 건너뛰도록 선택할 수도 있습니다.

범주 확장

확장은 기존 범주를 '키우기'위해서 디스크립터가 추가되거나 자동으로 개선되는 프로세스입니다. 목적은 해당 범주에 원래 지정되지 않은 관련 레코드 또는 문서를 캡처하는 더 나은 범주를 생성하는 것입니다.

선택하는 자동 그룹화 기술은 기존 범주 디스크립터와 관련된 개념, TLA 패턴 및 범주 규칙을 식별하려고 시도합니다. 이러한 새 개념, 패턴 및 범주 규칙은 그런 다음 새 디스크립터로 추가되거나 기존 디스크립터에 추가됩니다. 확장을 위한 그룹화 기술에는 *개념 루트 파생*(일본어에는 사용할 수 없음), *개념 포함*, *시맨틱 네트*

워크(영어만 해당) 및 *동시 발생 규칙*이 포함됩니다. **빈 범주를 범주 이름에서 생성된 디스크립터를 사용하여 확장** 방법은 범주 이름의 단어를 사용하여 디스크립터를 생성합니다. 그러므로 범주 이름이 설명적인 이름일수록 더 좋은 결과가 나옵니다.

참고: 빈도 기술은 범주를 확장할 때에는 사용할 수 없습니다.

확장은 범주를 대화식으로 개선하는 좋은 방법입니다. 다음은 범주를 확장할 수 있는 몇몇 예제입니다.

- 범주 분할창에서 범주를 작성하기 위해 개념 패턴을 끌어서 놓은 후
- 단순 범주 규칙 및 디스크립터를 추가하여 범주를 작성한 후
- 범주에 설명적인 이름이 있는 사전 정의된 범주 파일 을 가져온 후
- 선택한 TAP으로부터 나온 범주를 세분화한 후

범주를 여러 번 확장할 수 있습니다. 예를 들어, 매우 설명적인 이름이 있는 사전 정의된 범주 파일을 가져온 경우에는 **범주 이름에서 생성된 디스크립터를 사용하여 빈 범주 확장** 옵션을 사용하여 확장하여 첫 번째 디스크립터 세트를 획득한 다음 이러한 범주를 다시 확장할 수 있습니다. 그러나 다른 경우에는 여러 번 확장하면 디스크립터가 점점 더 넓게 확장되어 너무 일반화된 범주가 생길 수도 있습니다. 그룹 작성 및 확장 기술은 유사한 기반 알고리즘을 사용하므로 범주를 작성한 직후에 확장하면 흥미로운 결과가 나올 가능성이 적습니다.

팁:

- 확장을 시도하지만 결과를 사용하지 않으려는 경우에는 확장한 직후에 작업(**편집 > 실행 취소**)을 언제든지 취소할 수 있습니다.
- 확장하면 범주에 문서 세트가 정확하게 매치하는 둘 이상의 범주 규칙이 생성될 수 있습니다. 규칙은 프로세스와는 별개로 작성되기 때문입니다. 원하는 경우 범주를 검토하고 범주 설명을 수동으로 편집하여 중복을 제거할 수 있습니다. 자세한 정보는 151 페이지의 『범주 디스크립터 편집』의 내용을 참조하십시오.

범주 확장

1. 범주 분할창에서 확장하려는 범주를 선택하십시오.
2. 메뉴에서 **범주 > 범주 확장**을 선택하십시오. 프롬프트하지 않기로 선택하지 않은 한 메시지 상자가 나타납니다.
3. 지금 작성할지 또는 설정을 먼저 편집할지를 선택하십시오.

- 현재 설정을 사용하여 범주 확장을 시작하려면 **지금 확장**을 클릭하십시오. 프로세스가 시작되고 진행률 대화 상자가 나타납니다.
- 설정을 검토하고 수정하려면 **편집**을 클릭하십시오.

확장하려고 시도한 후에 새 디스크립터가 발견된 모든 범주는 범주 분할창에서 빠르게 식별할 수 있도록 **확장됨** 단어가 플래그되어 있습니다. 확장된 텍스트는 다시 확장하거나, 다른 방법으로 범주를 편집하거나 컨텍스트 메뉴를 통해 이를 지울 때까지 그대로 남아 있습니다.

참고: 표시할 수 있는 최대 범주 수는 10,000개입니다. 이 숫자에 도달했거나 초과되면 경고가 표시됩니다. 이 경우에는 범주 작성 또는 확장 옵션을 변경하여 작성된 범주 수를 줄여야 합니다.

범주를 작성하거나 확장할 때 사용 가능한 각 기술은 특정 데이터 유형과 상황에 적합하지만 종종 문서 또는 레코드의 전체 범위를 캡처하기 위해 동일한 분석에서 기술을 결합하는 것이 도움이 됩니다. 대화식 워크벤치에서 범주로 그룹화된 개념과 유형은 다음 번에 범주를 작성할 때에도 여전히 사용 가능합니다. 즉 여러 범주에서 개념을 보거나 중복된 범주를 찾을 수도 있습니다.

다음 영역과 필드는 범주 확장: 설정 대화 상자에서 사용 가능합니다.

확장 방법. 범주를 확장하는 데 사용될 입력을 선택하십시오.

- **미사용 추출 결과**. 이 옵션은 기존 범주에 사용되지 않는 추출 결과로부터 범주를 작성할 수 있게 해줍니다. 이는 레코드가 여러 범주에 매치하는 경향을 최소화하고 생성된 범주의 수를 제한합니다.
- **모든 추출 결과**. 이 옵션을 사용하면 추출 결과를 사용하여 범주를 작성할 수 있습니다. 이는 범주가 없거나 몇몇 범주만이 이미 존재하는 경우에 매우 유용합니다.

그룹화 기술

이러한 각 기술의 간단한 설명은 118 페이지의 『고급 언어학적 설정』의 내용을 참조하십시오. 이러한 기술은 다음을 포함합니다.

- **개념 루트 파생***(일본어에는 사용할 수 없음)*
- **시맨틱 네트워크** (*영어 텍스트에만 사용 가능하고 일반화 전용 옵션이 선택된 경우에는 사용되지 않습니다.*)
- **개념 포함**
- **동시 발생**과 **최소 문서 수** 하위 옵션.

많은 유형이 시맨틱 네트워크 기술에서 영구적으로 제외되었습니다. 이러한 유형이 관련 결과를 생성하지 않기 때문입니다. 여기에는 <Positive>, <Negative>, <IP>, 기타 비언어학적 유형 등이 포함됩니다.

최대 검색 거리. 범주를 생성하기 전에 기술이 얼마나 멀리까지 검색하기를 원하는지를 선택하십시오. 값이 낮을수록 더 적은 수의 결과를 얻게 되지만, 이러한 결과는 잡음이 적고 서로 간에 의미 있게 링크되어 있거나 연관되어 있을 수 있습니다. 값이 높을수록 더 많은 수의 결과를 얻게 되지만 이러한 결과는 믿을 만하지 않거나 적절하지 않을 수 있습니다. 이러한 옵션은 모든 기술에 글로벌하게 적용되지만 동시 발생과 시맨틱 네트워크에 미치는 영향은 상당합니다.

특정 개념 쌍 방지. 프로세스가 출력에 두 개의 개념을 함께 그룹화하거나 쌍으로 만드는 프로세스를 중지하려면 이 선택란을 선택하십시오. 개념 쌍을 작성하거나 관리하려면 **쌍 관리**를 클릭하십시오. 자세한 정보는 121 페이지의 『링크 예외 쌍 관리』 주제를 참조하십시오.

가능한 경우: 단순히 확장할지, 와일드카드를 사용하여 디스크립터를 일반화할지 또는 둘 모두를 사용할지 선택하십시오.

- **확장 및 일반화.** 이 옵션은 선택된 범주를 확장한 다음 디스크립터를 일반화합니다. 일반화하기로 선택하면 제품은 별표 와일드카드를 사용하여 범주에서 일반 범주 규칙을 작성합니다. 예를 들어, [apple tart + .] 및 [apple sauce + .] 등과 같은 여러 디스크립터를 생성하는 대신에 와일드카드를 사용하면 [apple * + .]을 생성할 수 있습니다. 와일드카드를 사용하여 일반화하면 종종 앞서 한 것과 똑같은 수의 레코드 또는 문서 수를 얻게 됩니다. 그러나 이 옵션은 숫자를 줄이고 범주 디스크립터를 단순화하는 장점이 있습

니다. 또한 이 옵션은 새 텍스트 데이터(예를 들어, 세로/파동 연구에서)에서 이러한 범주를 사용하여 더 많은 레코드 또는 문서를 범주화하는 기능을 증가시킵니다.

- **확장만.** 이 옵션은 일반화없이 범주를 확장합니다. 먼저 수동으로 작성된 범주에 대해 **확장만** 옵션을 선택한 다음 **확장 후 일반화** 옵션을 사용하여 동일한 범주를 다시 확장하는 것이 도움이 됩니다.
- **일반화만.** 이 옵션은 범주를 다른 방법으로 확장하지 않고 디스크립터를 일반화합니다.

참고: 이 옵션을 선택하면 **시맨틱 네트워크** 옵션이 사용 안함으로 설정됩니다. 이는 설명을 확장하려고 할 때 **시맨틱 네트워크** 옵션만이 사용 가능하기 때문입니다.

범주 확장을 위한 기타 옵션

적용할 기술을 선택하는 것에 추가로 다음 옵션을 편집할 수 있습니다.

디스크립터를 확장할 기준이 되는 최대 항목 수. 항목(개념, 유형 및 기타 표현식)과 함께 디스크립터를 확장할 때 단일 디스크립터에 추가할 수 있는 최대 항목 수를 정의하십시오. 이 한계를 10으로 설정하면 10개가 넘는 추가 항목은 기존 디스크립터에 추가할 수 없습니다. 추가할 항목이 10개가 넘으면 기술은 10번째가 추가된 후에는 새 항목 추가를 중지합니다. 이를 수행하면 디스크립터 목록이 더 짧아지지만 가장 흥미로운 항목이 먼저 사용되게 하지는 못합니다. **가능한 경우 와일드카드로 일반화** 옵션을 사용하여 품질을 저하시키지 않고도 확장의 크기를 줄이는 것을 선호할 수 있습니다. 이 옵션은 부울 &(AND) 또는 !(NOT)을 포함하는 디스크립터에만 적용됩니다.

하위 범주도 확장. 이 옵션은 선택한 범주 아래에 있는 모든 하위 범주를 확장합니다.

범주 이름에서 생성된 디스크립터를 사용하여 빈 범주 확장. 이 방법은 0개의 디스크립터가 있는 빈 범주에만 적용됩니다. 범주에 이미 디스크립터가 포함된 경우에는 이 방법으로 확장되지 않습니다. 이 옵션은 범주의 이름을 구성하는 단어를 기반으로 각 범주의 디스크립터를 자동으로 작성하려고 시도합니다. 범주 이름은 이름에 있는 단어가 추출된 개념과 매치하는지를 확인하기 위해 스캔됩니다. 개념이 인식된 경우에는 매치하는 개념 패턴을 찾는 데 사용되고 이들 모두는 범주의 디스크립터를 형성하는 데 사용됩니다. 이 옵션은 범주 이름이 둘 모두 길고 설명적인 경우에 최상의 결과를 생성합니다. 이는 범주가 이러한 디스크립터를 포함하는 레코드를 캡처할 수 있도록 범주 디스크립터를 생성하기 위한 빠른 방법입니다. 이 옵션은 다른 곳에서 범주를 가져오거나 긴 설명 이름을 사용하여 수동으로 범주를 작성할 때 가장 유용합니다.

디스크립터를 다른 이름으로 생성. 이 옵션은 앞의 옵션이 선택된 경우에만 적용됩니다.

- **개념.** 소스 텍스트에서 추출되었는지 여부와는 관계없이 결과로 나오는 디스크립터를 개념 양식으로 생성하려면 이 옵션을 선택하십시오.
- **패턴.** 결과로 나오는 패턴 또는 임의의 패턴이 추출되었는지 여부와 관계없이 결과로 나오는 디스크립터를 패턴 양식으로 생성하려면 이 옵션을 선택하십시오.

수동으로 범주 작성

자동화된 범주 작성 기술, 및 규칙 편집기를 사용하여 범주를 작성하는 데에 추가로 범주를 수동으로 작성할 수도 있습니다. 다음과 같은 수동 방법이 존재합니다.

- 요소를 하나씩 추가할 빈 범주 작성. 자세한 정보는 『범주 새로 작성 또는 이름 변경』의 내용을 참조하십시오.
- 용어, 유형 및 패턴을 범주 분할창으로 끌기. 자세한 정보는 『끌어서 놓기로 범주 작성』의 내용을 참조하십시오.

범주 새로 작성 또는 이름 변경

개념과 유형을 추가하기 위한 빈 범주를 작성할 수 있습니다. 범주의 이름을 변경할 수도 있습니다.

빈 범주를 새로 작성

1. 범주 분할창으로 이동하십시오.
2. 메뉴에서 **범주 > 빈 범주 작성**을 선택하십시오. 범주 특성 대화 상자가 열립니다.
3. 이름 필드에 이 범주의 이름을 입력하십시오.
4. 이름을 승인하려면 **확인**을 클릭하고 대화 상자를 닫으십시오. 대화 상자가 닫히고 새 범주 이름이 분할창에 나타납니다.

이제 이 범주에 추가를 시작할 수 있습니다. 자세한 정보는 151 페이지의 『디스크립터를 범주에 추가』의 내용을 참조하십시오.

범주 이름 변경

1. 범주를 선택하고 **범주 > 범주 이름 변경**을 선택하십시오. 범주 특성 대화 상자가 열립니다.
2. 이름 필드에 이 범주의 새 이름을 입력하십시오.
3. 이름을 승인하려면 **확인**을 클릭하고 대화 상자를 닫으십시오. 대화 상자가 닫히고 새 범주 이름이 분할창에 나타납니다.

끌어서 놓기로 범주 작성

끌어서 놓기 기술은 수동이며 알고리즘을 기반으로 하지 않습니다. 다음을 끌어서 범주 분할창에서 범주를 작성할 수 있습니다.

- 추출된 개념, 유형 또는 패턴을 추출 결과 분할창에서 범주 분할창으로.
- 추출된 개념을 데이터 분할창에서 범주 분할창으로.
- 전체 행을 데이터 분할창에서 범주 분할창으로. 그러면 추출된 모든 개념과 해당 행에 포함된 패턴으로 구성된 범주가 작성됩니다.

참고: 추출 결과 분할창은 여러 요소의 끌어서 놓기를 쉽게 할 수 있도록 다중 선택을 지원합니다.

중요! 텍스트로부터 추출되지 않았던 데이터 분할창에서는 개념을 끌어서 놓을 수 없습니다. 데이터에서 발견한 개념의 추출을 강제 실행하려면 유형에 이 개념을 추가해야 합니다. 그런 다음 추출을 다시 실행하십시오. 새 추출 결과에는 방금 추가한 개념이 포함됩니다. 그런 다음 이를 범주에 사용할 수 있습니다. 자세한 정보는 102 페이지의 『유형에 개념 추가』의 내용을 참조하십시오.

끌어서 놓기를 사용하여 범주 작성:

1. 추출 결과 분할창 또는 데이터 분할창에서 하나 이상의 개념, 패턴, 유형, 레코드 또는 부분 레코드를 선택하십시오.
2. 마우스 단추를 누르고 있는 동안 요소를 기존 범주 또는 분할창 영역으로 끌어서 새 범주를 작성하십시오.
3. 요소를 놓으려는 영역에 도달하면 마우스 단추를 놓으십시오. 요소가 범주 분할창에 추가됩니다. 수정된 범주가 특수 배경 색상과 함께 나타납니다. 이 색상은 **범주 피드백 배경**이라 부릅니다. 자세한 정보는 86 페이지의 『옵션 설정』 주제를 참조하십시오.

참고: 결과로 나오는 범주가 자동으로 이름이 지정되었습니다. 이름을 변경하려면 변경할 수 있습니다. 자세한 정보는 131 페이지의 『범주 새로 작성 또는 이름 변경』의 내용을 참조하십시오.

범주에 어떤 레코드가 지정되는지 보려면 범주 분할창에서 해당 범주를 선택하십시오. 데이터 분할창이 자동으로 새로 고쳐지고 해당 범주의 모든 레코드를 표시합니다.

범주 규칙 사용

여러 가지 방법으로 범주를 작성할 수 있습니다. 이러한 방법 중 하나는 아이디어를 표현하기 위해 범주 규칙을 정의하는 것입니다. 범주 규칙은 문서 또는 레코드를 추출된 개념, 유형 및 패턴뿐만 아니라 부울 연산자를 사용하여 논리적 표현식을 기반으로 범주에 자동으로 분류하는 명령문입니다. 예를 들어, *추출된 개념* `embassy` *을 포함하지만* `argentina`*는 포함하지 않는 모든 레코드를 이 범주에 포함*을 의미하는 표현식을 작성할 수 있습니다.

동시 발생 및 *개념 루트 파생*(**범주 > 작성 설정 > 고급 설정: 언어학적**) 등과 같은 그룹화 기술을 사용하여 범주를 작성할 때 몇몇 범주 규칙은 자동으로 생성되지만 규칙 편집기에서 데이터와 컨텍스트의 범주 이해를 사용하여 범주 규칙을 수동으로 작성할 수도 있습니다. 각 규칙은 규칙과 매치하는 각 문서 또는 레코드가 해당 범주에 기록될 수 있도록 단일 범주에 첨부됩니다.

범주 규칙은 반응을 특이도를 사용하여 범주화할 수 있게 하여 텍스트 마이닝 결과의 품질과 생산성 및 보다 양적인 분석을 개선하는 데 도움을 줍니다. 사용자의 경험과 비즈니스 지식은 데이터와 컨텍스트의 특정 이해를 제공할 수 있습니다. 추출된 요소를 부울 논리와 결합하여 문서 또는 레코드를 보다 효율적이고 정확하게 범주화하기 위해 이 이해를 활용하여 해당 지식을 범주 규칙으로 변환할 수 있습니다.

이러한 규칙을 작성하는 기능은 비즈니스 지식을 제품의 추출 기술로 계층화할 수 있도록 허용하여 코드 정확도, 효율성 및 생산성을 개선합니다.

참고: 규칙이 텍스트와 매치하는 방법에 대한 예제는 138 페이지의 『범주 규칙 예제』의 내용을 참조하십시오.

범주 규칙 구문

동시 발생 및 *개념 루트 파생*(**범주 > 작성 설정 > 고급 설정: 언어학적**) 등과 같은 그룹화 기술을 사용하여 범주를 작성할 때 몇몇 범주 규칙은 자동으로 생성되지만 규칙 편집기에서 범주 규칙을 수동으로 작성할 수도 있습니다. 각 규칙은 단일 범주의 디스크립터입니다. 그러므로 규칙과 매치하는 각 문서 또는 레코드가 해당 범주에 자동으로 기록됩니다.

참고: 규칙이 텍스트와 매치하는 방법에 대한 예제는 138 페이지의 『범주 규칙 예제』의 내용을 참조하십시오.

규칙을 작성하거나 편집할 때 규칙이 규칙 편집기에서 열려 있어야 합니다. 개념, 유형 또는 패턴을 추가하거나 와일드카드를 사용하여 매치를 확장할 수 있습니다. 추출된 개념, 유형 및 패턴을 사용하면 이는 관련된 모든 개념을 찾으므로 장점이 있습니다.

중요! 일반적인 오류를 피하려면 개념을 추출 결과 분할창, 텍스트 링크 분석 분할창 또는 데이터 분할창에서 직접 규칙 편집기로 끌어다 놓거나 가능한 경우 컨텍스트 메뉴를 통해 이를 추가하는 것이 좋습니다.

개념, 유형 및 패턴이 인식되면 아이콘이 텍스트 옆에 나타납니다.

표 18. 추출 아이콘

아이콘	설명
	추출된 개념
	추출된 유형
	추출된 패턴

규칙 구문 및 연산자

다음 테이블은 규칙 구문을 정의할 때 사용할 문자를 포함합니다. 이러한 문자를 개념, 유형 및 패턴과 함께 사용하여 사용자만의 규칙을 작성하십시오.

표 19. 지원되는 구문

문자	설명

&	"and" 부울. 예를 들어, `a & b`에는 다음과 같은 `a` 및 `b` 둘 모두가 포함됩니다.
	`- invasion & united states`
	`- 2016 & olympics`
	`- good & apple`

|	"or" 부울은 포함이며 이는 요소의 일부 또는 모두가 발견되면 매치가 이뤄짐을 의미합니다. 예를 들어, `a | b`에는 다음과 같은 `a` 또는 `b`가 포함됩니다.
	`- attack | france`
	`- condominium | apartment`

!()	"not" 부울. 예를 들어, `!(a)`에는 `a`를 포함하지 않습니다. 예: `!(good & hotel) , assassination & !(austria)` 또는 `!(gold) & !(copper)`
*	사용 방법에 따라 단일 문자부터 전체 단어까지 모든 것을 표현하는 와일드카드. 자세한 정보는 136 페이지의 『범주 규칙에서 와일드카드 사용』의 내용을 참조하십시오.
()	표현식 구분자. 괄호 안에 있는 표현식이 먼저 평가됩니다.
+	순서 특정 패턴을 형성하는 데 사용된 패턴 연결자. 이 패턴 연결자가 있으면 꺾쇠 대괄호가 사용되어야 합니다. 자세한 정보는 134 페이지의 『범주 규칙에서 TLA 패턴 사용』의 내용을 참조하십시오.

표 19. 지원되는 구문 (계속)

문자	설명
[]	범주 규칙 내에서 추출된 TLA 패턴을 기반으로 매치를 찾고 있는 경우에는 패턴 구분자는 필수입니다. 대괄호 내의 컨텐츠는 TLA 패턴을 가리키고 단순 동시 발생을 기반으로 하는 개념이나 유형과는 매치하지 않습니다. 이 TLA 패턴을 추출하지 않은 경우에는 매치가 가능하지 않습니다. 자세한 정보는 『범주 규칙에서 TLA 패턴 사용』의 내용을 참조하십시오. 패턴 대신에 개념과 유형 매치를 찾고 있다면 꺾쇠 대괄호를 사용하지 마십시오. 참고: 이전 버전에서는 범주 작성 기술을 사용하여 생성된 동시 발생 및 동의어 규칙은 꺾쇠 대괄호로 둘러싸여 있었습니다. 모든 신규 버전에서는 꺾쇠 대괄호는 TLA 패턴의 존재를 나타냅니다. 대신, 동시 발생 기술과 동의어를 사용하여 생성된 규칙은 괄호로 캡슐화되어 있습니다(예: `(speaker systems|speakers)`).

`&` 및 `|` 연산자는 가환적입니다(예: `a & b = b & a` 및 `a | b = b | a`).

문자를 백슬래시로 이스케이프

마찬가지로 구문 문자인 모든 문자를 포함하는 개념이 있는 경우에는 규칙이 제대로 해석되게 하려면 해당 문자 앞에 백슬래시를 놓아야 합니다. 백슬래시(\) 문자는 백슬래시를 사용하지 않을 경우에는 특별한 의미를 가지고 있는 문자를 이스케이프 처리하는 데 사용됩니다. 편집기에 끌어서 놓으면 백슬래시가 자동으로 추가됩니다.

규칙 구문 문자를 규칙 구문이 아닌 것처럼 처리하려면 규칙 구문 문자 앞에 백슬래시가 와야 합니다.

`& ! | + < > ( ) [ ] *`

예를 들어, 개념 `r&d`에 "and" 연산자(&)가 포함되므로 이를 규칙 편집기에 입력할 때 백슬래시가 필요합니다(예: `r\&d`).

범주 규칙에서 TLA 패턴 사용

텍스트 링크 분석 패턴은 보다 특정적이고 컨텍스트상 결과를 얻을 수 있도록 범주 규칙에 명시적으로 정의될 수 있습니다. 범주 규칙에서 패턴을 정의할 때 더 많은 단순 개념 추출 결과와 추출된 텍스트 링크 분석 패턴 결과를 기반으로 매치하는 문서와 레코드만을 무시하고 있습니다.

중요! 범주 규칙에서 TLA 패턴을 사용하여 문서를 매치시키려면 텍스트 링크 분석이 사용 가능한 상태에서 추출을 실행해야 합니다. 범주 규칙은 해당 프로세스 동안에 발견된 매치를 찾습니다. 텍스트 마이닝 노드의 모델 탭에서 TLA 결과를 탐색하려고 선택하지 않은 경우에는 대화식 세션 내에서 추출 설정에서 TLA 추출을 사용 가능으로 선택한 다음에 다시 추출할 수 있습니다. 자세한 정보는 92 페이지의 『데이터 추출』의 내용을 참조하십시오.

꺾쇠 대괄호로 구분. 범주 규칙 내에서 TLA 패턴을 사용 중이라면 이를 꺾쇠 대괄호 [ ]로 둘러싸야 합니다. 추출된 TLA 패턴을 기반으로 매치를 찾고 있는 경우에는 패턴 구분자는 필수입니다. 범주 규칙은 유형, 개념 또는 패턴을 포함할 수 있으므로, 대괄호는 대괄호 내의 컨텐츠가 추출된 TLA 패턴을 가리킨다는 점을 규칙에 분명히 합니다. 이 TLA 패턴을 추출하지 않은 경우에는 매치가 가능하지 않습니다. 범주 분할창에서 `apple + good` 등과 같이 대괄호가 없는 패턴을 보는 경우에는 패턴이 범주 규칙 편집기 외부의 범주에 직접 추가되었음을 의미합니다. 예를 들어, 개념 패턴을 텍스트 링크 분석 보기 에서 범주에 직접 추가하는 경우에는 이는

꺾쇠 대괄호와 함께 나타나지 않습니다. 그러나 범주 규칙 내의 패턴을 사용할 때는 범주 규칙 내에서 패턴을 꺾쇠 대괄호 내에 캡슐화해야 합니다(예: [banana + !(good)]).

패턴에서 + 부호 사용. IBM SPSS Modeler Text Analytics에서 최대 6개의 파트 또는 -슬롯, 패턴이 있을 수 있습니다. 순서가 중요함을 나타내려면 + 부호를 사용하여 각 요소를 연결하십시오(예: [company1 + acquired + company2]). 회사가 획득하고 있는 의미를 변경할 수 있으므로 여기에서는 순서가 중요합니다. 순서는 문장 구조로 결정되지 않지만 TLA 패턴 출력이 구조화되는 방법으로 결정됩니다. 예를 들어, "*I love Paris*" 텍스트가 있고 이 아이디어를 추출하려면, TLA 패턴은 [<Positive> + <Location>]가 아니라 [paris + like] 또는 [<Location> + <Positive>]일 수 있습니다. 기본 의견 자원은 일반적으로 의견을 2개의 파트로 된 패턴에서 두 번째 위치에 놓기 때문입니다. 따라서 문제를 피하기 위해서는 범주에서 패턴을 디스크립터로 직접 사용하는 것이 도움이 될 수 있습니다. 그러나 패턴을 보다 복잡한 명령문의 일부로 사용해야 하는 경우에는 텍스트 링크 분석 보기에 있는 패턴 내의 요소의 순서에 특히 주의를 기울이십시오. 순서는 매치를 발견할 수 있는지 여부에 커다란 역할을 하기 때문입니다.

예를 들어, 표현식 "I *like pineapple*"와 "I hate *pineapple*. However, I *like* strawberries"가 있는 두 개의 샘플 텍스트가 있다고 가정해 봅시다. like & pineapple 표현식은 개념 표현식이고 텍스트 링크 규칙이 아니므로(대괄호로 둘러싸이지 않음) 두 텍스트 모두와 매치합니다. 표현식 pineapple + like는 "I *like pineapple*"와만 매치합니다. 두 번째 텍스트에서 단어 *like*는 대신 *strawberries*와 연관되기 때문입니다.

패턴으로 그룹화. 사용자만의 패턴을 사용하여 규칙을 단순화할 수 있습니다. 다음과 같이 cayenne peppers + like, chili peppers + like 및 peppers + like의 세 개의 표현식을 캡처한다고 가정합니다. 이를 단일 범주 그룹으로 그룹화할 수 있습니다(예: [* peppers & like]). 또 다른 표현식 hot peppers + good이 있는 경우에는 이들 네 개를 규칙으로 그룹화할 수 있습니다(예: [* peppers + <Positive>]).

패턴에서의 순서. 출력을 보다 잘 구성하기 위해서 제품과 함께 설치한 템플리트에 제공된 텍스트 링크 분석 규칙은 문장에서의 단어 순서와는 관계 없이 동일한 순서로 기본 패턴을 출력하려고 시도합니다. 예를 들어, "*Good presentations.*" 텍스트를 포함하는 레코드와 "*the presentations were good*"을 포함하는 또 다른 레코드가 있는 경우에는 두 텍스트 모두 동일한 규칙으로 매치하고 개념 패턴 결과에서 presentation + good 및 good + presentation이 아닌 presentation + good과 동일한 순서로 출력됩니다. 예제에서처럼 두 개의 슬롯 패턴에서는 Opinions 라이브러리에서 유형에 지정된 개념은 기본적으로 apple + bad에서와 같이 출력에서 마지막에 제공됩니다.

표 20. 패턴 구문 및 부울 사용법

표현식	문서 또는 레코드와 매치합니다.
[ ]	모든 TLA 패턴을 포함합니다. 추출된 TLA 패턴을 기반으로 매치를 찾고 있는 경우에는 패턴 구분자는 *범주 규칙에서* 필수입니다. 대괄호 내의 컨텐츠는 단순 개념과 유형이 아니라 TLA 패턴을 가리킵니다. 이 TLA 패턴을 추출하지 않은 경우에는 매치가 가능하지 않습니다. 패턴을 포함하지 않는 규칙을 작성하려는 경우에는 !([ ])를 사용할 수 있습니다.
[a]	패턴에서의 위치와 관계 없이 하나 이상의 요소가 a인 패턴을 포함합니다. 예를 들어, [deal]은 [deal + good] 또는 단지 [deal + .]와 매치할 수 있습니다.
[a + b]	개념 패턴을 포함합니다. 예를 들어, [deal + good]입니다. *참고:* 다른 요소를 추가하지 않고 이 패턴을 캡처만 하려는 경우에는 패턴을 사용하여 규칙을 작성하는 대신 범주에 직접 추가하는 것이 좋습니다.

표 20. 패턴 구문 및 부울 사용법 (계속)

표현식	문서 또는 레코드와 매치합니다.
[a + b + c]	개념 패턴을 포함합니다. + 부호는 매치하는 요소의 순서가 중요함을 나타냅니다. 예를 들어, [company1 + acquired + company2]입니다.
[<A> + <B>]	첫 번째 슬롯에 <A> 유형이 있는 패턴과 두 번째 슬롯에 <B> 유형이 있는 패턴을 포함하고 정확히 두 개의 슬롯이 있습니다. + 부호는 매치하는 요소의 순서가 중요함을 나타냅니다. 예를 들어, [<Budget> + <Negative>]입니다. *참고:* 다른 요소를 추가하지 않고 이 패턴을 캡처만 하려는 경우에는 패턴을 사용하여 규칙을 작성하는 대신 범주에 직접 추가하는 것이 좋습니다.
[<A> & <B>]	<A> 유형 및 <B> 유형이 있는 모든 유형 패턴을 포함합니다. 예를 들어, [<Budget> & <Negative>]입니다. 이 TLA 패턴은 결코 추출되지 않습니다. 그러나 이렇게 작성되면 정말로 [<Budget> + <Negative>]|[<Negative> + <Budget>]과 동등합니다. 매치 요소의 순서는 중요하지 않습니다. 또한 다른 요소는 패턴에 있을 수 있지만 하나 이상의 <Budget> 및 <Negative>가 있어야 합니다.
[a + .]	a가 유일한 개념이고 해당 패턴의 다른 슬롯에 아무것도 없는 패턴을 포함합니다. 예를 들어, [deal + .]는 유일한 출력이 deal인 개념 패턴과 매치합니다. 개념 deal을 범주 디스크립터로서 추가한 경우에는 deal에 대한 긍정적인 명령문을 포함하여 deal이 있는 모든 레코드를 개념으로서 얻을 수 있습니다. 그러나 [deal + .]을 사용하면 deal을 표현하는 레코드 패턴 결과와만 매치하고 다른 관계나 의견과는 매치하지 않고 deal + fantastic과도 매치하지 않습니다. *참고:* 다른 요소를 추가하지 않고 이 패턴을 캡처만 하려는 경우에는 패턴을 사용하여 규칙을 작성하는 대신 범주에 직접 추가하는 것이 좋습니다.
[<A> + <>]	<A>가 유일한 유형인 패턴을 포함합니다. 예를 들어, [<Budget> + <>]는 유일한 출력이 <Budget> 유형의 개념인 패턴과 매치합니다. *참고:* <> 를 사용하여 유형 패턴에서 패턴 + 기호 뒤에 이를 배치할 때만 빈 유형을 나타낼 수 있습니다 (예: [<Budget> + <>], [price + <>]는 아님). *참고:* 다른 요소를 추가하지 않고 이 패턴을 캡처만 하려는 경우에는 패턴을 사용하여 규칙을 작성하는 대신 범주에 직접 추가하는 것이 좋습니다.
[a + !(b)]	개념 a를 포함하지만 개념 b를 포함하지 않는 하나 이상의 패턴을 포함합니다. 하나 이상의 패턴을 포함해야 합니다. 예를 들어, [price + !(high)] 또는 유형의 경우 [!(<Fruit>|<Vegetable>) + <Positive>]
!([<A> & <B>])	특정 패턴을 포함하지 않습니다. 예를 들어, !([<Budget> & <Negative>])입니다.

참고: 규칙이 텍스트와 매치하는 방법에 대한 예제는 138 페이지의 『범주 규칙 예제』의 내용을 참조하십시오.

범주 규칙에서 와일드카드 사용

와일드카드는 매치하는 기능을 확장하기 위해 규칙에서 개념에 추가할 수 있습니다. 별표 * 와일드카드는 개념이 어떻게 매치하는지를 표시하기 위해 단어 앞이나 뒤에 놓을 수 있습니다. 와일드카드 사용에는 두 개의 유형이 있습니다.

- **첨부 와일드카드**. 이러한 와일드카드는 문자열과 별표를 구분하는 공백 없이 접두문자나 접미문자를 바로 붙입니다. 예를 들어, operat*는 *operat*, *operate*, *operates*, *operations*, *operational* 등과 매치할 수 있습니다.
- **단어 와일드카드**. 이러한 와일드카드는 개념과 별표 사이의 공백을 사용하여 개념에 접두문자나 접미문자를 추가합니다. 예를 들어, * operation은 *operation*, *surgical operation*, *post operation* 등과 매치할 수

있습니다. 또한 단어 와일드카드는 첨부 와일드카드와 나란히 사용할 수 있습니다. 예를 들어, * operat* *는 *operation*, *surgical operation*, *telephone operator*, *operatic aria* 등과 매치할 수 있습니다. 이 마지막 예에서 볼 수 있듯이 그물을 너무 넓게 던져서 원하지 않는 매치를 캡처하지 않도록 와일드카드를 주의깊게 사용하는 것이 좋습니다.

예외!

- 와일드카드는 단독으로 사용될 수 없습니다. 예를 들어, (apple | *)는 허용되지 않습니다.
- 와일드카드는 유형 이름과 매치시키는 데 사용할 수 없습니다. <Negative*>는 어떤 유형 이름과도 매치하지 않습니다.
- 특정 유형을 와일드카드를 통해 발견된 개념과 매치하지 않도록 필터링할 수 없습니다. 개념이 지정된 유형이 자동으로 사용됩니다.
- 와일드카드는 단어의 끝이나 시작이거나 관계 없이 단어 순서의 중간에 오거나(open* account) 독립된 구성요소(open * account)일 수 없습니다. 와일드카드는 유형 이름에도 사용할 수 없습니다. 예를 들어, word* word(예: apple* recipe)는 applesauce recipe 또는 다른 어떤 것과도 매치하지 않습니다. 그러나 apple* *는 *applesauce recipe*, *apple pie*, *apple* 등과 매치합니다. 다른 예제에서 word * word(예: apple * toast)는 *apple cinnamon toast* 또는 다른 어떤 것과도 매치하지 않습니다. 별표가 두 개의 다른 단어 사이에 나타나기 때문입니다. 그러나 apple *는 *apple cinnamon toast*, *apple*, *apple pie* 등과 매치합니다.

표 21. 와일드카드 사용법

표현식	문서 또는 레코드와 매치합니다.
*apple	작성된 글자로 끝나지만 다른 여러 글자가 접두문자로 있을 수 있는 개념을 포함합니다. 예를 들어, *apple은 *apple* 글자로 끝나지만 다음과 같은 접두문자를 사용할 수 있습니다. - apple - pineapple - crabapple
apple*	작성된 글자로 시작하지만 다른 여러 글자가 접미문자로 있을 수 있는 개념을 포함합니다. 예를 들어, apple*는 글자 *apple*로 시작하지만 접미문자를 사용하거나 사용하지 않을 수 있습니다. 예: - apple - applesauce - applejack 예를 들어, apple* & !(pear* | quince)은 글자 apple로 시작하는 개념을 포함하지만 글자 *pear*로 시작하는 개념이나 quince 개념을 포함하지 않고 apple & quince와 매치하지 않습니다. 그러나 다음과는 매치할 수 있습니다. - applesauce - apple & orange
product	작성된 글자 product를 포함하지만 접두문자나 접미문자 또는 둘 모두로 여러 글자가 사용되고 있을 수 있는 개념을 포함합니다. 예를 들어, *product*는 다음과 매치할 수 있습니다. - product - byproduct - unproductive

표 21. 와일드카드 사용법 (계속)

표현식	문서 또는 레코드와 매치합니다.
* loan	단어 loan을 포함하지만 앞에 다른 단어가 있는 복합어일 수 있는 개념을 포함합니다. 예를 들어, * loan은 다음과 매치할 수 있습니다. - loan - car loan - home equity loan 예를 들어, [* delivery + <Negative>]는 첫 번째 위치에서 단어 delivery로 끝나는 개념을 포함하고 두 번째 위치에서 <Negative> 유형을 포함하고 다음 개념 패턴과 매치할 수 있습니다. - package delivery + slow - overnight delivery + late
event *	단어 event를 포함하지만 다른 단어가 따라오는 복합어일 수 있는 개념을 포함합니다. 예를 들어, event *는 다음과 매치할 수 있습니다. - event - event location - event planning committee
* apple *	또 다른 단어가 따라올 가능성이 있고 단어 apple이 따라오는 단어로 시작할 수 있는 개념을 포함합니다. * 는 0 또는 n을 의미하므로 이는 또한 apple과 매치합니다. 예를 들어, * apple *는 다음과 매치할 수 있습니다. - gala applesauce - granny smith apple crumble - famous apple pie - apple 예를 들어, [* reservation* * + <Positive>]는 단어 reservation(개념에 있는지 여부와 관계 없이)이 첫 번째 위치에 있는 개념을 포함하고 두 번째 위치에 유형 <Positive>를 포함하고 개념 패턴과 매치할 수 있습니다. - reservation system + good - online reservation + good

참고: 규칙이 텍스트와 매치하는 방법에 대한 예제는 『범주 규칙 예제』의 내용을 참조하십시오.

범주 규칙 예제

규칙이 이를 표현하는 데 사용된 구문에 따라 다르게 레코드에 어떻게 매치하는지를 시연하려면 다음 예제를 고려하십시오.

예제 레코드

두 개의 레코드가 있다고 상상해 보십시오.

- **레코드 A**: *"when I checked my wallet, I saw I was missing 5 dollars."*
- **레코드 B**: *"$5 was found at the picnic area, but the blanket was missing."*

다음 두 개의 테이블은 개념과 유형뿐만 아니라 개념 패턴과 유형 패턴에서 무엇이 추출될 수 있는지를 보여줍니다.

예제에서 추출된 개념 및 유형

표 22. 추출된 개념 및 유형 예제

추출된 개념	개념 유형
wallet	<Unknown>
missing	<Negative>
USD5	<Currency>
blanket	<Unknown>
picnic area	<Unknown>

예제에서 추출된 TLA 패턴

표 23. 예제 추출된 TLA 패턴 출력

추출된 개념 패턴	추출된 유형 패턴	시작 레코드
picnic area + .	<Unknown> + <>	레코드 B
wallet + .	<Unknown> + <>	레코드 A
blanket + missing	<Unknown> + <Negative>	레코드 B
USD5 + .	<Currency> + <>	레코드 B
USD5 + missing	<Currency> + <Negative>	레코드 A

범주 규칙이 매치하는 방법

다음 테이블에는 범주 규칙 편집기에 입력할 수 있는 몇몇 구문이 포함됩니다. 여기에 있는 모든 규칙이 작동하는 것은 아니며 모두 동일 레코드와 매치하는 것은 아닙니다. 서로 다른 구문이 매치된 레코드에 어떤 영향을 미치는지를 확인하십시오.

표 24. 표본 규칙

규칙 구문	결과
USD5 & missing	레코드 A와 B 둘 모두 추출된 개념 missing과 추출된 개념 USD5를 포함하므로 둘 모두 매치합니다. 이는 다음과 동등합니다. (USD5 & missing)
missing & USD5	레코드 A와 B 둘 모두 추출된 개념 missing과 추출된 개념 USD5를 포함하므로 둘 모두 매치합니다. 이는 다음과 동등합니다. (missing & USD5)
missing & <Currency>	레코드 A와 B 둘 모두 추출된 개념 missing과 <Currency> 유형과 매치하는 개념을 포함하므로 둘 모두 매치합니다. 이는 다음과 동등합니다. (missing & <Currency>)
<Currency> & missing	레코드 A와 B 둘 모두 추출된 개념 missing과 <Currency> 유형과 매치하는 개념을 포함하므로 둘 모두 매치합니다. 이는 다음과 동등합니다. (<Currency> & missing)
[USD5 + missing]	레코드 B는 USD5 + missing을 포함하는 TLA 패턴 출력을 생성하지 않으므로 A와는 매치하지만 B와는 매치하지 **않습니다**(이전 테이블 참조). 이는 TLA 패턴 출력과 동등합니다. USD5 + missing

표 24. 표본 규칙 (계속)

규칙 구문	결과
[missing + USD5]	추출된 TLA 패턴이 여기에 첫 번째 위치에 missing으로 표현된 순서와 매치하지 않으므로(이전 테이블 참조) 레코드 A나 B와 매치하지 않습니다. 이는 TLA 패턴 출력과 동등합니다. USD5 + missing
[missing & USD5]	이러한 TLA 패턴이 레코드 B로부터 추출되지 않았으므로 A와는 매치하지만 B와는 매치하지 **않습니다**. & 문자를 사용하면 매치할 때 순서가 중요하지 않음을 나타냅니다. 그러므로 이 규칙은 [missing + USD5] 또는 [USD5 + missing]에 대한 패턴 매치를 찾습니다. 레코드 A의 [USD5 + missing]만이 매치합니다.
[missing + <Currency>]	추출된 TLA 패턴이 이 순서와 매치하지 않았으므로 레코드 A나 B와 매치하지 않습니다. TLA 출력은 용어 (USD5 + missing) 또는 유형 (<Currency> + <Negative>)만을 기반으로 하지만 개념과 유형을 혼합하지 않으므로 동등한 것이 없습니다.
[<Currency> + <Negative>]	TLA 패턴이 레코드 B에서 추출되지 않았으므로 레코드 A와는 매치하지만 B와는 매치하지 않습니다. 이는 TLA 출력과 동등합니다. <Currency> + <Negative>
[<Negative> + <Currency>]	추출된 TLA 패턴이 이 순서와 매치하지 않았으므로 레코드 A나 B와 매치하지 않습니다. Opinions 템플리트에서 기본적으로 *주제*가 *의견*을 사용하여 발견되면 *주제*(<Currency>)는 첫 번째 슬롯 위치를 차지하고 *의견*(<Negative>)은 두 번째 슬롯 위치를 차지합니다.

범주 규칙 작성

규칙을 작성하거나 편집할 때 규칙이 규칙 편집기에서 열려 있어야 합니다. 개념, 유형 또는 패턴을 추가하거나 와일드카드를 사용하여 매치를 확장할 수 있습니다. 인식된 개념, 유형 및 패턴을 사용하면 이는 관련된 모든 개념을 찾으므로 장점이 있습니다. 예를 들어, 개념을 사용하면 연관된 모든 용어, 복수 형식 및 동의어는 또한 규칙과 매치합니다. 마찬가지로, 유형을 사용하면 모든 해당 개념 또한 규칙에 의해 캡처됩니다.

기존 규칙을 편집하거나 범주 이름을 마우스 오른쪽 단추로 클릭하고 **규칙 작성**을 선택하여 규칙 편집기를 열 수 있습니다.

컨텍스트 메뉴, 끌어다 놓기를 사용하거나 개념, 유형 및 패턴을 편집기에 수동으로 입력할 수 있습니다. 그런 다음 이들을 부울 연산자(&, !(), |) 및 대괄호와 결합하여 규칙 표현식을 작성하십시오. 일반적인 오류를 피하려면 개념을 추출 결과 분할창 또는 데이터 분할창에서 직접 규칙 편집기로 끌어다 놓는 것이 좋습니다. 오류를 피하려면 규칙의 구문에 세심한 주의를 기울이십시오. 자세한 정보는 132 페이지의 『범주 규칙 구문』의 내용을 참조하십시오.

참고: 규칙이 텍스트와 매치하는 방법에 대한 예제는 138 페이지의 『범주 규칙 예제』의 내용을 참조하십시오.

규칙 작성

1. 아직 데이터를 추출하지 않았거나 추출이 오래된 경우에는 지금 수행하십시오. 자세한 정보는 92 페이지의 『데이터 추출』의 내용을 참조하십시오.

참고: 더 이상 표시되는 개념이 없는 방식으로 추출을 필터링하는 경우에는 범주 규칙을 작성하거나 편집하려고 시도할 때 오류 메시지가 표시됩니다. 이를 방지하려면 개념을 사용 가능하도록 추출 필터를 수정하십시오.

2. 범주 분할창에서 규칙을 추가하려는 범주를 선택하십시오.
3. 메뉴에서 **범주 > 규칙 작성**을 선택하십시오. 범주 규칙 편집기 분할창이 창에서 열립니다.
4. 규칙 이름 필드에 규칙의 이름을 입력하십시오. 이름을 제공하지 않으면 표현식이 자동으로 이름으로 사용됩니다. 나중에 이 규칙의 이름을 변경할 수 있습니다.
5. 더 큰 표현식 텍스트 필드에서 다음을 수행할 수 있습니다.
   - 필드에 텍스트를 직접 입력하거나 다른 분할창에서 끌어다 놓으십시오. 추출된 개념, 유형 및 패턴만을 사용하십시오. 예를 들어, 단어 `cats`를 입력했는데 단수형 `cat`만이 추출 결과 분할창에 나타나는 경우에는 편집기는 `cats`를 인식할 수 없습니다. 이 마지막 케이스에서 단수형은 자동으로 복수형을 포함할 수도 있지만 그렇지 않은 경우에는 와일드카드를 사용할 수 있습니다. 자세한 정보는 132 페이지의 『범주 규칙 구문』의 내용을 참조하십시오.
   - 규칙에 추가하려는 개념, 유형 또는 패턴을 선택하고 메뉴를 사용하십시오.
   - 규칙의 요소를 서로 링크하려면 부울 연산자를 추가하십시오. 도구 모음 단추를 사용하여 "and" 부울 **&**, "or" 부울 **|**, "not" 부울 **!**(), 괄호 () 및 패턴의 대괄호 **[ ]**를 규칙에 추가하십시오.
6. **규칙 테스트** 단추를 클릭하여 규칙이 잘 형성되었는지 확인하십시오. 자세한 정보는 132 페이지의 『범주 규칙 구문』의 내용을 참조하십시오. 발견된 문서 또는 레코드 수가 텍스트 **테스트 결과** 옆의 괄호에 나타납니다. 이 텍스트 오른쪽에는 규칙에서 인식되었던 요소 또는 오류 메시지를 볼 수 있습니다. 유형, 패턴 또는 개념 옆의 그래픽이 빨간색 물음표와 함께 나타나는 경우에는 이는 요소가 알려진 추출과 매치하지 않음을 나타냅니다. 매치하지 않으면 규칙은 레코드를 발견하지 않습니다.
7. 규칙의 일부를 테스트하려면 해당 파트를 선택하고 **테스트 선택**을 클릭하십시오.
8. 문제점을 발견한 경우 필요한 변경사항을 수행하고 규칙을 다시 테스트하십시오.
9. 완료되면 **저장 & 닫기**를 클릭하여 규칙을 다시 저장하고 편집기를 닫으십시오. 새 규칙 이름이 범주에 나타납니다.

규칙 편집 및 삭제

규칙을 작성하고 저장한 후에는 언제든지 그 규칙을 편집할 수 있습니다. 자세한 정보는 132 페이지의 『범주 규칙 구문』의 내용을 참조하십시오.

규칙을 더 이상 원하지 않으면 이를 삭제할 수 있습니다.

규칙 편집하기

1. 범주 정의 대화 상자의 디스크립터 테이블에서 규칙을 선택하십시오.
2. 메뉴에서 **범주 > 규칙 편집**을 선택하거나 규칙 이름을 두 번 클릭하십시오. 편집기가 선택된 규칙이 열린 상태로 열립니다.
3. 추출 결과와 도구 모음 단추를 사용하여 규칙을 변경하십시오.

4. 예측한 결과를 리턴하게 하려면 규칙을 다시 테스트하십시오.
5. **저장 & 닫기**를 클릭하여 규칙을 다시 저장하고 편집기를 닫으십시오.

규칙 삭제하기

1. 범주 정의 대화 상자의 디스크립터 테이블에서 규칙을 선택하십시오.
2. 메뉴에서 **편집 > 삭제**를 선택하십시오. 규칙이 범주에서 삭제됩니다.

사전 정의된 범주 가져오기 및 내보내기

Microsoft Excel(*.*xls*, *.*xlsx*) 파일에 사용자 고유의 범주가 저장된 경우에는 이를 IBM SPSS Modeler Text Analytics 로 가져올 수 있습니다.

열려 있는 대화식 워크벤치 세션 에 있는 범주를 Microsoft Excel(*.*xls*, *.*xlsx*) 파일로 내보낼 수도 있습니다. 범주를 내보낼 때 디스크립터 및 스코어 등과 같은 몇몇 추가 정보를 포함하거나 제외할지를 선택할 수 있습니다. 자세한 정보는 146 페이지의 『범주 내보내기』의 내용을 참조하십시오.

사전 정의된 범주에 코드가 없거나 새 코드를 원하는 경우에는 메뉴에서 **범주 > 범주 관리 > 코드 자동 생성**을 선택하여 범주 분할창에서 범주 세트에 대한 새 코드 세트를 자동으로 생성할 수 있습니다. 그러면 기존 코드가 제거되고 모두 자동으로 다시 번호가 지정됩니다.

사전 정의된 범주 가져오기

사전 정의된 범주를 IBM SPSS Modeler Text Analytics 로 가져올 수 있습니다. 가져오기 전에 사전 정의된 범주 파일이 Microsoft Excel(*.*xls*, *.*xlsx*) 파일에 있고 지원 형식 중 하나로 구조화되었는지를 확인하십시오. 제품이 자동으로 형식을 발견하도록 선택할 수도 있습니다. 다음 형식이 지원됩니다.

- **평면 목록 형식**: 자세한 정보는 144 페이지의 『평면 목록 형식』 주제를 참조하십시오.
- **최소 형식**: 자세한 정보는 144 페이지의 『최소 형식』 주제를 참조하십시오.
- **들여쓰기 형식**: 자세한 정보는 145 페이지의 『들여쓰기 형식』 주제를 참조하십시오.

사전에 정의된 범주 가져오기

1. 대화식 워크벤치 메뉴에서 **범주 > 범주 관리 > 사전 정의된 범주 가져오기**를 선택하십시오. 사전 정의된 범주 가져오기 마법사가 표시됩니다.
2. 찾아보기 드롭 다운 목록에서 파일이 있는 드라이브와 폴더를 선택하십시오.
3. 목록에서 파일을 선택하십시오. 파일 이름이 파일 이름 텍스트 상자에 나타납니다.
4. 목록에서 사전 정의된 범주를 포함하는 워크시트를 선택하십시오. 워크시트 이름이 워크시트 필드에 나타납니다.
5. 데이터 형식 선택을 시작하려면 **다음**을 클릭하십시오.
6. 파일의 형식을 선택하거나 제품이 자동으로 형식을 발견하게 하는 옵션을 선택하십시오. 자동 발견은 대부분의 공통 형식에서 잘 작동합니다.
   - **평면 목록 형식**: 자세한 정보는 144 페이지의 『평면 목록 형식』 주제를 참조하십시오.

- **최소 형식**: 자세한 정보는 144 페이지의 『최소 형식』 주제를 참조하십시오.
- **들여쓰기 형식**: 자세한 정보는 145 페이지의 『들여쓰기 형식』 주제를 참조하십시오.

7. 추가로 가져오기 옵션을 정의하려면 **다음**을 클릭하십시오. 형식을 자동으로 발견하도록 선택하면 최종 단계로 이동됩니다.
8. 하나 이상의 행에 열 헤더 또는 다른 관련 없는 정보가 포함된 경우에는 **행에서 가져오기 시작** 옵션에서 가져오기를 시작할 행 번호를 선택하십시오. 예를 들어, 범주 이름이 7행에서 시작하는 경우에는 파일을 올바르게 가져오려면 이 옵션에 숫자 7을 입력해야 합니다.
9. 파일에 범주 코드가 포함된 경우에는 **범주 코드 포함** 옵션을 선택하십시오. 이를 수행하면 마법사가 데이터를 제대로 인식하는 데 도움이 됩니다.
10. 색상 코딩된 셀과 범례를 검토하여 데이터가 올바르게 식별되었는지 확인하십시오. 파일에서 발견된 오류는 빨간색으로 표시되고 형식 미리보기 테이블 아래에 표시됩니다. 잘못된 형식이 선택된 경우에는 뒤로 돌아가서 또 다른 형식을 선택하십시오. 파일을 수정해야 하는 경우에는 변경을 수행하고 파일을 다시 선택하여 마법사를 다시 시작하십시오. 마법사를 마치기 전에 모든 오류를 수정해야 합니다.
11. 가져올 범주 및 하위 범주 세트를 검토하고 이러한 범주에 대한 디스크립터를 작성하는 방법을 정의하려면 **다음**을 클릭하십시오.
12. 테이블에서 가져올 범주 세트를 검토하십시오. 디스크립터로 표시될 것으로 예상한 키워드가 보이지 않으면 가져오기 중에 인식되지 않은 것일 수도 있습니다. 이러한 키워드에 접두문자가 제대로 추가되었는지와 올바른 셀에 나타나는지를 확인하십시오.
13. 세션에서 사전에 존재하는 범주를 처리하는 방법을 선택하십시오.
    - **기존 범주를 모두 대체**. 이 옵션은 기존 범주를 모두 제거한 다음 그 자리에 새로 가져온 범주가 단독으로 사용됩니다.
    - **기존 범주에 추가**. 이 옵션은 범주를 가져오고 공통된 범주를 기존 범주와 병합합니다. 기존 범주에 추가할 때에는 중복이 처리되는 방법을 결정해야 합니다. 한 가지 선택사항(옵션: **병합**)은 가져오는 범주를 기존 범주와 병합하는 것입니다(범주 이름을 공유하는 경우). 또 다른 선택사항(옵션: **가져오기에서 제외**)은 같은 이름의 범주가 존재하는 경우 범주 가져오기를 금지하는 것입니다.
14. **키워드를 디스크립터로서 가져오기**는 데이터에서 식별된 키워드를 연관된 범주의 디스크립터로서 가져오는 것입니다.
15. **디스크립터를 파생하여 범주 확장**은 범주 이름 또는 하위 범주를 나타내는 단어 및/또는 주석을 구성하는 단어에서 디스크립터를 생성하는 옵션입니다. 단어가 추출된 결과와 매치하면 이들은 디스크립터로서 범주에 추가됩니다. 이 옵션은 범주 이름 또는 주석이 둘 모두 길고 설명적인 경우에 최상의 결과를 생성합니다. 이는 범주가 이러한 디스크립터를 포함하는 레코드를 캡처할 수 있도록 범주 디스크립터를 생성하기 위한 빠른 방법입니다.
    - **시작** 필드를 사용하면 디스크립터가 어떤 텍스트에서 파생될지, 이름 또는 범주 및 하위 범주인지 주석에 있는 단어인지 또는 둘 모두인지를 선택할 수 있습니다.
    - **양식** 필드를 사용하면 이러한 디스크립터를 개념 또는 TLA 패턴의 양식으로 작성할지를 선택할 수 있습니다. TLA 추출이 발생하지 않으면 이 마법사에서 **패턴** 옵션은 사용 안함으로 설정됩니다.
16. 사전 정의된 범주를 범주 분할창으로 가져오려면 **마침**을 클릭하십시오.

평면 목록 형식

평면 목록 형식에는 계층 구조가 없는 단 하나의 최상의 수준 범주만이 있습니다. 즉 하위 범주나 서브넷이 없습니다. 범주 이름은 단일 열에 있습니다.

다음 정보가 이 형식의 파일에 포함될 수 있습니다.

- 선택적 **코드** 열에는 각 범주를 고유하게 식별하는 숫자 값이 포함됩니다. 데이터 파일에 코드가 포함되는 경우에는(**컨텐츠 설정** 단계에서 **범주 코드 포함** 옵션), 각 범주의 고유한 코드가 범주 이름의 바로 왼쪽에 있는 셀에 존재해야 합니다. 데이터에 코드가 포함되지 않지만 나중에 몇몇 코드를 작성하려는 경우에는 나중에 언제든지 코드를 생성할 수 있습니다(**범주 > 범주 관리 > 코드 자동 생성**).
- *필수* **범주 이름** 열에는 범주의 모든 이름이 포함됩니다. 이 열은 이 형식을 사용하여 가져오는 데 필요합니다.
- 범주 이름의 바로 오른쪽에 있는 셀에 있는 선택적 **주석**. 이 주석은 범주/하위 범주를 설명하는 텍스트로 구성됩니다.
- 선택적 **키워드**는 범주의 디스크립터로서 가져올 수 있습니다. 이러한 키워드가 인식되기 위해서는 연관된 범주/하위 범주의 바로 아래에 있는 셀에 있어야 하며 키워드 목록에는 밑줄(_) 문자가 접두문자로 추가되어야 합니다(예: `_firearms, weapons / guns`). 키워드 셀에는 각 범주를 설명하는 데 사용되는 하나 이상의 단어가 포함될 수 있습니다. 이러한 단어는 디스크립터로서 가져오거나 마법사의 마지막 단계에 지정하는 내용에 따라 무시됩니다. 나중에, 디스크립터는 텍스트에서 추출된 결과와 비교됩니다. 매치가 발견되면 해당 레코드나 문서는 이 디스크립터를 포함하는 범주에 기록됩니다.

표 25. 코드, 키워드 및 주석이 있는 평면 목록 형식

열 A	**열 B**	**열 C**
범주 코드(*선택적*)	범주 이름	주석
	_Descriptor/keyword 목록(*선택적*)	

최소 형식

최소 형식은 계층적인 범주와 함께 사용된다는 점을 제외하고는 평면 목록 형식과 유사하게 구조화되어 있습니다. 그러므로 각 범주와 하위 범주의 계층적 수준을 정의하려면 코드 수준 열이 필요합니다.

다음 정보가 이 형식의 파일에 포함될 수 있습니다.

- *필수* **코드 수준** 열에는 해당 행에서 후속 정보의 계층적 위치를 나타내는 번호가 포함됩니다. 예를 들어, 값 1, 2 또는 3이 지정되고 범주와 하위 범주 둘 모두가 있는 경우에는, 1은 범주용이고, 2는 하위 범주용이고, 3은 하위 하위 범주용입니다. 범주와 하위 범주만이 있는 경우에는 1은 범주용이고, 2는 하위 범주용입니다. 원하는 범주 깊이까지 이런 방식입니다.
- 선택적 **코드** 열에는 각 범주를 고유하게 식별하는 값을 포함합니다. 데이터 파일에 코드가 포함되는 경우에는(**컨텐츠 설정** 단계에서 **범주 코드 포함** 옵션), 각 범주의 고유한 코드가 범주 이름의 바로 왼쪽에 있는 셀에 존재해야 합니다. 데이터에 코드가 포함되지 않지만 나중에 몇몇 코드를 작성하려는 경우에는 나중에 언제든지 코드를 생성할 수 있습니다(**범주 > 범주 관리 > 코드 자동 생성**).
- *필수* **범주 이름** 열에는 범주와 하위 범주의 모든 이름이 포함됩니다. 이 열은 이 형식을 사용하여 가져오는 데 필요합니다.

- 범주 이름의 바로 오른쪽에 있는 셀에 있는 선택적 **주석**. 이 주석은 범주/하위 범주를 설명하는 텍스트로 구성됩니다.
- 선택적 **키워드**는 범주의 디스크립터로서 가져올 수 있습니다. 이러한 키워드가 인식되기 위해서는 연관된 범주/하위 범주의 바로 아래에 있는 셀에 있어야 하며 키워드 목록에는 밑줄(_) 문자가 접두문자로 추가되어야 합니다(예: `_firearms, weapons / guns`). 키워드 셀에는 각 범주를 설명하는 데 사용되는 하나 이상의 단어가 포함될 수 있습니다. 이러한 단어는 디스크립터로서 가져오거나 마법사의 마지막 단계에 지정하는 내용에 따라 무시됩니다. 나중에, 디스크립터는 텍스트에서 추출된 결과와 비교됩니다. 매치가 발견되면 해당 레코드나 문서는 이 디스크립터를 포함하는 범주에 기록됩니다.

표 26. 코드를 포함하는 소형 형식 예제

열 A	**열 B**	**열 C**
계층적 코드 수준	범주 코드(*선택적*)	범주 이름
계층적 코드 수준	하위 범주 코드(*선택적*)	하위 범주 이름

표 27. 코드가 없는 소형 형식 예제

열 A	**열 B**
계층적 코드 수준	범주 이름
계층적 코드 수준	하위 범주 이름

들여쓰기 형식

들여쓰기 파일 형식에서는 컨텐츠는 계층적이며 이는 범주와 하나 이상의 하위 범주 수준이 포함됨을 의미합니다. 또한 구조는 이 계층 구조를 표시하기 위해 들여쓰기되어 있습니다. 파일의 각 행에는 범주 또는 하위 범주가 포함되어 있지만 하위 범주는 범주로부터 들여쓰기되어 있고 모든 하위 하위 범주는 하위 범주로부터 들여쓰기되어 있습니다. 이 구조를 Microsoft Excel에서 수동으로 작성하거나 또 다른 제품에서 내보내었고 Microsoft Excel 형식으로 저장된 구조를 사용할 수 있습니다.

- **상위 수준 범주 코드 및 범주 이름**은 각각 열 A와 B를 차지합니다. 또는 코드가 없는 경우에는 범주 이름은 열 A에 있습니다.
- **하위 범주 코드 및 하위 범주 이름**은 각각 열 B와 C를 차지합니다. 또는 코드가 없는 경우에는 하위 범주 이름은 열 B에 있습니다. 하위 범주는 범주의 멤버입니다. 최상위 수준 범주가 없는 경우에는 하위 범주를 가질 수 없습니다.

표 28. 코드를 포함하는 들여쓰기 구조

열 A	**열 B**	**열 C**	**열 D**
범주 코드(*선택적*)	범주 이름		
	하위 범주 코드(*선택적*)	하위 범주 이름	
		하위 하위 범주 코드(*선택적*)	하위 하위 범주 이름

표 29. 코드를 포함하지 않는 들여쓰기 구조

열 A	**열 B**	**열 C**
범주 이름		
	하위 범주 이름	

표 29. 코드를 포함하지 않는 들여쓰기 구조 (계속)

열 A	열 B	열 C
		하위 하위 범주 이름

다음 정보가 이 형식의 파일에 포함될 수 있습니다.

- 선택적 **코드**는 각 범주 또는 하위 범주를 고유하게 식별하는 값이어야 합니다. 데이터 파일에 코드가 포함됨을 지정한 경우에는(**컨텐츠 설정** 단계에서 **범주 코드 포함** 옵션), 각 범주 또는 하위 범주의 고유한 코드가 범주/하위 범주 이름의 바로 왼쪽에 있는 셀에 존재해야 합니다. 데이터에 코드가 포함되지 않지만 나중에 몇몇 코드를 작성하려는 경우에는 나중에 언제든지 코드를 생성할 수 있습니다(**범주 > 범주 관리 > 코드 자동 생성**).
- 각 범주 및 하위 범주의 *필수* **이름**입니다. 하위 범주는 개별 행에서 범주로부터 오른쪽으로 한 셀씩 들여써야 합니다.
- 범주 이름의 바로 오른쪽에 있는 셀에 있는 선택적 **주석**. 이 주석은 범주/하위 범주를 설명하는 텍스트로 구성됩니다.
- 선택적 **키워드**는 범주의 디스크립터로서 가져올 수 있습니다. 이러한 키워드가 인식되기 위해서는 연관된 범주/하위 범주의 바로 아래에 있는 셀에 있어야 하며 키워드 목록에는 밑줄(_) 문자가 접두문자로 추가되어야 합니다(예: `_firearms, weapons / guns`). 키워드 셀에는 각 범주를 설명하는 데 사용되는 하나 이상의 단어가 포함될 수 있습니다. 이러한 단어는 디스크립터로서 가져오거나 마법사의 마지막 단계에 지정하는 내용에 따라 무시됩니다. 나중에, 디스크립터는 텍스트에서 추출된 결과와 비교됩니다. 매치가 발견되면 해당 레코드나 문서는 이 디스크립터를 포함하는 범주에 기록됩니다.

중요! 코드를 한 수준에서 사용하는 경우에는 각 범주와 하위 범주의 코드를 포함해야 합니다. 그렇지 않으면 가져오기 프로세스가 실패합니다.

범주 내보내기

열려 있는 대화식 워크벤치 세션 에 있는 범주를 Microsoft Excel(**.xls*, **.xlsx*) 파일 형식으로 내보낼 수도 있습니다. 내보낼 데이터는 대부분 범주 분할창의 현재 내용 또는 범주 특성에서 나옵니다. 그러므로 **Docs.** 스코어 값 또한 내보낼 계획이면 다시 스코어링하는 것이 좋습니다.

표 30. 범주 내보내기 옵션

항상 내보냄...	선택적으로 내보냄...
• 범주 코드(있는 경우)	• **Docs.** 스코어
• 범주(및 하위 범주) 이름	• 범주 주석
• 코드 수준(있는 경우)(*평면/최소* 형식)	• 디스크립터 이름
• 열 머리말(*평면/최소* 형식)	• 디스크립터 수

중요! 디스크립터를 내보낼 때 이들은 텍스트 문자열로 변환되고 밑줄이 접두문자로 추가됩니다. 이 제품으로 다시 가져오면 패턴인 디스크립터와 범주 규칙인 디스크립터 및 일반 개념인 디스크립터를 구분하는 기능이 유실됩니다. 이러한 범주를 이 제품에서 다시 사용하려면, 대신 텍스트 분석 패키지(TAP) 파일을 작성할 것을 권장합니다. TAP 형식은 모든 디스크립터뿐만 아니라 모든 범주, 코드 및 사용된 언어학적 자원까지 모두 현

재 정의된 대로 보존하기 때문입니다. TAP 파일은 IBM SPSS Modeler Text Analytics 및 IBM SPSS Text Analytics for Surveys 둘 모두에서 사용할 수 있습니다. 자세한 정보는 『텍스트 분석 패키지 사용』의 내용을 참조하십시오.

사전에 정의된 범주 내보내기

1. 대화식 워크벤치 메뉴에서 **범주 > 범주 관리 > 범주 내보내기**를 선택하십시오. 범주 내보내기 마법사가 표시됩니다.
2. 위치를 선택하고 내보낼 파일의 이름을 입력하십시오.
3. 파일 이름 텍스트 상자에 출력 파일의 이름을 입력하십시오.
4. 범주 데이터를 내보낼 형식을 선택하려면 **다음**을 클릭하십시오.
5. 다음에서 형식을 선택하십시오.
   - **평면 또는 최소 목록 형식**: 자세한 정보는 144 페이지의 『평면 목록 형식』 주제를 참조하십시오. 평면 목록에는 하위 범주가 없습니다. 자세한 정보는 144 페이지의 『최소 형식』의 내용을 참조하십시오. 최소 목록 형식에는 계층적 범주가 포함됩니다.
   - **들여쓰기 형식**: 자세한 정보는 145 페이지의 『들여쓰기 형식』 주제를 참조하십시오.
6. 내보낼 내용을 선택하기 시작하고 제안된 데이터를 검토하려면 **다음**을 클릭하십시오.
7. 내보낸 파일의 내용을 검토하십시오.
8. **주석** 또는 **디스크립터 이름** 등과 같이 내보낼 추가적인 내용 설정을 선택하거나 선택 취소하십시오.
9. 범주를 내보내려면 **마침**을 클릭하십시오.

텍스트 분석 패키지 사용

TAP라고도 불리는 텍스트 분석 패키지는 텍스트 반응 범주화를 위한 템플리트의 역할을 합니다. TAP를 사용하는 것은 최소한의 개입만으로 텍스트 데이터를 범주화하는 쉬운 방법입니다. 여기에는 방대한 수의 레코드를 빠르게 자동으로 코딩하는 데 필요한 사전에 작성된 범주 세트 및 언어학적 자원이 포함되어 있기 때문입니다. 언어학적 자원을 사용하여 텍스트 데이터는 주요 개념을 추출하기 위해 분석되고 마이닝됩니다. 텍스트에서 발견된 주요 개념과 패턴을 기반으로 레코드는 사용자가 TAP에서 선택한 범주 세트로 범주화될 수 있습니다. 사용자만의 TAP를 작성하거나 이를 업데이트할 수 있습니다.

TAP는 다음 요소로 구성됩니다.

- **범주 세트**. 범주 세트는 본질적으로 사전 정의된 범주, 범주 코드, 각 범주의 디스크립터, 마지막으로 전체 범주 세트의 이름으로 구성됩니다. 디스크립터는 용어 *저렴한* 또는 패턴 *좋은 가격* 등과 같은 언어학적 요소(개념, 유형, 패턴 및 규칙)입니다. 디스크립터는 텍스트가 범주 디스크립터와 매치하고, 문서 또는 레코드가 범주에 들어갈 수 있도록 범주를 정의하는 데 사용됩니다.
- **언어학적 자원**. 언어학적 자원은 주요 개념과 패턴을 추출하기 위해 조정된 라이브러리와 고급 자원의 세트입니다. 이러한 추출 개념 및 패턴은 그런 다음 레코드를 범주 세트의 범주에 배치할 수 있게 해주는 디스크립터로서 사용됩니다.

사용자만의 TAP를 작성하고, 이를 업데이트하거나 텍스트 분석 패키지를 로드할 수 있습니다.

TAP를 선택하고 범주 세트를 선택한 후에는 IBM SPSS Modeler Text Analytics 는 사용자의 레코드를 추출하고 범주화할 수 있습니다.

참고: TAP는 IBM SPSS Text Analytics for Surveys와 IBM SPSS Modeler Text Analytics 사이에 상호 교환적으로 작성 및 사용될 수 있습니다.

텍스트 분석 패키지 작성

하나 이상의 범주와 몇몇 자원이 있는 세션이 있을 때마다 열려 있는 대화식 워크벤치 세션의 컨텐츠에서 텍스트 분석 패키지(TAP)를 작성할 수 있습니다. 범주 및 디스크립터(개념, 유형, 규칙 또는 TLA 패턴 출력) 세트는 자원 편집기에 열려 있는 모든 언어학적 자원과 함께 TAP에 작성할 수 있습니다.

자원 작성 시 사용된 언어를 볼 수 있습니다. 언어는 템플리트 편집기 또는 자원 편집기의 고급 자원 탭에 설정됩니다.

텍스트 분석 패키지 작성

1. 메뉴에서 **파일 > 텍스트 분석 패키지 > 패키지 작성**을 선택하십시오. 패키지 작성 대화 상자가 나타납니다.
2. TAP을 저장할 디렉토리로 이동하십시오. 기본적으로 TAP은 제품 설치 디렉토리의 `\TAP` 하위 디렉토리에 저장됩니다.
3. **파일 이름** 필드에 TAP의 이름을 입력하십시오.
4. **패키지 레이블** 필드에 레이블을 입력하십시오. 파일 이름을 입력하면 이 이름은 레이블로 자동으로 나타나지만 이 레이블은 변경할 수 있습니다.
5. TAP에서 범주 세트를 제외하려면 **포함** 선택란을 선택 취소하십시오. 그러면 이는 패키지에 추가되지 않게 됩니다. 기본적으로 질문당 하나의 범주 세트가 TAP에 포함됩니다. TAP에는 항상 하나 이상의 범주 세트가 있어야 합니다.
6. 범주 세트의 이름을 변경하십시오. **새 범주 세트** 열에는 기본적으로 일반 이름이 포함됩니다. 이는 텍스트 변수 이름에 `Cat_` 접두문자를 추가하여 생성됩니다. 셀을 한 번 클릭하면 이름이 편집 가능하게 됩니다. Enter를 누르거나 다른 곳을 클릭하면 이름 변경이 적용됩니다. 범주 세트의 이름을 변경하는 경우에는 이름은 TAP에서만 변경되고 열려 있는 세션에서 변수 이름을 변경하지 않습니다.
7. 원하는 경우 범주 세트 테이블 오른쪽의 화살표 키를 사용하여 범주 세트를 다시 정렬하십시오.
8. 텍스트 분석 패키지를 작성하려면 **저장**을 클릭하십시오. 대화 상자가 닫힙니다.

텍스트 분석 패키지 로드

텍스트 마이닝 모델링 노드를 구성할 때 추출 중에 사용될 자원을 지정해야 합니다. 자원 템플리트를 선택하는 대신에 자원뿐만 아니라 범주 세트를 노드에 복사하기 위해 텍스트 분석 패키지(TAP)를 선택할 수 있습니다.

TAP는 범주 모델을 대화식으로 작성할 때 가장 적절합니다. 범주 세트를 범주화의 시작점으로 사용할 수 있기 때문입니다. 스트림을 실행할 때 대화식 워크벤치 세션이 실행되고 이 범주 세트가 범주 분할창에 나타납니

다. 이 방법으로 즉시 이러한 범주를 사용하여 문서와 레코드를 기록한 다음 사용자의 요구를 충족시킬 때까지 이러한 범주를 계속해서 세분화, 작성 및 확장하십시오. 자세한 정보는 108 페이지의 『범주 작성을 위한 방법 및 전략』의 내용을 참조하십시오.

버전 14부터 시작하여, **로드**를 클릭하고 TAP를 선택하면 이 TAP의 자원이 정의되는 언어를 볼 수도 있습니다.

텍스트 분석 패키지 로드

1. 텍스트 마이닝 모델링 노드를 편집하십시오.
2. 모델 탭에서, **자원 복사 시작** 섹션에서 *텍스트 분석 패키지*를 선택하십시오.
3. **로드**를 클릭하십시오. 텍스트 분석 패키지 로드 대화 상자가 열립니다.
4. 노드를 복사하려는 자원과 범주 세트를 포함하는 TAP의 위치로 이동하십시오. 기본적으로 TAP은 제품 설치 디렉토리의 `\TAP` 하위 디렉토리에 저장됩니다.
5. **파일 이름** 필드에 TAP의 이름을 입력하십시오. 레이블이 자동으로 표시됩니다.
6. 사용하려는 범주 세트를 선택하십시오. 이는 대화식 워크벤치 세션에 나타나는 범주 세트입니다. 그런 다음 이러한 범주를 수동으로나 범주 작성 또는 확장 옵션을 사용하여 수정하고 향상시킬 수 있습니다.
7. 텍스트 분석 패키지의 컨텐츠를 노드에 복사하려면 **로드**를 클릭하십시오. 대화 상자가 닫힙니다. TAP이 로드되면 TAP의 사본이 노드에 복사됩니다. 그러므로 자원과 범주의 변경사항은 TAP을 명시적으로 업데이트하고 이를 다시 로드하기 전까지는 TAP에 반영되지 않습니다.

텍스트 분석 패키지 업데이트

범주 세트, 언어학적 자원을 개선하거나 전체 새 범주 세트를 작성하는 경우에는 텍스트 분석 패키지(TAP)를 업데이트하여 이러한 개선사항을 나중에 쉽게 다시 사용할 수 있습니다. 이를 수행하려면 TAP에 넣으려는 정보를 포함하는 열려 있는 세션에 있어야 합니다. 업데이트할 때 범주 세트를 추가하고, 자원을 재배치하고, 패키지 레이블을 변경하거나 범주 세트의 이름을 변경하거나 순서를 다시 정렬하기를 선택할 수 있습니다.

텍스트 분석 패키지 업데이트

1. 메뉴에서 **파일 > 텍스트 분석 패키지 > 패키지 업데이트**를 선택하십시오. 패키지 업데이트 대화 상자가 나타납니다.
2. 업데이트하려는 텍스트 분석 패키지를 포함하는 디렉토리로 이동하십시오.
3. **파일 이름** 필드에 TAP의 이름을 입력하십시오.
4. TAP 내부에 있는 언어학적 자원을 현재 세션에 있는 자원으로 대체하려면 **이 패키지의 자원을 열려 있는 세션의 자원으로 대체** 옵션을 선택하십시오. 이는 범주 정의를 작성하는 데 사용된 주요 개념과 패턴을 추출하는 데 사용되었으므로 일반적으로 언어학적 자원을 업데이트하는 것이 맞습니다. 가장 최신 언어학적 자원을 가지고 있으면 레코드를 범주화할 때 최상의 결과를 얻을 수 있습니다. 이 옵션을 선택하지 않으면 패키지에 이미 있는 언어학적 자원은 변경되지 않은 상태로 있습니다.
5. 언어학적 자원만을 업데이트하려면 **이 패키지의 자원을 열려 있는 세션의 자원으로 대체** 옵션을 선택하고 TAP에 이미 있는 현재 범주 세트만을 선택했는지 확인하십시오.

6. 열려 있는 세션에서 새 범주 세트를 TAP에 포함시키려면 추가할 각 범주 세트의 선택란을 선택하십시오. 범주 세트를 하나, 여러 개 추가하거나 하나도 추가하지 않을 수 있습니다.
7. TAP에서 범주 세트를 제거하려면 해당하는 **포함** 선택란을 선택 취소하십시오. 개선된 범주 세트를 추가 중이므로 TAP에 이미 있는 범주 세트를 제거하기로 선택할 수도 있습니다. 이를 수행하려면 현재 범주 세트 열에서 해당하는 범주 세트의 **포함** 선택란을 선택 취소하십시오. TAP에는 항상 하나 이상의 범주 세트가 있어야 합니다.
8. 필요한 경우 범주 세트의 이름을 변경하십시오. 셀을 한 번 클릭하면 이름이 편집 가능하게 됩니다. Enter를 누르거나 다른 곳을 클릭하면 이름 변경이 적용됩니다. 범주 세트의 이름을 변경하는 경우에는 이름은 TAP에서만 변경되고 열려 있는 세션에서 변수 이름을 변경하지 않습니다. 두 개의 범주 세트가 같은 이름을 가지고 있으면 이름은 중복을 수정할 때까지 빨간색으로 나타납니다.
9. 세션 컨텐츠가 선택된 TAP의 컨텐츠와 병합된 상태로 새 패키지를 작성하려면 **다른 이름으로 새로 저장**을 클릭하십시오. 텍스트 분석 패키지를 다른 이름으로 저장 대화 상자가 나타납니다. 다음 지시사항을 참조하십시오.
10. TAP에 선택한 변경사항을 저장하려면 **업데이트**를 클릭하십시오.

텍스트 분석 패키지 저장

1. TAP 파일을 저장할 디렉토리로 이동하십시오. 기본적으로 TAP 파일은 설치 디렉토리의 TAP 서브디렉토리에 저장됩니다.
2. 파일 이름 필드에 TAP 파일의 이름을 입력하십시오.
3. 패키지 레이블 필드에 레이블을 입력하십시오. 파일 이름을 입력하면 이 이름은 자동으로 레이블로 사용됩니다. 그러나 이 레이블의 이름을 변경할 수 있습니다. 레이블이 있어야 합니다.
4. 새 패키지를 작성하려면 **저장**을 클릭하십시오.

범주 편집 및 세분화

일부 범주를 작성한 후에는 이를 예외없이 검사한 다음에 조정하려고 할 수 있습니다. 언어학적 자원을 세분화하는 데 추가로 정의를 결합하거나 정리하는 방법을 찾고 일부 범주화된 문서 또는 레코드를 확인하여 범주를 검토해야 합니다. 범주에서 문서 또는 레코드를 검토하고 범주가 뉘앙스와 차이가 캡처되는 방식으로 정의될 수 있도록 조정할 수도 있습니다.

내장된 자동화된 범주 작성 기술을 사용하여 범주를 작성할 수 있습니다. 그러나 이러한 범주에 몇몇 조정 작업을 수행하려고 할 수도 있습니다. 하나 이상의 기술을 사용한 후에는 많은 새 범주가 창에 나타납니다. 그런 다음 범주에서 데이터를 검토하고 범주 정의에 만족할 때까지 조정할 수 있습니다. 자세한 정보는 113 페이지의 『범주 정보』의 내용을 참조하십시오.

다음은 범주를 세분화하기 위한 몇몇 옵션입니다. 대부분은 다음 페이지에 설명되어 있습니다.

디스크립터를 범주에 추가

자동화된 기술을 사용한 후에 범주 정의에 사용되지 않은 추출 결과가 여전히 남아 있을 수 있습니다. 확장 결과 분할창에서 이 목록을 검토해야 합니다. 범주로 이동하려는 요소를 찾으면 이를 기존 범주 또는 신규 범주에 추가할 수 있습니다.

개념이나 유형을 범주에 추가

1. 추출 결과 및 데이터 분할창 내에서 신규 또는 기존 범주에 추가하려는 요소를 선택하십시오.
2. 메뉴에서 **범주 > 범주에 추가**를 선택하십시오. 모든 범주 대화 상자가 범주 세트를 표시합니다. 선택한 요소를 추가하려는 범주를 선택하십시오. 요소를 새 범주에 추가하려면 **새 범주**를 선택하십시오. 새 범주가 첫 번째 선택된 요소의 이름을 사용하여 범주 분할창에 나타납니다.

범주 디스크립터 편집

몇몇 범주를 작성한 후에는 각 범주를 열어서 해당 정의를 구성하는 모든 디스크립터를 볼 수 있습니다. 범주 정의 대화 상자 내에서 범주 디스크립터를 여러 번 편집할 수 있습니다. 또한 범주가 범주 트리에 표시되면 여기에서 이에 대해 작업할 수도 있습니다.

범주 편집

1. 범주 분할창에서 편집할 범주를 선택하십시오.
2. 메뉴에서 **보기 > 범주 정의**를 선택하십시오. 범주 정의 대화 상자가 열립니다.
3. 편집하려는 디스크립터를 선택하고 해당하는 도구 모음 단추를 클릭하십시오.

다음 테이블은 범주 정의를 편집하기 위해 사용할 수 있는 각 도구 모음 단추를 설명합니다.

표 31. 도구 모음 단추 및 설명.

아이콘	설명
	범주에서 선택한 디스크립터를 삭제합니다.
	선택한 디스크립터를 신규 또는 기존 범주로 이동합니다.
	선택한 디스크립터를 & 범주 규칙의 양식으로 범주로 이동합니다. 자세한 정보는 132 페이지의 『범주 규칙 사용』의 내용을 참조하십시오.
	선택한 각 디스크립터를 자체 신규 범주로 이동합니다.
 표시	데이터 분할창과 시각화 분할창에 표시된 내용을 선택한 디스크립터에 따라 업데이트합니다.

범주 이동

범주를 또 다른 기존 범주에 놓거나 디스크립터를 또 다른 범주로 이동하려는 경우에는 이를 이동할 수 있습니다.

범주 이동

1. 범주 분할창에서 또 다른 범주로 이동하려는 범주 를 선택하십시오.
2. 메뉴에서 **범주 > 범주로 이동**을 선택하십시오. 메뉴는 범주 세트에 목록의 맨 위에 있는 가장 최근에 작성된 범주를 제공합니다. 선택한 개념을 이동하려는 범주 이름을 선택하십시오.

- 찾고 있는 이름이 보이면, 이를 선택하면 선택된 요소가 해당 범주에 추가됩니다.
- 보이지 않으면 **기타**를 선택하여 모든 범주 대화 상자를 표시하고 목록에서 범주를 선택하십시오.

범주 평면화

범주와 하위 범주가 있는 계층적 범주 구조가 있는 경우 구조를 평면화할 수 있습니다. 범주를 평면화할 때 해당 범주의 하위 범주에 있는 모든 디스크립터가 선택한 범주로 이동되고 현재 비어 있는 하위 범주는 삭제됩니다. 이런 방식으로 하위 범주와 매치시키는 데 사용된 모든 문서는 이제 선택된 범주로 범주화됩니다.

범주 평면화

1. 범주 분할창에서 평면화하려는 범주(최상위 수준 또는 하위 범주)를 선택하십시오.
2. 메뉴에서 **범주 > 범주 평면화**를 선택하십시오. 하위 범주는 제거되고 디스크립터가 선택된 범주로 병합됩니다.

범주 병합 또는 결합

둘 이상의 기존 범주를 새 범주로 결합하려는 경우에는 이를 병합하면 됩니다. 범주를 병합할 때에는 새 범주가 일반 이름으로 작성됩니다. 범주 디스크립터에 사용된 모든 개념, 유형 및 패턴은 이 새 범주로 이동됩니다. 나중에 범주 특성을 편집하여 이 범주의 이름을 변경할 수 있습니다.

범주 또는 범주의 일부 병합

1. 범주 분할창에서 병합하려는 요소를 선택하십시오.
2. 메뉴에서 **범주 > 범주 병합**을 선택하십시오. 새로 작성된 범주의 이름을 입력하는 범주 특성 대화 상자가 표시됩니다. 선택된 범주가 새 범주에 하위 범주로서 병합됩니다.

범주 삭제

범주를 유지하지 않으려면 이를 삭제하면 됩니다.

범주 삭제

1. 범주 분할창에서 삭제하려는 범주를 선택하십시오.
2. 메뉴에서 **편집 > 삭제**를 선택하십시오.

제 11 장 군집 분석

군집 보기(**보기 > 군집**)에서 개념 군집을 작성하고 탐색할 수 있습니다. **군집**은 이러한 개념이 문서/레코드 세트에서 발생하는 빈도와 동일한 문서에서 함께 나타나는 빈도(**동시 발생**이라고도 함)를 기반으로 군집화 알고리즘에 의해 생성된 관련 개념 집단입니다. 군집의 각 개념은 군집에서 하나 이상의 다른 개념과 동시 발생합니다. 범주의 목적은 범주에 포함된 텍스트가 각 범주에 대해 디스크립터(개념, 규칙, 패턴)를 매치하는 방법을 기반으로 문서 또는 레코드를 그룹화하는 것인 반면 군집의 목적은 함께 동시 발생하는 개념을 그룹화하는 것입니다.

좋은 군집은 강하게 링크되고 자주 동시 발생하는 개념이 있으며 다른 군집의 개념에 대한 링크가 거의 없는 군집입니다. 큰 데이터 세트에 대한 작업 시 이 기술로 인해 처리 시간이 상당히 길어질 수 있습니다.

참고: 모든 문서 또는 레코드 서브세트만으로 작성하려면 군집 작성 대화 상자에서 **군집 계산에 사용할 최대 문서 수** 옵션을 사용하십시오.

군집화는 개념 세트를 분석하고 문서에서 자주 발생하는 개념을 검색함으로써 시작되는 프로세스입니다. 한 문서에서 동시 발생하는 두 개의 개념을 개념 쌍으로 간주합니다. 그 다음에 군집화 프로세스는 쌍이 함께 발생하는 문서 수를 각 개념이 발생하는 문서 수와 비교하여 각 개념 쌍의 **유사성 값**을 평가합니다. 자세한 정보는 156 페이지의 『유사성 링크 값 계산』의 내용을 참조하십시오.

마지막으로 군집화 프로세스는 통합을 통해 유사한 개념을 군집으로 그룹화하고 해당 링크 값과 군집 작성 대화 상자에 정의된 설정을 고려합니다. 통합을 통해 군집이 포화 상태가 될 때까지 개념이 추가되거나 작은 군집이 더 큰 군집으로 병합됩니다. 개념 또는 작은 군집을 추가로 합쳐 군집 작성 대화 상자에 정의된 설정(개념, 내부 링크 또는 외부 링크 수)을 초과하게 되면 군집이 **포화 상태**가 됩니다. 군집은 군집 내에서 군집 내 다른 개념에 대한 전체 링크 수가 가장 많은 개념의 이름을 사용합니다.

다른 군집에 더 강한 링크가 있거나 포화로 인해 개념 쌍이 발생하는 군집을 병합할 수 없기 때문에 결국 모든 개념 쌍이 동일한 군집에 함께 있게 되는 것은 아닙니다. 이러한 이유로 내부 및 외부 링크가 모두 있습니다.

- **내부 링크**는 군집 내 개념 쌍 간의 링크입니다. 모든 개념이 군집에서 서로 링크되는 것은 아닙니다. 그러나 각 개념은 군집 내에서 하나 이상의 다른 개념에 링크됩니다.
- **외부 링크**는 별도 군집의 개념 쌍(한 군집 내 개념과 다른 군집에서 외부의 개념) 간의 링크입니다.

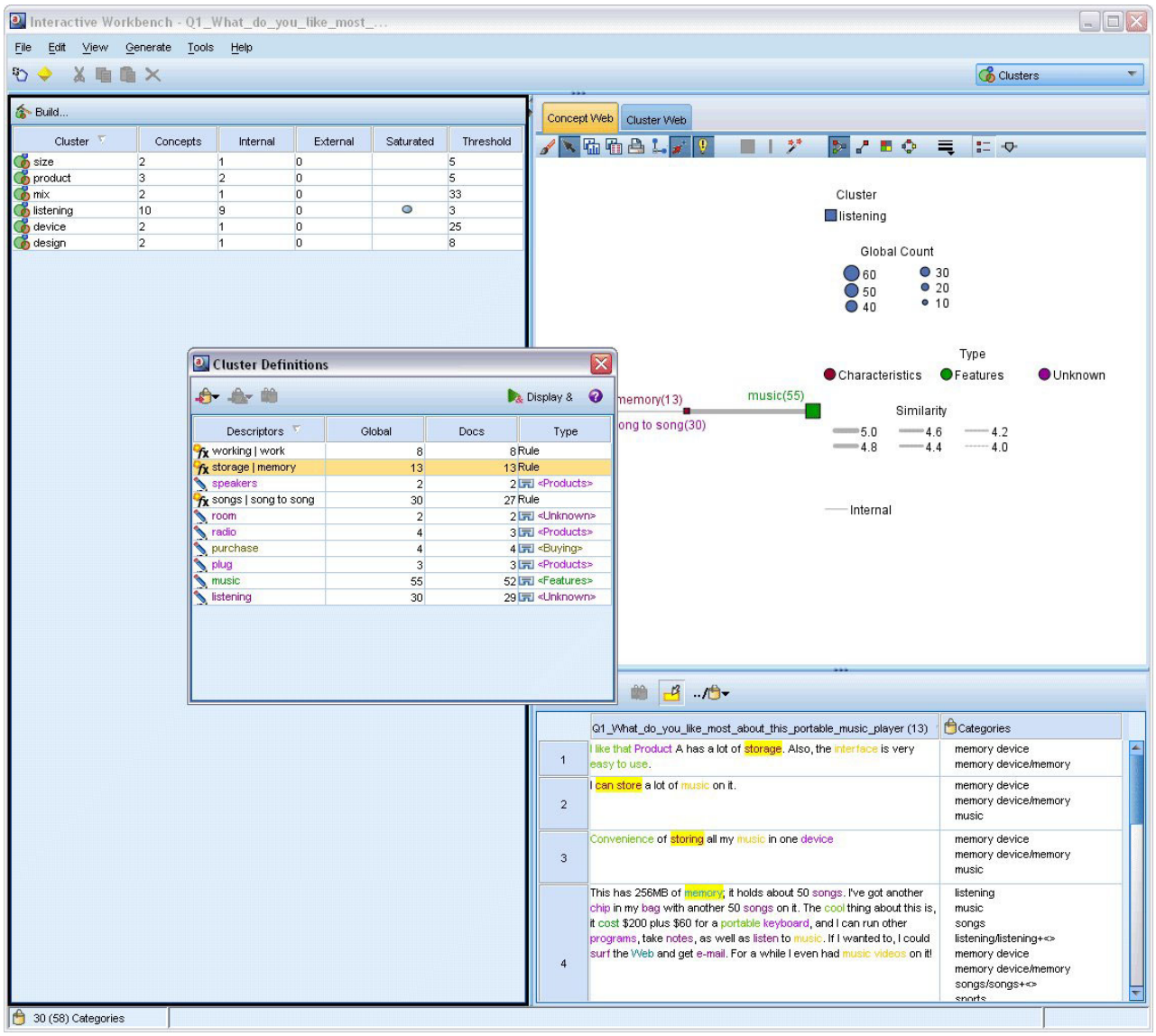


그림 30. 군집 보기

군집 보기는 세 개의 분할창으로 구성되며, 보기 메뉴에서 해당 이름을 선택하여 각각 숨기거나 표시할 수 있습니다.

- **군집 분할창.** 이 분할창에서 군집을 작성하고 관리할 수 있습니다. 자세한 정보는 157 페이지의 『군집 탐색』의 내용을 참조하십시오.
- **시각화 분할창.** 이 분할창에서 군집 및 군집 상호작용 방법을 시각적으로 탐색할 수 있습니다. 자세한 정보는 167 페이지의 『군집 그래프』의 내용을 참조하십시오.
- **데이터 분할창.** 군집 정의 대화 상자의 선택사항에 해당하는 문서 및 레코드 내에 포함된 텍스트를 탐색하고 검토할 수 있습니다. 자세한 정보는 157 페이지의 『군집 정의』의 내용을 참조하십시오.

군집 작성

처음 군집 보기에 액세스할 때, 어떤 군집도 볼 수 없습니다. 메뉴를 통하거나(**도구 >군집 작성**) 도구 모음에서 **작성...** 단추를 클릭하여 군집을 작성할 수 있습니다. 이 조치는 군집 작성을 위한 설정 및 한계를 정의할 수 있는 군집 작성 대화 상자를 엽니다.

참고: 추출 결과가 더 이상 자원과 매치하지 않는 경우, 이 분할창은 추출 결과 분할창에서와 같이 노란색이 됩니다. 최근 추출 결과를 확보하기 위해 다시 추출할 수 있으며, 노란색이 사라집니다. 그러나 추출이 수행될 때마다 군집 분할창이 지워지므로 군집을 다시 작성해야 합니다. 마찬가지로 군집은 한 세션에서 다른 세션으로 저장되지 않습니다.

군집 작성 대화 상자에서 다음 영역 및 필드를 사용할 수 있습니다.

입력

입력 테이블. 군집은 특정 유형에서 파생된 디스크립터에서 작성됩니다. 테이블에서, 작성 프로세스에 포함할 유형을 선택할 수 있습니다. 대부분의 레코드 또는 문서를 캡처하는 유형은 기본적으로 미리 선택됩니다.

군집에 대한 개념: 군집에 사용할 개념을 선택하는 방법을 선택하십시오. 개념 수를 줄여서, 군집 프로세스를 가속화할 수 있습니다. 상위 개념 수, 상위 개념 퍼센트 또는 모든 개념 사용을 통해 군집할 수 있습니다.

- **문서 개수를 기준으로 한 수. 상위 개념 수**를 선택하는 경우, 군집에 고려될 개념 수를 입력하십시오. 개념은 최상위 문서 수 값을 가지고 있는 개념을 기준으로 선택됩니다. 문서 개수는 개념이 나타나는 문서 또는 레코드의 수입니다.
- **문서 개수를 기준으로 한 퍼센트. 개념의 상위 퍼센트**를 선택하는 경우, 군집에 고려될 개념의 퍼센트를 입력하십시오. 개념은 최상위 문서 개수 값을 가지고 있는 개념의 퍼센트를 기준으로 선택됩니다.

군집 계산에 사용할 최대 문서 수. 기본적으로 링크 값은 전체 문서 또는 레코드 세트를 사용하여 계산됩니다. 그러나 어떤 경우에는 링크를 계산하기 위해 사용되는 문서 또는 레코드 수를 제한하여 군집 프로세스를 가속화할 수 있습니다. 문서를 제한하면 군집의 품질이 떨어질 수 있습니다. 이 옵션을 사용하려면, 왼쪽에 있는 확인 상자를 선택하고 사용할 최대 문서 또는 레코드 수를 입력하십시오.

출력 한계

작성할 최대 군집 수. 이 값은 군집 분할창에서 생성하고 표시할 최대 군집 수입니다. 군집 프로세스 동안, 포화모형 군집은 불포화모형 군집 이전에 제시되므로, 결과로 생성되는 많은 군집이 포화모형이 됩니다. 불포화모형 군집을 더 보려면, 이 설정을 포화모형 군집 수보다 큰 값으로 변경하면 됩니다.

군집의 최대 개념 수. 이 값은 군집이 포함할 수 있는 최대 개념 수입니다.

군집의 최소 개념 수. 이 값은 군집을 작성하기 위해 링크해야 하는 최소 개념 수입니다.

최대 내부 링크 수. 이 값은 군집이 포함할 수 있는 최대 내부 링크 수입니다. 내부 링크는 군집 내의 개념 쌍 사이에 있는 링크입니다.

최대 외부 링크 수. 이 값은 군집 외부에서 개념에 대한 최대 링크 수입니다. 외부 링크는 별도의 군집에서 개념 쌍 사이에 있는 링크입니다.

최소 링크 값. 이 값은 군집에 고려할 개념 쌍에 대해 승인된 가장 작은 링크 값입니다. 링크 값은 유사성 수식을 사용하여 계산됩니다. 자세한 정보는 『유사성 링크 값 계산』의 내용을 참조하십시오.

특정 개념 쌍 방지. 프로세스가 출력에 두 개의 개념을 함께 그룹화하거나 쌍으로 만드는 프로세스를 중지하려면 이 선택란을 선택하십시오. 개념 쌍을 작성하거나 관리하려면 **쌍 관리**를 클릭하십시오. 자세한 정보는 121 페이지의 『링크 예외 쌍 관리』의 내용을 참조하십시오.

유사성 링크 값 계산

개념 쌍이 동시 발생하는 문서 수만 알면 본질적으로 두 개념이 어느 정도 유사한지 알 수 없습니다. 이러한 경우 유사성 값이 유용합니다. 유사성 링크 값은 동시 발생 문서 개수를 관계의 각 개념에 대한 개별 문서 수와 비교하여 측정됩니다. 유사성을 계산할 때 측정 단위는 개념 또는 개념 쌍을 찾은 문서 수(문서 개수)입니다. 문서에서 *최*소 한 번 발생하면 문서에서 개념 또는 개념 쌍을 "찾을 수 있습니다". 개념 그래프에서 선 굵기로 그래프에 유사성 링크 값을 표시하도록 선택할 수 있습니다.

알고리즘은 가장 강력한 해당 관계를 표시하는데, 텍스트 데이터에 함께 나타날 개념에 대한 경향이 독립적으로 발생할 경향보다 훨씬 높음을 의미합니다. 내부적으로 알고리즘은 0 - 1 범위의 유사성 계수를 산출합니다. 여기서 1 값은 두 개념이 항상 동시에 나타나며 별도로 나타나지 않음을 의미합니다. 유사성 계수 결과는 100을 곱하여 가장 가까운 정수로 반올림됩니다. 유사성 계수는 다음 그림에 표시된 수식을 사용하여 계산됩니다.

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

그림 31. 유사성 계수 수식

여기서,

- C_I는 개념 I가 발생하는 문서 또는 레코드 수입니다.
- C_J는 개념 J가 발생하는 문서 또는 레코드 수입니다.
- C_{IJ}는 개념 쌍 I와 J가 문서 세트에서 동시 발생하는 문서 또는 레코드 수입니다.

예를 들어, 5,000개의 문서가 있다고 가정하십시오. I와 J는 추출된 개념이고 IJ는 I와 J의 개념 쌍 동시 발생입니다. 다음 테이블에서는 계수 및 링크 값 계산 방법을 보여주는 두 개의 시나리오를 제안합니다.

표 32. 개념 빈도 예제

개념/쌍	시나리오 A	시나리오 B
개념: I	20개 문서에 발생	30개 문서에 발생
개념: J	20개 문서에 발생	60개 문서에 발생
개념 쌍: IJ	20개 문서에 동시 발생	20개 문서에 동시 발생
유사성 계수	1	0.22222
유사성 링크 값	100	22

시나리오 A에서 개념 I와 J는 물론 쌍 IJ는 20개 문서에서 발생하며 유사성 계수 1을 산출합니다(개념이 항상 함께 발생함을 의미). 이 쌍의 유사성 링크 값은 100입니다.

시나리오 B에서 개념 I는 30개 문서에서 발생하고 개념 J는 60개 문서에서 발생하지만 쌍 IJ는 20개 문서에서만 발생합니다. 따라서 유사성 계수는 0.22222입니다. 이 쌍의 유사성 링크 값은 22로 반내림됩니다.

군집 탐색

군집을 작성하면, 군집 분할창에서 결과 세트를 볼 수 있습니다. 군집마다, 테이블에서 다음 정보를 사용할 수 있습니다.

- **군집.** 군집의 이름입니다. 군집 이름은 내부 링크 수가 가장 많은 개념 뒤에 지정됩니다.
- **개념.** 군집의 개념 수입니다. 자세한 정보는 『군집 정의』의 내용을 참조하십시오.
- **내부.** 군집의 내부 링크 수입니다. 내부 링크는 군집 내의 개념 쌍 사이에 있는 링크입니다.
- **외부.** 군집에 있는 외부 링크 수입니다. 외부 링크는 하나의 개념이 하나의 군집에 있고 다른 개념이 다른 군집에 있는 경우 개념 쌍 사이에 있는 링크입니다.
- **포화.** 기호가 존재하면, 이 군집이 더 커질 수 있지만, 하나 이상의 한계가 초과될 수 있어서, 군집 프로세스가 해당 군집에 대해 종료되고 *포화모형*인 것으로 간주됩니다. 군집 프로세스 끝에서, 포화모형 군집은 불포화모형 군집 이전에 제시되므로, 결과로 생성되는 많은 군집이 포화모형이 됩니다. 불포화모형 군집을 보려면, **작성할 최대 군집 수** 설정을 포화모형 군집 수보다 더 큰 값으로 변경하거나 **최소 링크 값**을 감소시킬 수 있습니다. 자세한 정보는 155 페이지의 『군집 작성』의 내용을 참조하십시오.
- **임계값.** 군집에서 발생하는 모든 개념 쌍에 대해, 군집에서 유사성이 가장 낮은 링크 값입니다. 자세한 정보는 156 페이지의 『유사성 링크 값 계산』의 내용을 참조하십시오. 임계값이 높은 군집은 군집의 개념이 전반적으로 높은 유사성을 가지고 있고 임계값이 낮은 군집의 개념보다 훨씬 근접하게 관련됨을 나타냅니다.

지정된 군집에 대해 더 자세히 알아보기 위해 군집을 선택할 수 있으며, 오른쪽의 시각화 분할창이 군집을 탐색할 수 있도록 두 개의 그래프를 표시합니다. 자세한 정보는 167 페이지의 『군집 그래프』의 내용을 참조하십시오. 또한 테이블의 내용을 다른 애플리케이션으로 잘라내어 붙여넣을 수도 있습니다.

추출 결과가 더 이상 자원과 매치하지 않는 경우, 이 분할창은 추출 결과 분할창에서와 같이 노란색이 됩니다. 최근 추출 결과를 확보하기 위해 다시 추출할 수 있으며, 노란색이 사라집니다. 그러나 추출이 수행될 때마다 군집 분할창이 지워지므로 군집을 다시 작성해야 합니다. 마찬가지로 군집은 한 세션에서 다른 세션으로 저장되지 않습니다.

군집 정의

군집 분할창에서 선택하고 군집 정의 대화 상자를 열어서(**보기 > 군집 정의**) 군집 내에서 모든 개념을 볼 수 있습니다.

선택된 군집의 모든 개념은 군집 정의 대화 상자에 나타납니다. 군집 정의 대화 상자에서 하나 이상의 개념을 선택하고 **표시 &**를 클릭하면, 데이터 분할창에 *선택된 모든 개념이 함께 표시되는* 모든 문서 또는 레코드가 표시됩니다. 그러나 군집 분할창에서 군집을 선택할 때 데이터 분할창에 텍스트 레코드 또는 문서가 표시되지 않습니다. 데이터 분할창에 관한 일반 정보는 in의 내용을 참조하십시오.

이 대화 상자에서 개념을 선택하면 개념 웹 그래프도 변경됩니다. 자세한 정보는 167 페이지의 『군집 그래프』의 내용을 참조하십시오. 마찬가지로, 군집 정의 대화 상자에서 하나 이상의 개념을 선택할 때 해당 개념의 모든 외부 및 내부 링크가 시각화 분할창에 표시됩니다.

열 설명

각 디스크립터를 쉽게 식별할 수 있도록 아이콘이 표시됩니다.

표 33. 열 및 디스크립터 아이콘

열	설명
디스크립터	개념의 이름.
 **전역값**	해당 디스크립터가 전체 데이터 세트에 나타나는 횟수를 표시하며, 전역 빈도라고도 합니다.
 **문서 수**	이 디스크립터가 나타나는 문서 또는 레코드 수로, 문서 빈도라고도 합니다.
유형	디스크립터가 속하는 유형을 표시합니다. 디스크립터가 범주 규칙인 경우, 어떤 유형 이름도 이 열에 표시되지 않습니다.

도구 모음 조치

이 대화 상자에서, 범주에 사용할 하나 이상의 개념을 선택할 수도 있습니다. 이를 수행하기 위한 몇 가지 방법이 있지만, 군집에 발생하는 개념을 선택하고 범주 규칙으로 추가하는 것이 가장 좋습니다. 자세한 정보는 125 페이지의 『동시 발생 규칙』의 내용을 참조하십시오. 도구 모음 단추를 사용하여 범주에 개념을 추가할 수 있습니다.

표 34. 범주에 개념을 추가할 도구 모음 단추

아이콘	설명
	선택된 개념을 기존 또는 새 범주에 추가합니다.
	& 범주 규칙 양식으로 선택된 개념을 기존 또는 새 범주에 추가합니다. 자세한 정보는 132 페이지의 『범주 규칙 사용』의 내용을 참조하십시오.
	선택된 각 개념을 자체의 고유한 새 범주로 추가합니다.
	데이터 분할창과 시각화 분할창에 표시된 내용을 선택한 디스크립터에 따라 업데이트합니다.

참고: 컨텍스트 메뉴를 사용하여 동의어나 제외 항목으로 유형에 개념을 추가할 수도 있습니다.

제 12 장 텍스트 링크 분석 탐색

텍스트 링크 분석(TLA) 보기에서 텍스트 링크 분석 패턴 결과를 탐색할 수 있습니다. 텍스트 링크 분석은 패턴 규칙을 정의하고 이를 텍스트에서 찾은 실제로 추출된 개념 및 관계와 비교할 수 있게 하는 패턴 매치 기술입니다.

예를 들어, 조직에 대한 아이디어 추출이 충분히 흥미롭지 않을 수 있습니다. TLA를 사용하면 이 조직과 다른 조직 또는 조직 내 사용자 간의 링크에 대해서도 알 수 있습니다. TLA를 사용하여 제품에 대한 의견 또는 일부 언어의 경우 유전자 간의 관계도 추출할 수 있습니다.

일부 TLA 패턴 결과를 추출했으면 텍스트 링크 분석 보기의 유형 및 개념 패턴 분할창에서 검토할 수 있습니다. 자세한 정보는 161 페이지의 『유형 및 개념 패턴』의 내용을 참조하십시오. 이 보기의 데이터 또는 시각화 분할창에서 추가로 탐색할 수 있습니다. 아마 가장 중요하게는 범주에 추가할 수 있습니다.

아직 그렇게 하도록 선택하지 않았으면 **추출**을 클릭하고 추출 설정 대화 상자에서 **텍스트 링크 분석 패턴 추출 사용**을 선택할 수 있습니다. 자세한 정보는 160 페이지의 『TLA 패턴 결과 추출』의 내용을 참조하십시오.

TLA 패턴 결과를 추출하려면 사용 중인 자원 템플리트 또는 라이브러리에 일부 TLA 패턴 규칙이 정의되어 있어야 합니다. IBM SPSS Modeler Text Analytics와 함께 제공되는 일정 자원 템플리트에서 TLA 패턴을 사용할 수 있습니다. 추출할 수 있는 관계 및 패턴 종류는 자원에 정의된 TLA 규칙에 전적으로 좌우됩니다. 일본어를 *제외한* 모든 텍스트 언어에 대해 직접 TLA 규칙을 정의할 수 있습니다. 패턴은 입력 텍스트와 비교되는 부울 쿼리 또는 규칙을 형성하기 위한 매크로, 단어 목록 또는 단어 간격으로 구성됩니다. 자세한 정보는 223 페이지의 제 19 장 『텍스트 링크 규칙에 대한 정보』의 내용을 참조하십시오.

TLA 패턴 규칙이 텍스트와 매치할 때마다 이 텍스트를 패턴으로 추출하고 출력 데이터로 재구성할 수 있습니다. 그러면 텍스트 링크 분석 보기 분할창에 결과가 표시됩니다. 보기 메뉴에서 해당 이름을 선택하여 각 분할창을 숨기거나 표시할 수 있습니다.

- **유형 및 개념 패턴 분할창.** 이 두 개의 분할창에서 패턴을 작성하고 탐색할 수 있습니다. 자세한 정보는 161 페이지의 『유형 및 개념 패턴』의 내용을 참조하십시오.
- **시각화 분할창.** 이 분할창에서 패턴의 개념 및 유형이 상호작용하는 방법을 시각적으로 탐색할 수 있습니다. 자세한 정보는 168 페이지의 『텍스트 링크 분석 그래프』의 내용을 참조하십시오.
- **데이터 분할창.** 다른 분할창의 선택사항에 해당하는 문서 및 레코드 내에 포함된 텍스트를 탐색하고 검토할 수 있습니다. 자세한 정보는 163 페이지의 『데이터 분할창』의 내용을 참조하십시오.

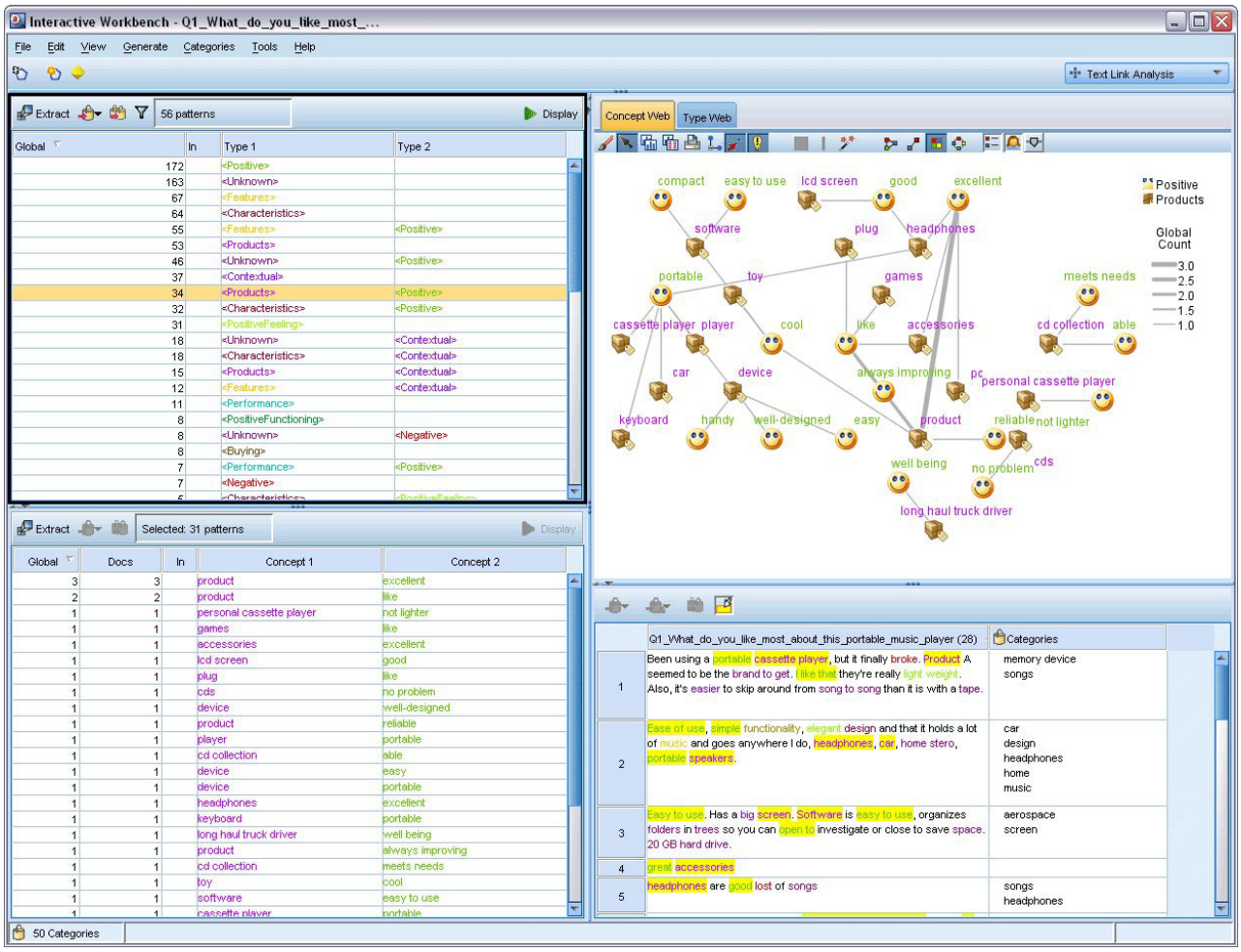


그림 32. 텍스트 링크 분석 보기

TLA 패턴 결과 추출

추출 프로세스는 개념 및 유형 세트뿐만 아니라 텍스트 링크 분석(TLA) 패턴(사용 가능한 경우)을 결과로 생성합니다. TLA 패턴을 추출한 경우 텍스트 링크 분석 보기에서 이를 볼 수 있습니다. 추출 결과가 자원과 동기화되지 않을 때마다 패턴 분할창은 재추출이 다른 결과를 생성함을 나타내는 노랑 색상이 됩니다.

노드 설정 또는 추출 대화 상자에서 **텍스트 링크 분석 패턴 추출 사용** 옵션을 사용하여 이러한 패턴을 추출하도록 선택해야 합니다. 자세한 정보는 92 페이지의 『데이터 추출』의 내용을 참조하십시오.

참고: 데이터 세트의 크기와 추출 프로세스를 완료하는 데 걸리는 시간 간의 관계가 있습니다. 성능 통계 및 권장사항은 설치 지시사항을 참조하십시오. 표본 노드 업스트림 삽입 또는 시스템 구성 최적화를 항상 고려할 수 있습니다.

데이터 추출 방법

1. 메뉴에서 **도구 > 추출**을 선택하십시오. 또는 **추출** 도구 모음 단추를 클릭하십시오.

2. 사용할 옵션을 변경하십시오. TLA 패턴 결과를 추출하려면 이 탭에서 **텍스트 링크 분석 패턴 추출 사용** 옵션을 선택해야 하는 것은 물론 템플리트에 TLA 규칙이 있어야 함을 명심하십시오. 자세한 정보는 92 페이지의 『데이터 추출』의 내용을 참조하십시오.
3. **추출**을 클릭하여 추출 프로세스를 시작하십시오.

추출이 시작되면 진행 대화 상자가 열립니다. 추출을 중단하려면 **취소**를 클릭하십시오. 추출이 완료되면 대화 상자가 닫히고 분할창에 결과가 나타납니다. 자세한 정보는 『유형 및 개념 패턴』의 내용을 참조하십시오.

유형 및 개념 패턴

패턴은 두 개의 파트 즉, 개념과 유형을 조합하여 구성됩니다. 패턴은 특정 주제에 대한 의견 또는 개념 간의 관계를 발견하려고 시도할 때 가장 유용합니다. 예를 들어, 경쟁자의 제품 이름을 추출하는 것이 충분히 흥미롭지 않을 수 있습니다. 이 경우, 추출된 패턴을 살펴 문서 또는 레코드에 제품이 좋음, 나쁨 또는 비쌈을 표현하는 텍스트가 포함된 예제를 찾을 수 있는지 여부를 확인할 수 있습니다.

패턴은 최대 6개의 유형 또는 6개의 개념으로 구성될 수 있습니다. 이러한 이유로 두 패턴 분할창 모두의 행은 최대 6개의 슬롯 또는 위치를 포함합니다. 언어학적 자원에서 정의된 대로 각 슬롯은 TLA 패턴 규칙에서 요소의 특정 위치에 해당합니다. 대화형 워크벤치에서는 슬롯에 값이 포함되지 않은 경우 테이블에 슬롯이 표시되지 않습니다. 예를 들어, 가장 긴 패턴 결과에 단지 4개의 슬롯이 포함된 경우 마지막 2개는 표시되지 않습니다. 자세한 정보는 223 페이지의 제 19 장 『텍스트 링크 규칙에 대한 정보』의 내용을 참조하십시오.

패턴 결과를 추출할 때 패턴 결과는 먼저 유형 수준에서 그룹화된 후 개념 패턴으로 나뉩니다. 이러한 이유로 두 개의 다른 결과 분할창 즉, **유형 패턴**(왼쪽 상단)과 **개념 패턴**(왼쪽 하단)이 있습니다. 리턴된 개념 패턴을 모두 보려면 모든 유형 패턴을 선택하십시오. 그러면 맨 아래 개념 패턴 분할창이 필터 대화 상자에 정의된 대로 순위 최대값까지 개념 패턴을 모두 표시합니다.

유형 패턴. 이 분할창은 TLA 패턴 규칙과 매치하는 하나 이상의 관련 유형으로 구성되는 패턴 결과를 제공합니다. 유형 패턴은 특정 위치의 조직에 대한 긍정적 피드백을 제공할 수 있는 <Organization> + <Location> + <Positive>로 표시됩니다. 구문은 다음과 같습니다.

```
<Type1> + <Type2> + <Type3> + <Type4> + <Type5> + <Type6>
```

개념 패턴. 이 분할창은 그 위의 유형 패턴 분할창에서 현재 선택된 모든 유형 패턴에 대한 패턴 결과를 개념 수준에서 제공합니다. 개념 패턴은 hotel + paris + wonderful과 같은 구조를 따릅니다. 구문은 다음과 같습니다.

```
concept1 + concept2 + concept3 + concept4 + concept5 + concept6
```

패턴 결과가 최대 6개 미만의 슬롯을 사용하는 경우 필요한 수의 슬롯(또는 열)만 표시됩니다. 채워진 두 개의 슬롯 사이에 발견된 빈 슬롯은 버리므로 <Type1>+<>+<Type2>+<>+<>+<> 패턴을 <Type1>+<Type3>으로 나타낼 수 있습니다. 개념 패턴의 경우, 이는 concept1+.+concept2입니다(여기서 .는 널값을 나타냄).

범주 및 개념 보기에서의 추출 결과의 경우와 마찬가지로 여기서 결과를 검토할 수 있습니다. 이러한 패턴을 구성하는 유형 및 개념에 대해 수행할 세분화가 있으면 범주 및 개념 보기의 추출 결과 분할창에서 또는 자원 편집기에서 직접 세분화를 수행하고 패턴을 재추출할 수 있습니다. 개념, 유형 또는 패턴이 범주 정의에서 있는 그대로 또는 규칙의 일부로 사용될 때마다 범주 또는 규칙 아이콘이 패턴 또는 추출 결과 테이블의 **위치** 열에 나타납니다.

TLA 결과 필터링

매우 큰 데이터 세트에 대한 작업 시 추출 프로세스는 수백만 개의 결과를 생성할 수 있습니다. 많은 사용자가 이 양으로 인해 결과를 효과적으로 검토하기가 더 어렵습니다. 그러나 가장 흥미로운 해당 결과에 주목하려면 이러한 결과를 필터링할 수 있습니다. 필터 대화 상자의 설정을 변경하여 표시되는 패턴을 제한할 수 있습니다. 이러한 설정은 모두 함께 사용됩니다.

TLA 보기에서 필터 대화 상자는 다음 영역과 필드를 포함합니다.

빈도별 필터. 일정 글로벌 또는 문서 빈도 값을 가진 해당 결과만 표시하도록 필터링할 수 있습니다.

- **글로벌 빈도**는 전체 문서 또는 레코드 세트에 패턴이 나타나는 총 횟수이며 **글로벌** 열에 표시됩니다.
- **문서 빈도**는 패턴이 나타나는 총 문서 또는 레코드 수이며 **문서** 열에 표시됩니다.

예를 들어, 패턴이 500개 레코드에 300번 나타났으면 이 패턴의 글로벌 빈도는 300이고 문서 빈도는 500입니다.

매치 텍스트별. 여기에 정의하는 규칙과 매치하는 해당 결과만 표시하도록 필터링할 수도 있습니다. **매치 텍스트** 필드에 매치될 문자 세트를 입력한 후 슬롯 번호 또는 모두를 식별하여 개념 또는 유형 이름 내에서 이 텍스트를 검색할 것인지 여부를 선택하십시오. 그리고 나서 매치를 적용할 조건을 선택하십시오(꺾쇠괄호를 사용하여 유형 이름의 시작 또는 끝을 표시하지 않아도 됨). 규칙이 두 명령문 모두 또는 이 중 하나와만 매치하도록 **And** 또는 **Or**을 선택하고 첫 번째와 동일한 방식으로 두 번째 텍스트 매치 명령문을 정의하십시오.

표 35. 매치 텍스트 조건

조건	설명
포함	문자열이 어딘가에 발생하면 텍스트가 매치됩니다(기본 선택사항).
시작 문자	개념 또는 유형이 지정된 텍스트로 시작하는 경우에만 텍스트가 매치됩니다.
끝 문자	개념 또는 유형이 지정된 텍스트로 끝나는 경우에만 텍스트가 매치됩니다.
정확히 일치	전체 문자열이 개념 또는 유형 이름과 매치해야 합니다.

순위별. 또한 글로벌 빈도(**글로벌**) 또는 문서 빈도(**문서**)에 따라 오름차순이나 내림차순으로 순위가 높은 패턴만 표시하도록 필터링할 수 있습니다. 이 최대 순위 값은 표시를 위해 리턴된 총 패턴 수를 제한합니다.

필터가 적용되면 최대 총 개념 패턴 수(순위 최대값)를 초과할 때까지 제품이 유형 패턴을 추가합니다. 순위가 가장 높은 유형 패턴을 살펴보는 것으로 시작한 후 해당 개념 패턴 합계를 사용합니다. 이 합계가 순위 최대값을 초과하지 않으면 패턴이 보기에 표시됩니다. 그리고 나서 다음 유형 패턴의 개념 패턴 수를 합산합니다. 해당 숫자에 이전 유형 패턴의 총 개념 패턴 수를 더한 값이 순위 최대값보다 작으면 해당 패턴도 보기에 표시됩니다. 순위 최대값을 초과하지 않고 가능한 많은 패턴이 표시될 때까지 이러한 작업이 계속됩니다.

패턴 분할창에 표시된 결과

영어 버전의 소프트웨어를 사용 중이라고 가정하십시오. 필터를 기반으로 패턴 분할창 도구 모음에 결과가 표시되는 방법의 몇 가지 예제는 다음과 같습니다.

그림 33. 필터 결과 예제 1

이 예제에서 도구 모음은 순위 최대값이 필터에 지정되기 때문에 리턴된 패턴 수가 제한되었음을 보여줍니다. 보라색 아이콘이 있는 경우 이는 최대 패턴 수가 충족되었음을 의미합니다. 자세한 정보를 보려면 아이콘 위에 마우스를 올려 놓으십시오. **순위별** 필터에 대한 이전 설명을 참조하십시오.

그림 34. 필터 결과 예제 2

이 예제에서 도구 모음은 매치 텍스트 필터를 사용하여 결과가 제한되었음을 표시합니다(돋보기 아이콘 참조). 아이콘 위에 마우스를 올려 놓아 매치 텍스트 내용을 볼 수 있습니다.

결과 필터링 방법

1. 메뉴에서 **도구 > 필터**를 선택하십시오. 필터 대화 상자가 열립니다.
2. 사용할 필터를 선택하고 세분화하십시오.
3. **확인**을 클릭하여 필터를 적용하고 새 결과를 확인하십시오.

데이터 분할창

텍스트 링크 분석 패턴을 추출하고 탐색할 때 작업 중인 일부 데이터를 검토할 수 있습니다. 예를 들어, 패턴 그룹이 발견된 실제 레코드를 볼 수 있습니다. 오른쪽 하단에 있는 데이터 분할창에서 레코드 또는 문서를 검토할 수 있습니다. 기본적으로 표시되지 않으면 메뉴에서 **보기 > 분할창 > 데이터**를 선택하십시오.

데이터 분할창은 일정 표시 한계까지 보기의 선택사항에 해당하는 문서 또는 레코드당 1행을 제공합니다. 기본적으로 데이터 분할창에 표시된 문서 또는 레코드 수는 데이터를 보다 빨리 볼 수 있도록 제한됩니다. 그러나 옵션 대화 상자에서 이를 조정할 수 있습니다. 자세한 정보는 86 페이지의 『옵션: 세션 탭』의 내용을 참조하십시오.

데이터 분할창 표시 및 새로 고침

큰 데이터 세트의 경우 자동 데이터 새로 고침을 완료하려면 약간의 시간이 걸리기 때문에 데이터 분할창은 자동으로 표시를 새로 고치지 않습니다. 따라서 이 보기에서 유형 또는 개념을 선택할 때마다 **표시**를 클릭하여 데이터 분할창의 컨텐츠를 새로 고칠 수 있습니다.

텍스트 문서 또는 레코드

텍스트 데이터가 레코드 양식으로 되어 있고 텍스트의 길이가 비교적 짧으면, 데이터 분할창의 텍스트 필드는 텍스트 데이터를 전부 표시합니다. 그러나 레코드와 큰 데이터 세트에 대한 작업을 할 때 텍스트 필드 열은 텍스트의 짧은 조각을 표시하고 테이블에서 선택한 레코드의 텍스트를 모두 또는 더 많이 표시할 수 있도록 오른쪽에 텍스트 미리보기 분할창을 엽니다. 텍스트 데이터가 개별 문서 양식으로 되어 있으면 데이터 분할창이 문서의 파일 이름을 표시합니다. 문서를 선택하면 선택된 문서의 텍스트와 함께 텍스트 미리보기 분할창이 열립니다.

색상 및 강조표시

데이터를 표시할 때마다 텍스트에서 쉽게 식별할 수 있도록 해당 문서 또는 레코드에서 찾은 개념 및 디스크립터가 색상으로 강조표시됩니다. 색상 코딩은 개념이 속한 유형에 해당합니다. 색상 코드화된 항목 위에 마우스를 올려 놓아 추출된 개념과 지정된 유형도 표시할 수 있습니다. 추출되지 않은 텍스트는 검은색으로 나타납니다. 일반적으로 추출되지 않은 이러한 단어는 접속사(*and* 또는 *with*), 대명사(*me* 또는 *they*), 동사(*is*, *have* 또는 *take*)인 경우가 많습니다.

데이터 분할창 열

텍스트 필드 열이 항상 표시되는 동안에는 다른 열도 표시할 수 있습니다. 다른 열을 표시하려면 메뉴에서 **보기 > 데이터 분할창**을 선택한 후 데이터 분할창에 표시할 열을 선택하십시오. 표시할 수 있는 열은 다음과 같습니다.

- **"텍스트 필드 이름" (#)/문서.** 개념과 유형이 추출된 텍스트 데이터에 열을 추가합니다. 데이터가 문서에 있는 경우, 열을 문서라고 하며 문서 파일 이름 또는 전체 경로만 표시됩니다. 해당 문서에 대한 텍스트를 보려면 텍스트 미리보기 분할창에서 보아야 합니다. 데이터 분할창의 행 수는 이 열 이름 다음에 괄호로 표시됩니다. 옵션 대화 상자에서 로드 속도를 늘리는 데 사용되는 한계 때문에 모든 문서 또는 레코드가 표시되지는 않는 경우가 있습니다. 최대값에 도달하면 숫자 뒤에 **- Max**가 옵니다. 자세한 정보는 86 페이지의 『옵션: 세션 탭』의 내용을 참조하십시오.
- **범주.** 레코드가 속한 범주를 각각 나열합니다. 이 열이 표시될 때마다 데이터 분할창을 새로 고치면 최신 정보를 표시하기 위해 시간이 약간 오래 걸립니다.
- **관련성 순위.** 단일 범주의 각 레코드에 대한 순위를 제공합니다. 이 순위는 해당 범주의 다른 레코드와 비교하여 레코드가 범주에 얼마나 잘 맞는지 보여줍니다. 순위를 보려면 범주 분할창(왼쪽 상단 분할창)에서 범주를 선택하십시오. 자세한 정보는 116 페이지의 『범주 관련성』의 내용을 참조하십시오.
- **범주 수.** 레코드가 속한 범주 수를 나열합니다.

제 13 장 그래프 시각화

범주 및 개념 보기, 군집 보기, 텍스트 링크 분석 보기에는 모두 창의 오른쪽 상단 모서리에 시각화 분할창이 있습니다. 이 분할창을 사용하여 데이터를 시각적으로 탐색할 수 있습니다. 사용 가능한 그래프 및 차트는 다음과 같습니다.

- **범주 및 개념 보기.** 이 보기에는 세 개의 그래프 및 차트(*범주 막대*, *범주 웹*, *범주 웹 테이블*)가 있습니다. 이 보기에서는 **표시**를 클릭하는 경우에만 그래프가 업데이트됩니다. 자세한 정보는 『범주 그래프 및 도표』의 내용을 참조하십시오.
- **군집 보기.** 이 보기에는 두 개의 웹 그래프(*개념 웹 그래프*와 *군집 웹 그래프*)가 있습니다. 자세한 정보는 167 페이지의 『군집 그래프』의 내용을 참조하십시오.
- **텍스트 링크 분석 보기.** 이 보기에는 두 개의 그래프(*개념 웹 그래프*와 *유형 웹 그래프*)가 있습니다. 자세한 정보는 168 페이지의 『텍스트 링크 분석 그래프』의 내용을 참조하십시오.

그래프 편집에 사용되는 모든 일반 도구 모음 및 팔레트에 대한 정보는 온라인 도움말 또는 IBM SPSS Modeler DVD의 *\Documentation\en* 폴더에 있는 *modeler_nodes_general_book.pdf* 파일의 그래프 편집 섹션을 참조하십시오.

범주 그래프 및 도표

범주를 작성할 때 범주 정의, 포함하는 문서 또는 레코드 , 범주가 겹치는 방법을 검토하는 것이 중요합니다. 시각화 분할창은 사용자의 범주에 대한 몇몇 퍼스펙티브를 제공합니다. 시각화 분할창은 범주 및 개념 보기 의 오른쪽 상단 구석에 있습니다. 아직 표시되지 않은 경우에는 보기 메뉴(**보기 > 분할창 > 시각화**)에서 이 분할창에 액세스할 수 있습니다.

이 보기에서는 시각화 분할창은 문서 또는 레코드 범주화에서 일반성에 대한 세 개의 퍼스펙티브를 제공합니다. 이 분할창의 차트와 그래프는 범주화 결과를 분석하는 데 사용하고 범주 또는 보고를 세부 조정하는 데 도움을 줄 수 있습니다. 범주를 세분화할 때 이 분할창을 사용하여 범주 정의를 검토하여 너무 유사하거나(예를 들어, 문서 또는 레코드의 75% 이상을 공유함) 너무 다른 범주를 찾아낼 수 있습니다. 두 개의 범주가 너무 유사한 경우에는 두 개의 범주를 결합하기로 결정하는 것이 도움이 될 수 있습니다. 또는 특정 디스크립터를 한 범주 또는 다른 범주에서 제거하여 범주 정의를 세분화하기로 결정할 수도 있습니다.

추출 결과 분할창 , 범주 분할창 또는 범주 정의 대화 상자에서 선택된 내용에 따라서 이 분할창의 각 탭에서 문서/레코드 및 범주 간의 해당하는 상호작용을 볼 수 있습니다. 각각은 유사한 정보를 제공하지만 서로 다른 방식으로 보여주거나 세부사항의 수준이 다릅니다. 그러나 현재 선택의 그래프를 새로 고치기 위해서는 선택한 분할창 또는 대화 상자의 도구 모음에서 **표시**를 클릭하십시오.

범주 및 개념 보기에서 시각화 분할창은 다음 그래프와 도표를 제공합니다.

- **범주 막대형 차트.** 테이블과 막대형 차트는 사용자의 선택 및 연관된 범주에 해당하는 문서/레코드 간의 겹침을 제공합니다. 막대형 차트는 또한 범주에서 문서/레코드 과 문서/레코드 자세한 정보는 『범주 막대형 차트』의 내용을 참조하십시오.
- **범주 웹 그래프.** 이 그래프는 다른 분할창에서 선택사항에 따라 which the 문서/레코드 이 속하는 범주의 문서/레코드 겹침을 제공합니다. 자세한 정보는 『범주 웹 그래프』의 내용을 참조하십시오.
- **범주 웹 테이블.** 이 테이블은 범주 웹과 동일한 정보를 테이블 형식으로 제공합니다. 테이블에는 열 헤더를 클릭하면 정렬할 수 있는 세 개의 열을 포함합니다. 자세한 정보는 167 페이지의 『범주 웹 테이블』의 내용을 참조하십시오.

자세한 정보는 105 페이지의 제 10 장 『텍스트 데이터 범주화』의 내용을 참조하십시오.

범주 막대형 차트

이 탭은 사용자 선택에 대응하는 문서/레코드와 연관된 범주 사이의 겹침을 표시하는 테이블과 막대형 차트를 표시합니다. 막대형 차트는 또한 문서 또는 레코드 의 총 수에 대한 범주에 있는 문서/레코드 의 비율을 표시합니다. 이 차트의 레이아웃은 편집할 수 없습니다. 그러나 열 헤더를 클릭하여 열을 정렬할 수 있습니다.

차트에는 다음 열이 들어 있습니다.

- **범주.** 이 열은 사용자 선택의 범주 이름을 표시합니다. 기본적으로, 선택에서 가장 일반적인 범주가 처음 나열됩니다.
- **막대.** 이 열은 문서 또는 레코드의 총 수에 대한 주어진 범주에 있는 문서 또는 레코드의 비율을 시각적인 방식으로 표시합니다.
- **선택 %.** 이 열은 선택에서 나타나는 문서 또는 레코드의 총 수에 대한 범주에 대한 문서 또는 레코드의 총 수의 비율을 기반으로 백분율을 표시합니다.
- **문서.** 이 열은 주어진 범주에 대한 선택에서 문서 또는 레코드 의 수를 나타냅니다.

범주 웹 그래프

이 탭은 범주 웹 그래프를 표시합니다. 웹은 다른 분할창에서의 선택에 따라서 문서 또는 레코드 이 속하는 범주에 대한 문서 또는 레코드 겹침을 표시합니다. 범주 레이블이 있으면 이들 레이블이 그래프에 나타납니다. 이 분할창의 도구 모음 단추를 사용하여 그래프 레이아웃(네트워크, 원형, 지시 또는 눈금)을 선택할 수 있습니다.

웹에서 각 노드는 범주를 나타냅니다. 마우스를 사용하여 분할창 안에서 노드를 선택하고 이동할 수 있습니다. 노드의 *크기*는 사용자가 선택한 범주에 대한 문서 또는 레코드 수를 바탕으로 하는 상대 *크기*를 나타냅니다. 두 범주 사이의 선의 두께와 색상이 범주가 갖는 공통 문서 또는 레코드의 수를 나타냅니다. 탐색 모드에서 마우스를 노드 위에서 움직이면 도구팁이 범주의 이름(또는 레이블) 및 범주에 있는 문서 또는 레코드의 전체 수를 표시합니다.

참고: 기본적으로 탐색 모드는 노드를 이동할 수 있는 그래프에서 사용 가능합니다. 그러나 편집 모드로 전환하여 색상, 글꼴, 범례 등을 포함한 그래프 레이아웃을 편집할 수 있습니다. 자세한 정보는 169 페이지의 『그래프 도구 모음 및 팔레트 사용』의 내용을 참조하십시오.

범주 웹 테이블

이 탭은 범주 웹 탭과 동일한 정보를 테이블 형식으로 표시합니다. 테이블에는 열 헤더를 클릭하여 정렬할 수 있는 세 개의 열이 들어 있습니다.

- **개수.** 이 열은 두 범주 사이의 공유 또는 공통 문서 또는 레코드의 수를 표시합니다.
- **범주 1.** 이 열은 첫 번째 범주의 이름과 소괄호 안에 표시된 범주의 문서 또는 레코드의 총 수를 표시합니다.
- **범주 2.** 이 열은 두 번째 범주의 이름과 소괄호 안에 표시된 범주의 문서 또는 레코드의 총 수를 표시합니다.

군집 그래프

군집을 작성한 후에는 시각화 분할창의 웹 그래프에서 시각적으로 탐색할 수 있습니다. 시각화 분할창은 군집화에 대한 두 개의 퍼스펙티브 즉, 개념 웹 그래프와 군집 웹 그래프를 제공합니다. 이 분할창의 웹 그래프를 사용하여 군집 결과를 분석하고 범주에 추가할 일부 개념과 규칙 발견 시 도움을 받을 수 있습니다. 시각화 분할창은 군집 보기의 오른쪽 상단 모서리에 있습니다. 아직 표시되지 않으면 보기 메뉴에서 이 분할창에 액세스할 수 있습니다(**보기 > 분할창 > 시각화**). 군집 분할창에서 군집을 선택하여 시각화 분할창에 해당 그래프를 자동으로 표시할 수 있습니다.

참고: 기본적으로 그래프는 노드를 이동할 수 있는 대화형/선택 모드에 있습니다. 그러나 편집 모드에서 색상과 글꼴, 범례 등을 포함하여 그래프 레이아웃을 편집할 수 있습니다. 자세한 정보는 169 페이지의 『그래프 도구 모음 및 팔레트 사용』의 내용을 참조하십시오.

군집 보기에는 두 개의 웹 그래프가 있습니다.

- **개념 웹 그래프**. 이 그래프는 선택된 군집 내의 모든 개념은 물론 군집 외부의 링크된 개념도 제공합니다. 이 그래프를 통해 군집 내의 개념이 링크되는 방법과 외부 링크를 볼 수 있습니다. 자세한 정보는 『개념 웹 그래프』의 내용을 참조하십시오.
- **군집 웹 그래프**. 이 그래프는 점선으로 표시된 선택된 군집 간의 모든 외부 링크가 있는 선택된 군집을 제공합니다. 자세한 정보는 168 페이지의 『군집 웹 그래프』의 내용을 참조하십시오.

자세한 정보는 153 페이지의 제 11 장 『군집 분석』의 내용을 참조하십시오.

개념 웹 그래프

이 탭은 선택된 군집 내의 개념은 물론 군집 외부의 링크된 개념도 모두 표시합니다. 이 그래프를 통해 군집 내의 개념이 링크되는 방법과 외부 링크를 볼 수 있습니다. 군집의 각 개념은 노드로 표시되며 유형 색상에 따라 색상 코드화됩니다. 자세한 정보는 199 페이지의 『유형 작성』의 내용을 참조하십시오.

군집 내 개념 간의 내부 링크가 그려지며 각 링크의 선 굵기는 그래프 도구 모음에서의 선택사항에 따라 각 개념 쌍의 동시 발생에 대한 문서 개수 또는 유사성 링크 값과 직접 관련됩니다. 군집의 개념과 군집 외부의 해당 개념 간의 외부 링크도 표시됩니다.

군집 정의 대화 상자에서 개념을 선택하면 개념 웹 그래프가 해당 개념 및 해당 개념과 연관된 내부 및 외부 링크를 표시합니다. 선택된 개념 중 하나를 포함하지 않는 다른 개념 간의 링크는 그래프에 나타나지 않습니다.

참고: 기본적으로 그래프는 노드를 이동할 수 있는 대화형/선택 모드에 있습니다. 그러나 편집 모드에서 색상과 글꼴, 범례 등을 포함하여 그래프 레이아웃을 편집할 수 있습니다. 자세한 정보는 169 페이지의 『그래프 도구 모음 및 팔레트 사용』의 내용을 참조하십시오.

군집 웹 그래프

이 탭은 선택된 군집를 보여주는 웹 그래프를 표시합니다. 선택된 군집 간의 외부 링크는 물론 다른 군집 간의 링크도 모두 점선으로 표시됩니다. 군집 웹 그래프에서 각 노드는 전체 군집을 나타내며 이들 사이에 그려진 선 굵기는 두 군집 간의 외부 링크 수를 나타냅니다.

중요! 군집 웹 그래프를 표시하려면 외부 링크가 있는 군집을 이미 작성했어야 합니다. 외부 링크는 별도 군집의 개념 쌍(한 군집 내 개념과 다른 군집에서 외부의 개념) 간의 링크입니다.

예를 들어, 두 개의 군집이 있습니다. `Cluster A`에는 새 개의 개념(`A1`, `A2`, `A3`)이 있습니다. `Cluster B`에는 두 개의 개념(`B1`, `B2`)이 있습니다. 다음 개념이 링크됩니다. `A1-A2`, `A1-A3`, `A2-B1`(외부), `A2-B2`(외부), `A1-B2`(외부), `B1-B2`. 이는 군집 웹 그래프에서 선 굵기가 세 개의 외부 링크를 나타냄을 의미합니다.

참고: 기본적으로 그래프는 노드를 이동할 수 있는 대화형/선택 모드에 있습니다. 그러나 편집 모드에서 색상과 글꼴, 범례 등을 포함하여 그래프 레이아웃을 편집할 수 있습니다. 자세한 정보는 169 페이지의 『그래프 도구 모음 및 팔레트 사용』의 내용을 참조하십시오.

텍스트 링크 분석 그래프

텍스트 링크 분석(TLA) 패턴을 추출한 후에는 시각화 분할창의 웹 그래프에서 시각적으로 탐색할 수 있습니다. 시각화 분할창은 TLA에 대한 두 개의 퍼스펙티브 즉, 개념 (패턴) 웹 그래프와 유형 (패턴) 웹 그래프를 제공합니다. 이 분할창의 웹 그래프를 사용하여 패턴을 시각적으로 표시할 수 있습니다. 시각화 분할창은 텍스트 링크 분석의 오른쪽 상단 모서리에 있습니다. 아직 표시되지 않으면 보기 메뉴에서 이 분할창에 액세스할 수 있습니다(**보기 > 분할창 > 시각화**). 선택사항이 없으면 그래프 영역이 비어 있습니다.

참고: 기본적으로 그래프는 노드를 이동할 수 있는 대화형/선택 모드에 있습니다. 그러나 편집 모드에서 색상과 글꼴, 범례 등을 포함하여 그래프 레이아웃을 편집할 수 있습니다. 자세한 정보는 169 페이지의 『그래프 도구 모음 및 팔레트 사용』의 내용을 참조하십시오.

텍스트 링크 분석 보기에는 두 개의 웹 그래프가 있습니다.

- **개념 웹 그래프**. 이 그래프는 선택된 패턴의 모든 개념을 제공합니다. 개념 그래프에서 선 굵기와 노드 크기(유형 아이콘이 표시되지 않은 경우)는 선택된 테이블에서 글로벌 발생 수를 표시합니다. 자세한 정보는 169 페이지의 『개념 웹 그래프』의 내용을 참조하십시오.

- **유형 웹 그래프**. 이 그래프는 선택된 패턴의 모든 유형을 제공합니다. 그래프에서 선 굵기와 노드 *크기*(유형 아이콘이 표시되지 않은 경우)는 선택된 테이블에서 글로벌 발생 수를 표시합니다. 노드는 유형 색상 또는 아이콘으로 표시됩니다. 자세한 정보는 『유형 웹 그래프』의 내용을 참조하십시오.

자세한 정보는 159 페이지의 제 12 장 『텍스트 링크 분석 탐색』의 내용을 참조하십시오.

개념 웹 그래프

이 웹 그래프는 현재 선택에 표시된 모든 개념을 제공합니다. 예를 들어, 세 개의 매치하는 개념 패턴이 있는 유형 패턴을 선택한 경우 이 그래프는 세 세트의 링크된 개념을 표시합니다. 개념 그래프에서 선 굵기와 노드 크기는 글로벌 빈도 수를 표시합니다. 그래프는 패턴 분할창에서 선택된 것과 동일한 정보를 시각적으로 표시합니다. 각 개념의 유형은 그래프 도구 모음에서의 선택사항에 따라 색상 또는 아이콘으로 제공됩니다. 자세한 정보는 『그래프 도구 모음 및 팔레트 사용』의 내용을 참조하십시오.

유형 웹 그래프

이 웹 그래프는 현재 선택에 대한 각 유형 패턴을 제공합니다. 예를 들어, 두 개의 개념 패턴을 선택한 경우 이 그래프는 선택된 패턴의 유형별 하나의 노드와 동일한 패턴에서 찾은 개념 간의 링크를 표시합니다. 선 굵기와 노드 크기는 세트의 글로벌 빈도 수를 표시합니다. 그래프는 패턴 분할창에서 선택된 것과 동일한 정보를 시각적으로 표시합니다. 그래프에 나타나는 유형 이름 이외에 유형은 그래프 도구 모음에서의 선택사항에 따라 해당 색상 또는 유형 아이콘으로도 식별됩니다. 자세한 정보는 『그래프 도구 모음 및 팔레트 사용』의 내용을 참조하십시오.

그래프 도구 모음 및 팔레트 사용

각 그래프에 대해, 그래프를 사용하여 많은 조치를 수행할 수 있는 몇 가지 공통 팔레트에 대한 빠른 액세스를 제공하는 도구 모음이 있습니다. 각 보기(범주 및 개념, 군집, 텍스트 링크 분석)는 약간 다른 도구 모음을 갖고 있습니다. *탐색* 보기 모드와 *편집* 보기 모드 사이에서 선택할 수 있습니다.

탐색 모드에서는 시각화로 표시되는 데이터 및 값을 분석적으로 탐색할 수 있는 반면, 편집 모드에서는 시각화의 레이아웃 및 모양을 변경할 수 있습니다. 예를 들어, 조직의 스타일 가이드에 맞게 글꼴 및 색상을 변경할 수 있습니다. 이 모드를 선택하려면 메뉴에서 **보기 > 시각화 분할창 > 편집 모드**를 선택하십시오(또는 도구 모음 아이콘을 클릭).

편집 모드에서는 시각화 레이아웃의 다양한 측면에 영향을 주는 여러 도구 모음이 제공됩니다. 사용하지 않는 도구 모음이 있는 경우 이러한 도구 모음을 숨겨 대화 상자에서 그래프가 표시되는 공간을 늘릴 수 있습니다. 도구 모음을 선택 또는 선택 취소하려면 보기 메뉴에서 관련 도구 모음이나 팔레트 이름을 클릭하십시오.

그래프 편집에 사용되는 모든 일반 도구 모음 및 팔레트에 대한 자세한 정보는 온라인 도움말이나 IBM SPSS Modeler DVD의 *\Documentation\en* 폴더에서 볼 수 있는 *modeler_nodes_general_book.pdf* 파일에서 그래프 편집에 대한 절을 참조하십시오.

표 36. 텍스트 분석 도구 모음 단추.

단추/목록	설명
	편집 모드를 활성화합니다. 글꼴을 확대하거나 기업 스타일 가이드와 매치하도록 색상을 변경하거나 레이블 및 범례를 제거하는 등 그래프의 모양을 변경하려면 편집 모드로 전환하십시오.
	탐색 모드를 활성화합니다. 기본적으로 탐색 모드가 켜지는데, 이것은 노드를 그래프 주위로 이동하고 끌고갈 수 있으며 그래프 오브젝트 위에서 움직여서 추가 도구팁 정보를 볼 수 있음을 의미합니다.
	범주 및 개념 보기뿐 아니라 텍스트 링크 분석 보기에서 그래프에 대한 웹 표시의 유형을 선택하십시오. • **원형 레이아웃.** 임의의 그래프에 적용할 수 있는 일반 레이아웃입니다. 링크가 방향이 지정되지 않고 모든 노드를 동일하게 취급한다고 가정하고 그래프를 배치합니다. 노드는 원의 주변에만 배치됩니다. • **네트워크 레이아웃.** 임의의 그래프에 적용할 수 있는 일반 레이아웃입니다. 링크가 방향이 지정되지 않고 모든 노드를 동일하게 취급한다고 가정하고 그래프를 배치합니다. 노드는 레이아웃 안에서 자유롭게 배치됩니다. • **방향이 있는 레이아웃.** 방향이 있는 그래프에만 사용해야 하는 레이아웃입니다. 이 레이아웃은 루트 노드에서 리프 노드를 향하는 트리형 구조를 생성하며 색상으로 구성됩니다. 계층 구조 데이터는 이 레이아웃으로 잘 표시되는 경향이 있습니다. • **눈금 레이아웃.** 임의의 그래프에 적용할 수 있는 일반 레이아웃입니다. 링크가 방향이 지정되지 않고 모든 노드를 동일하게 취급한다고 가정하고 그래프를 배치합니다. 노드는 공간 내의 격자점에만 배치됩니다.
	링크 크기 표시입니다. 그래프에서 선의 두께가 표시하는 것을 선택하십시오. 이것은 군집 보기에만 적용됩니다. 군집 웹 그래프는 군집 사이의 외부 링크 수만 표시합니다. 다음 중에서 선택할 수 있습니다. • **유사성.** 두께는 두 군집 사이의 외부 링크 수를 표시합니다. • **동시 발생.** 두께가 디스크립터의 동시 발생이 발생하는 문서 수를 표시합니다.
	누르면 범례를 표시하는 전환 단추입니다. 이 단추를 누르지 않으면 범례가 표시되지 않습니다.
	누르면 그래프에 유형 색상이 아니라 유형 아이콘을 표시하는 전환 단추입니다. 이것은 텍스트 링크 분석 보기에만 적용됩니다.
	누르면 그래프 아래에 링크 슬라이더를 표시하는 전환 단추입니다. 화살표를 밀어서 결과를 필터링할 수 있습니다.
	하위 범주가 아니라 선택된 범주의 최상위 레벨에 대한 그래프를 표시합니다.
	선택된 범주의 최하위 레벨에 대한 그래프를 표시합니다.
	이 옵션은 하위 범주의 이름이 출력에 표시되는 방법을 제어합니다. • **전체 범주 경로.** 이 옵션은 적용 가능한 경우 슬래시를 사용하여 범주 이름을 하위 범주 이름과 구분하여 범주의 이름 및 상위 범주의 전체 경로를 출력합니다. • **짧은 범주 경로.** 이 옵션은 범주의 이름만 출력하지만 생략 기호를 사용하여 문제가 되는 범주에 대한 상위 범주의 수를 표시합니다. • **최하위 레벨 범주.** 이 옵션은 전체 경로 또는 상위 범주가 표시되지 않으면서 범주의 이름만 출력합니다.

제 14 장 세션 자원 편집기

IBM SPSS Modeler Text Analytics는 텍스트 데이터에서 주요 개념을 신속하고 정확하게 캡처하고 추출합니다. 이 추출 프로세스는 주로 언어학적 자원에 의존하여 텍스트 데이터에서 정보를 추출하는 방법을 지시합니다. 기본적으로 이러한 자원은 자원 템플리트에서 비롯됩니다.

IBM SPSS Modeler Text Analytics는 데이터 처리 및 추출 방법을 쉽게 정의할 수 있도록 라이브러리 및 고급 자원 양식으로 언어 및 비언어학적 자원 세트를 포함하는 **자원 템플리트** 세트와 함께 제공됩니다. 자세한 정보는 175 페이지의 제 15 장 『템플리트 및 자원』의 내용을 참조하십시오.

노드 대화 상자에서 템플리트의 자원 사본을 노드에 로드할 수 있습니다. 대화형 워크벤치 세션 안에 있으면 원할 경우 이 노드의 데이터에 대해 특별히 이러한 자원을 사용자 정의할 수 있습니다. 대화형 워크벤치 세션 동안 자원 편집기 보기에서 자원에 대한 작업을 할 수 있습니다. 대화형 세션이 실행될 때마다, 노드에 데이터 및 추출 결과를 캐시하지 않았으면 노드 대화 상자에 로드된 자원을 사용하여 추출이 수행됩니다.

자원 편집기에서 자원 편집

자원 편집기는 대화형 워크벤치 세션에 대한 추출 결과(개념, 유형, 패턴)를 생성하는 데 사용되는 자원 세트에 대한 액세스를 제공합니다. 자원 편집기에서는 이 세션에 대한 자원을 편집한다는 점을 제외하고 이 편집기는 템플리트 편집기와 매우 유사합니다. 자원 및 완료한 다른 작업에 대한 작업을 완료했으면 모델링 노드를 업데이트하여 후속 대화형 워크벤치 세션에서 복원할 수 있도록 이 작업을 저장할 수 있습니다. 자세한 정보는 88 페이지의 『모델링 노드 업데이트 및 저장』의 내용을 참조하십시오.

노드에 자원을 로드하는 데 사용되는 템플리트에 대해 직접 작업하려면 템플리트 편집기를 사용하는 것이 좋습니다. 자원 편집기에서 수행할 수 있는 많은 태스크가 템플리트 편집기에서와 마찬가지로 수행되는데, 다음과 같습니다.

- **라이브러리에 대한 작업**. 자세한 정보는 187 페이지의 제 16 장 『라이브러리에 대한 작업』의 내용을 참조하십시오.
- **유형 사전 작성**. 자세한 정보는 199 페이지의 『유형 작성』의 내용을 참조하십시오.
- **사전에 용어 추가**. 자세한 정보는 200 페이지의 『용어 추가』의 내용을 참조하십시오.
- **동의어 작성**. 자세한 정보는 206 페이지의 『동의어 정의』의 내용을 참조하십시오.
- **템플리트 가져오기 및 내보내기**. 자세한 정보는 183 페이지의 『템플리트 가져오기 및 내보내기』의 내용을 참조하십시오.
- **라이브러리 출판**. 자세한 정보는 194 페이지의 『라이브러리 출판』의 내용을 참조하십시오.

네덜란드어, 영어, 프랑스어, 독일어, 이탈리아어, 포르투갈어 및 스페인어 텍스트의 경우

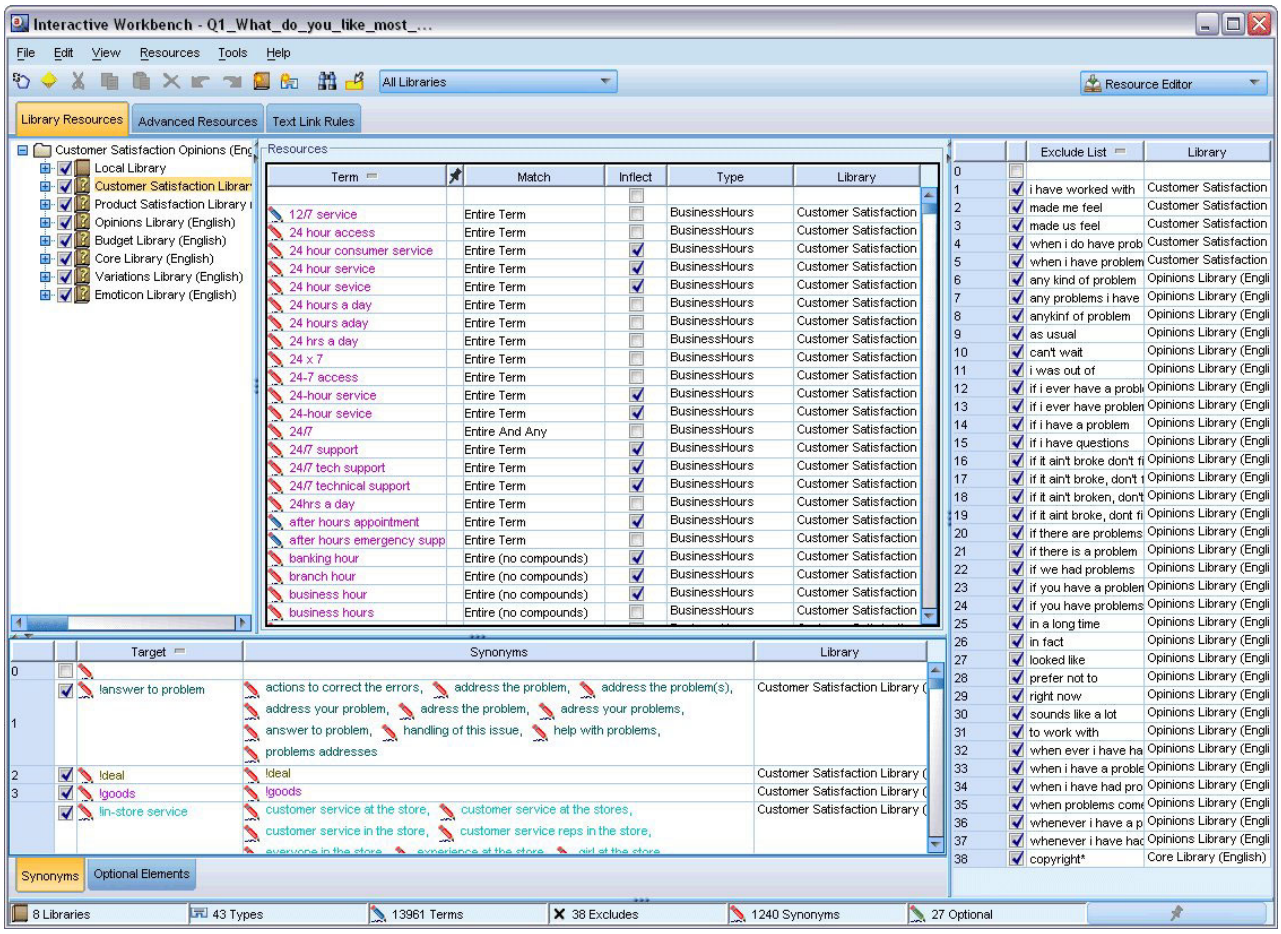


그림 35. 비일본어용 자원 편집기 보기

일본어 텍스트의 경우

일본어 텍스트 언어용 편집기 인터페이스는 다른 텍스트 언어와 다릅니다.

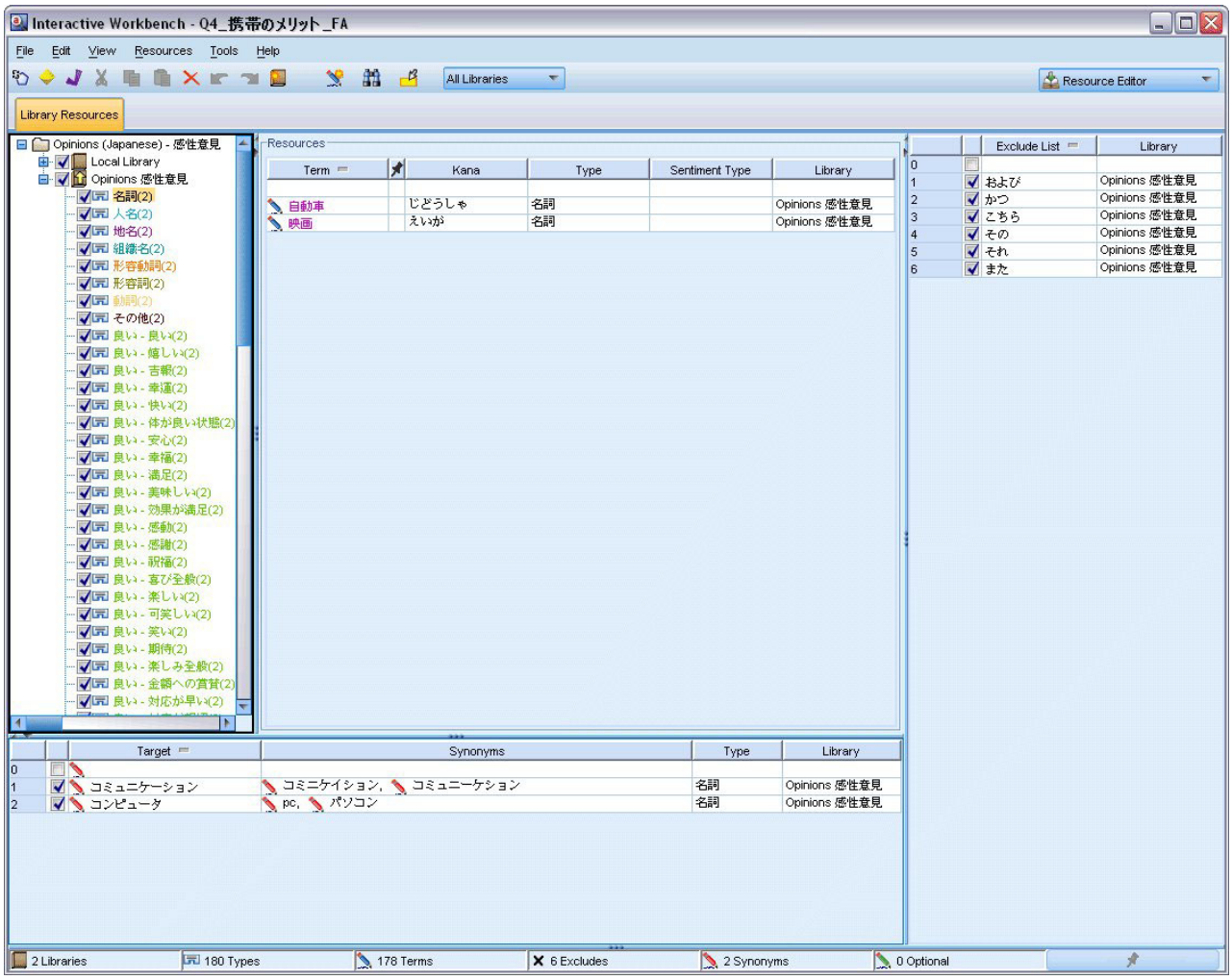


그림 36. 일본어 텍스트용 자원 편집기 보기

템플리트 작성 및 업데이트

자원을 변경하고 나중에 재사용하기 원할 때마다 자원을 템플리트로 저장할 수 있습니다. 그렇게 할 때 기존 템플리트 이름을 사용하거나 새 이름을 제공하여 저장할 것을 선택할 수 있습니다. 그러면, 나중에 이 템플리트를 로드할 때마다 동일한 자원을 얻을 수 있습니다. 자세한 정보는 27 페이지의 『템플리트 및 TAP에서 자원 복사』 주제를 참조하십시오.

참고: 라이브러리를 출판하고 공유할 수도 있습니다. 자세한 정보는 192 페이지의 『라이브러리 공유』의 내용을 참조하십시오.

템플리트를 작성(또는 업데이트)하려면 다음을 수행하십시오.

1. 자원 편집기 보기의 메뉴에서 **자원 > 자원 템플리트 작성**을 선택하십시오. 자원 템플리트 작성 대화 상자가 열립니다.

2. 새 템플리트를 작성하려는 경우 템플리트 이름 필드에 새 이름을 입력하십시오. 기존 템플리트를 현재 로드된 자원으로 덮어쓰려면 테이블에서 템플리트를 선택하십시오.
3. 템플리트를 작성하려면 **저장**을 클릭하십시오.

중요! 템플리트는 노드에서 선택할 때 로드되고 스트림이 실행될 때는 로드되지 않으므로, 최신 변경사항을 얻으려는 경우 템플리트가 사용되는 다른 모든 노드에서 자원 템플리트를 다시 로드하십시오. 자세한 정보는 181 페이지의 『로드 후 노드 자원 업데이트』의 내용을 참조하십시오.

자원 템플리트 전환

세션에 있는 현재 로드된 자원을 다른 템플리트의 자원 사본으로 바꾸려는 경우 해당 자원으로 전환할 수 있습니다. 그렇게 하면 세션에서 현재 로드된 모든 자원을 덮어씁니다. 몇 가지 사전 정의된 텍스트 링크 분석(TLA) 패턴 규칙을 갖기 위해 자원을 전환하려는 경우, 반드시 TLA 열에서 표시된 템플리트를 선택하십시오.

중요! 일본어 템플리트에서 비일본어 템플리트로 또는 그 반대로 전환할 수 없습니다.

자원 전환은 특히 세션 작업(범주, 패턴, 자원)을 복원하지만 다른 세션 작업을 잃지 않고 템플리트로부터 자원의 업데이트된 사본을 로드하려는 경우에 유용합니다. 내용을 자원 편집기에 복사하려는 템플리트를 선택하고 **확인**을 클릭할 수 있습니다. 그러면 이 세션에서 갖고 있는 자원이 대체됩니다. 다음에 대화형 워크벤치 세션을 시작할 때 이들 변경을 보존하려는 경우 세션 종료 시에 모델링 노드를 업데이트하십시오.

참고: 대화형 세션 동안 다른 템플리트의 컨텐츠로 전환하면 노드에 나열된 템플리트의 이름이 여전히 로드 및 복사된 마지막 템플리트의 이름이 됩니다. 이들 자원이나 다른 세션 작업을 활용하기 위해서, 세션을 종료하기 전에 모델링 노드를 업데이트하고 노드에서 **세션 작업 사용** 옵션을 선택하십시오. 자세한 정보는 88 페이지의 『모델링 노드 업데이트 및 저장』의 내용을 참조하십시오.

자원을 전환하려면 다음을 수행하십시오.

1. 자원 편집기 보기의 메뉴에서 **자원 > 자원 템플리트 전환**을 선택하십시오. 자원 전환 대화 상자가 열립니다.
2. 테이블에 표시된 템플리트에서 사용할 템플리트를 선택하십시오.
3. **확인**을 클릭하여 현재 로드된 자원을 중단하고 그 자리에 선택된 템플리트에 있는 자원의 사본을 로드하십시오. 자원을 변경했고 나중에 사용하기 위해 라이브러리를 저장하려는 경우, 전환하기 전에 자원을 출판, 업데이트 및 공유할 수 있습니다. 자세한 정보는 192 페이지의 『라이브러리 공유』 주제를 참조하십시오.

제 15 장 템플리트 및 자원

IBM SPSS Modeler Text Analytics 는 텍스트 데이터에서 주요 개념을 신속하고 정확하게 캡처하고 추출합니다. 이 추출 프로세스는 주로 언어학적 자원에 의존하여 텍스트 데이터에서 정보를 추출하는 방법을 지시합니다. 자세한 정보는 5 페이지의 『추출 작동 방법』의 내용을 참조하십시오. 자원 편집기 보기에서 이러한 자원을 미세 조정할 수 있습니다.

소프트웨어를 설치하면 특수 자원 세트도 얻게 됩니다. 이렇게 제공된 자원을 사용하여 특정 언어 및 특정 애플리케이션에 대한 수년 간의 연구 및 미세 조정을 통해 도움을 받을 수 있습니다. 제공된 자원이 항상 데이터 컨텍스트에 맞게 완벽하게 조정된 것은 아니므로 이러한 자원 템플리트를 편집하거나 조직의 데이터에 맞게 고유하게 미세 조정된 사용자 정의 라이브러리를 작성하여 사용할 수 있습니다. 이러한 자원은 다양한 양식으로 제공되며 각각 세션에서 사용할 수 있습니다. 다음 위치에서 자원을 찾을 수 있습니다.

- **자원 템플리트**. 템플리트는 특정 도메인이나 컨텍스트(예: 제품 의견)에 맞게 조정된 특수 자원 세트를 함께 형성하는 일부 고급 자원, 라이브러리, 유형 세트로 구성됩니다.
- **텍스트 분석 패키지(TAP)**. 템플리트에 저장된 자원 이외에 TAP도 해당 자원을 사용하여 생성된 하나 이상의 특수 범주 세트를 함께 번들화하므로 범주와 자원 모두 함께 저장되고 재사용 가능합니다. 자세한 정보는 147 페이지의 『텍스트 분석 패키지 사용』의 내용을 참조하십시오.
- **라이브러리**. 라이브러리는 TAP과 템플리트 모두에 구성 요소로 사용됩니다. 세션 의 자원에 개별적으로 추가할 수도 있습니다. 각 라이브러리는 유형, 동의어, 제외 목록을 정의하고 관리하는 데 사용되는 몇 개의 사전으로 구성됩니다. 라이브러리는 개별적으로도 제공되지만 템플리트와 TAP에서는 함께 패키지됩니다. 자세한 정보는 187 페이지의 제 16 장 『라이브러리에 대한 작업』의 내용을 참조하십시오.

참고: 추출 중에는 컴파일된 일부 내부 자원도 사용됩니다. 컴파일된 이러한 자원은 코어 라이브러리의 유형을 보완하는 상당수의 정의를 포함합니다. 컴파일된 이러한 자원은 편집할 수 없습니다.

자원 편집기는 추출 결과(개념, 유형, 패턴)를 생성하는 데 사용되는 자원 세트에 대한 액세스를 제공합니다. 자원 편집기에서 수행할 수 있는 다수의 태스크가 있으며 다음과 같습니다.

- **라이브러리에 대한 작업**. 자세한 정보는 187 페이지의 제 16 장 『라이브러리에 대한 작업』의 내용을 참조하십시오.
- **유형 사전 작성**. 자세한 정보는 199 페이지의 『유형 작성』의 내용을 참조하십시오.
- **사전에 용어 추가**. 자세한 정보는 200 페이지의 『용어 추가』의 내용을 참조하십시오.
- **동의어 작성**. 자세한 정보는 206 페이지의 『동의어 정의』의 내용을 참조하십시오.
- **TAP의 자원 업데이트**. 자세한 정보는 149 페이지의 『텍스트 분석 패키지 업데이트』의 내용을 참조하십시오.
- **템플리트 작성**. 자세한 정보는 173 페이지의 『템플리트 작성 및 업데이트』의 내용을 참조하십시오.
- **템플리트 가져오기 및 내보내기**. 자세한 정보는 183 페이지의 『템플리트 가져오기 및 내보내기』의 내용을 참조하십시오.

- **라이브러리 출판**. 자세한 정보는 194 페이지의 『라이브러리 출판』의 내용을 참조하십시오.

템플리트 편집기 vs. 자원 편집기

템플리트, 라이브러리 및 해당 자원에 대한 작업과 편집을 위한 두 가지 주요 방법이 있습니다. 템플리트 편집기 또는 자원 편집기에서 언어학적 자원에 대해 작업할 수 있습니다.

템플리트 편집기

템플리트 편집기를 사용하면 대화형 워크벤치 세션 없이 특정 노드 또는 스트림과 관계 없이 자원 템플리트를 작성하고 편집할 수 있습니다. 이 편집기를 사용하여 텍스트 링크 분석 노드 및 텍스트 마이닝 모델링 노드에 로드하기 전에 자원 템플리트를 작성하거나 편집할 수 있습니다.

도구 > 텍스트 분석 템플리트 편집기 메뉴에서 기본 IBM SPSS Modeler 도구 모음을 통해 템플리트 편집기에 액세스할 수 있습니다.

자원 편집기

대화형 워크벤치 세션에서 액세스 가능한 자원 편집기를 사용하면 특정 노드 및 데이터 세트의 컨텍스트에서 자원에 대한 작업을 할 수 있습니다. 스트림에 텍스트 마이닝 모델링 노드를 추가할 때 자원 템플리트의 컨텐츠 사본 또는 텍스트 분석 패키지 사본(범주 세트 및 자원)을 로드하여 텍스트 마이닝을 위한 텍스트 추출 방법을 제어할 수 있습니다. 대화형 워크벤치 세션을 실행할 때 범주 작성, 텍스트 링크 분석 패턴 추출, 범주 모델 작성 이외에 통합된 자원 편집기 보기에서 해당 세션의 데이터에 대한 자원을 미세 조정할 수도 있습니다. 자세한 정보는 171 페이지의 『자원 편집기에서 자원 편집』의 내용을 참조하십시오.

대화형 워크벤치 세션에서 자원에 대한 작업을 할 때마다 해당 변경사항은 해당 세션에만 적용됩니다. 후속 세션에서 계속할 수 있도록 작업(자원, 범주, 패턴 등)을 저장하려면 모델링 노드를 업데이트해야 합니다. 자세한 정보는 88 페이지의 『모델링 노드 업데이트 및 저장』의 내용을 참조하십시오.

업데이트된 이 템플리트를 다른 노드에 로드할 수 있도록 변경사항을 컨텐츠가 모델링 노드에 복사된 원래 템플리트에 다시 저장하려면 자원으로부터 템플리트를 작성할 수 있습니다. 자세한 정보는 173 페이지의 『템플리트 작성 및 업데이트』의 내용을 참조하십시오.

편집기 인터페이스

템플리트 편집기 또는 자원 편집기에서 수행하는 작업은 언어학적 자원 관리 및 미세 조정을 중심으로 돌아갑니다. 이러한 자원은 템플리트 및 라이브러리 양식으로 저장됩니다. 자세한 정보는 197 페이지의 『유형 사전』의 내용을 참조하십시오.

라이브러리 자원 탭

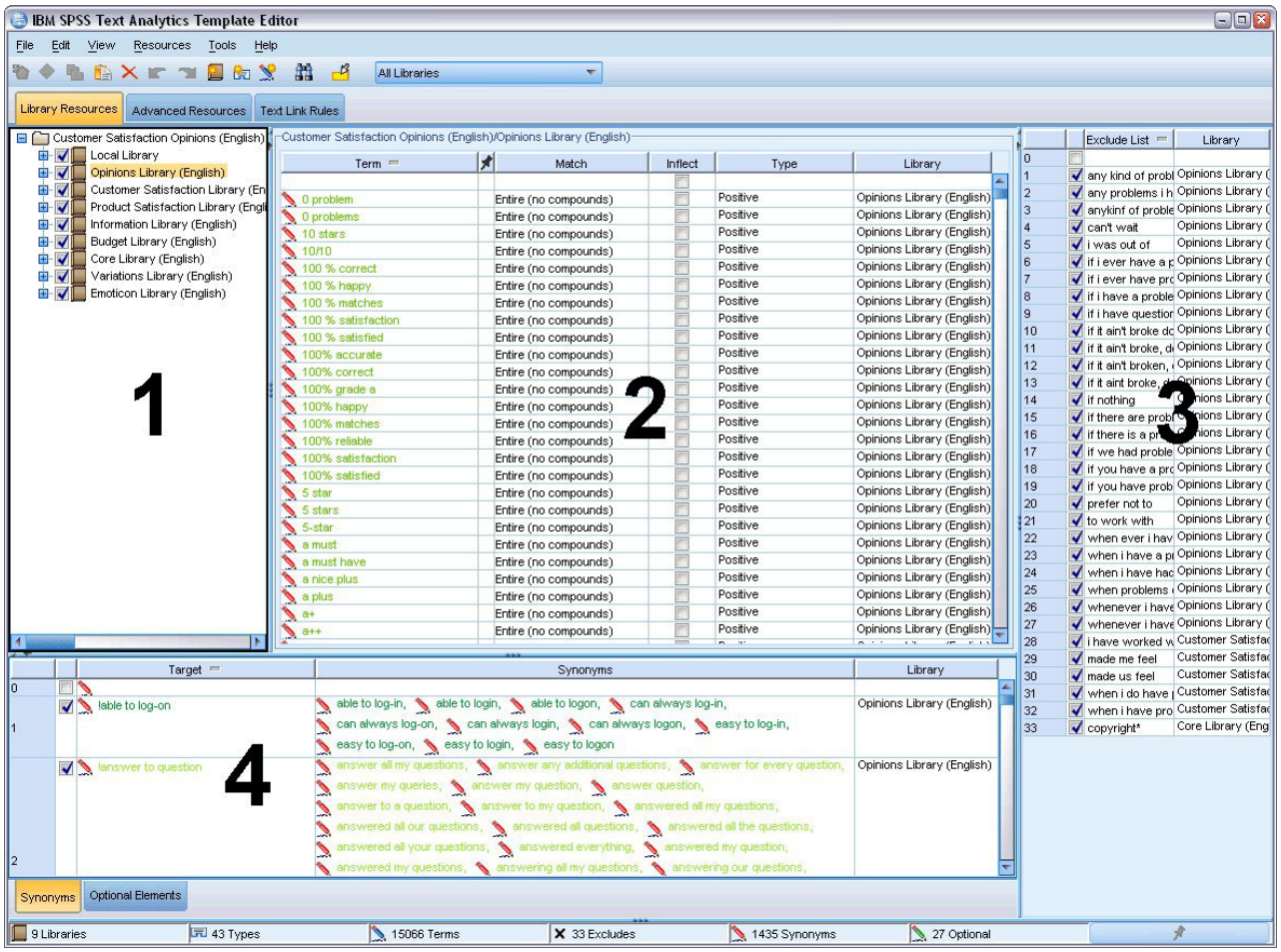


그림 37. 텍스트 마이닝 템플리트 편집기

인터페이스는 다음과 같이 네 개의 파트로 구성됩니다.

1. 라이브러리 트리 분할창. 왼쪽 상단 구석에 있는 이 계획은 라이브러리의 트리를 표시합니다. 이 트리에서 라이브러리를 사용 및 사용 안함으로 설정하고 트리에서 라이브러리를 선택하여 다른 분할창에서 보기를 필터링할 수 있습니다. 컨텍스트 메뉴를 사용하여 이 트리에서 여러 작업을 수행할 수 있습니다. 트리에서 라이브러리를 펼치면 포함된 유형 세트를 볼 수 있습니다. 특정 라이브러리에만 초점을 맞추고 싶으면 **보기** 메뉴를 통해 이 목록을 필터링할 수도 있습니다.

2. 유형 사전 분할창의 용어 목록. 라이브러리 트리의 오른쪽에 있는 이 분할창은 트리에서 선택된 라이브러리의 유형 사전의 용어 목록을 표시합니다. **유형 사전**은 하나의 레이블 또는 유형, 이름 아래에 그룹화될 유형 콜렉션입니다. 추출 엔진은 텍스트 데이터를 읽고 텍스트에서 찾은 단어를 유형 사전의 용어와 비교합니다. 추출된 개념이 유형 사전에 용어로 나타나 있는 경우에는 해당 유형 이름이 지정됩니다. 유형 사전을 공통점이 있는 개별 용어 사전으로 간주할 수 있습니다. 예를 들어, 코어 라이브러리의 `<Location>` 유형에는 `new orleans`, `great britain` 및 `new york` 등과 같은 개념이 포함됩니다. 이러한 용어는 모두 지리적 위치를 나타냅니다. 라이브러리에는 하나 이상의 유형 사전을 포함할 수 있습니다. 자세한 정보는 197 페이지의 『유형 사전』의 내용을 참조하십시오.

3. 제외 사전 분할창. 오른쪽에 있는 이 분할창은 최종 추출 결과에서 제외될 용어 콜렉션을 표시합니다. 이 제외 사전에 나타나는 용어는 추출 결과 분할창에 나타나지 않습니다. 제외된 용어는 선택하는 라이브러리에 저장될 수 있습니다. 그러나, 제외 사전 분할창은 라이브러리 트리에 표시 가능한 모든 라이브러리의 제외된 용어를 모두 표시합니다. 자세한 정보는 208 페이지의 『제외 사전』 주제를 참조하십시오.

4. 대체 사전 분할창. 왼쪽 하단에 위치한 이 분할창에는 동의어 및 선택적 요소가 각자의 탭에 표시됩니다. 동의어 및 선택적 요소는 유사한 용어를 하나의 리드 또는 대상, 최종 추출 결과의 개념 아래에 그룹화합니다. 이 사전에는 알려진 동의어 및 사용자 정의 동의어 및 요소뿐만 아니라 올바른 맞춤법과 쌍을 이룬 자주 틀리는 맞춤법을 포함할 수 있습니다. 동의어 정의 및 선택적 요소는 사용자가 선택하는 라이브러리에 저장될 수 있습니다. 그러나 대체 사전 분할창은 라이브러리 트리에 표시 가능한 모든 라이브러리의 모든 컨텐츠를 표시합니다. 이 분할창은 모든 라이브러리의 모든 동의어 또는 선택적 요소를 표시하지만 트리에 있는 모든 라이브러리의 대체가 이 분할창에 함께 표시됩니다. 라이브러리는 단 하나의 대체 사전만을 포함할 수 있습니다. 자세한 정보는 205 페이지의 『대체/동의어 사전』 주제를 참조하십시오. 선택적 요소 탭은 일본어 텍스트 언어 자원에는 적용되지 않음을 유의하십시오.

참고:

- 단일 라이브러리와 관련된 정보만을 볼 수 있도록 필터링하려는 경우에는 도구 모음의 드롭 다운 목록을 사용하여 라이브러리 보기를 변경할 수 있습니다. 여기에는 **모든 라이브러리**라고 불리는 최상위 수준 항목뿐만 아니라 각 개별 라이브러리의 추가 항목이 포함됩니다. 자세한 정보는 190 페이지의 『라이브러리 보기』의 내용을 참조하십시오.
- 일본어 텍스트 언어용 편집기 인터페이스는 다른 텍스트 언어와 다릅니다.

고급 자원 탭

편집기 보기의 두 번째 탭에서 고급 자원 탭을 사용할 수 있습니다. 이 탭에서 고급 자원을 검토하고 편집할 수 있습니다. 자세한 정보는 211 페이지의 제 18 장 『고급 자원에 대한 정보』의 내용을 참조하십시오.

중요! 일본어 텍스트에 맞게 조정된 자원에는 이 탭을 사용할 수 없습니다.

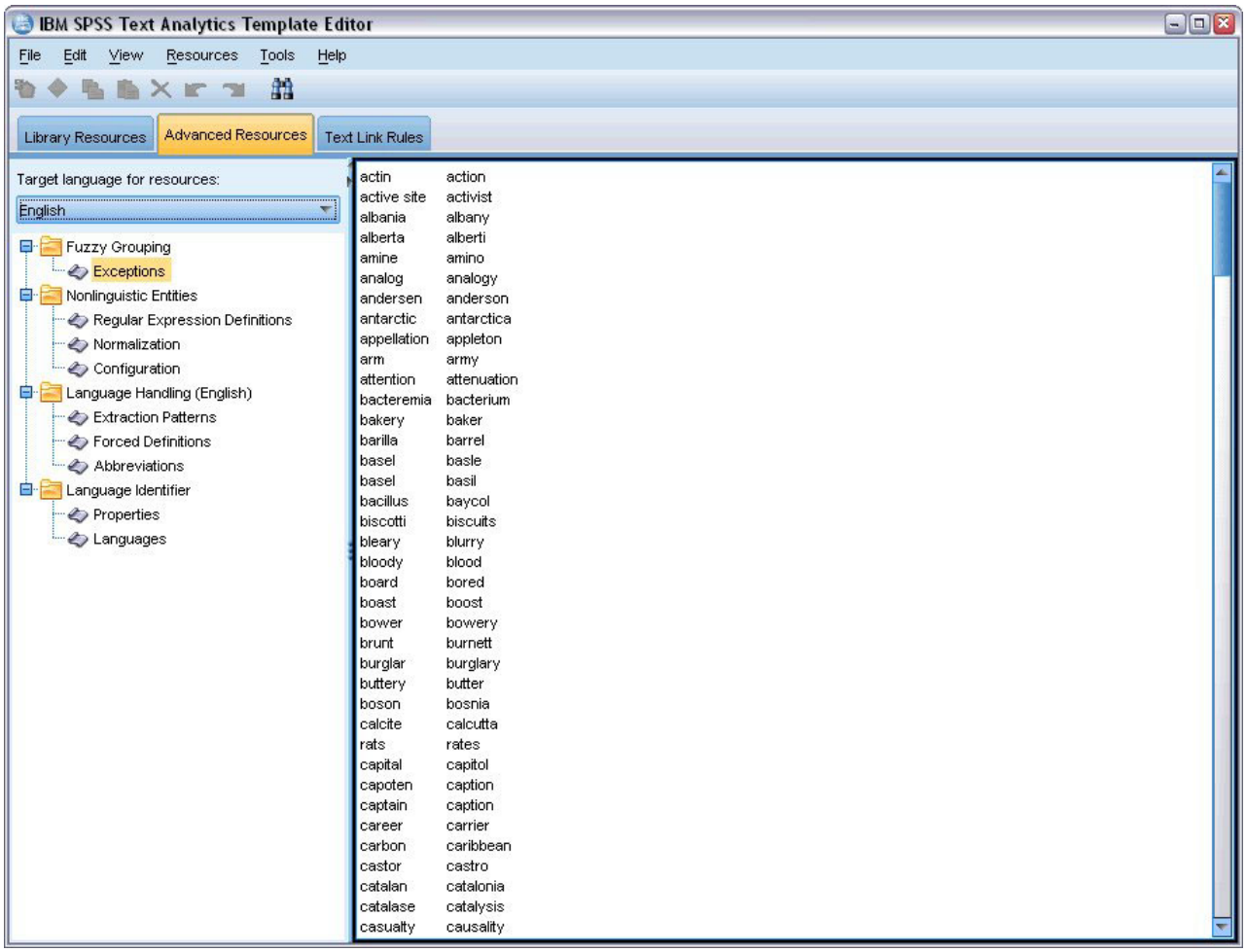


그림 38. 텍스트 마이닝 템플리트 편집기 - 고급 자원 탭

텍스트 링크 규칙 탭

버전 14부터 텍스트 링크 분석 규칙을 편집기 보기의 자체 탭에서 편집할 수 있습니다. 규칙 편집기에서 작업하고 자체 규칙을 작성하며, 시뮬레이션을 실행하여 규칙이 TLA 결과에 어떻게 영향을 주는지도 확인할 수 있습니다. 자세한 정보는 223 페이지의 제 19 장 『텍스트 링크 규칙에 대한 정보』의 내용을 참조하십시오.

중요! 일본어 텍스트에 맞게 조정된 자원에는 이 탭을 사용할 수 없습니다.

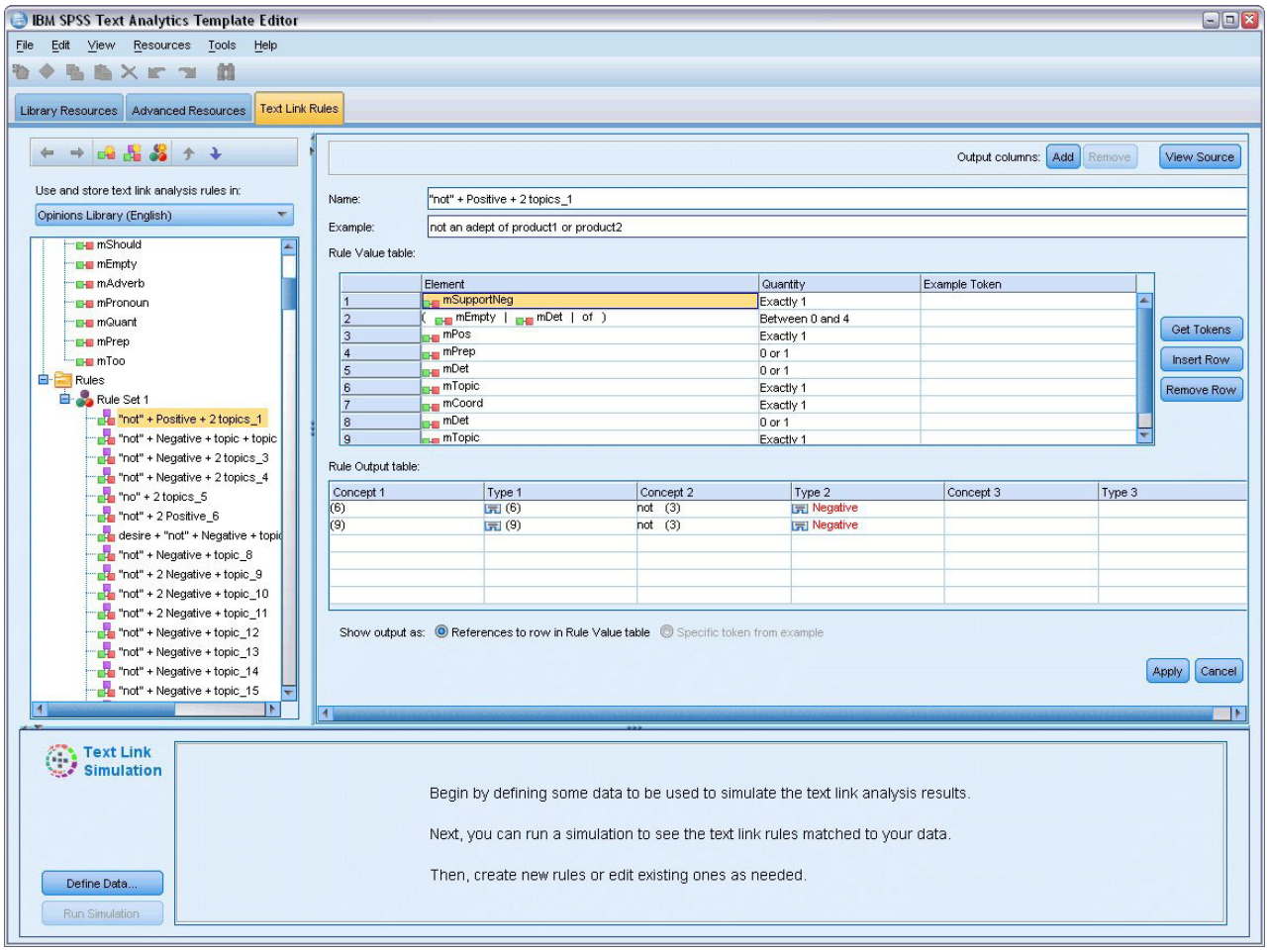


그림 39. 텍스트 마이닝 템플리트 편집기 - 텍스트 링크 규칙 탭

템플리트 열기

템플리트 편집기를 실행하면 템플리트를 열도록 프롬프트가 표시됩니다. 마찬가지로 파일 메뉴에서 템플리트를 열 수 있습니다. 일부 텍스트 링크 분석(TLA) 규칙을 포함하는 템플리트를 원할 경우 TLA 열에 아이콘이 있는 템플리트를 선택하십시오. 템플리트가 작성된 언어가 언어 열에 표시됩니다.

테이블에 표시되지 않은 템플리트를 가져오거나 템플리트를 내보내려면 템플리트 열기 대화 상자의 단추를 사용할 수 있습니다. 자세한 정보는 183 페이지의 『템플리트 가져오기 및 내보내기』의 내용을 참조하십시오.

템플리트를 여는 방법

1. 템플리트 편집기의 메뉴에서 **파일 > 자원 템플리트 열기**를 선택하십시오. 템플리트 열기 대화 상자가 열립니다.
2. 테이블에 표시된 템플리트에서 사용할 템플리트를 선택하십시오.
3. **확인**을 클릭하여 이 템플리트를 여십시오. 편집기에서 다른 템플리트가 현재 열려 있는 경우, 확인을 클릭하면 해당 템플리트를 포기하고 여기서 선택한 템플리트를 표시합니다. 자원을 변경했고 향후 사용을 위해

라이브러리를 저장하려면 다른 템플리트를 열기 전에 라이브러리를 출판, 업데이트, 공유할 수 있습니다. 자세한 정보는 192 페이지의 『라이브러리 공유』의 내용을 참조하십시오.

템플리트 저장

템플리트 편집기에서는 템플리트 변경사항을 저장할 수 있습니다. 기존 템플리트 이름을 사용하거나 새 이름을 제공하여 저장하도록 선택할 수 있습니다.

이전 시간에 노드에 이미 로드한 템플리트를 변경하는 경우, 최신 변경사항을 가져오려면 노드에 템플리트 컨텐츠를 재로드해야 합니다. 자세한 정보는 27 페이지의 『템플리트 및 TAP에서 자원 복사』의 내용을 참조하십시오.

또는 텍스트 마이닝 노드의 모델 탭에서 **저장된 대화형 작업 사용** 옵션을 사용하는 경우(이전 대화형 워크벤치 세션의 자원을 사용함을 의미) 대화형 워크벤치 세션에서 이 템플리트의 자원으로 전환해야 합니다. 자세한 정보는 174 페이지의 『자원 템플리트 전환』의 내용을 참조하십시오.

참고: 라이브러리를 출판하고 공유할 수도 있습니다. 자세한 정보는 192 페이지의 『라이브러리 공유』의 내용을 참조하십시오.

템플리트 저장 방법

1. 템플리트 편집기의 메뉴에서 **파일 > 자원 템플리트 저장**을 선택하십시오. 자원 템플리트 저장 대화 상자가 열립니다.
2. 이 템플리트를 새 템플리트로 저장하려면 템플리트 이름 필드에 새 이름을 입력하십시오. 기존 템플리트를 현재 로드된 자원으로 덮어쓰려면 테이블에서 템플리트를 선택하십시오.
3. 원하면 테이블에 주석 또는 주석(Annotation)을 표시할 설명을 입력하십시오.
4. **저장**을 클릭하여 템플리트를 저장하십시오.

중요! 템플리트 또는 TAP의 자원이 노드에 로드/복사되기 때문에 템플리트를 변경하고 기존 스트림에서 이러한 변경의 도움을 받으려면 재로드하여 자원을 업데이트해야 합니다. 자세한 정보는 『로드 후 노드 자원 업데이트』의 내용을 참조하십시오.

로드 후 노드 자원 업데이트

기본적으로 스트림에 노드를 추가하면 기본 템플리트의 자원 세트가 노드에 로드되어 임베드됩니다. 템플리트를 변경하거나 TAP을 사용하는 경우 로드하면 해당 자원 사본이 자원을 덮어씁니다. 템플리트와 TAP은 노드에 직접적으로 링크되지 않기 때문에 템플리트 또는 TAP 변경사항을 기존 노드에서 자동으로 사용할 수 없습니다. 해당 변경의 도움을 받으려면 해당 노드에서 자원을 업데이트해야 합니다. 두 가지 방법 중 하나로 자원을 업데이트할 수 있습니다.

방법 1: 모델 탭에서 자원 재로드

새 또는 업데이트된 템플리트나 TAP을 사용하여 노드의 자원을 업데이트하려면 노드의 모델 탭에서 재로드할 수 있습니다. 재로드하면 노드의 자원 사본이 최신 사본으로 대체됩니다. 편의를 위해 업데이트된 시간 및 날짜가 원래 템플리트의 이름과 함께 모델 탭에 나타납니다. 자세한 정보는 27 페이지의 『템플리트 및 TAP에서 자원 복사』의 내용을 참조하십시오.

그러나 텍스트 마이닝 모델링 노드에서 대화형 세션 데이터에 대해 작업 중이고 모델 탭에서 **세션 작업 사용** 옵션을 선택한 경우에는 저장된 세션 작업과 자원이 사용되고 **로드** 단추를 사용할 수 없습니다. 대화형 워크벤치 세션 동안 일찍이 **모델링 노드 업데이** 옵션을 선택하고 범주, 자원, 다른 세션 작업을 유지했기 때문에 사용할 수 없습니다. 그런 경우, 해당 자원을 변경하거나 업데이트하려면 자원 편집기에서 자원을 전환하는 다음 방법을 시도할 수 있습니다.

방법 2: 자원 편집기에서 자원 전환

대화형 세션 동안 다른 자원을 사용하려면 언제든 자원 전환 대화 상자를 사용하여 해당 자원을 교환할 수 있습니다. 이는 기존 범주 작업을 재사용하지만 자원을 대체할 경우 특히 유용합니다. 이 경우, 텍스트 마이닝 모델링 노드의 모델 탭에서 **세션 작업 사용** 옵션을 선택할 수 있습니다. 이렇게 하면 노드 대화 상자를 통한 템플리트 재로드 기능을 사용할 수 없으며 대신에 세션 동안 수행된 변경사항과 설정을 유지합니다. 그러면 스트림을 실행하여 대화형 워크벤치 세션을 실행하고 자원 편집기에서 자원을 전환할 수 있습니다. 자세한 정보는 174 페이지의 『자원 템플리트 전환』의 내용을 참조하십시오.

자원을 포함하여 후속 세션에 대해 세션 작업을 유지하려면 자원(및 다른 데이터)이 노드에 다시 저장되도록 대화형 워크벤치 세션 내에서 모델링 노드를 업데이트해야 합니다. 자세한 정보는 88 페이지의 『모델링 노드 업데이트 및 저장』의 내용을 참조하십시오.

참고: 대화형 세션 동안 다른 템플리트의 컨텐츠로 전환하면 노드에 나열된 템플리트의 이름이 여전히 로드 및 복사된 마지막 템플리트의 이름이 됩니다. 이러한 자원 또는 다른 세션 작업의 도움을 받으려면 세션을 종료하기 전에 모델링 노드를 업데이트하십시오.

템플리트 관리

템플리트에 대해 이따금씩 수행할 몇 가지 기본 관리 태스크(예: 템플리트 이름 변경, 템플리트 가져오기 및 내보내기 또는 사용하지 않는 템플리트 삭제)도 있습니다. 이러한 태스크는 템플리트 관리 대화 상자에서 수행됩니다. 템플리트 가져오기 및 내보내기를 사용하여 다른 사용자와 템플리트를 공유할 수 있습니다. 자세한 정보는 183 페이지의 『템플리트 가져오기 및 내보내기』의 내용을 참조하십시오.

참고: 이 제품과 함께 설치된(또는 제공된) 템플리트를 이름 변경하거나 삭제할 수 없습니다. 대신에, 이름을 변경하려면 설치된 템플리트를 열고 선택한 이름을 가진 새 템플리트를 작성할 수 있습니다. 사용자 정의 템플리트를 삭제할 수 있습니다. 그러나 제공된 템플리트를 삭제하려고 하면 원래 설치된 버전으로 재설정됩니다.

템플리트 이름 변경 방법

1. 메뉴에서 **자원 > 자원 템플리트 관리**를 선택하십시오. 템플리트 관리 대화 상자가 열립니다.

2. 이름을 변경할 템플리트를 선택하고 **이름 변경**을 클릭하십시오. 이름 상자가 테이블에서 편집 가능한 필드가 됩니다.
3. 새 이름을 입력하고 Enter 키를 누르십시오. 확인 대화 상자가 열립니다.
4. 이름 변경에 만족하면 **예**를 클릭하십시오. 그렇지 않으면 **아니오**를 클릭하십시오.

템플리트 삭제 방법

1. 메뉴에서 **자원 > 자원 템플리트 관리**를 선택하십시오. 템플리트 관리 대화 상자가 열립니다.
2. 템플리트 관리 대화 상자에서 삭제할 템플리트를 선택하십시오.
3. **삭제**를 클릭하십시오. 확인 대화 상자가 열립니다.
4. **예**를 클릭하여 삭제하거나 **아니오**를 클릭하여 요청을 취소하십시오. **예**를 클릭하면 템플리트가 삭제됩니다.

템플리트 가져오기 및 내보내기

템플리트를 가져오고 내보내 다른 사용자 또는 시스템과 공유할 수 있습니다. 템플리트는 내부 데이터베이스에 저장되지만 하드 드라이브에 **.lrt* 파일로 내보낼 수 있습니다.

템플리트를 가져오거나 내보낼 상황이 있기 때문에 해당 기능을 제공하는 몇 개의 대화 상자가 있습니다.

- 템플리트 편집기의 템플리트 열기 대화 상자
- 텍스트 마이닝 모델링 노드 및 텍스트 링크 분석 노드의 자원 로드 대화 상자
- 템플리트 편집기 및 자원 편집기의 템플리트 관리 대화 상자

템플리트를 가져오는 방법

1. 대화 상자에서 **가져오기**를 클릭하십시오. 템플리트 가져오기 대화 상자가 열립니다.
2. 가져올 자원 템플리트 파일(**.lrt*)을 선택하고 **가져오기**를 클릭하십시오. 가져올 템플리트를 다른 이름으로 저장하거나 기존 템플리트를 덮어쓸 수 있습니다. 대화 상자가 닫히고 이제 템플리트가 테이블에 나타납니다.

템플리트를 내보내는 방법

1. 대화 상자에서 내보낼 템플리트를 선택하고 **확인**을 클릭하십시오. 디렉토리 선택 대화 상자가 열립니다.
2. 내보낼 디렉토리를 선택하고 **내보내기**를 클릭하십시오. 이 대화 상자가 닫히며, 템플리트를 내보내고 파일 확장자(**.lrt*)를 수반합니다.

템플리트 편집기 종료

템플리트 편집기에서 작업을 완료했으면 작업을 저장하고 편집기를 종료할 수 있습니다.

템플리트 편집기 종료 방법

1. 메뉴에서 **파일 > 닫기**를 선택하십시오. 저장 및 닫기 대화 상자가 열립니다.
2. 편집기를 닫기 전에 열린 템플리트를 저장하려면 **템플리트 변경사항 저장**을 선택하십시오.

3. 편집기를 닫기 전에 열린 템플리트의 라이브러리를 출판하려면 **라이브러리 출판**을 선택하십시오. 이 옵션을 선택하면 출판할 라이브러리를 선택하도록 프롬프트가 표시됩니다. 자세한 정보는 194 페이지의 『라이브러리 출판』의 내용을 참조하십시오.

자원 백업

보안 조치로 이따금씩 자원을 백업할 수 있습니다.

중요! 복원할 때는 자원의 전체 컨텐츠를 깨끗하게 지우고 제품에서 백업 파일의 컨텐츠에만 액세스할 수 있습니다. 여기에는 열린 작업이 포함됩니다.

참고: 동일한 주 버전의 소프트웨어만 백업하고 복원할 수 있습니다. 예를 들어, 버전 15에서 백업하는 경우에는 해당 백업을 버전 16으로 복원할 수 없습니다.

자원 백업 방법

1. 메뉴에서 **자원 > 백업 도구 > 자원 백업**을 선택하십시오. 백업 대화 상자가 열립니다.
2. 백업 파일의 이름을 입력하고 **저장**을 클릭하십시오. 대화 상자가 닫히고 백업 파일이 작성됩니다.

자원 복원 방법

1. 메뉴에서 **자원 > 백업 도구 > 자원 복원**을 선택하십시오. 경고가 표시되어 복원하면 데이터베이스의 현재 컨텐츠를 덮어쓰게 됨을 경고합니다.
2. **예**를 클릭하여 진행하십시오. 대화 상자가 열립니다.
3. 복원할 백업 파일을 선택하고 **열기**를 클릭하십시오. 대화 상자가 닫히고 애플리케이션에서 자원이 복원됩니다.

자원 파일 가져오기

이 제품 외부에서 자원 파일에 직접 변경사항을 작성한 경우, 해당 라이브러리를 선택하고 가져오기로 진행하여 선택된 라이브러리로 파일을 가져올 수 있습니다. 디렉토리를 가져올 때, 모든 지원 파일을 특정의 열린 라이브러리로 가져올 수 있습니다. **.txt* 파일만 가져올 수 있습니다.

중요! 일본어 파일의 경우, 가져오려는 .txt 파일은 UTF8로 인코딩해야 합니다. 또한, 일본어에 대한 제외 목록을 가져올 수 없습니다.

가져온 각 파일에는 행당 하나의 항목만 포함하며, 내용이 다음과 같이 구조화되는 경우

- 목록 단어 또는 구문(행마다 하나). 파일은 유형 사전의 용어 목록으로 가져옵니다. 유형 사전은 파일 이름에서 확장자를 제외하고 사용합니다.
- *term1* <TAB> *term2*와 같은 항목 목록은 동의어 목록으로 가져옵니다. 여기서 *term1*은 기본적인 용어 세트이며 *term2*는 대상 용어입니다.

단일 자원 파일 가져오기

1. 메뉴에서, **자원 > 파일 가져오기 > 단일 파일 가져오기**를 선택하십시오. 파일 가져오기 대화 상자가 열립니다.
2. 가져오려는 파일을 선택하고 **가져오기**를 클릭하십시오. 파일 내용은 내부 형식으로 변환되고 라이브러리에 추가됩니다.

디렉토리에서 모든 파일 가져오기

1. 메뉴에서 **자원 > 파일 가져오기 > 전체 디렉토리 가져오기**를 선택하십시오. 디렉토리 가져오기 대화 상자가 열립니다.
2. 모든 자원 파일을 **가져오기** 목록에서 가져오려고 하는 라이브러리를 선택하십시오. **기본값** 옵션을 선택하면, 새 라이브러리는 해당 이름으로 디렉토리의 이름을 사용하여 작성됩니다.
3. 파일을 가져올 디렉토리를 선택하십시오. 서브디렉토리는 읽지 않습니다.
4. **가져오기**를 클릭하십시오. 대화 상자가 닫히고 가져온 자원 파일의 내용이 사전 및 고급 자원 파일 양식으로 편집기에 나타납니다.

제 16 장 라이브러리에 대한 작업

텍스트 데이터에서 용어를 추출하고 그룹화하기 위해 추출 엔진에서 사용되는 자원에는 항상 하나 이상의 라이브러리가 포함됩니다. 템플리트 편집기 및 자원 편집기의 상단 왼쪽 부분에 있는 라이브러리 트리에 라이브러리 세트가 표시될 수 있습니다. 라이브러리는 세 가지 종류의 사전인 유형, 대체 및 제외로 구성됩니다. 자세한 정보는 197 페이지의 제 17 장 『라이브러리 사전 정보』 주제를 참조하십시오.

선택한 TAP의 자원 또는 자원 템플리트에는 텍스트 데이터에서 개념 추출을 즉시 시작할 수 있도록 하는 몇 개의 라이브러리가 포함되어 있습니다. 그러나, 자신의 고유 라이브러리를 작성하고 재사용할 수 있도록 이 라이브러리를 출판할 수도 있습니다. 자세한 정보는 194 페이지의 『라이브러리 출판』의 내용을 참조하십시오.

예를 들어, 자동차 산업에 관련된 텍스트 데이터에 대해 자주 작업한다고 가정해 보십시오. 데이터를 분석한 후, 산업 특정의 어휘 또는 전문어를 처리하기 위해 일부 사용자 정의 자원을 작성할 것을 결정합니다. 템플리트 편집기 를 사용하여, 새 템플리트를 작성하고, 이 템플리트에서 자동차 용어를 추출하고 그룹화하기 위해 라이브러리를 작성할 수 있습니다. 이 라이브러리의 정보를 다시 필요하게 되므로, **라이브러리 관리** 대화 상자에서 액세스 가능한 중앙 저장소에 라이브러리를 출판하여, 다른 스트림 세션 에서 독립적으로 재사용할 수 있도록 합니다.

또한 전자 장치, 엔진, 냉각 시스템 또는 특정 제조업체나 시장과 같은 다양한 하위 산업에 특정한 용어를 그룹화하는 데 관심이 있다고 가정해 보십시오. 그룹마다 라이브러리를 작성한 후 여러 텍스트 데이터 세트에 사용할 수 있도록 라이브러리를 출판할 수 있습니다. 이 방식에서, 텍스트 데이터에 가장 잘 맞는 라이브러리를 추가할 수 있습니다.

참고: 고급 자원 탭에서 추가 자원을 구성하고 관리할 수 있습니다. 일부는 모든 라이브러리에 적용되고, 비언어 엔티티, 퍼지 그룹화 예외 등을 관리합니다. 또한, 텍스트 링크 분석 패턴 규칙을 편집할 수 있습니다. 이 규칙은 텍스트 링크 규칙 탭에서 라이브러리에 특정한 규칙입니다. 자세한 정보는 211 페이지의 제 18 장 『고급 자원에 대한 정보』의 내용을 참조하십시오.

제공된 라이브러리

기본적으로 IBM SPSS Modeler Text Analytics 사전 형식화된 이러한 라이브러리를 사용하여 사전 정의된 수천 개의 용어와 동의어는 물론 다수의 많은 유형에 액세스할 수 있습니다. 제공된 이러한 라이브러리는 여러 도메인에 맞게 미세 조정되며 여러 언어로 사용 가능합니다.

라이브러리는 많지만 가장 일반적으로 사용되는 라이브러리는 다음과 같습니다.

- **로컬 라이브러리.** 사용자 정의 사전을 저장하는 데 사용됩니다. 기본적으로 모든 자원에 추가된 빈 라이브러리입니다. 빈 유형 사전도 포함합니다. 범주 및 개념 보기, 군집 보기, 텍스트 링크 분석 보기 에서 자원을 직접 변경하거나 세분화할 경우(예: 유형에 단어 추가) 가장 유용합니다. 이 경우, 변경사항 및 세분화 사항

은 자원 편집기의 라이브러리 트리에 나열된 첫 번째 라이브러리(기본적으로 이는 *로컬 라이브러리*임)에 자동으로 저장됩니다. 세션 데이터에 특정하기 때문에 이 라이브러리를 출판할 수는 없습니다. 컨텐츠를 출판하려면 먼저 라이브러리 이름을 변경해야 합니다.

- **코어 라이브러리.** 사용자, 위치, 조직, 제품, 알 수 없음을 나타내는 기본적인 5개의 내장 유형으로 구성되기 때문에 대부분의 경우에 사용됩니다. 유형 사전 중 하나에 나열된 몇 개의 용어만 볼 수 있긴 하지만 코어 라이브러리에 표시된 유형은 텍스트 마이닝 제품과 함께 제공되는 컴파일된 내부 자원에서 찾은 로버스트 유형을 실제로 보완합니다. 이러한 컴파일된 내부 자원은 각 유형에 대해 수천 개의 용어를 포함합니다. 이러한 이유로 유형 사전 용어 목록에 용어가 표시되지 않더라도 여전히 추출하여 코어 유형으로 유형을 지정할 수 있습니다. 이는 코어 라이브러리의 `<Person>` 유형 사전에 *John*이 나타나는 경우에만 *George*와 같은 이름을 추출하고 `<Person>`으로 유형을 지정할 수 있는 방법을 설명합니다. 마찬가지로 코어 라이브러리를 포함시키지 않으면 이러한 유형이 포함된 컴파일된 자원을 추출 엔진이 여전히 사용하기 때문에 추출 결과에서 이러한 유형을 여전히 볼 수 있습니다.
- **Opinions 라이브러리.** 텍스트 데이터에서 의견과 정서를 추출하는 데 가장 일반적으로 사용됩니다. 이 라이브러리는 주제에 대한 의견을 표시하는 태도, 규정자, 기본 설정(다른 용어와 함께 사용되는 경우)을 나타내는 수천 개의 단어를 포함합니다. 이 라이브러리는 다수의 내장 유형, 동의어, 제외를 포함합니다. 텍스트 링크 분석에 사용되는 큰 패턴 규칙 세트도 포함합니다. 이 라이브러리의 텍스트 링크 분석 규칙과 이 규칙이 생성하는 패턴 결과가 도움이 되려면 이 라이브러리를 텍스트 링크 규칙 탭에 지정해야 합니다. 자세한 정보는 223 페이지의 제 19 장 『텍스트 링크 규칙에 대한 정보』 주제를 참조하십시오.
- **예산 라이브러리.** 어떤 것의 비용을 참조하는 용어를 추출하는 데 사용됩니다. 이 라이브러리는 어떤 것의 가격이나 품질에 관한 형용사, 규정자, 판단을 나타내는 많은 단어와 문구를 포함합니다.
- **변형 라이브러리.** 일정 언어 변형을 적절히 그룹화하려면 동의어 정의가 필요한 경우를 포함시키는 데 사용됩니다. 이 라이브러리는 동의어 정의만 포함합니다.

템플리트 외부에 제공된 일부 라이브러리가 일부 템플리트의 컨텐츠와 유사하긴 하지만 템플리티는 특정 애플리케이션에 맞게 특별히 조정되었으며 추가 고급 자원을 포함합니다. 작업할 텍스트 데이터 종류에 맞게 설계된 템플리트를 사용하고 보다 일반적인 템플리트에 단순히 개별 라이브러리를 추가하기 보다는 해당 자원을 변경하는 것이 좋습니다.

컴파일된 자원은 또한 IBM SPSS Modeler Text Analytics 항상 추출 프로세스 중에 사용되며 기본 라이브러리에 내장 유형 사전에 대한 상당수의 보완 정의를 포함합니다. 이러한 자원은 컴파일되기 때문에 보거나 편집할 수 없습니다. 그러나 이러한 컴파일된 자원이 유형 지정한 용어를 다른 사전에 강제 실행할 수 있습니다. 자세한 정보는 203 페이지의 『용어 강제 실행』의 내용을 참조하십시오.

라이브러리 작성

라이브러리를 얼마든지 작성할 수 있습니다. 새 라이브러리를 작성한 후 이 라이브러리에서 유형 사전 작성을 시작하고 용어, 동의어, 제외를 입력할 수 있습니다.

라이브러리 작성 방법

1. 메뉴에서 **자원 > 새 라이브러리**를 선택하십시오. 라이브러리 특성 대화 상자가 열립니다.

2. 이름 텍스트 상자에 라이브러리 이름을 입력하십시오.
3. 원하면 주석(Annotation) 텍스트 상자에 주석을 입력하십시오.
4. 라이브러리에 무언가를 입력하기 전에 지금 이 라이브러리를 출판하려면 **출판**을 클릭하십시오. 자세한 정보는 192 페이지의 『라이브러리 공유』의 내용을 참조하십시오. 나중에 언제든지 출판할 수도 있습니다.
5. **확인**을 클릭하여 라이브러리를 작성하십시오. 대화 상자가 닫히고 라이브러리가 트리 보기에 나타납니다. 트리에서 라이브러리를 펼치면 빈 유형 사전이 라이브러리에 자동으로 추가되었음을 알 수 있습니다. 여기에서 용어 추가를 즉시 시작할 수 있습니다. 자세한 정보는 200 페이지의 『용어 추가』의 내용을 참조하십시오.

공용 라이브러리 추가

다른 세션 데이터의 라이브러리를 재사용하는 경우, 공용 라이브러리이면 현재 자원에 추가할 수 있습니다. **공용 라이브러리**는 출판된 라이브러리입니다. 자세한 정보는 194 페이지의 『라이브러리 출판』의 내용을 참조하십시오.

중요! 일본어 라이브러리를 비일본어 자원에 추가하거나 비일본어 자원에 일본어 라이브러리를 추가할 수 없습니다.

공용 라이브러리를 추가할 때 **로컬** 사본이 세션 데이터에 임베드됩니다. 이 라이브러리를 변경할 수 있습니다. 그러나 변경사항을 공유하려면 공용 버전의 라이브러리를 재출판해야 합니다.

공용 라이브러리를 추가할 때 한 라이브러리의 용어 및 유형과 다른 로컬 라이브러리의 용어 및 유형 간에 충돌이 발견되면 충돌 해결 대화 상자가 나타날 수 있습니다. 이 작업을 완료하려면 이러한 충돌을 해결하거나 제안된 해결책을 승인해야 합니다. 자세한 정보는 195 페이지의 『충돌 해결』의 내용을 참조하십시오.

참고: 대화형 워크벤치 세션을 실행하거나 세션을 닫을 때 출판 하는 경우 라이브러리를 항상 업데이트하면 라이브러리가 동기화되지 않을 가능성이 낮습니다. 자세한 정보는 192 페이지의 『라이브러리 공유』의 내용을 참조하십시오.

라이브러리 추가 방법

1. 메뉴에서 **자원 > 라이브러리 추가**를 선택하십시오. 라이브러리 추가 대화 상자가 열립니다.
2. 목록에서 라이브러리를 선택하십시오.
3. **추가**를 클릭하십시오. 새로 추가된 라이브러리와 이미 거기에 있던 라이브러리 간에 충돌이 발생하는 경우, 작업을 완료하기 전에 충돌 해결책을 확인하거나 라이브러리를 변경하라는 요청을 받습니다. 자세한 정보는 195 페이지의 『충돌 해결』의 내용을 참조하십시오.

용어 및 유형 찾기

찾기 기능을 사용하여 편집기의 여러 분할창에서 검색할 수 있습니다. 편집기의 메뉴에서 **편집 > 찾기**를 선택할 수 있으며 찾기 도구 모음이 나타납니다. 이 도구 모음을 사용하여 한 번에 하나의 발생을 찾을 수 있습니다. **찾기**를 다시 클릭하여 검색어의 후속 발생을 찾을 수 있습니다.

검색 시 편집기는 찾기 도구 모음의 드롭 다운 목록에 나열된 라이브러리만 검색합니다. **모든 라이브러리**가 선택된 경우 프로그램은 편집기에서 모든 것을 검색합니다.

검색을 시작할 때 검색은 초점이 있는 영역에서 시작됩니다. 검색은 각 섹션을 통해 계속되며 활성 셀로 돌아갈 때까지 루프백됩니다. 방향 화살표를 사용하여 검색 순서를 반대로 할 수 있습니다. 검색이 대소문자를 구분하는지 여부도 선택할 수 있습니다.

보기에서 문자열을 찾는 방법

1. 메뉴에서 **편집 > 찾기**를 선택하십시오. 찾기 도구 모음이 나타납니다.
2. 검색할 문자열을 입력하십시오.
3. **찾기** 단추를 클릭하여 검색을 시작하십시오. 용어 또는 유형의 다음 발생이 강조표시됩니다.
4. 단추를 다시 클릭하여 발생 간에 이동하십시오.

라이브러리 보기

하나의 특정 라이브러리 또는 모든 라이브러리의 컨텐츠를 표시할 수 있습니다. 이는 많은 라이브러리를 처리하거나 출판하기 전에 특정 라이브러리의 컨텐츠를 검토하려는 경우 유용합니다. 보기 변경은 이 라이브러리 자원 탭에 표시되는 내용에만 영향을 주지만 추출 중에 라이브러리를 사용할 수 없게 하지는 않습니다. 자세한 정보는 191 페이지의 『로컬 라이브러리 사용 안함』의 내용을 참조하십시오.

기본 보기는 트리에 모든 라이브러리를 표시하고 다른 분할창에 해당 컨텐츠를 표시하는 **모든 라이브러리**입니다. 도구 모음의 드롭 다운 목록을 사용하여 또는 메뉴 선택(**보기 > 라이브러리**)을 통해 이 선택을 변경할 수 있습니다. 단일 라이브러리를 보는 경우에는 다른 라이브러리의 모든 항목이 보기에서 사라지지만 추출 중에 여전히 읽을 수 있습니다.

라이브러리 보기 변경 방법

1. 라이브러리 자원 탭의 메뉴에서 **보기 > 라이브러리**를 선택하십시오. 모든 로컬 라이브러리가 있는 메뉴가 열립니다.
2. 볼 라이브러리를 선택하거나, 모든 라이브러리의 컨텐츠를 보려면 **모든 라이브러리** 옵션을 선택하십시오. 보기 컨텐츠는 선택사항에 따라 필터링됩니다.

로컬 라이브러리 관리

로컬 라이브러리는 공용 라이브러리와 대조적으로 대화형 워크벤치 세션 안에 있거나 템플리트 안에 있는 라이브러리입니다. 자세한 정보는 192 페이지의 『공용 라이브러리 관리』의 내용을 참조하십시오. 다음을 로컬 라이브러리 이름 변경, 사용 안함 또는 삭제를 포함하여 수행할 몇 가지 기본 로컬 라이브러리 관리 태스크도 있습니다.

로컬 라이브러리 이름 변경

로컬 라이브러리의 이름을 변경할 수 있습니다. 로컬 라이브러리의 이름을 변경하는 경우 공용 버전이 존재하면 공용 버전과 분리됩니다. 이는 후속 변경사항을 공용 버전과 더 이상 공유할 수 없음을 의미합니다. 이 로컬 라이브러리를 새 이름으로 재출판할 수 있습니다. 이는 원래 공용 버전을 이 로컬 버전에 수행하는 변경사항으로 업데이트할 수 없음도 의미합니다.

참고: 공용 라이브러리의 이름을 변경할 수 없습니다.

1. 메뉴에서 **편집 > 라이브러리 특성**을 선택하십시오. 라이브러리 특성 대화 상자가 열립니다.

로컬 라이브러리 이름 변경 방법

1. 트리 보기에서 이름을 변경할 라이브러리를 선택하십시오.
2. 이름 텍스트 상자에 라이브러리의 새 이름을 입력하십시오.
3. **확인**을 클릭하여 라이브러리의 새 이름을 승인하십시오. 대화 상자가 닫히고 라이브러리 이름이 트리 보기에서 업데이트됩니다.

로컬 라이브러리 사용 안함

추출 프로세스에서 일시적으로 라이브러리를 제외하려면 트리 보기에서 라이브러리 이름 왼쪽의 확인 상자를 선택 취소할 수 있습니다. 이는 라이브러리를 유지하지만 충돌 검사 시와 추출 중에는 컨텐츠를 무시함을 나타냅니다.

라이브러리를 사용 안함으로 설정하는 방법

1. 라이브러리 트리 분할창에서 사용 안함으로 설정할 라이브러리를 선택하십시오.
2. 스페이스바를 클릭하십시오. 이름 왼쪽에 있는 확인 상자가 지워집니다.

로컬 라이브러리 삭제

공용 버전의 라이브러리를 삭제하지 않고 라이브러리를 제거할 수 있으며 그 반대도 가능합니다. 로컬 라이브러리를 삭제하면 세션 에서만 라이브러리와 모든 컨텐츠가 삭제됩니다. 로컬 버전의 라이브러리를 삭제하면 다른 세션 또는 공용 버전에서 해당 라이브러리가 제거되지 않습니다. 자세한 정보는 192 페이지의 『공용 라이브러리 관리』의 내용을 참조하십시오.

로컬 라이브러리 제거 방법

1. 트리 보기에서 삭제할 라이브러리를 선택하십시오.
2. 메뉴에서 **편집 > 삭제**를 선택하여 라이브러리를 삭제하십시오. 라이브러리가 제거됩니다.
3. 전에 이 라이브러리를 출판한 적이 없으면 이 라이브러리를 삭제할 것인지 유지할 것인지 여부를 묻는 메시지가 열립니다. **삭제**를 클릭하여 계속하거나 이 라이브러리를 유지하려면 **유지**를 클릭하십시오.

참고: 항상 한 개의 라이브러리가 남아 있어야 합니다.

공용 라이브러리 관리

로컬 라이브러리를 재사용하려면 로컬 라이브러리를 출판한 후 이에 대한 작업을 수행하고 라이브러리 관리 대화 상자(**자원 > 라이브러리 관리**)를 통해 볼 수 있습니다. 자세한 정보는 『라이브러리 공유』의 내용을 참조하십시오. 수행할 몇 가지 기본 공용 라이브러리 관리 태스크로는 공용 라이브러리 가져오기, 내보내기 또는 삭제가 있습니다. 공용 라이브러리의 이름을 변경할 수 없습니다.

공용 라이브러리 가져오기

1. 라이브러리 관리 대화 상자에서 **가져오기...**를 클릭하십시오. 라이브러리 가져오기 대화 상자가 열립니다.
2. 가져올 라이브러리 파일(*.*lib*)을 선택하고 이 라이브러리를 로컬로 추가하려면 **현재 프로젝트에 라이브러리 추가**를 선택하십시오.
3. **가져오기**를 클릭하십시오. 대화 상자가 닫힙니다. 동일한 이름을 가진 공용 라이브러리가 이미 존재하는 경우, 가져올 라이브러리의 이름을 변경하거나 현재 공용 라이브러리를 덮어쓰라는 요청을 받습니다.

공용 라이브러리 내보내기

공유할 수 있도록 공용 라이브러리를 *.lib* 형식으로 내보낼 수 있습니다.

1. 라이브러리 관리 대화 상자의 목록에서 내보낼 라이브러리를 선택하십시오.
2. **내보내기**를 클릭하십시오. 디렉토리 선택 대화 상자가 열립니다.
3. 내보낼 디렉토리를 선택하고 **내보내기**를 클릭하십시오. 대화 상자가 닫히고 라이브러리 파일(*.*lib*)을 내보냅니다.

공용 라이브러리 삭제

공용 버전의 라이브러리를 삭제하지 않고 로컬 라이브러리를 제거할 수 있으며 그 반대도 가능합니다. 그러나 라이브러리가 이 대화 상자에서 삭제되면 로컬 버전이 다시 출판될 때까지 세션 자원 에 더 이상 추가할 수 없습니다.

제품과 함께 설치된 라이브러리를 삭제하면 원래 설치된 버전이 복원됩니다.

1. 라이브러리 관리 대화 상자에서 삭제할 라이브러리를 선택하십시오. 해당 헤더를 클릭하여 목록을 정렬할 수 있습니다.
2. **삭제**를 클릭하여 라이브러리를 삭제하십시오. IBM SPSS Modeler Text Analytics 는 로컬 버전의 라이브러리가 공용 라이브러리와 동일한지 여부를 확인합니다. 동일하면 경고 없이 라이브러리가 제거됩니다. 라이브러리 버전이 다르면 공용 버전을 유지할 것인지 제거할 것인지 여부를 묻는 경고가 열려 발행됩니다.

라이브러리 공유

라이브러리를 사용하여 다중 대화형 워크벤치 세션 간에 공유하기 쉬운 방법으로 자원에 대한 작업을 할 수 있습니다. 라이브러리가 두 개의 상태 또는 버전으로 존재할 수 있습니다. 편집기에서 편집 가능하고 대화형 워크벤치 세션의 일부인 라이브러리를 **로컬 라이브러리**라고 합니다. 대화형 워크벤치 세션에서 작업하는 동안 예를 들어 *채소류* 라이브러리에서 많은 변경을 수행할 수 있습니다. 변경사항이 다른 데이터에 유용한 경우 *채*

소류 라이브러리의 **공용 라이브러리** 버전을 작성하여 이러한 자원을 사용 가능하게 할 수 있습니다. 이름이 나타내듯 공용 라이브러리는 다른 대화형 워크벤치 세션의 자원에 사용 가능합니다.

라이브러리 관리 대화 상자에서 공용 라이브러리를 볼 수 있습니다. 일단 이 공용 라이브러리 버전이 존재하면 이러한 사용자 정의 언어학적 자원을 공유할 수 있도록 다른 컨텍스트에서 자원에 추가할 수 있습니다.

제공된 라이브러리는 처음에 공용 라이브러리입니다. 이러한 라이브러리의 자원을 편집한 후 새 공용 버전을 작성할 수 있습니다. 그런 다음 다른 대화형 워크벤치 세션에서 해당 새 버전에 액세스할 수 있습니다.

라이브러리에 대한 작업을 계속하고 변경을 수행하면 라이브러리 버전이 비동기화됩니다. 로컬 버전이 공용 버전보다 최신인 경우가 있고 공용 버전이 로컬 버전보다 최신인 경우도 있습니다. 다른 대화형 워크벤치 세션 내에서 공용 버전이 업데이트된 경우 공용 및 로컬 버전 모두 다른 하나가 포함하지 않는 변경사항을 포함하는 것도 가능합니다. 라이브러리 버전이 비동기화되게 되면 다시 동기화할 수 있습니다. 라이브러리 버전 동기화는 로컬 라이브러리 재출판 및/또는 업데이트로 구성됩니다.

대화형 워크벤치 세션을 실행하거나 세션을 닫을 때마다 업데이트 또는 재출판이 필요한 라이브러리를 동기화하도록 프롬프트가 표시됩니다. 또한 트리 보기에서 라이브러리 이름 옆에 나타나는 아이콘을 통해 또는 라이브러리 특성 대화 상자를 보고 로컬 라이브러리의 동기화 상태를 쉽게 식별할 수 있습니다. 또한 메뉴 선택을 통해 언제든지 동기화를 수행하도록 선택할 수 있습니다. 다음 테이블에서는 5개의 가능한 상태 및 연관된 아이콘을 설명합니다.

표 37. 로컬 라이브러리 동기화 상태.

아이콘	로컬 라이브러리 상태 설명
	출판되지 않음 - 로컬 라이브러리가 출판되지 않았습니다.
	동기화됨 - 로컬 및 공용 라이브러리 버전이 동일합니다. 세션 특정 자원만 포함하기 때문에 출판할 수 없는 *로컬 라이브러리*에도 적용됩니다.
	오래됨 - 공용 라이브러리 버전이 로컬 버전보다 최신입니다. 변경사항으로 로컬 버전을 업데이트할 수 있습니다.
	더 최신임 - 로컬 라이브러리 버전이 공용 버전보다 최신입니다. 공용 버전에 로컬 버전을 재출판할 수 있습니다.
	동기화되지 않음 - 로컬 및 공용 라이브러리 모두 다른 하나가 포함하지 않는 변경사항을 포함합니다. 로컬 라이브러리를 업데이트할 것인지 출판할 것인지 여부를 결정해야 합니다. 업데이트하는 경우 지난 번 업데이트하거나 출판한 이후로 수행된 변경사항이 유실됩니다. 공개하기로 선택하면 공용 버전의 변경사항을 덮어씁니다.

참고: 대화형 워크벤치 세션을 실행하거나 세션을 닫을 때 출판하는 경우 라이브러리를 항상 업데이트하면 라이브러리가 동기화되지 않을 가능성이 낮습니다.

라이브러리의 변경사항이 이 라이브러리도 포함할 수 있는 다른 스트림 에 도움이 된다고 생각하면 언제든지 라이브러리를 재출판할 수 있습니다. 그런 다음 변경사항이 다른 스트림 에 도움이 되면 해당 스트림에서 로컬 버전을 업데이트할 수 있습니다. 이런 방법으로 새 라이브러리 작성 및/또는 자원에 임의 수의 공용 라이브러리 추가를 통해 데이터에 적용되는 각 컨텍스트 또는 도메인에 대해 스트림을 작성할 수 있습니다.

공용 버전의 라이브러리가 공유되면 로컬 버전과 공용 버전 간에 차이가 발생할 가능성이 커집니다. 대화형 워크벤치 세션에서 실행 또는 닫고 출판하거나 템플리트 편집기에서 템플리트를 열거나 닫을 때마다 버전이 라이브러리 관리 대화 상자의 해당 버전과 동기화되지 않은 라이브러리를 출판 및/또는 업데이트할 수 있도록 메시지가 표시됩니다. 공용 라이브러리 버전이 로컬 버전보다 최신이면 업데이트할 것인지 여부를 묻는 대화 상자가 열립니다. 공용 버전으로 업데이트하는 대신 로컬 버전을 있는 그대로 유지할 것인지 업데이트를 로컬 라이브러리에 병합할 것인지 여부를 선택할 수 있습니다.

라이브러리 출판

특정 라이브러리를 출판한 적이 없는 경우 출판하면 데이터베이스에 로컬 라이브러리 공용 사본이 작성됩니다. 라이브러리를 재출판하는 경우, 로컬 라이브러리의 컨텐츠가 기존 공용 버전의 컨텐츠를 대체합니다. 재출판한 후에는 로컬 버전이 공용 버전과 동기화되도록 다른 스트림 세션에서 이 라이브러리를 업데이트할 수 있습니다. 라이브러리를 출판할 수 있긴 하지만 로컬 버전은 항상 세션 에 저장됩니다.

중요! 로컬 라이브러리를 변경하고 도중에 공용 버전의 라이브러리도 변경된 경우 라이브러리는 동기화되지 않은 것으로 간주됩니다. 로컬 버전을 공용 변경사항으로 업데이트하는 것으로 시작하고 원하는 변경을 수행한 후 두 버전이 모두 동일하도록 로컬 버전을 다시 출판하는 것이 좋습니다. 먼저 변경을 수행하고 출판하는 경우 공용 버전의 변경사항을 덮어씁니다.

데이터베이스에 로컬 라이브러리를 출판하는 방법

1. 메뉴에서 **자원 > 라이브러리 출판**을 선택하십시오. 출판이 필요한 모든 라이브러리가 기본적으로 선택된 상태에서 라이브러리 출판 대화 상자가 열립니다.
2. 출판하거나 재출판할 각 라이브러리의 왼쪽에 있는 확인 상자를 선택하십시오.
3. **출판**을 클릭하여 라이브러리 관리 데이터베이스에 라이브러리를 출판하십시오.

라이브러리 업데이트

대화형 워크벤치 세션을 실행하거나 닫을 때마다 공용 버전과 더 이상 동기화되지 않은 라이브러리를 업데이트하거나 출판할 수 있습니다. 공용 라이브러리 버전이 로컬 버전보다 최신이면 라이브러리를 업데이트할 것인지 여부를 묻는 대화 상자가 열립니다. 공용 버전으로 업데이트하거나 로컬 버전을 공용 버전으로 대체하는 대신 로컬 버전을 유지할 것인지 여부를 선택할 수 있습니다. 공용 버전의 라이브러리가 로컬 버전보다 최신이면 로컬 버전을 업데이트하여 컨텐츠를 공용 버전의 컨텐츠와 동기화할 수 있습니다. 업데이트는 공용 버전에서 찾은 변경사항을 로컬 버전에 통합함을 의미합니다.

참고: 대화형 워크벤치 세션을 실행하거나 세션을 닫을 때 출판 하는 경우 라이브러리가 동기화되지 않을 가능성이 낮습니다. 자세한 정보는 192 페이지의 『라이브러리 공유』의 내용을 참조하십시오.

로컬 라이브러리 업데이트 방법

1. 메뉴에서 **자원 > 라이브러리 업데이트**를 선택하십시오. 업데이트가 필요한 모든 라이브러리가 기본적으로 선택된 상태에서 라이브러리 업데이트 대화 상자가 열립니다.
2. 출판하거나 재출판할 각 라이브러리의 왼쪽에 있는 확인 상자를 선택하십시오.
3. **업데이트**를 클릭하여 로컬 라이브러리를 업데이트하십시오.

충돌 해결

로컬 대 공용 라이브러리 충돌

스트림 세션을 시작할 때마다 IBM SPSS Modeler Text Analytics 는 로컬 라이브러리를 라이브러리 관리 대화 상자에 나열된 라이브러리와 비교합니다. 세션 의 로컬 라이브러리가 출판된 버전과 동기화되지 않은 경우에는 라이브러리 동기화 경고 대화 상자가 열립니다. 다음 옵션 중에서 선택하여 여기서 사용할 라이브러리 버전을 선택할 수 있습니다.

- **파일에 로컬인 모든 라이브러리.** 이 옵션은 모든 로컬 라이브러리를 있는 그대로 유지합니다. 나중에 재출판하거나 업데이트할 수 있습니다.
- **이 시스템에 출판된 모든 라이브러리.** 이 옵션은 표시된 로컬 라이브러리를 데이터베이스에서 발견된 버전으로 대체합니다.
- **모든 최신 라이브러리.** 이 옵션은 오래된 로컬 라이브러리를 데이터베이스의 최신 공용 버전으로 대체합니다.
- **기타.** 이 옵션을 사용하여 테이블에서 버전을 선택하여 원하는 버전을 수동으로 선택할 수 있습니다.

강제 실행된 용어 충돌

공용 라이브러리를 추가하거나 로컬 라이브러리를 업데이트할 때마다 이 라이브러리의 용어 및 유형과 자원에 있는 다른 라이브러리의 용어 및 유형 간에 충돌과 중복 항목이 발견될 수 있습니다. 이런 상황이 발생하면 강제 실행된 용어 편집 대화 상자에서 작업을 완료하기 전에 제안된 충돌 해결책을 확인하거나 용어를 변경하라는 요청을 받습니다. 자세한 정보는 203 페이지의 『용어 강제 실행』의 내용을 참조하십시오.

강제 실행된 용어 편집 대화 상자는 충돌하는 각 용어 또는 유형 쌍을 포함합니다. 대체 배경 색상은 각 충돌 쌍을 시각적으로 구별하는 데 사용됩니다. 이러한 색상을 옵션 대화 상자에서 변경할 수 있습니다. 자세한 정보는 86 페이지의 『옵션: 표시 탭』 주제를 참조하십시오. 강제 실행된 용어 편집 대화 상자에는 다음 두 개의 탭이 있습니다.

- **중복.** 이 탭은 라이브러리에서 발견된 중복 용어를 포함합니다. 용어 뒤에 푸시핀 아이콘이 나타나는 경우 이는 이 용어 발생이 강제 실행되었음을 의미합니다. 검은색 X 아이콘이 나타나는 경우 이는 다른 곳에 강제 실행되었기 때문에 추출 중에 이 용어 발생이 무시됨을 의미합니다.
- **사용자 정의.** 이 탭은 충돌을 통해서가 아니라 유형 사전 용어 분할창에서 수동으로 강제 실행된 용어 목록을 포함합니다.

참고: 라이브러리를 추가하거나 업데이트한 후에는 강제 실행된 용어 편집 대화 상자가 열립니다. 이 대화 상자에서 취소하는 경우 라이브러리 업데이트 또는 추가를 취소하지 않습니다.

충돌 해결 방법

1. 강제 실행된 용어 편집 대화 상자에서 강제 실행할 용어에 대한 사용 열에서 단일 선택 단추를 선택하십시오.
2. 완료했으면 **확인**을 클릭하여 강제 실행된 용어를 적용하고 대화 상자를 닫으십시오. **취소**를 클릭하면 이 대화 상자에서 수행한 변경사항을 취소합니다.

제 17 장 라이브러리 사전 정보

텍스트 데이터를 추출하는 데 사용되는 자원은 템플리트 및 라이브러리 양식으로 저장됩니다. 라이브러리는 세 개의 사전으로 구성될 수 있습니다.

- **유형 사전**은 하나의 레이블 또는 유형 이름 아래에 그룹화된 용어 콜렉션을 포함합니다. 추출 엔진은 텍스트 데이터를 읽을 때 텍스트에서 찾은 단어를 유형 사전에서 정의된 용어와 비교합니다. 추출 중에 굴절된 양식의 유형의 용어와 동의어는 개념이라는 대상 용어 아래에 그룹화됩니다. 추출된 개념은 용어로 나타나는 유형 사전에 지정됩니다. 편집기의 왼쪽 상단 및 가운데 분할창(라이브러리 트리 및 용어 분할창)에서 유형 사전을 관리할 수 있습니다. 자세한 정보는 『유형 사전』의 내용을 참조하십시오.
- **대체 사전**은 하나의 대상 용어(최종 추출 결과에서는 개념이라고 함) 아래에 유사한 용어를 그룹화하는 데 사용되는 동의어로 또는 선택적 요소로 정의된 단어 콜렉션을 포함합니다. 동의어 탭 및 선택사항 탭을 사용하여 편집기의 왼쪽 하단 분할창에서 대체 사전을 관리할 수 있습니다. 자세한 정보는 205 페이지의 『대체/동의어 사전』의 내용을 참조하십시오.
- **제외 사전**은 최종 추출 결과에서 제거될 용어 및 유형 콜렉션을 포함합니다. 편집기의 맨 오른쪽 분할창에서 제외 사전을 관리할 수 있습니다. 자세한 정보는 208 페이지의 『제외 사전』의 내용을 참조하십시오.

자세한 정보는 187 페이지의 제 16 장 『라이브러리에 대한 작업』의 내용을 참조하십시오.

유형 사전

유형 사전은 유형 이름 또는 레이블과 용어 목록으로 구성됩니다. 유형 사전은 편집기에 있는 라이브러리 자원 탭의 상단 왼쪽 및 중앙 분할창에서 관리됩니다.메뉴에서 **보기 > 자원 편집기**를 사용하여 이 보기에 액세스할 수 있습니다(대화형 워크벤치 세션에 있는 경우). 그렇지 않으면, 템플리트 편집기에서 특정 템플리트에 대한 사전을 편집할 수 있습니다.

추출 엔진이 텍스트 데이터를 읽을 때, 유형 사전에 정의된 용어와 텍스트에서 발견된 단어를 비교합니다. 용어는 언어학적 자원의 유형 사전에 있는 단어 또는 문구입니다.

단어가 용어와 매치될 때, 이 단어는 해당 용어에 대한 유형 이름에 지정됩니다. 자원이 추출 동안 읽혀지면, 텍스트에서 발견된 용어는 추출 결과 분할창에서 개념이 되기 전에 몇 개의 처리 단계를 거칩니다. 동일한 유형 사전에 속하는 여러 용어가 추출 엔진에 의해 동의어인 것으로 판별되면, 가장 자주 발생하는 용어 아래에서 그룹화되고 추출 결과 분할창에서 **개념**이라고 합니다. 예를 들어, 용어 `question`과 `query`가 끝에서 개념 이름 `question` 아래에 나타날 수 있습니다.

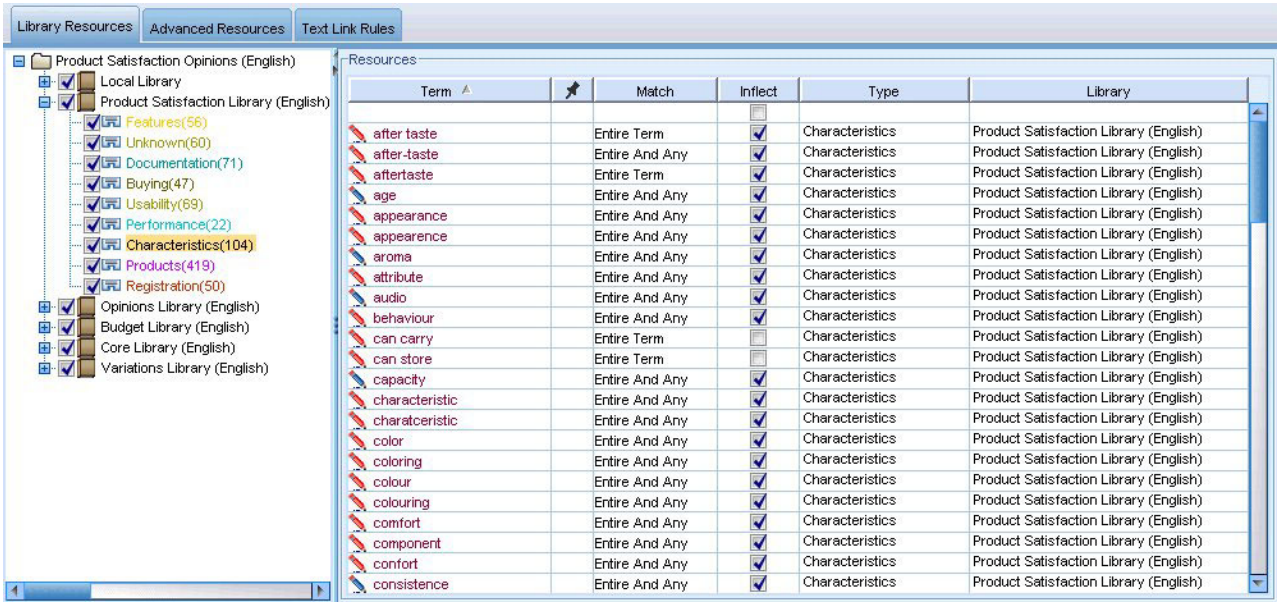


그림 40. 라이브러리 트리 및 용어 분할창

유형 사전 목록은 왼쪽에 있는 라이브러리 트리 분할창에 표시됩니다. 각 유형 사전의 내용은 중앙 분할창에 나타납니다. 유형 사전은 용어 목록보다 많은 용어로 구성됩니다. 텍스트 데이터의 단어 및 단어 문구가 유형 사전에 정의된 용어에 매치되는 방식은 정의된 매치 옵션으로 판별됩니다. **매치 옵션**은 텍스트 데이터의 후보 단어 또는 문구에 대해 용어가 고정(anchor)되는 방법을 지정합니다. 자세한 정보는 200 페이지의 『용어 추가』 주제를 참조하십시오.

참고: 모든 옵션(예: 매치 옵션 및 굴절된 양식)이 일본어 텍스트에 적용되는 것은 아닙니다.

또한, 자동으로 용어의 굴절된 양식을 생성하고 사전에 추가할 것인지 여부를 지정하여 유형 사전에서 용어를 확장할 수 있습니다. 굴절된 양식을 생성하여, 자동으로 단수 용어의 복수 양식, 복수 용어의 단수 양식, 형용사를 유형 사전에 추가합니다. 자세한 정보는 200 페이지의 『용어 추가』의 내용을 참조하십시오.

참고: 유형 사전에는 없고 텍스트에서는 추출된 대부분의 언어의 경우, 개념은 자동으로 <Unknown>으로 유형이 지정됩니다.

내장 유형

IBM SPSS Modeler Text Analytics는 제공된 라이브러리와 컴파일된 자원 양식으로 언어학적 자원 세트와 함께 제공됩니다. 제공된 라이브러리는 <Location>, <Organization>, <Person>, <Product>와 같은 내장 유형 사전 세트를 포함합니다.

참고: 일본어 텍스트의 경우 기본 내장 유형 세트가 다릅니다.

추출 엔진은 이러한 유형 사전을 사용하여 추출하는 개념에 유형을 지정합니다(예: paris 개념에 <Location> 유형 지정). 내장 유형 사전에 상당수의 용어가 정의되어 있긴 하지만 모든 가능성을 다루지는 않습니다. 따라서 직접 추가하거나 작성할 수 있습니다. 제공된 특정 유형 사전의 컨텐츠에 대한 설명은 유형 특성 대화 상자의 주석을 읽으십시오. 트리에서 유형을 선택하고 컨텍스트 메뉴에서 **편집 > 특성**을 선택하십시오.

참고: 제공된 라이브러리 이외에 컴파일된 자원(역시 추출 엔진이 사용함)은 내장 유형 사전을 보완하는 상당 수의 정의를 포함하지만 해당 컨텐츠는 제품에 표시되지 않습니다. 그러나 컴파일된 사전이 유형 지정한 용어를 다른 사전에 강제 실행할 수 있습니다. 자세한 정보는 203 페이지의 『용어 강제 실행』의 내용을 참조하십시오.

유형 작성

유사한 용어를 쉽게 그룹화할 수 있도록 유형 사전을 작성할 수 있습니다. 이 사전에 나타나는 용어가 추출 프로세스 중에 발견되면 이 유형 이름에 지정되고 개념 이름으로 추출됩니다. 라이브러리를 작성할 때마다 용어 입력을 즉시 시작할 수 있도록 빈 유형 라이브러리가 항상 포함됩니다.

중요!: 일본어 자원의 경우에는 새 유형을 작성할 수 없습니다.

식품에 대한 텍스트를 분석하고 채소류에 관한 용어를 그룹화하려면 직접 `<Vegetables>` 유형 사전을 작성할 수 있습니다. 그리고 나서 텍스트에 나타날 중요한 용어라고 생각되면 `carrot`, `broccoli`, `spinach`와 같은 용어를 추가할 수 있습니다. 그런 다음 추출 동안 이러한 용어가 발견되면 개념으로 추출되어 `<Vegetables>` 유형에 지정됩니다.

용어의 굴절된 양식을 생성하도록 선택할 수 있기 때문에 단어 또는 표현식의 모든 양식을 정의하지 않아도 됩니다. 이 옵션을 선택하면 추출 엔진이 이 유형에 속한 다른 양식 중에서 용어의 단수 또는 복수 양식을 자동으로 인식합니다. 동사나 형용사의 굴절된 양식을 원할 가능성이 없기 때문에 이 옵션은 유형에 주로 명사가 포함된 경우 특히 유용합니다.

유형 특성 대화 상자는 다음 필드를 포함합니다.

이름. 작성할 유형 사전에 제공하는 이름입니다. 유형 이름에(특히, 두 개 이상의 유형 이름이 같은 단어로 시작하는 경우) 공백을 사용하지 말 것을 권장합니다.

참고: 유형 이름과 기호 사용에 대한 몇 가지 제약조건이 있습니다. 예를 들어, "@" 또는 "!"와 같은 기호를 이름에서 사용하지 마십시오.

기본 매치. 기본 매치 속성은 이 용어를 텍스트 데이터와 매치시킬 방법을 추출 엔진에 알려 줍니다. 이 유형 사전에 용어를 추가할 때마다 이는 자동으로 지정되는 매치 속성입니다. 용어 목록에서 수동으로 항상 매치 선택을 변경할 수 있습니다. 옵션은 **전체 용어**, **시작**, **끝**, **모두**, **시작 또는 끝**, **전체 및 시작**, **전체 및 끝**, **전체 및 (시작 또는 끝)**, **전체(복합어 없음)**입니다. 자세한 정보는 200 페이지의 『용어 추가』 주제를 참조하십시오. 일본어 자원에는 이 옵션이 적용되지 않습니다.

추가 대상. 이 필드는 새 유형 사전을 작성할 라이브러리를 표시합니다.

기본적으로 굴절된 양식 생성. 이 옵션은 문법적 형태론을 사용하여 이 사전에 추가하는 용어의 유사 양식(예: 용어의 단수 또는 복수 양식)을 캡처하고 그룹화하도록 추출 엔진에 알립니다. 이 옵션은 유형에 주로 명사가 포함된 경우 특히 유용합니다. 이 옵션을 선택하면, 목록에서 수동으로 변경할 수 있긴 하지만 이 유형에 추가된 모든 새 용어가 자동으로 이 옵션을 갖게 됩니다. 일본어 자원에는 이 옵션이 적용되지 않습니다.

글꼴 색상. 이 필드를 사용하여 이 유형의 결과를 인터페이스의 다른 유형의 결과와 구별할 수 있습니다. **상위 색상 사용**을 선택하는 경우, 이 유형 사전에도 기본 유형 색상이 사용됩니다. 이 기본 색상은 옵션 대화 상자에서 설정됩니다. 자세한 정보는 86 페이지의 『옵션: 표시 탭』 주제를 참조하십시오. **사용자 정의**를 선택하는 경우, 드롭 다운 목록에서 색상을 선택하십시오.

주석. 이 필드는 선택사항이며 임의의 주석이나 설명에 사용할 수 있습니다.

유형 사전 작성 방법

1. 새 유형 사전을 작성할 라이브러리를 선택하십시오.
2. 메뉴에서 **편집 > 새 유형**을 선택하십시오. 유형 특성 대화 상자가 열립니다.
3. **이름** 텍스트 상자에 유형 사전의 이름을 입력하고 원하는 옵션을 선택하십시오.
4. **확인**을 클릭하여 유형 사전을 작성하십시오. 새 유형이 라이브러리 트리 분할창에 표시되며 가운데 분할창에 나타납니다. 용어 추가를 즉시 시작할 수 있습니다. 자세한 정보는 『용어 추가』의 내용을 참조하십시오.

참조: 이러한 지시사항은 자원 편집기 보기 또는 템플리트 편집기에서 변경하는 방법을 보여줍니다. 추출 결과 분할창 , 데이터 분할창, 범주 분할창 또는 다른 보기의 군집 정의 대화 상자에서도 직접 이런 종류의 미세 조정을 수행할 수 있음을 명심하십시오. 자세한 정보는 99 페이지의 『추출 결과 세분화』의 내용을 참조하십시오.

용어 추가

라이브러리 트리 분할창은 라이브러리를 표시하며 펼쳐서 포함된 유형 사전을 표시할 수 있습니다. 가운데 분할창에서 용어 목록은 트리에서의 선택사항에 따라 선택된 라이브러리 또는 유형 사전의 용어를 표시합니다.

중요! 일본어 자원의 경우 용어가 다르게 정의됩니다.

자원 편집기에서, 용어 분할창에서 또는 새 용어 추가 대화 상자를 통해 유형 사전에 용어를 직접 추가할 수 있습니다. 추가하는 용어는 단일 단어 또는 복합 단어입니다. 새 용어를 추가할 수 있도록 항상 목록 맨 위에서 공백 행을 찾습니다.

참조: 이러한 지시사항은 자원 편집기 보기 또는 템플리트 편집기에서 변경하는 방법을 보여줍니다. 추출 결과 분할창 , 데이터 분할창, 범주 분할창 또는 다른 보기의 군집 정의 대화 상자에서도 직접 이런 종류의 미세 조정을 수행할 수 있음을 명심하십시오. 자세한 정보는 99 페이지의 『추출 결과 세분화』의 내용을 참조하십시오.

용어 열

이 열에서 단일 또는 복합 단어를 셀에 입력하십시오. 용어가 나타나는 색상은 용어가 저장되거나 강제 실행되는 유형의 색상에 따라 다릅니다. 유형 특성 대화 상자에서 유형 색상을 변경할 수 있습니다. 자세한 정보는 199 페이지의 『유형 작성』의 내용을 참조하십시오.

강제 실행 열

이 열에서 이 셀에 푸시핀 아이콘을 두면 추출 엔진이 다른 라이브러리에서 이 동일 용어의 다른 발생을 무시함을 알게 됩니다. 자세한 정보는 203 페이지의 『용어 강제 실행』의 내용을 참조하십시오.

매치 열

이 열에서 이 용어를 텍스트 데이터와 매치시킬 방법을 추출 엔진에 알려 줄 매치 옵션을 선택하십시오. 예제는 테이블을 참조하십시오. 유형 특성을 편집하여 기본값을 변경할 수 있습니다. 자세한 정보는 199 페이지의 『유형 작성』의 내용을 참조하십시오. 메뉴에서 **편집 > 매치 변경**을 선택하십시오. 이러한 조합도 가능하기 때문에 기본 매치 옵션은 다음과 같습니다.

- **시작**. 사전의 용어가 텍스트에서 추출된 개념의 첫 번째 단어와 매치하면 이 유형이 지정됩니다. 예를 들어, apple을 입력하면 apple tart가 매치됩니다.
- **끝**. 사전의 용어가 텍스트에서 추출된 개념의 마지막 단어와 매치하면 이 유형이 지정됩니다. 예를 들어, apple을 입력하면 cider apple이 매치됩니다.
- **모두**. 사전의 용어가 텍스트에서 추출된 개념의 임의 단어와 매치하면 이 유형이 지정됩니다. 예를 들어, apple을 입력하면 **모두** 옵션은 apple tart, cider apple, cider apple tart를 동일한 방법으로 유형 지정합니다.
- **전체 용어**. 텍스트에서 추출된 전체 개념이 사전의 정확한 용어와 매치하면 이 유형이 지정됩니다. 용어를 **전체 용어**로 추가하면 **전체 및 시작**, **전체 및 끝**, **전체 및 모두** 또는 **전체(복합어 없음)**가 용어 추출을 강제 실행합니다.

  뿐만 아니라 <Person> 유형은 두 개의 파트 이름(예: *edith piaf* 또는 *mohandas gandhi*)만 추출하기 때문에 성이 언급되지 않을 때 이름을 추출하려는 경우 이 유형 사전에 명시적으로 이름을 추가할 수 있습니다. 예를 들어, *edith*의 인스턴스를 이름으로 모두 포착하려면 **전체 용어** 또는 **전체 및 시작**을 사용하여 <Person> 유형에 edith를 추가해야 합니다.
- **전체(복합어 없음)**. 텍스트에서 추출된 전체 개념이 사전의 정확한 용어와 매치하면 이 유형이 지정되며, 추출이 중지되어 추출이 용어를 더 이상 복합어와 매치하지 못합니다. 예를 들어, apple을 입력하면 **전체(복합어 없음)** 옵션이 apple을 유형 지정하고 어딘가 다른 곳에서 강제 실행되지 않는 한 복합어 apple sauce를 추출하지 않습니다.

다음 테이블에서는 apple 용어가 유형 사전에 있다고 가정하십시오. 매치 옵션에 따라 이 테이블은 텍스트에서 발견되면 추출되고 유형 지정되는 개념을 보여줍니다.

표 38. 매치 예제.

용어에 대한 매치 옵션  **apple**	**추출된 개념**			
	apple	**apple** *tart*	*ripe* **apple**	*homemade* **apple** *tart*
전체 용어	✔			
시작(**Start**)		✔		

표 38. 매치 예제 (계속).

용어에 대한 매치 옵션  **apple**	추출된 개념			
	apple	**apple** *tart*	*ripe* **apple**	*homemade* **apple** *tart*
끝			✔	
시작 또는 끝		✔	✔	
전체 및 시작	✔	✔		
전체 및 끝	✔		✔	
전체 및 (시작 또는 끝)	✔	✔	✔	
모두		✔	✔	✔
전체 및 모두	✔	✔	✔	✔
전체(복합어 없음)	✔	*추출되지 않음*	*추출되지 않음*	*추출되지 않음*

굴절 열

이 열에서 모두 함께 그룹화되도록 추출 엔진이 추출 중에 이 용어의 굴절된 양식을 생성해야 하는지 여부를 선택하십시오. 이 열의 기본값은 유형 특성에 정의되어 있지만 이 열에서 직접 케이스별로 이 옵션을 변경할 수 있습니다. 메뉴에서 **편집 > 굴절 변경**을 선택하십시오.

유형 열

이 열의 드롭 다운 목록에서 유형 사전을 선택하십시오. 유형 목록은 라이브러리 트리 분할창에서 사용자의 선택에 따라 필터링됩니다. 목록의 첫 번째 유형은 항상 라이브러리 트리 분할창에서 선택된 기본 유형입니다. 메뉴에서 **편집 > 유형 변경**을 선택하십시오.

라이브러리 열

이 열에 용어가 저장된 라이브러리가 나타납니다. 용어를 라이브러리 트리 분할창의 다른 유형으로 끌어서 놓아 라이브러리를 변경할 수 있습니다.

유형 사전에 단일 용어 추가 방법

1. 라이브러리 트리 분할창에서 용어를 추가할 유형 사전을 선택하십시오.

2. 가운데 분할창의 용어 목록에서 사용 가능한 첫 번째 빈 셀에 용어를 입력하고 이 용어에 원하는 옵션을 설정하십시오.

유형 사전에 다중 용어 추가 방법

1. 라이브러리 트리 분할창에서 용어를 추가할 유형 사전을 선택하십시오.
2. 메뉴에서 **도구 > 새 용어**를 선택하십시오. 새 용어 추가 대화 상자가 열립니다.
3. 용어를 입력하거나 용어 세트를 복사해서 붙여넣어 선택된 유형 사전에 추가할 용어를 입력하십시오. 여러 용어를 입력하는 경우, 옵션 대화 상자에 정의된 구분자를 사용하여 구분하고 새 행에서 각 용어를 추가해야 합니다. 자세한 정보는 86 페이지의『옵션 설정』 주제를 참조하십시오.
4. **확인**을 클릭하여 용어를 사전에 추가하십시오. 매치 옵션은 이 유형 라이브러리의 기본 옵션으로 자동으로 설정됩니다. 대화 상자가 닫히고 새 용어가 사전에 나타납니다.

용어 강제 실행

용어를 특정 유형에 지정하려면 해당 유형 사전에 추가할 수 있습니다. 그러나 이름이 동일한 용어가 여러 개 있으면 추출 엔진이 사용해야 하는 유형을 알아야 합니다. 따라서 사용해야 하는 유형을 선택하도록 프롬프트가 표시됩니다. 이를 유형에 용어를 **강제 실행**한다고 합니다. 이 옵션은 컴파일된 (내부, 편집 불가능) 사전의 유형 할당을 대체하는 경우 가장 유용합니다. 일반적으로 중복 용어를 전적으로 피하는 것이 좋습니다.

강제 실행은 이 용어의 다른 발생을 *제거*하지 않습니다. 오히려 추출 엔진이 이를 무시합니다. 용어를 강제 실행하거나 강제 실행을 해제하여 사용해야 하는 발생을 나중에 변경할 수 있습니다. 공용 라이브러리를 추가하거나 공용 라이브러리를 업데이트하는 경우에도 유형 사전에 용어를 강제 실행해야 합니다.

용어 분할창의 두 번째 열인 강제 실행 열에서 강제 실행되거나 무시된 용어를 볼 수 있습니다. 푸시핀 아이콘이 나타나는 경우 이는 이 용어 발생이 강제 실행되었음을 의미합니다. 검은색 X 아이콘이 나타나는 경우 이는 다른 곳에 강제 실행되었기 때문에 추출 중에 이 용어 발생이 무시됨을 의미합니다. 또한 용어를 강제 실행하면 강제 실행된 유형에 대한 색상으로 용어가 나타납니다. 이는 Type 1과 Type 2 모두에 있는 용어를 Type 1에 강제 실행한 경우 창에 이 용어가 표시될 때는 언제든지 Type 1에 대해 정의된 글꼴 색상으로 나타남을 의미합니다.

아이콘을 두 번 클릭하여 상태를 변경할 수 있습니다. 용어가 다른 곳에 나타나는 경우, 사용해야 하는 발생을 선택할 수 있도록 충돌 해결 대화 상자가 열립니다.

유형 이름 변경

유형 사전의 이름을 변경하거나 유형 특성을 편집하여 다른 사전 설정을 변경할 수 있습니다.

중요! 유형 이름에(특히, 두 개 이상의 유형 이름이 같은 단어로 시작하는 경우) 공백을 사용하지 말 것을 권장합니다. 코어 또는 Opinions 라이브러리의 유형 이름을 변경하거나 기본 매치 속성을 변경하지 말 것을 권장합니다.

유형 이름 변경 방법

1. 라이브러리 트리 분할창에서 이름을 변경할 유형 사전을 선택하십시오.

2. 마우스 오른쪽 단추를 클릭하고 컨텍스트 메뉴에서 **유형 특성**을 선택하십시오. 유형 특성 대화 상자가 열립니다.
3. 이름 텍스트 상자에 유형 사전의 새 이름을 입력하십시오.
4. **확인**을 클릭하여 새 이름을 승인하십시오. 새 유형 이름이 라이브러리 트리 분할창에 표시됩니다.

유형 이동

라이브러리 내 다른 위치 또는 트리의 다른 라이브러리로 유형 사전을 끌 수 있습니다.

라이브러리에서 유형을 다시 정렬하는 방법

1. 라이브러리 트리 분할창에서 이동할 유형 사전을 선택하십시오.
2. 메뉴에서 **편집 > 위로 이동**을 선택하여 유형 사전을 라이브러리 트리 분할창에서 한 위치 위로 이동하거나 **편집 > 아래로 이동**을 선택하여 한 위치 아래로 이동하십시오.

다른 라이브러리로 유형을 이동하는 방법

1. 라이브러리 트리 분할창에서 이동할 유형 사전을 선택하십시오.
2. 마우스 오른쪽 단추를 클릭하고 컨텍스트 메뉴에서 **유형 특성**을 선택하십시오. 유형 특성 대화 상자가 열립니다. (다른 라이브러리로 유형을 끌어서 놓을 수도 있습니다.)
3. 추가 대상 목록 상자에서 유형 사전을 이동할 라이브러리를 선택하십시오.
4. **확인**을 클릭하십시오. 대화 상자가 닫히며, 유형은 이제 사용자가 선택한 라이브러리에 있습니다.

유형 사용 안함 및 삭제

일시적으로 유형 사전을 제거하려면 라이브러리 트리 보기에서 사전 이름 왼쪽의 확인 상자를 선택 취소하여 사용 안함으로 설정할 수 있습니다. 이는 라이브러리에 사전을 유지하지만 충돌 검사 중 및 추출 프로세스 중에는 컨텐츠를 무시함을 나타냅니다.

라이브러리에서 유형 사전을 영구적으로 삭제할 수도 있습니다.

유형 사전을 사용 안함으로 설정하는 방법

1. 라이브러리 트리 분할창에서 사용 안함으로 설정할 유형 사전을 선택하십시오.
2. 스페이스바를 클릭하십시오. 유형 이름 왼쪽에 있는 확인 상자가 지워집니다.

유형 사전 삭제 방법

1. 라이브러리 트리 분할창에서 삭제할 유형 사전을 선택하십시오.
2. 메뉴에서 **편집 > 삭제**를 선택하여 유형 사전을 삭제하십시오.

대체/동의어 사전

대체 사전은 하나의 대상 용어에서 유사한 용어를 그룹화하는 데 도움이 되는 용어 콜렉션입니다. 대체 사전은 라이브러리 자원 탭의 맨 아래 분할창에서 관리됩니다. 메뉴에서 **보기 > 자원 편집기**를 사용하여 이 보기에 액세스할 수 있습니다(대화형 워크벤치 세션에 있는 경우). 그렇지 않으면, 템플리트 편집기에서 특정 템플리트에 대한 사전을 편집할 수 있습니다.

이 사전에서 두 가지 대체 양식인 **동의어** 및 **선택적 요소**를 정의합니다. 이 분할창에서 탭을 클릭하여 두 양식 사이에 전환할 수 있습니다.

텍스트 데이터에서 추출을 실행한 후, 동의어나 다른 개념의 굴절된 양식인 몇 개의 개념을 찾을 수 있습니다. 선택적 요소 및 동의어를 식별하여, 추출 엔진이 이를 단일 대상 용어에 맵핑하도록 할 수 있습니다.

동의어 및 선택적 요소를 사용하여 대체하면 빈도 문서 개수가 높은 더 유의하고 대표적인 개념으로 개념이 결합되어 추출 결과 분할창에서 개념 수가 감소됩니다.

참고: 일본어 자원의 경우, 선택적 요소는 적용되지 않으므로 사용할 수 없습니다. 또한 동의어는 일본어 텍스트의 경우 약간 다르게 처리됩니다.

동의어

동의어는 동일한 의미를 가지고 있는 두 개 이상의 단어를 연관시킵니다. 동의어를 사용하여 해당 약어가 있는 용어를 그룹화하거나 일반적으로 맞춤법이 틀린 용어를 올바른 맞춤법으로 그룹화할 수도 있습니다. 동의어 탭에서 이러한 동의어를 정의할 수 있습니다.

동의어 정의는 두 부분으로 구성됩니다. 첫 번째는 **대상** 용어로, 추출 엔진이 그 아래에서 모든 동의어 용어를 그룹화하도록 할 용어입니다. 이 대상 용어가 다른 대상 용어의 동의어로 사용되거나 제외되지 않으면, 추출 결과 분할창에 나타내는 개념이 됩니다. 두 번째는 대상 용어 아래에서 그룹화될 동의어 목록입니다.

예를 들어, `automobile`을 `vehicle`로 바꾸려는 경우, `automobile`은 동의어이고 `vehicle`은 대상 용어입니다.

동의어 열에 단어를 입력할 수 있지만, 추출 동안 단어가 발견되지 않고 용어의 매치 옵션이 `Entire`인 경우 대체는 발생할 수 없습니다. 그러나 대상 용어는 이 용어 아래에서 그룹화될 동의어에 대해 추출되지 않아도 됩니다.

선택적 요소

선택적 요소는 텍스트에서 약간 다르게 나타나더라도 유사한 용어를 함께 유지하기 위해 추출 동안 무시될 수 있는 선택적 단어를 복합 용어에서 식별합니다. 선택적 요소는 복합 용어에서 제거된 경우 다른 용어와의 매치를 작성할 수 있는 단일 단어입니다. 이 단일 단어는 복합 용어 어디에서나(시작, 중간 또는 끝에) 나타날 수 있습니다. 선택사항 탭에서 선택적 요소를 정의할 수 있습니다.

예를 들어, 용어 ibm 및 ibm corp를 함께 그룹화하려면, corp가 이 경우에 선택적 요소로 처리되도록 선언해야 합니다. 다른 예에서, 용어 access를 선택적 요소가 되도록 지정하고 추출 동안 internet access speed 및 internet speed 둘 다가 발견되는 경우, 가장 자주 발생하는 용어 아래에서 함께 그룹화됩니다.

참고: 일본어 텍스트 자원의 경우, 선택적 요소가 적용되지 않으므로 선택적 요소 탭이 없습니다.

동의어 정의

동의어 탭에서 테이블의 맨 위에 있는 빈 행에 동의어 정의를 입력할 수 있습니다. 대상 용어 및 동의어를 정의하여 시작하십시오. 또한 이 정의를 저장하려는 라이브러리를 선택할 수도 있습니다. 추출 중에 모든 동의어 발생은 최종 추출에서 대상 용어 아래에 그룹화됩니다. 자세한 정보는 200 페이지의 『용어 추가』의 내용을 참조하십시오.

예를 들어, 텍스트 데이터에 많은 원격 통신 정보가 포함된 경우에는 cellular phone, wireless phone, mobile phone 용어가 있습니다. 이 예제에서는 cellular와 mobile을 wireless의 동의어로 정의하려고 합니다. 이러한 동의어를 정의하는 경우, 추출된 모든 cellular phone 및 mobile phone 발생은 wireless phone과 동일한 용어로 간주되며 용어 목록에 함께 나타납니다.

유형 사전을 작성할 때 용어를 입력한 후 해당 용어에 대해 3개 또는 4개의 동의어를 생각할 수 있습니다. 이러한 경우, 모든 용어를 입력한 후 대상 용어를 대체 사전에 입력한 후 동의어를 끌어갈 수 있습니다.

참고: 일본어 텍스트의 경우 동의어가 약간 다르게 처리됩니다.

동의어 대체는 동의어의 굴절된 양식(예: 복수 양식)에도 적용됩니다. 컨텍스트에 따라 용어 대체 방법에 대한 제약조건을 둘 수 있습니다. 일정 문자를 사용하여 동의어 처리가 진행되어야 하는 정도에 대한 제한을 둘 수 있습니다.

- **느낌표(!).** 느낌표가 동의어 바로 앞에 오는 경우(!synonym) 이는 동의어의 굴절된 양식이 대상 용어로 대체되지 않음을 표시합니다. 그러나 대상 용어 바로 앞에 오는 느낌표(!target-term)는 복합 대상 용어 또는 변량의 일부가 추가 대체를 수신하지 않음을 의미합니다.
- **별표(*).** 동의어 바로 뒤에 위치한 별표(예: synonym*)는 이 단어가 대상 용어로 대체됨을 의미합니다. 예를 들어, manage*를 동의어로, management를 대상으로 정의한 경우 associate managers는 대상 용어 associate management로 대체됩니다. 또한 단어 뒤에 공백과 별표를 추가(synonym *)할 수 있습니다(예: internet *). 대상을 internet으로, 동의어를 internet * * 및web *로 정의한 경우 internet access card 및 web portal은 internet으로 대체됩니다. 이 사전에서는 별표 와일드카드로 단어나 문자열을 시작할 수 없습니다.
- **캐럿(^).** 동의어 앞에 오는 캐럿과 공백(예: ^ synonym)은 용어가 동의어로 시작하는 경우에만 동의어 집단이 적용됨을 의미합니다. 예를 들어, ^ wage를 동의어로, income을 대상으로 정의하고 두 용어 모두 추출된 경우에는 income 용어 아래에 함께 그룹화됩니다. 그러나 minimum wage 및 income이 추출된 경우에는 minimum wage가 wage로 시작하지 않기 때문에 함께 그룹화되지 않습니다. 공백은 이 기호와 동의어 사이에 위치해야 합니다.
- **달러 부호($).** 동의어 다음에 오는 공백과 달러 부호(예: synonym $)는 용어가 동의어로 끝나는 경우에만 동의어 집단이 적용됨을 의미합니다. 예를 들어, cash $를 동의어로, money를 대상으로 정의하고 두 용어

모두 추출된 경우에는 money 용어 아래에 함께 그룹화됩니다. 그러나 cash cow 및 money가 추출된 경우에는 cash cow가 cash로 끝나지 않기 때문에 함께 그룹화되지 않습니다. 공백은 이 기호와 동의어 사이에 위치해야 합니다.

- **캐럿(^) 및 달러 부호($).** 캐럿 및 달러 부호가 함께 사용되는 경우(예: ^ synonym $) 정확히 일치하는 경우에만 용어가 동의어와 매치됩니다. 이는 동의어 집단이 발생하려면 추출된 용어에서 동의어 앞이나 뒤에 단어가 나타날 수 없음을 의미합니다. 예를 들어, van만 truck과 함께 그룹화되는 반면 marie van guerin은 변경되지 않은 채로 남도록 ^ van $를 동의어로, truck을 대상으로 정의하려고 합니다. 또한 캐럿 및 달러 부호를 사용하여 동의어를 정의하고 이 단어가 소스 텍스트 어딘가에 나타날 때마다 동의어가 자동으로 추출됩니다.

참고: 이러한 특수 문자와 와일드카드는 일본어 텍스트에는 지원되지 않습니다.

동의어 항목 추가 방법

1. 대체 분할창이 표시된 상태에서 왼쪽 하단 모서리에 있는 **동의어** 탭을 클릭하십시오.
2. 테이블의 맨 위의 빈 줄에서 대상 열에 대상 용어를 입력하십시오. 입력한 대상 용어는 색상으로 나타납니다. 이 색상은 용어가 나타나거나 강제 실행되는(해당 경우) 유형을 나타냅니다. 용어가 검은색으로 나타나는 경우, 어떤 유형 사전에도 나타나지 않음을 의미합니다.
3. 대상의 오른쪽에 있는 두 번째 셀에서 클릭하고 동의어 세트를 입력하십시오. 옵션 대화 상자에 정의된 대로 글로벌 구분자를 사용하여 각 항목을 분리하십시오. 자세한 정보는 86 페이지의 『옵션 설정』 주제를 참조하십시오. 입력하는 용어는 색상으로 나타납니다. 이 색상은 용어가 나타나는 유형을 나타냅니다. 용어가 검은색으로 나타나는 경우, 어떤 유형 사전에도 나타나지 않음을 의미합니다.
4. 마지막 셀에서 클릭하여 이 동의어 정의를 저장하려는 라이브러리를 선택하십시오.

참조: 이러한 지시사항은 자원 편집기 보기 또는 템플리트 편집기에서 변경하는 방법을 보여줍니다. 추출 결과 분할창 , 데이터 분할창, 범주 분할창 또는 다른 보기의 군집 정의 대화 상자에서도 직접 이런 종류의 미세 조정을 수행할 수 있음을 명심하십시오. 자세한 정보는 99 페이지의 『추출 결과 세분화』의 내용을 참조하십시오.

선택적 요소 정의

선택사항 탭에서 원하는 라이브러리에 대해 선택적 요소를 정의할 수 있습니다. 이러한 항목은 각 라이브러리에 대해 함께 그룹화됩니다. 라이브러리가 라이브러리 트리 창에 추가되는 즉시 비어 있는 선택적 요소 행이 선택사항 탭에 추가됩니다.

모든 항목은 자동으로 소문자 단어로 변환됩니다. 추출 엔진은 항목을 텍스트에서 소문자와 대문자 단어 모두와 매치시킵니다.

참고: 일본어 자원의 경우, 선택적 요소는 적용되지 않으므로 사용할 수 없습니다.

참고: 용어는 옵션 대화 상자에 정의된 구분자를 사용하여 구분됩니다. 자세한 정보는 86 페이지의 『옵션 설정』 주제를 참조하십시오. 입력하는 선택적 요소에 용어의 일부와 동일한 구분자가 포함된 경우 앞에 백슬래시가 와야 합니다.

항목 추가 방법

1. 대체 분할창이 표시된 상태에서 편집기 왼쪽 하단 모서리에 있는 선택사항 탭을 클릭하십시오.
2. 이 항목을 추가할 라이브러리에 대해 선택적 요소 열에서 셀을 클릭하십시오.
3. 선택적 요소를 입력하십시오. 옵션 대화 상자에 정의된 대로 글로벌 구분자를 사용하여 각 항목을 분리하십시오. 자세한 정보는 86 페이지의 『옵션 설정』 주제를 참조하십시오.

대체 사용 안함 및 삭제

사전에서 사용 안함으로 설정하여 일시적인 방법으로 항목을 제거할 수 있습니다. 항목을 사용 안함으로 설정하면 추출 중에 항목을 무시합니다.

대체 사전에서 더 이상 사용하지 않는 항목도 삭제할 수 있습니다.

항목을 사용 안함으로 설정하는 방법

1. 사전에서 사용 안함으로 설정할 항목을 선택하십시오.
2. 스페이스바를 클릭하십시오. 항목 왼쪽에 있는 확인 상자가 지워집니다.

참고: 항목 왼쪽에 있는 확인 상자를 선택 취소하여 사용 안함으로 설정할 수도 있습니다.

동의어 항목 삭제 방법

1. 사전에서 삭제할 항목을 선택하십시오.
2. 메뉴에서 **편집 > 삭제**를 선택하거나 키보드의 **Delete** 키를 누르십시오. 항목이 더 이상 사전에 없습니다.

선택적 요소 항목 삭제 방법

1. 사전에서 삭제할 항목을 두 번 클릭하십시오.
2. 용어를 수동으로 삭제하십시오.
3. Enter를 눌러 변경사항을 적용하십시오.

제외 사전

제외 사전은 단어, 문구 또는 부분 문자열 목록입니다. 제외 사전의 항목과 매치하거나 이를 포함하는 용어는 추출에서 무시되거나 제외됩니다. 제외 사전은 편집기의 오른쪽 분할창에서 관리됩니다. 일반적으로 이 목록에 추가하는 용어는 연속성을 위해 텍스트에서 사용되지만 텍스트에 중요한 것을 실제로 추가하지 않으며 추출 결과를 혼란스럽게 할 수 있는 기입 단어 또는 문구입니다. 이러한 용어를 제외 사전에 추가하여 추출되지 않게 할 수 있습니다.

제외 사전은 편집기에서 라이브러리 자원 탭의 오른쪽 상단 분할창에서 관리됩니다. 메뉴에서 **보기 > 자원 편집기**를 사용하여 이 보기에 액세스할 수 있습니다(대화형 워크벤치 세션에 있는 경우). 그렇지 않으면, 템플리트 편집기에서 특정 템플리트에 대한 사전을 편집할 수 있습니다.

제외 사전에서 테이블 맨 위의 빈 줄에 단어, 문구 또는 부분 문자열을 입력할 수 있습니다. 하나 이상의 단어 또는 별표를 와일드카드로 사용하는 부분 단어로 제외 사전에 문자열을 추가할 수 있습니다. 제외 사전에 선언된 항목은 개념이 추출되지 않도록 하는 데 사용됩니다. 인터페이스의 어딘가 다른 곳(예: 유형 사전)에도

항목이 선언된 경우, 취소선으로 표시되어 현재 제외됨을 나타냅니다. 이 문자열은 텍스트 데이터에 나타나거나 적용될 유형 사전의 일부로 선언되지 않아도 됩니다.

참고: 동의어 항목에서 대상의 역할도 하는 개념을 제외 사전에 추가하면 대상과 모든 동의어도 제외됩니다. 자세한 정보는 206 페이지의 『동의어 정의』의 내용을 참조하십시오.

와일드카드 사용(*)

일본어를 제외한 모든 텍스트 언어의 경우, 별표 와일드카드를 사용하여 제외 항목을 부분 문자열로 간주하도록 표시할 수 있습니다. 추출 엔진이 찾은 제외 사전에 입력된 문자열로 시작하거나 끝나는 단어가 포함된 용어가 최종 추출에서 제외됩니다. 그러나 와일드카드 사용이 허용되지 않는 두 가지 경우가 있습니다.

- 별표 와일드카드가 앞에 오는 대시 문자(-)(예: *-)
- 별표 와일드카드가 앞에 오는 어포스트로피(')(예: *'*s*)

표 39. 제외 항목 예제.

항목	예제	결과
단어	*next*	*next* 단어가 포함된 경우 개념(또는 용어)이 추출되지 않습니다.
문구	*for example*	*for example* 문구가 포함된 경우 개념(또는 용어)이 추출되지 않습니다.
부분	*copyright**	*copyright* 단어의 변형과 매치하거나 이를 포함하는 개념(또는 용어)(예: `copyrighted`, `copyrighting`, `copyrights` 또는 `copyright 2010`)을 제외합니다.
부분	**ware*	*ware* 단어의 변형과 매치하거나 이를 포함하는 개념(또는 용어)(예: `freeware`, `shareware`, `software`, `hardware`, `beware` 또는 `silverware`)을 제외합니다.

항목 추가 방법

1. 테이블의 맨 위의 빈 줄에 용어를 입력하십시오. 입력하는 용어는 색상으로 나타납니다. 이 색상은 용어가 나타나는 유형을 나타냅니다. 용어가 검은색으로 나타나는 경우, 어떤 유형 사전에도 나타나지 않음을 의미합니다.

항목을 사용 안함으로 설정하는 방법

제외 사전에서 사용 안함으로 설정하여 일시적으로 항목을 제거할 수 있습니다. 항목을 사용 안함으로 설정하면 추출 중에 항목을 무시합니다.

1. 제외 사전에서 사용 안함으로 설정할 항목을 선택하십시오.
2. 스페이스바를 클릭하십시오. 항목 왼쪽에 있는 확인 상자가 지워집니다.

참고: 항목 왼쪽에 있는 확인 상자를 선택 취소하여 사용 안함으로 설정할 수도 있습니다.

항목 삭제 방법

제외 사전에서 필요하지 않은 항목을 삭제할 수 있습니다.

1. 제외 사전에서 삭제할 항목을 선택하십시오.
2. 메뉴에서 **편집 > 삭제**를 선택하십시오. 항목이 더 이상 사전에 없습니다.

제 18 장 고급 자원에 대한 정보

또한 유형, 제외 및 대체 사전 외에도, 퍼지 그룹화 설정 또는 비언어 유형 정의와 같은 다양한 고급 자원 설정에 대해 작업할 수 있습니다. 템플리트 편집기 또는 자원 편집기 보기에서 고급 자원 탭에서 이 자원에 대해 작업할 수 있습니다.

중요! 일본어 텍스트에 맞게 조정된 자원에는 이 탭을 사용할 수 없습니다.

고급 자원 탭으로 이동할 때, 다음 정보를 편집할 수 있습니다.

- **자원에 대한 대상 언어.** 자원이 작성되고 조정될 언어를 선택하기 위해 사용됩니다. 자세한 정보는 213 페이지의 『자원의 대상 언어』의 내용을 참조하십시오.
- **퍼지 그룹화(예외).** 퍼지 그룹화(맞춤법 오류 정정) 알고리즘에서 단어 쌍을 제외하기 위해 사용됩니다. 자세한 정보는 214 페이지의 『퍼지 그룹화』의 내용을 참조하십시오.
- **비언어 엔티티.** 추출될 수 있는 비언어 항목과, 추출 동안 적용되는 정규식 및 정규화 규칙을 사용하거나 사용하지 않도록 설정하기 위해 사용됩니다. 자세한 정보는 215 페이지의 『비언어 엔티티』의 내용을 참조하십시오.
- **언어 처리** 선택된 언어에 대해 문장을 구조화하고(추출 패턴 및 강제실행된 정의) 약어를 사용하는 특수 방식을 선언하기 위해 사용됩니다. 자세한 정보는 220 페이지의 『언어 처리』의 내용을 참조하십시오.
- **언어 식별자.** 언어가 **모두**로 설정될 때 자동 언어 식별자를 구성하기 위해 사용됩니다. 자세한 정보는 221 페이지의 『언어 식별자』의 내용을 참조하십시오.

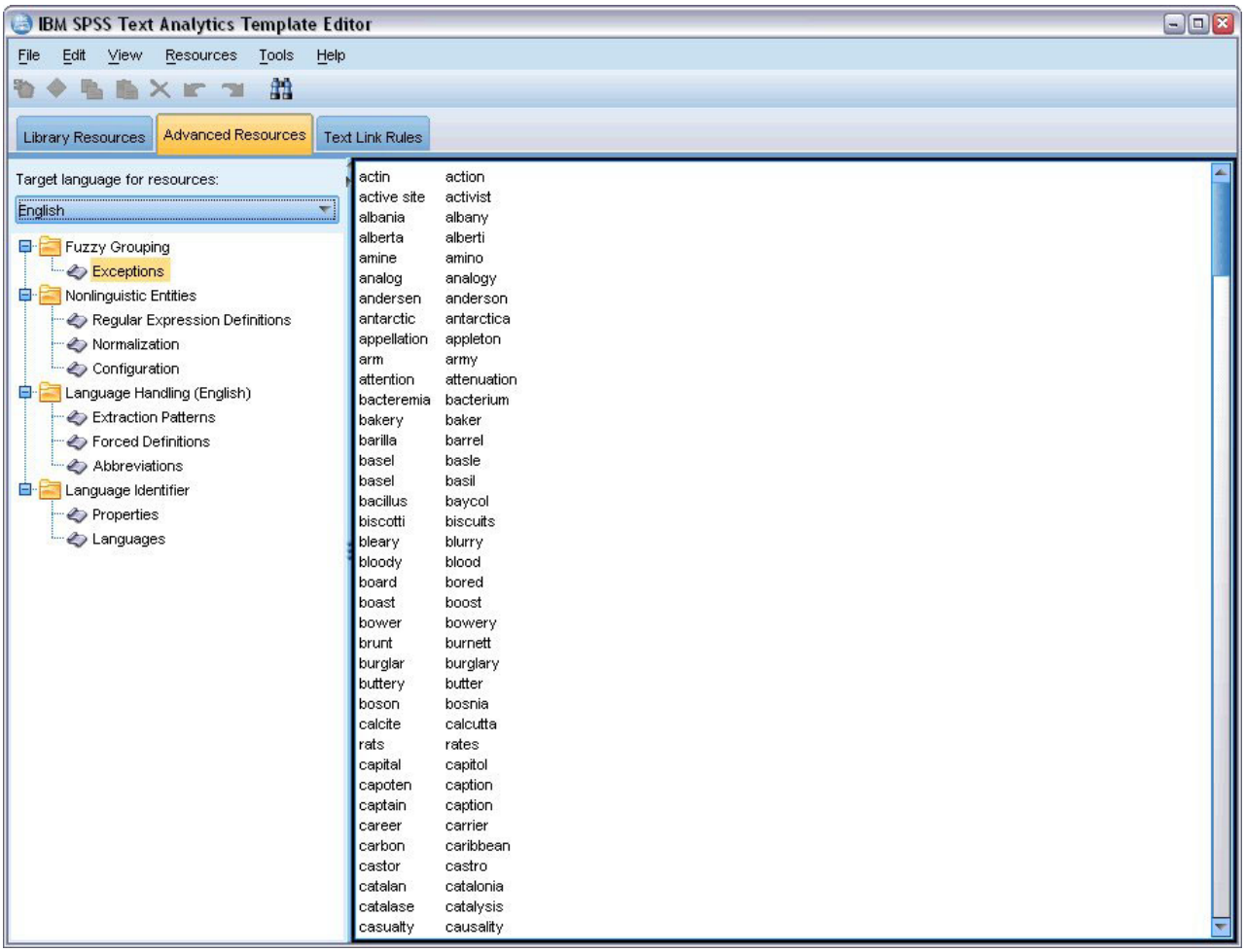


그림 41. 텍스트 마이닝 템플리트 편집기 - 고급 자원 탭

참고: 정보를 빠르게 찾거나 섹션에 대해 일정한 변경사항을 작성하기 위해 찾기/바꾸기 도구 모음을 사용할 수 있습니다. 자세한 정보는 213 페이지의 『바꾸기』의 내용을 참조하십시오.

고급 자원 편집

1. 편집할 자원 섹션을 찾고 선택하십시오. 내용은 오른쪽 분할창에 표시됩니다.
2. 필요한 경우, 내용을 자르거나, 복사하거나, 붙여넣기 위해 메뉴 또는 도구 모음을 사용하십시오.
3. 이 섹션에서 형식화 규칙을 사용하여 변경하려는 파일을 편집하십시오. 변경사항은 작성하는 즉시 저장됩니다. 이전 변경사항으로 되돌리려면 도구 모음에서 실행 취소 또는 다시 실행 화살표를 사용하십시오.

찾기

일부 경우 특정 섹션에서 빨리 정보를 찾아야 합니다. 예를 들어, 텍스트 링크 분석을 수행하는 경우 수백 개의 매크로 및 패턴 정의를 가지고 있을 수 있습니다. 찾기 기능을 사용하여, 특정 규칙을 신속하게 찾을 수 있습니다. 섹션에서 정보를 검색하기 위해 찾기 도구 모음을 사용할 수 있습니다.

찾기 기능을 사용하려면 다음을 수행하십시오.

1. 검색하려고 하는 자원 섹션을 찾아서 선택하십시오. 편집기의 오른쪽 분할창에 내용이 표시됩니다.
2. 메뉴에서 **편집 > 찾기**를 선택하십시오. 편집 고급 자원 대화 상자의 상단 오른쪽에 찾기 도구 모음이 나타납니다.
3. 텍스트 상자에 검색할 단어 문자열을 입력하십시오. 도구 모음 단추를 사용하여 대소문자 구분, 부분 매치 및 검색 방향을 제어할 수 있습니다.
4. **찾기**를 클릭하여 검색을 시작하십시오. 매치가 발견되면 창에서 텍스트가 강조표시됩니다.
5. 다음 매치를 찾으려면 다시 **찾기**를 클릭하십시오.

참고: 텍스트 링크 규칙 탭에서 작업할 때, 찾기 옵션은 소스 코드를 보고 있을 때만 사용할 수 있습니다.

바꾸기

일부 경우에, 고급 자원에 대해 광범위하게 업데이트를 작성해야 할 수 있습니다. 바꾸기 기능은 내용에 대해 일정한 업데이트를 작성하는 데 도움이 될 수 있습니다.

바꾸기 기능을 사용하려면 다음을 수행하십시오.

1. 검색하고 바꿀 자원 섹션을 찾아서 선택하십시오. 편집기의 오른쪽 분할창에 내용이 표시됩니다.
2. 메뉴에서, **편집 > 바꾸기**를 선택하십시오. 바꾸기 대화 상자가 열립니다.
3. **찾을 문자열** 텍스트 상자에서 검색할 단어 문자열을 입력하십시오.
4. **바꿀 문자열** 텍스트 상자에서 발견된 텍스트 대신에 사용하려는 문자열을 입력하십시오.
5. 완전한 단어만 찾거나 바꾸려면 **전체 단어 매치만**을 선택하십시오.
6. 대소문자가 완전하게 매치하는 단어만 찾거나 바꾸려면 **대소문자 구분**을 선택하십시오.
7. 매치를 찾으려면 **다음 찾기**를 클릭하십시오. 매치가 발견되면 창에서 텍스트가 강조표시됩니다. 이 매치를 바꾸지 않으려면, 바꿀 매치를 찾을 때까지 다시 **다음 찾기**를 클릭하십시오.
8. 선택된 매치를 바꾸려면 **바꾸기**를 클릭하십시오.
9. 섹션에서 모든 매치를 바꾸려면 **바꾸기**를 클릭하십시오. 작성된 바꾸기 수와 함께 메시지가 열립니다.
10. 바꾸기 작성이 완료되면 **닫기**를 클릭하십시오. 대화 상자가 닫힙니다.

참고: 바꾸기 오류를 작성한 경우, 대화 상자를 닫고 메뉴에서 **편집 > 실행 취소**를 선택하여 바꾸기를 실행 취소할 수 있습니다. 실행 취소할 변경사항마다 한 번씩 이를 수행해야 합니다.

자원의 대상 언어

자원은 특정 텍스트 언어에 대해 작성됩니다. 이 자원이 조정되는 언어는 고급 자원 탭에서 정의됩니다. 필요한 경우 **자원에 대한 대상 언어** 콤보 상자에서 해당 언어를 선택하여 다른 언어로 전환할 수 있습니다. 또한 여기에 나열되는 언어는 자원으로 작성하는 텍스트 분석 패키지의 언어로 표시됩니다.

중요! 드물게는 자원에서 언어를 변경해야 합니다. 그렇게 하면 자원이 더 이상 추출 언어와 매치하지 않은 때 문제가 발생할 수 있습니다. 드물게 사용되지만, 두 개 이상의 언어로 된 텍스트가 있을 것으로 예상하여 추출

동안 ALL 언어 옵션을 사용하려고 한 경우, 언어를 변경할 수 있습니다. 예를 들어 언어를 변경하여, 관심이 있는 2차 언어에 대한 추출 패턴, 약어 및 강제 실행 정의에 대한 자원을 처리하는 언어에 액세스할 수 있습니다. 그러나 작성한 자원 변경사항을 저장하거나 출판하기 전에 추출 시 관심이 있는 1차 언어로 다시 언어를 설정해야 합니다.

퍼지 그룹화

텍스트 마이닝 노드 및 추출 설정에서 **최소 루트 문자 한계에 대한 맞춤법 수용**을 선택하면, 퍼지 그룹화 알고리즘을 사용할 수 있습니다.

퍼지 그룹화는 추출된 단어에서 모든 모음(첫 번째 모음 제외)과 이중 또는 삼중 자음을 임시로 스트리핑한 후 동일한지 보기 위해 비교하여 일반적으로 맞춤법이 틀린 단어나 거의 형성된 단어를 그룹화하는 데 도움이 됩니다. 추출 프로세스 동안, 퍼지 그룹화 기능은 추출된 용어에 적용되고 결과는 매치 발견 여부를 판별하기 위해 비교됩니다. 그러한 경우, 원래 용어는 최종 추출 목록에서 함께 그룹화됩니다. 데이터에서 가장 자주 발생하는 용어 아래에서 그룹화됩니다.

참고: 비교되는 두 개의 용어가 여러 유형에 지정되면(<Unknown> 유형 제외), 퍼지 그룹화 기술이 이 쌍에 적용되지 않습니다. 다시 말하면, 기술을 적용하기 위해 용어가 동일한 유형이나 <Unknown> 유형에 속해야 합니다.

이 기능을 사용 가능하게 하고, 유사한 맞춤법의 두 단어가 올바르지 않게 함께 그룹화된 경우 퍼지 그룹화에서 제외할 수 있습니다. 고급 자원 탭의 예외 섹션에 매치된 쌍을 올바르지 않게 입력하여 이를 수행할 수 있습니다. 자세한 정보는 211 페이지의 제 18 장 『고급 자원에 대한 정보』의 내용을 참조하십시오.

다음 예는 퍼지 그룹화 수행 방법을 보여줍니다. 퍼지 그룹화가 사용되면, 다음 단어는 동일하게 나타나고 다음 방식으로 매치됩니다.

```
color  -> colr                mountain -> montn
colour -> colr                montana  -> montn

modeling  -> modlng           furniture -> furntr
modelling -> modlng           furnature -> furntr
```

이전 예에서, mountain과 montana를 함께 그룹화하는 작업에서 가장 제외하려고 합니다. 따라서, 다음 방식으로 예외 섹션에서 이 단어를 입력할 수 있습니다.

```
mountain      montana
```

중요! 일부 경우에, 퍼지 그룹화 예외는 특정 동의어 규칙이 적용되기 때문에 2 단어가 쌍을 이루는 것을 중지하지 않습니다. 그러한 경우, 느낌표 와일드카드(!)와 함께 동의어를 입력하여 단어가 출력에서 동의어가 되는 것을 금지할 수 있습니다. 자세한 정보는 206 페이지의 『동의어 정의』의 내용을 참조하십시오.

퍼지 그룹화 예외에 대한 형식화 규칙

- 행당 단 하나의 예외 쌍만 정의합니다.
- 단순어 또는 복합어를 사용합니다.

- 단어의 소문자만 사용합니다. 대문자 단어는 무시됩니다.
- 쌍에서 각 단어를 구분하려면 TAB 문자를 사용하십시오.

비언어 엔티티

특정 종류의 데이터에 대해 작업할 때, 날짜, 주민등록번호, 퍼센트 또는 다른 비언어 엔티티 추출에 많은 흥미가 있을 수 있습니다. 이 엔티티는 엔티티를 사용하거나 사용하지 않도록 설정할 수 있는 구성 파일에서 명시적으로 선언됩니다. 자세한 정보는 219 페이지의 『구성』의 내용을 참조하십시오. 추출 엔진에서 출력을 최적화하려면, 비언어 처리의 입력이 사전정의된 형식에 따라 유사한 엔티티를 그룹화하도록 정규화됩니다. 자세한 정보는 218 페이지의 『정규화』의 내용을 참조하십시오.

참고: 추출 설정에서 비언어 엔티티 추출을 켜고 끌 수 있습니다.

사용 가능한 비언어 엔티티

다음 테이블의 비언어 엔티티를 추출할 수 있습니다. 유형 이름은 소괄호로 묶습니다.

표 40. 추출될 수 있는 비언어 엔티티

주소	(<Address>)
아미노산	(<Aminoacid>)
통화	(<Currency>)
날짜	(<Date>)
지연	(<Delay>)
숫자	(<Digit>)
이메일 주소	(<email>)
HTTP/URL 주소	(<url>)
IP 주소	(<IP>
조직	(<Organization>)
퍼센트	(<Percent>)
제품	(<Product>)
단백질	(<Gene>)
전화번호	(<PhoneNumber>)
시간	(<Time>)
주민등록번호	(<SocialSecurityNumber>)
가중값 및 측도	(<Weights-Measures>)

처리를 위해 텍스트 정리

비언어 엔티티 추출이 발생하기 전에, 입력 텍스트가 정리됩니다. 이 단계 동안, 다음 용어 변경사항이 작성되어, 비언어 엔티티가 식별되고 추출될 수 있습니다.

- 두 개 이상의 공백 시퀀스는 단일 공백으로 바뀝니다.
- 도표 작성은 공백으로 바뀝니다.

- 하나의 행 끝 문자 또는 시퀀스 문자는 공백으로 바뀌며, 여러 행 끝 시퀀스는 단락 끝으로 표시됩니다. 행 끝은 캐리지 리턴(CR) 및 줄 바꾸기(LF) 또는 둘 다로 표시될 수 있습니다.
- HTML 및 XML 태그는 임시로 스트립되고 무시됩니다.

정규식 정의

비언어 엔티티를 추출할 때, 정규식을 식별하는 데 사용되는 정규식 정의에 추가하거나 편집할 수 있습니다. 이는 고급 자원 탭의 **정규식 정의** 섹션에서 수행됩니다. 자세한 정보는 211 페이지의 제 18 장 『고급 자원에 대한 정보』의 내용을 참조하십시오.

파일은 고유 섹션으로 분리됩니다. 첫 번째 섹션을 [macros]라고 합니다. 해당 섹션 외에, 추가 섹션이 각 비언어 엔티티에 존재할 수 있습니다. 이 파일에 섹션을 추가할 수 있습니다. 각 섹션 내에서, 규칙에 번호가 매겨집니다(*regexp1*, *regexp2* 등). 이 규칙에는 1 – *n*으로 연속으로 번호를 매겨야 합니다. 번호 매김이 중단되면 이 파일의 처리도 함께 일시중단됩니다.

특정 경우에서 엔티티는 언어의 영향을 받습니다. 엔티티는 구성 파일에서 언어 매개변수에 대해 0 이외의 값을 사용하면 언어의 영향을 받는다고 간주됩니다. 자세한 정보는 219 페이지의 『구성』의 내용을 참조하십시오. 엔티티가 언어 종속 상태이면, 언어를 사용하여 섹션 이름에 접두문자를 붙여야 합니다(예: [english/PhoneNumber]). 해당 섹션에는 PhoneNumber 엔티티에 언어에 대해 2 값이 제공되는 경우 영어 전화 번호에만 적용되는 규칙이 포함됩니다.

중요! 편집기에서 이 파일 또는 다른 파일을 변경하고 추출 엔진이 더 이상 원하는 대로 작동하지 않는 경우, 파일을 원래 제공된 내용으로 재설정하기 위해 도구 모음에서 **원래값으로 재설정** 옵션을 사용하십시오. 이 파일에는 정규식과의 특정 수준의 친숙도가 필요합니다. 이 영역에서 추가 지원이 필요한 경우 IBM Corp.에 도움을 요청하십시오.

특수 문자. [] {} () \ * + ? | ^ $

모든 문자는 다음 특수 문자를 제외하고 자체와 매치됩니다. 특수 문자는 표현식에서 특정 목적으로 사용됩니다. .[{()*+?|^$ 특수 문자를 이와 같이 사용하려면, 정의에서 앞에 백슬래시(\)를 붙여야 합니다.

예를 들어, 웹 주소를 추출하기 위해 시도한 경우, 전체 중지 문자는 엔티티에 매우 중요하므로, 다음과 같이 백슬래시를 사용해야 합니다.

```
www\.[a-z]+\.[a-z]+
```

반복 연산자 및 수량사 ? + * {}

정의를 한층 융통성 있게 하려면, 정규식에 표준인 몇 개의 와일드카드를 사용할 수 있습니다. 와일드카드는 * ? +입니다.

- *별표 *는 0개 이상의* 이전 문자열이 있음을 표시합니다. 예: ab*c는 "*ac*", "*abc*", "*abbbc*" 등과 매치됩니다.
- *더하기 부호 +는 하나 이상의* 이전 문자열이 있음을 표시합니다. 예: ab+c는 "*abc*", "*abbc*", "*abbbc*"와 매치되지만 "*ac*"에는 매치되지 않습니다.

- *물음표 ?는 0개 또는 하나의* 이전 문자열이 있음을 표시합니다. 예: `modell?ing`은 "*modeling*" 및 "*modeling*" 둘 다와 매치됩니다.
- *대괄호 {}로 반복 제한*은 반복의 경계를 표시합니다. 예를 들어,

  `[0-9]{n}`은 정확히 *n*번 반복되는 숫자를 매치합니다. 예를 들어, `[0-9]{4}`는 "*1998*"을 매치하지만 "*33*" 또는 "*19983*"은 매치하지 않습니다.

  `[0-9]{n,}`은 *n번 이상* 반복되는 숫자를 매치합니다. 예를 들어, `[0-9]{3,}`은 "*199*" 또는 "*1998*"은 매치하지만, "*19*"는 매치하지 않습니다.

  `[0-9]{n,m}`은 *n 및 m번(n 및 m 포함)* 사이에 반복되는 숫자를 매치합니다. 예를 들어, `[0-9]{3,5}`는 "*199*", "*1998*" 또는 "*19983*"은 매치하지만 "*19*"또는 "*199835*"는 매치하지 않습니다.

선택적 공백 및 하이픈

어떤 경우에는 정의에 선택적 공백을 포함해야 합니다. 예를 들어, "*uruguayan pesos*", "*uruguayan peso*", "*uruguay pesos*", "*uruguay peso*", "*pesos*" 또는 "*peso*"와 같은 통화를 추출하려는 경우, 공백으로 구분되는 두 단어가 있다는 사실을 처리해야 합니다. 이러한 경우, 이 정의는 `(uruguayan |uruguay )?pesos?`와 같이 작성해야 합니다. *uruguayan* 또는 *uruguay*는 *pesos*/*peso*와 함께 사용될 때 공백이 뒤에 오므로, 선택적 공백을 선택적 시퀀스 `(uruguayan |uruguay )` 내에서 정의해야 합니다. 공백이 선택적 시퀀스에 없는 경우 (예: `(uruguayan|uruguay)? pesos?`), 공백이 필요하므로 "*pesos*" 또는 "*peso*"에 대해 매치되지 않습니다.

목록에서 하이픈 문자(-)를 포함하여 어떤 것의 시리즈를 찾고 있는 경우, 하이픈을 마지막으로 정의해야 합니다. 예를 들어, 콤마(,) 또는 하이픈(-)을 찾는 경우, `[,-]`를 사용하고 `[-,]`는 사용하지 마십시오.

목록 및 매크로에서 문자열 순서

짧은 시퀀스 이전에 가장 긴 시퀀스를 정의해야 합니다. 그렇지 않으면 매치가 짧은 시퀀스에 대해 발생하므로 가장 긴 시퀀스는 읽혀지지 않습니다. 예를 들어, 문자열 "*billion*" 또는 "*bill*"을 찾는 경우, "*billion*"이 "*bill*" 전에 정의되어야 합니다. 따라서, 예를 들어 `(billion|bill)`은 가능하지만 `(bill|billion)`은 안 됩니다. 이는 매크로에도 적용됩니다. 매크로는 문자열 목록이기 때문입니다.

정의 섹션에서 규칙의 순서

행마다 하나의 규칙을 정의합니다. 각 섹션 내에서, 규칙에 번호가 매겨집니다(*regexp1*, *regexp2* 등). 이 규칙에는 1 – *n*으로 연속으로 번호를 매겨야 합니다. 번호 매김이 중단되면 이 파일의 처리도 함께 일시중단됩니다. 항목을 사용하지 않으려면 정규식을 정의하기 위해 사용되는 각 행의 맨 앞에 # 기호를 놓으십시오. 항목을 사용하려면 해당 행 앞의 # 문자를 제거하십시오.

각 섹션에서, 가장 특정적인 규칙은 적절한 처리를 위해 가장 일반적인 규칙 이전에 정의해야 합니다. 예를 들어, "*month year*" 및 "*month*" 양식의 날짜를 찾는 경우 "*month year*" 규칙이 "*month*" 규칙 이전에 정의되어야 합니다. 다음은 정의하는 방법입니다.

```
#@# January 1932
regexp1=$(MONTH),? [0-9]{4}

#@# January
regexp2=$(MONTH)
```

다음은 아닙니다.

```
#@# January
regexp1=$(MONTH)

#@# January 1932
regexp2=$(MONTH),? [0-9]{4}
```

규칙에서 매크로 사용

특정 시퀀스가 여러 규칙에서 사용될 때마다, 매크로를 사용할 수 있습니다. 그러면, 이 시퀀스의 정의를 변경해야 하는 경우에 한 번만 변경해야 하고, 이를 참조하는 모든 규칙에서 변경하지는 않습니다. 예를 들어 다음 스크립트가 있다고 가정합니다.

```
MONTH=((january|february|march|april|june|july|august|september|october|
november|december)|(jan|feb|mar|apr|may|jun|jul|aug|sep|oct|nov|dec)(\.)?)
```

매크로의 이름을 참조할 때마다 `$()`로 묶어야 합니다(예: `regexp1=$(MONTH)`).

모든 매크로는 `[macros]` 섹션에서 정의해야 합니다.

정규화

비언어 엔티티를 추출할 때, 발견되는 엔티티는 사전정의된 형식에 따라 유사한 엔티티를 그룹화하도록 정규화됩니다. 예를 들어, 통화 기호와 해당되는 단어는 동일하게 처리됩니다. 정규화 항목은 고급 자원 탭의 **정규화** 섹션에서 저장됩니다. 자세한 정보는 211 페이지의 제 18 장 『고급 자원에 대한 정보』의 내용을 참조하십시오. 파일은 고유 섹션으로 분리됩니다.

중요! 이 파일은 고급 사용자를 위한 파일입니다. 이 파일을 변경해야 할 경우는 거의 없습니다. 이 영역에서 추가 지원이 필요한 경우 IBM Corp.에 도움을 요청하십시오.

정규화에 대한 형식화 규칙

- 행마다 하나의 정규화 항목만 추가하십시오.
- 반드시 이 파일에 있는 섹션에 따르십시오. 새 섹션을 추가할 수 없습니다.
- 항목을 사용하지 않으려면 해당 행의 맨 앞에 # 기호를 놓으십시오. 항목을 사용하려면 해당 행 앞에 # 문자를 제거하십시오.

정규화에서 영억 날짜

기본적으로 영어 템플리트의 날짜는 미국 스타일 날짜 형식(즉, 월, 일, 년)으로 인식됩니다. 이를 일, 월, 년 형식으로 변경해야 하는 경우, "format:US" 행을 사용하지 않도록 설정하고(행 앞에 # 추가) "format:UK"를 사용하도록 설정하십시오(해당 행에서 # 제거).

구성

비언어 엔티티 구성 파일에서 추출하려는 비언어 엔티티 유형을 사용하거나 사용하지 않도록 설정할 수 있습니다. 필요하지 않은 엔티티를 사용하지 않도록 설정하여, 필요한 처리 시간을 줄일 수 있습니다. 이는 고급 자원 탭의 **구성** 섹션에서 수행됩니다. 자세한 정보는 211 페이지의 제 18 장 『고급 자원에 대한 정보』의 내용을 참조하십시오. 비언어학적 추출이 사용되는 경우, 추출 엔진은 추출해야 하는 비언어 엔티티 유형을 판별하기 위해 추출 프로세스 동안 구성 파일을 읽습니다.

이 파일의 명령문은 다음과 같습니다.

#name<TAB>*Language*<TAB>*Code*

표 41. 구성 파일의 명령문.

열 레이블	설명
#name	비언어 엔티티가 비언어 엔티티 추출을 위한 다른 두 개의 필수 파일에서 참조될 어귀. 여기에서 사용되는 이름에서는 대소문자가 구분됩니다.
Language	문서 의 언어. 특정 언어를 선택하는 것이 최상이지만 **모두** 옵션이 있습니다. 가능한 옵션은 0 = 모두 (regexp가 언어에 특정하지 않고 IP/URL/이메일 주소와 같이 언어가 다른 여러 템플리트에서 사용될 수 있을 때마다 사용됨), 1 = 프랑스어, 2 = 영어, 4 = 독일어, 5 = 스페인어, 6 = 네덜란드어, 8 = 포르투갈어, 10 = 이탈리아어입니다.
Code	품사 코드. 대부분의 엔티티는 약간의 경우를 제외하고 “s” 값을 사용합니다. 가능한 값은 s(검색 엔진에서 제외되는 단어), a(형용사), n(명사)입니다. 사용되는 경우, 비언어 엔티티가 첫 번째로 추출되며 추출 패턴은 대형 컨텍스트에서 해당 역할을 식별하기 위해 적용됩니다. 예를 들어, 백분율에는 “a” 값이 제공됩니다. 30%가 비언어 엔티티로 추출되었다고 가정해 보십시오. 형용사로 식별됩니다. 그리고 나서 텍스트에 "30% salary increase"가 포함된 경우 “30%” 비언어 엔티티는 품사 패턴 “ann”(형용사 명사 명사)에 맞춰집니다.

정의 엔티티에서 순서

파일에서 엔티티가 선언되는 순서는 중요하며 추출 방식에 영향을 줍니다. 나열되는 순서에 적용됩니다. 순서를 변경하면 결과가 변경됩니다. 가장 특정한 비언어 엔티티는 더 일반적인 엔티티 이전에 정의해야 합니다.

예를 들어, 비언어 엔티티 “Aminoacid”는 다음과 같이 정의됩니다.

```
regexp1=($(AA)-?$(NUM))
```

여기서 $(AA)는 특정 아미노산에 해당되는 특정의 3자 시퀀스인
“(ala|arg|asn|asp|cys|gln|glu|gly|his|ile|leu|lys|met|phe|pro|ser)”에 해당됩니다.

다른 한편으로, 비언어 엔티티 "Gene"는 한층 일반적이며 다음과 같이 정의됩니다.

```
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?
```

"Gene"가 구성 섹션에서 "Aminoacid" 이전에 정의되는 경우, "Aminoacid"는 결코 매치되지 않습니다. "Gene"의 regexp3가 항상 먼저 매치됩니다.

구성에 대한 형식화 규칙

- 열의 각 항목을 구분하기 위해 TAB 문자를 사용하십시오.
- 행을 삭제하지 마십시오.
- 이전 테이블에 표시된 구문을 준수하십시오.
- 항목을 사용하지 않으려면 해당 행의 맨 앞에 # 기호를 놓으십시오. 엔티티를 사용하려면, 해당 행 앞에서 # 문자를 제거하십시오.

언어 처리

오늘날 사용되는 모든 언어에는 아이디어 표현, 문장 구조화 및 약어 사용에 대한 특수한 방식이 있습니다. 언어 처리 섹션에서, 추출 패턴을 편집하고, 해당 패턴에 대한 정의를 강제 실행하며, 언어 드롭 다운 목록에서 선택한 언어에 대한 약어를 선언할 수 있습니다.

- 추출 패턴
- 강제 실행된 정의
- 약어

추출 패턴

문서 에서 정보를 추출할 때 추출 엔진은 추출을 위한 후보 용어(단어 및 문구)를 식별하기 위해 텍스트의 단어 "스택"에 품사 추출 패턴 세트를 적용합니다. 추출 패턴을 추가하거나 수정할 수 있습니다.

품사에는 명사, 형용사, 과거 분사, 한정사, 전치사, 등위 접속사, 이름, 이니셜, 불변화사와 같은 문법적 요소가 포함됩니다. 이러한 일련의 요소가 품사 추출 패턴을 구성합니다. IBM Corp. 텍스트 마이닝 제품에서 각 품사는 패턴을 쉽게 정의할 수 있도록 1자로 표시됩니다. 예를 들어, 형용사는 소문자 *a*로 표시됩니다. 지원 코드 세트는 사용되는 각 코드를 쉽게 이해할 수 있도록 패턴 세트 및 각 패턴 예제와 함께 각각의 기본 추출 패턴 섹션 맨 위에 기본적으로 나타납니다.

추출 패턴 형식화 규칙

- 행당 하나의 패턴.
- 패턴을 사용하지 않으려면 행 처음에 #을 사용하십시오.

주어진 단어 시퀀스는 추출 엔진이 한 번만 읽고 엔진이 매치를 찾는 첫 번째 추출 패턴에 지정되기 때문에 추출 패턴을 나열하는 순서는 매우 중요합니다.

강제 실행된 정의

문서 에서 정보를 추출할 때 추출 엔진은 텍스트를 스캔하고 발생하는 모든 단어의 품사를 식별합니다. 일부 경우에는 컨텍스트에 따라 한 단어가 여러 개의 역할에 맞습니다. 강제로 단어가 특정 품사 역할을 갖게 하거나 단어를 처리에서 완전히 제외하려면 고급 자원 탭의 **강제 실행된 정의** 섹션에서 이렇게 할 수 있습니다. 자세한 정보는 211 페이지의 제 18 장 『고급 자원에 대한 정보』의 내용을 참조하십시오.

주어진 단어에 품사 역할을 강제 실행하려면 다음 구문을 사용하여 이 섹션에 행을 추가해야 합니다.

```
term:code
```

표 42. 구문 설명.

항목	설명
term	용어 이름입니다.
code	품사 역할을 나타내는 1자 코드입니다. 단일어당 최대 6개의 다른 품사 코드를 나열할 수 있습니다. 또한 소문자 코드 s를 사용하여(예: additional:s) 단어가 복합 단어/문구로 추출되지 않게 할 수 있습니다.

강제 실행된 정의 형식화 규칙

- 단어당 한 행.
- 용어는 콜론을 포함할 수 없습니다.
- 단어가 함께 추출되지 않게 하려면 s를 품사 코드로 사용하십시오.
- 행당 최대 6개의 품사 코드를 사용하십시오. 지원되는 품사 코드는 추출 패턴 섹션에 표시됩니다. 자세한 정보는 220 페이지의 『추출 패턴』의 내용을 참조하십시오.
- 부분 매치의 경우 문자열 끝에 별표 문자(*)를 와일드카드로 사용하십시오. 예를 들어, add*:s를 입력하면 add, additional, additionally, addendum, additive와 같은 단어가 용어 또는 복합 단어 용어의 일부로 추출되지 않습니다. 그러나 컴파일된 사전 또는 강제 실행된 정의에 단어 매치가 용어로 명시적으로 선언된 경우에는 여전히 추출됩니다. 예를 들어, add*:s와 addendum:n을 모두 입력하는 경우 텍스트에서 찾으면 addendum이 여전히 추출됩니다.

약어

추출 엔진은 텍스트를 처리할 때 일반적으로 찾은 마침표를 문장이 끝났다는 표시로 간주합니다. 이는 일반적으로 맞습니다. 그러나 텍스트에 약어가 포함된 경우에는 이 마침표 문자 처리가 적용되지 않습니다.

텍스트에서 용어를 추출하고 일정 약어가 잘못 처리되었음을 발견하면 이 섹션에서 해당 약어를 명시적으로 선언해야 합니다.

참고: 약어가 동의어 정의에 이미 나타나거나 유형 사전에 용어로 정의된 경우에는 여기서 약어 항목을 추가하지 않아도 됩니다.

약어 형식화 규칙

- 행당 하나의 약어를 정의하십시오.

언어 식별자

분석하고 있는 텍스트 데이터에 대해 맞는 특정 언어를 선택하는 것이 항상 최상이지만, 텍스트가 몇 가지의 다른 또는 알 수 없는 언어로 되어 있을 경우 **모두** 옵션을 지정할 수 있습니다. **모두** 언어 옵션은 언어 식별자라고 하는 언어 자동 인식 엔진을 사용합니다. 언어 식별자는 지원되는 언어로 되어 있는 문서를 식별하기 위해 문서 를 스캔하고 추출 동안 각 파일에 대한 최상의 내부 사전을 자동으로 적용합니다. **모두** 옵션은 특성 섹션에서 매개변수에 의해 다뤄집니다.

특성

언어 식별자는 이 섹션에서 매개변수를 사용하여 구성됩니다. 다음 테이블은 고급 자원 탭의 **언어 ID - 특성** 섹션에서 설정할 수 있는 매개변수를 설명합니다. 자세한 정보는 211 페이지의 제 18 장 『고급 자원에 대한 정보』의 내용을 참조하십시오.

표 43. 매개변수 설명

매개변수	설명
NUM_CHARS	텍스트의 언어를 판별하기 위해 추출 엔진이 읽어야 하는 문자 수를 지정합니다. 숫자가 낮을 수록 언어는 더 빠르게 식별됩니다. 숫자가 높을수록 더 정확하게 언어가 식별됩니다. 값을 0으로 설정하는 경우, 문서 의 전체 텍스트가 읽혀집니다.
USE_FIRST_SUPPORTED _LANGUAGE	추출 엔진이 언어 식별자에 의해 발견된 첫 번째 지원 언어를 사용해야 하는지 여부를 지정합니다. 값을 1로 설정하는 경우, 첫 번째 지원 언어가 사용됩니다. 값을 0으로 설정하면, 폴백 언어 값이 사용됩니다.
FALLBACK_LANGUAGE	식별자에 의해 리턴된 언어가 지원되지 않는 경우 사용할 언어를 지정합니다. 가능한 값은 english, french, german, spanish, dutch, italian 및 ignore입니다. 값을 ignore로 설정하는 경우, 지원되는 언어가 없는 문서가 무시됩니다.

언어

언어 식별자는 다양한 많은 언어를 지원합니다. 고급 자원 탭의 **언어 식별자- 언어** 섹션에서 언어 목록을 편집할 수 있습니다.

이 목록에서 사용되지 않을 것 같은 언어는 제거하는 것이 좋습니다. 더 많은 언어가 존재하면 긍정 오류 가능성이 높아지고 성능이 느려질 수 있습니다. 그러나 이 파일에 새 언어를 추가할 수는 없습니다. 언어 식별자가 문서에 매치되는 언어를 더 빨리 찾을 수 있도록 가장 가능한 언어를 목록의 맨 위에 놓도록 하십시오.

제 19 장 텍스트 링크 규칙에 대한 정보

텍스트 링크 분석(TLA)은 규칙 세트를 사용하여 텍스트에서 발견된 관계를 추출하기 위해 사용되는 패턴 매치 기법입니다. 추출에 대해 텍스트 링크 분석이 사용되는 경우, 텍스트 데이터는 이 규칙에 대해 비교됩니다. 매치가 발견되면, 텍스트 링크 분석 패턴이 추출되어 제시됩니다. 이 규칙은 텍스트 링크 규칙 탭에서 정의됩니다.

예를 들어, 조직에 대한 단순한 아이디어를 나타내는 개념을 추출하는 것은 사용자에게 충분히 흥미롭지 않을 수도 있지만, TLA를 사용하여 다양한 조직 또는 조직과 연관된 사람들 사이의 링크에 대해 배울 수도 있습니다. TLA는 또한 제공된 제품이나 경험에 대해 사람들이 어떻게 느끼는지와 같은 주제에 대한 의견을 추출하기 위해 사용할 수도 있습니다.

TLA의 이득을 얻기 위해서는 텍스트 링크(TLA) 규칙을 포함하는 자원을 가지고 있어야 합니다. 템플리트를 선택할 때, TLA 열에 아이콘을 가지고 있는지 여부에 의해 TLA 규칙이 있는 템플리트를 알 수 있습니다.

텍스트 링크 분석 패턴은 추출 프로세스의 패턴 매치 단계 동안 텍스트 데이터에서 발견됩니다. 이 단계 동안, 규칙은 텍스트 데이터와 비교되고 매치가 발견되면 해당 정보가 패턴으로 추출됩니다. 텍스트 링크 분석에서 더 많은 것을 가져오거나 매치 방법을 변경하고자 할 경우가 있습니다. 이러한 경우, 규칙을 세분화하여 특정 필요성에 적용하십시오. 이는 텍스트 링크 규칙 탭에서 수행됩니다.

노트: 변수에 대한 지원은 버전 13에서 더 이상 사용되지 않습니다. 대신 매크로를 사용하십시오. 자세한 정보는 229 페이지의 『매크로에 대한 작업』의 내용을 참조하십시오.

텍스트 링크 규칙에 대해 작업할 위치

템플리트 편집기 또는 자원 편집기 보기의 텍스트 링크 탭에서 직접 규칙을 편집하고 작성할 수 있습니다. 규칙이 텍스트와 매치될 수 있는 방법을 알기 위해 이 탭에서 시뮬레이션을 실행할 수 있습니다. 시뮬레이션 동안, 추출은 샘플 시뮬레이션 데이터에 대해서만 실행되고 텍스트 링크 규칙은 패턴이 매치되는지 보기 위해 적용됩니다. 텍스트와 매치되는 규칙이 시뮬레이션 분할창에 표시됩니다. 매치를 기반으로, 텍스트가 매치되는 방법을 변경할 규칙 및 매크로를 편집할 것을 선택할 수 있습니다.

다른 고급 자원과 달리, TLA 규칙은 라이브러리에 고유하므로, 한 번에 하나의 라이브러리에서만 TLA 규칙을 사용할 수 있습니다. 템플리트 편집기 또는 자원 편집기에서, **텍스트 링크 규칙** 탭으로 이동하십시오. 이 탭에서, 사용하거나 편집하려는 TLA 규칙을 포함하는 라이브러리를 템플리트에서 지정할 수 있습니다. 이러한 이유로, 특정한 이유가 없으면 모든 규칙을 하나의 라이브러리에 저장할 것을 권장합니다.

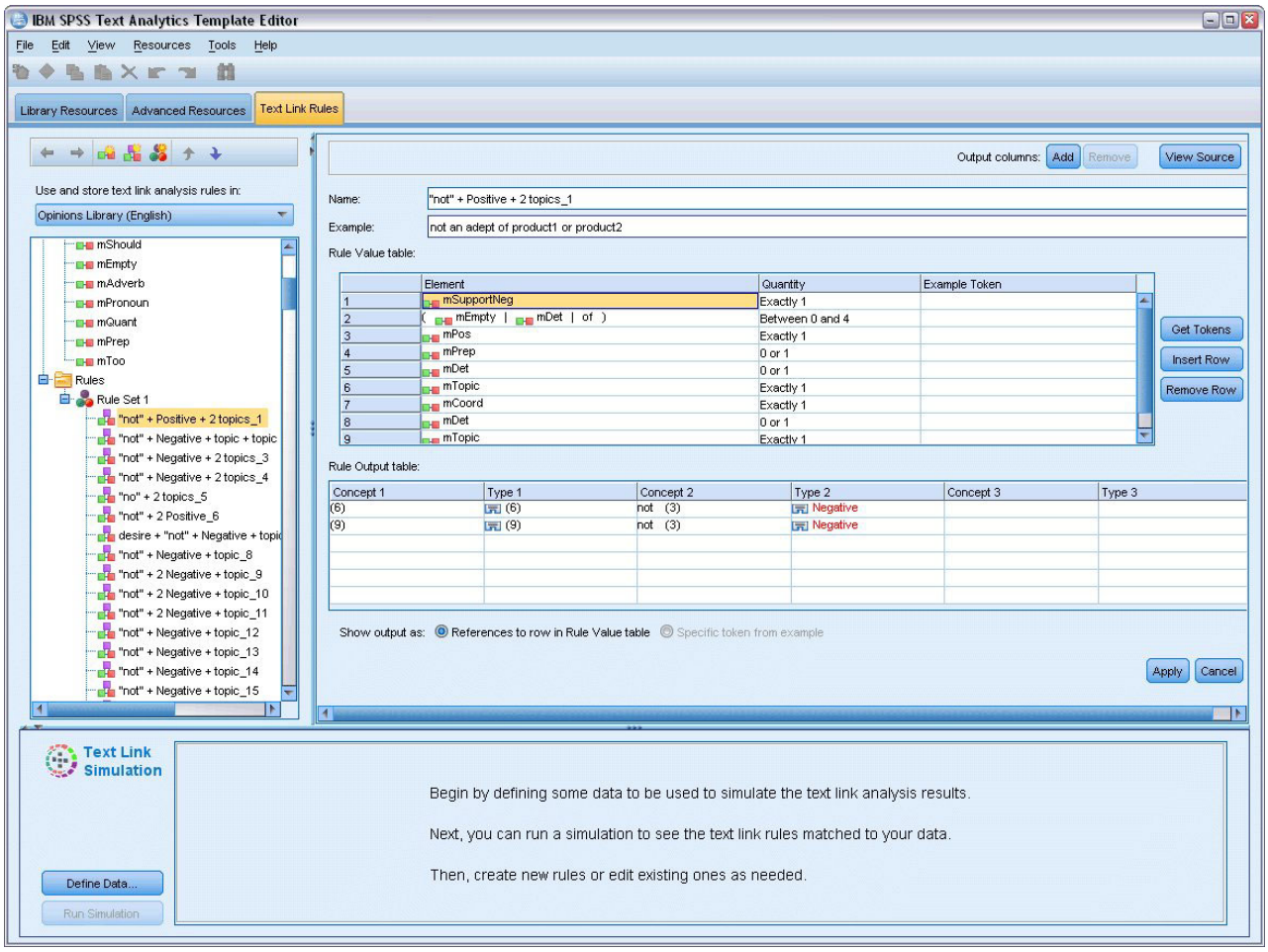


그림 42. 텍스트 링크 규칙 탭

중요! 이 탭은 일본어 자원에 사용할 수 없습니다.

시작 위치

텍스트 링크 규칙 탭 편집기에서 작업을 시작하기 위한 많은 방법이 있습니다.

- 일부 샘플 텍스트로 결과를 시뮬레이션하는 것으로 시작하여 시뮬레이션 데이터에서 현재 규칙 세트가 패턴을 추출하는 방법을 기반으로 매치 규칙을 편집하거나 작성합니다.
- 스크래치에서 새 규칙을 작성하거나 기존 규칙을 편집합니다.
- 소스 보기에서 직접 작업합니다.

규칙 편집 또는 작성 시기

각 템플리트와 함께 전달된 텍스트 링크 분석 규칙이 종종 많은 단순하거나 복잡한 관계를 텍스트에서 추출하는 데 적합한 반면, 이 규칙을 수정하거나 사용자 자신의 규칙을 작성하고자 하는 경우가 있습니다. 예를 들어,

- 새 규칙 또는 매크로를 작성하여 기존 규칙에서 추출되지 않는 아이디어 또는 관계를 캡처하려는 경우.

- 자원에 추가한 유형의 기본 작동을 변경하려는 경우. 이때 보통 mTopic 또는 mNonLingEntities와 같은 매크로를 편집해야 합니다. 자세한 정보는 231 페이지의 『특수 매크로: mTopic, mNonLingEntities, SEP』의 내용을 참조하십시오.
- 기존 텍스트 링크 분석 규칙 및 매크로에 새 유형을 추가하는 경우. 예를 들어, 유형 <Organization>이 너무 광범위하다고 생각하면, <Pharmaceuticals>, <Car Manufacturing>, <Finance> 등과 같은 여러 다른 업무 부문에 조직을 위한 새 유형을 작성할 수 있습니다. 이러한 경우, 텍스트 링크 분석 규칙을 편집하고(하거나) 이 새 유형을 이용하고 각각의 경우에 이를 처리하기 위해 매크로를 작성해야 합니다.
- 유형을 기존 텍스트 링크 분석 규칙에 추가하는 경우. 예를 들어, 다음 텍스트 john doe called jane doe를 캡처하는 규칙을 가지고 있지만 이 규칙이 이메일 교환을 캡처하기 위해 전화 통신을 캡처하는 것도 원할 수 있습니다. 규칙에 이메일에 대한 비언어 엔티티 유형을 추가할 수 있으므로, johndoe@ibm.com emailed janedoe@ibm.com과 같은 텍스트도 캡처합니다.
- 새 규칙을 작성하는 대신 기존 규칙을 약간 수정하는 경우. 예를 들어, 다음 텍스트 xyz is very good doe와 매치되는 규칙을 가지고 있지만 이 규칙이 xyz is very, very good도 캡처하기를 원할 수 있습니다.

텍스트 링크 분석 결과 시뮬레이션

새 텍스트 링크 규칙을 쉽게 정의하거나 특정 문장이 텍스트 링크 분석 동안 매치되는 방법을 쉽게 이해하기 위해, 샘플 텍스트 조각을 사용하여 시뮬레이션을 실행하는 것이 종종 유용합니다. 시뮬레이션 동안, 추출은 현재 언어학적 자원 세트 및 현재 추출 설정을 사용하여 샘플 시뮬레이션 데이터에 대해서만 실행됩니다. 목적은 시뮬레이트된 결과를 확보하고 이 결과를 사용하여 규칙을 개선하거나, 새 규칙을 작성하거나, 매치 발생 방법을 더 잘 이해하기 위한 것입니다. 텍스트 조각(컨텍스트에 따라 문장, 단어 또는 절) 각각에 대해, 시뮬레이션 출력은 해당 텍스트에서 패턴을 포함하지 않은 TLA 규칙과 토큰 콜렉션을 표시합니다. **토큰**은 추출 프로세스 동안 식별되는 단어 또는 단어 구문으로 정의됩니다.

다른 고급 자원과 달리, TLA 규칙은 라이브러리에 고유하므로, 한 번에 하나의 라이브러리에서만 TLA 규칙을 사용할 수 있습니다. 템플리트 편집기 또는 자원 편집기에서, **텍스트 링크 규칙** 탭으로 이동하십시오. 이 탭에서, 사용하거나 편집하려는 TLA 규칙을 포함하는 라이브러리를 템플리트에서 지정할 수 있습니다. 이러한 이유로, 특정한 이유가 없으면 모든 규칙을 하나의 라이브러리에 저장할 것을 권장합니다.

중요! 데이터 파일을 사용하는 경우, 처리 시간을 최소화하기 위해 이 파일에 포함되는 텍스트가 간단한지 확인하십시오. 시뮬레이션의 목적은 텍스트 조각이 해석되는 방법을 보고 규칙이 이 텍스트와 매치되는 방법을 이해하는 것입니다. 이 정보는 규칙을 작성하고 편집하는 데 도움이 됩니다. 한층 완전한 데이터 세트에 대한 결과를 확보하려면 TLA 추출이 사용되는 대화형 세션으로 스트림을 실행하거나 텍스트 링크 분석 노드를 사용하십시오. 이 시뮬레이션은 단지 테스트 및 규칙 작성 목적을 위한 것입니다.

시뮬레이션에 대한 데이터 정의

규칙이 텍스트와 매치될 수 있는 방법을 쉽게 알려면, 표본 데이터를 사용하여 시뮬레이션을 실행할 수 있습니다. 첫 번째 단계는 데이터를 정의하는 것입니다.

데이터 정의

1. **텍스트 링크 규칙** 탭의 맨 아래에 있는 시뮬레이션 분할창에서 **데이터 정의**를 클릭하십시오. 또는 이전에 데이터를 정의하지 않은 경우, 메뉴에서 **도구 > 시뮬레이션 실행**을 선택하십시오. 시뮬레이션 데이터 마법사가 열립니다.
2. 다음 중 하나를 선택하여 데이터 유형을 지정하십시오.
   - **직접 텍스트 붙여넣기 또는 입력**: 텍스트를 클립보드에서 붙여넣거나 처리할 텍스트를 수동으로 입력할 수 있는 텍스트 상자가 제공됩니다. 행마다 하나의 문장을 입력하거나 구두점(예: 마침표 또는 콤마)을 사용하여 문장을 분리할 수 있습니다. 텍스트를 입력한 경우, **시뮬레이션 실행**을 클릭하여 시뮬레이션을 시작할 수 있습니다.
   - **파일 데이터 소스 지정**: 이 옵션은 텍스트를 포함하는 파일을 처리할 것을 표시합니다. 처리될 파일을 정의할 수 있는 마법사 단계로 진행하려면 **다음**을 클릭하십시오. 파일이 선택되면, **시뮬레이션 실행**을 클릭하여 시뮬레이션을 시작할 수 있습니다. `.txt` 및 `.text` 파일 유형이 지원됩니다. 선택하는 데이터 파일은 시뮬레이션 동안 "있는 그대로" 읽혀집니다. 전체 파일은 사용자가 텍스트 마이닝 노드에 파일 목록 노드를 연결한 것과 동일한 방식으로 처리됩니다.

   중요사항: 데이터 파일을 사용하는 경우, 처리 시간을 최소화하기 위해 이 파일에 포함되는 텍스트가 간단한지 확인하십시오. 시뮬레이션의 목적은 텍스트 조각이 해석되는 방법을 보고 규칙이 이 텍스트와 매치되는 방법을 이해하는 것입니다. 이 정보는 규칙을 작성하고 편집하는 데 도움이 됩니다. 한층 완전한 데이터 세트에 대한 결과를 확보하려면 TLA 추출이 사용되는 대화형 세션으로 스트림을 실행하거나 텍스트 링크 분석 노드를 사용하십시오. 이 시뮬레이션은 단지 테스트 및 규칙 작성 목적을 위한 것입니다.
3. 시뮬레이션 프로세스를 시작하려면 **시뮬레이션 실행**을 클릭하십시오. 진행률 대화 상자가 나타납니다. 대화형 세션에 있으면, 시뮬레이션 동안 사용한 추출 설정은 이 대화형 세션에서 현재 선택되어 있는 설정입니다(개념 및 범주 보기의 **도구 > 추출 설정** 참조). 템플리트 편집기에 있는 경우에는, 시뮬레이션 동안 사용한 추출 설정이 기본 추출 설정이며, 이 설정은 텍스트 링크 분석 노드의 전문가 탭에 표시되는 것과 같습니다. 자세한 정보는 『시뮬레이션 결과 이해』의 내용을 참조하십시오.

시뮬레이션 결과 이해

규칙이 텍스트와 매치될 수 있는 방법을 쉽게 알려면, 표본 데이터 사용을 사용하여 시뮬레이션을 실행하고 결과를 검토할 수 있습니다. 여기에서 규칙 세트를 사용자 데이터에 잘 맞도록 변경할 수 있습니다. 추출 및 시뮬레이션 프로세스가 완료되면, 시뮬레이션 결과와 함께 표시됩니다.

추출 동안 식별된 “문장”마다, 정확한 "문장", 입력 텍스트 문장에서 발견된 토큰의 명세, 마지막으로 해당 문장에서 텍스트와 매치된 규칙을 포함하여 몇몇 정보 조각이 함께 표시됩니다. "**문장**"에 의해, 추출기가 텍스트를 읽을 수 있는 청크로 분리하는 방법에 따라 단어, 문장 또는 절을 의미합니다.

토큰은 추출 프로세스 중에 식별된 단어 또는 단어 문구로 정의됩니다. 예를 들어, *My uncle lives in New York* 문장에서는 추출 중에 *my*, *uncle*, *lives*, *in* 및 *new york* 토큰을 발견할 수 있습니다. 또한 *uncle*을 개념으로 추출하고 `<Unknown>`으로 유형 지정하며, *new york*을 개념으로 추출하고 `<Location>`으로 유형 지정할 수 있습니다. 모든 개념은 토큰이지만 모든 토큰이 개념인 것은 아닙니다. 토큰은 다른 매크로, 리터럴 문자열, 단어 간격일 수도 있습니다. 유형이 지정된 해당 단어 또는 단어 문구만 개념이 될 수 있습니다.

대화형 세션 또는 자원 편집기에서 작업할 때, 개념 수준에서 작업하는 것입니다. TLA 규칙은 한층 세부 단위여서, 추출되고 유형이 지정되지 않아도 문장의 개별 토큰이 규칙 정의에서 사용될 수 있습니다. 개념이 아닌 토큰을 사용할 수 있으면 텍스트에서 복잡한 관계 캡처 시 추가 융통성이 규칙에 제공됩니다.

시뮬레이션 데이터에 두 개 이상 문장이 있으면, **다음** 및 **이전**을 클릭하여 결과에서 앞뒤로 이동할 수 있습니다.

문장이 선택된 라이브러리(이 탭에서 트리 위체 있는 라이브러리 이름 참조)의 어떤 TLA 규칙과도 매치되지 않는 경우, 결과는 매치되지 않는 것으로 간주되고 어떤 규칙도 매치를 발견하지 못한 텍스트가 있음을 알 수 있도록 하고 해당 인스턴스를 신속하게 탐색할 수 있도록 하기 위해 **다음 매치되지 않음** 및 **이전 매치되지 않음** 단추가 사용됩니다.

새 규칙을 작성한 후, 규칙을 편집하거나 자원 또는 추출 설정을 변경하거나, 시뮬레이션을 다시 실행할 수 있습니다. 시뮬레이션을 다시 실행하려면 시뮬레이션 분할창에서 **시뮬레이션 실행**을 클릭하십시오. 동일한 입력 데이터가 다시 사용됩니다.

다음 필드 및 테이블이 시뮬레이션 결과에 표시됩니다.

입력 텍스트. 마법사에서 정의한 시뮬레이션 데이터로부터 추출 프로세스에 의해 식별된 실제 '문장'. 문장 기준으로, 추출기가 텍스트를 읽을 수 있는 청크로 분리하는 방법에 따라 단어, 문장 또는 절을 의미합니다.

시스템 보기. 추출 프로세스가 식별한 토큰의 콜렉션.

- **입력 텍스트 토큰.** 각 토큰은 입력 텍스트에서 발견되었습니다. 토큰은 이 주제의 앞에서 정의했습니다.
- **다음과 같이 유형 지정.** 토큰이 개념으로 식별되고 유형이 지정된 경우, 연관된 유형 이름(예: <Unknown>, <Person>, <Location>)이 이 열에 표시됩니다.
- **매치 매크로.** 토큰이 기존 매크로와 매치된 경우, 연관된 매크로 이름이 이 열에 표시됩니다.

입력 텍스트에 매치된 규칙. 이 테이블은 입력 텍스트에 대해 매치된 TLA 규칙을 보여줍니다. 매치된 규칙마다, **규칙 출력** 열에서 규칙의 이름이 표시되고 해당 규칙에 대해 연관된 출력 값(개념 + 유형 쌍)이 표시됩니다. 매치된 규칙 이름을 두 번 클릭하여 시뮬레이션 분할창 위의 편집기 분할창에서 규칙을 열 수 있습니다.

규칙 생성 단추. 시뮬레이션 분할창에서 이 단추를 클릭하면, 새 규칙이 시뮬레이션 분할창 위의 규칙 편집기 분할창에서 열립니다. 이 규칙은 입력 텍스트를 해당되는 예로 사용합니다. 마찬가지로, 시뮬레이션 동안 매크로에 대해 매치되거나 유형이 지정된 토큰은 **규칙 값 테이블**에서 요소 열에 자동으로 삽입됩니다. 토큰에 유형이 지정되고 토큰이 매크로에 매치된 경우, 매크로 값은 규칙을 단순화하기 위해 규칙에서 사용될 값입니다. 예를 들어, 문장 "*I like pizza*"눈 추출 동안 <Unknown>으로 유형이 지정되고 매크로 mTopic에 매치될 수 있습니다(기본 영어 자원을 사용하는 경우). 이러한 경우 mTopic은 생성된 규칙에서 요소로 사용됩니다. 자세한 정보는 232 페이지의 『텍스트 링크 규칙에 대한 작업』의 내용을 참조하십시오.

트리에서 규칙 및 매크로 탐색

추출 동안 텍스트 링크분석이 수행될 때, **텍스트 링크 규칙** 탭에서 선택된 라이브러리에 저장된 텍스트 링크 규칙이 사용됩니다.

다른 고급 자원과 달리, TLA 규칙은 라이브러리에 고유하므로, 한 번에 하나의 라이브러리에서만 TLA 규칙을 사용할 수 있습니다. 템플리트 편집기 또는 자원 편집기에서, **텍스트 링크 규칙** 탭으로 이동하십시오. 이 탭에서, 사용하거나 편집하려는 TLA 규칙을 포함하는 라이브러리를 템플리트에서 지정할 수 있습니다. 이러한 이유로, 특정한 이유가 없으면 모든 규칙을 하나의 라이브러리에 저장할 것을 강력하게 권장합니다.

이 탭의 **텍스트 링크 분석 규칙 사용 및 저장 위치:** 드롭다운 목록에서 해당 라이브러리를 선택하여 텍스트 링크 규칙 탭에서 작업하려고 하는 라이브러리를 지정할 수 있습니다. 추출 동안 텍스트 링크분석이 수행될 때, **텍스트 링크 규칙** 탭에서 선택된 라이브러리에 저장된 텍스트 링크 규칙이 사용됩니다. 따라서, 두 개 이상의 라이브러리에서 텍스트 링크 규칙(TLA 규칙)을 정의한 경우, TLA 규칙이 발견된 첫 번째 라이브러리만 텍스트 링크 분석에 사용됩니다. 이러한 이유로, 특정한 이유가 없으면 모든 규칙을 하나의 라이브러리에 저장할 것을 강력하게 권장합니다.

트리에서 매크로 또는 규칙을 선택할 때, 해당 내용은 오른쪽에 있는 편집기 창에 표시됩니다. 트리에서 항목을 마우스 오른쪽 단추로 클릭하면, 다음과 같이 가능한 다른 태스크를 표시하기 위해 컨텍스트 메뉴가 열립니다.

- 트리에서 새 매크로를 작성하고 오른쪽에 있는 편집기에서 엽니다.
- 트리에서 새 규칙을 작성하고 오른쪽에 있는 편집기에서 엽니다.
- 트리에서 새 규칙을 작성합니다.
- 편집을 단순화하기 위해 항목을 잘라내고, 복사하며, 붙여넣습니다.
- 자원에서 제거하기 위해 매크로, 규칙 및 규칙 세트를 삭제합니다.
- 처리 동안 무시되어야 함을 표시하기 위해 매크로, 규칙 및 규칙 세트를 사용하지 않도록 설정합니다.
- 처리 순서에 영향이 미치도록 위 또는 아래로 규칙을 이동합니다.

트리의 경고

경고는 트리에서 노란색 삼각형으로 표시되고, 문제점이 있을 수 있음을 알립니다. 팝업 설명을 표시하려면 결함 매크로 또는 규칙 위로 마우스 포인터를 움직이십시오. 대부분의 경우, **경고: 제공된 예가 없습니다. 예를 입력하십시오**가 표시되므로 예를 입력해야 합니다.

예가 없거나 예가 규칙과 매치되지 않는 경우, 토큰 가져오기 기능을 사용할 수 없으므로 규칙마다 하나의 예만 입력하도록 합니다.

규칙이 노랑색에서 강조표시될 때 유형 또는 매크로가 TLA 편집기에 알려지지 않음을 의미합니다. 메시지는 **경고: 알 수 없는 유형 또는 매크로**와 유사합니다. 소스 보기에서 `$something`에 의해 정의되는 항목(예: `$myType`)은 라이브러리에서 레거시 유형이 아니며 매크로도 아닙니다.

명령문 검사 프로그램을 업데이트하려면 다른 규칙 또는 매크로로 전환해야 합니다. 어떤 것도 다시 컴파일하지 않아도 됩니다. 따라서, 예를 들어 예가 없어서 규칙 A가 경고를 표시하는 경우, 예를 추가하고 상단 또는 하단 규칙을 클릭한 후 규칙 A로 다시 이동하여 현재 올바른지 확인해야 합니다.

매크로에 대한 작업

매크로는 유형, 다른 매크로 및 리터럴(단어) 문자열을 OR 연산자(|)로 함께 그룹화할 수 있도록 하여 텍스트 링크 분석 규칙의 형태를 단순화할 수 있습니다. 매크로 사용의 장점은 여러 텍스트 링크 분석 규칙에서 매크로를 재사용하여 단순화시키는 것 외에도, 모든 텍스트 링크 분석 규칙에서 업데이트를 수행하기 보다는 하나의 매크로에서 업데이트를 수행할 수 있도록 하는 것입니다. 제공되는 대부분의 TLA 규칙에는 사전 정의된 매크로가 포함됩니다. 매크로는 텍스트 링크 규칙 탭의 가장 왼쪽 분할창에서 트리 맨 위에 나타납니다.

다음 필드 및 테이블이 시뮬레이션 결과에 표시됩니다.

이름. 이 매크로를 식별하는 고유 이름. 규칙에서 신속하게 매크로를 식별할 수 있도록 소문자 m을 매크로 이름 앞에 붙일 것을 권장합니다. 수동으로 규칙에서 매크로를 참조할 때(인라인 편집 시 또는 소스 보기에서) 추출 프로세스에서 이 특수 이름을 찾을 수 있도록 $ 접두문자를 사용해야 합니다. 그러나 매크로 이름을 끌어서 놓거나 컨텍스트 메뉴를 통해 이 이름을 추가하는 경우, 제품은 자동으로 이를 매크로로 인식하고 $가 추가되지 않습니다.

매크로 값 테이블.

- 이 매크로가 표시할 수 있는 가능한 모든 값을 표시하는 여러 행. 이 값에서는 대소문자가 구분됩니다.
- 이 값에는 유형, 리터럴 문자열, 단어 간격 또는 매크로 중 하나이거나 이 유형의 조합이 포함될 수 있습니다. 자세한 정보는 239 페이지의 『규칙 및 매크로에 대해 지원되는 요소』의 내용을 참조하십시오.
- 매크로에서 요소의 값을 입력하려면 작업할 행을 두 번 클릭하십시오. 유형 참조, 매크로 참조, 리터럴 문자열 또는 단어 간격을 입력할 수 있는 편집 가능한 텍스트 상자가 나타납니다. 또는 공통 매크로, 유형 이름 및 비언어 유형 이름의 목록을 제공하는 컨텍스트 메뉴를 표시하기 위해 셀에서 마우스 오른쪽 단추를 클릭하십시오. 유형 또는 매크로를 참조하려면 '$' 문자를 매크로 또는 유형 이름 앞에 붙여야 합니다(예: 매크로 mTopic의 경우 $mTopic). 인수를 조합할 때, 괄호 ( )를 사용하여 인수를 그룹화하고 문자 |를 사용하여 부울 OR을 표시해야 합니다.
- 오른쪽에 있는 단추를 사용하여 매크로 값 테이블에서 행을 추가하거나 제거할 수 있습니다.
- 해당되는 행에서 각 요소를 입력하십시오. 예를 들어, am OR was OR is와 같은 3 리터럴 문자열 중 하나를 나타내는 매크로를 작성하려는 경우, 보기에서 별도의 행에 각 리터럴 문자열을 입력하고 매크로 테이블에 세 행을 포함합니다.

매크로 작성 및 편집

새 매크로를 작성하거나 기존 매크로를 편집할 수 있습니다. 매크로 편집기에 대한 지침과 설명을 따르십시오. 자세한 정보는 『매크로에 대한 작업』의 내용을 참조하십시오.

새 매크로 작성

1. 메뉴에서 **도구 > 새 매크로**를 선택하십시오. 또는 트리 도구 모음에서 새 매크로 아이콘을 클릭하여 편집기에서 새 매크로를 여십시오.
2. 고유한 이름을 입력하고 매크로 값 요소를 정의하십시오.
3. 오류 확인을 완료하면 **적용**을 클릭하십시오.

매크로 편집

1. 트리에서 매크로 이름을 클릭하십시오. 매크로는 오른쪽에 있는 편집기 분할창에서 열립니다.
2. 변경사항을 작성하십시오.
3. 오류 확인을 완료하면 **적용**을 클릭하십시오.

매크로 사용 안함 및 삭제

매크로 사용 안함

처리 동안 매크로가 무시되도록 하려면, 매크로를 사용하지 않도록 설정할 수 있습니다. 이렇게 하면 사용하지 않도록 설정한 매크로를 계속 참조하는 규칙에서 경고나 오류가 발생할 수 있습니다. 매크로를 삭제하고 사용하지 않도록 설정할 때 주의하십시오.

1. 트리에서 매크로 이름을 클릭하십시오. 매크로는 오른쪽에 있는 편집기 분할창에서 열립니다.
2. 이름을 마우스 오른쪽 단추로 클릭하십시오.
3. 컨텍스트 메뉴에서 **사용할 수 없음**을 선택하십시오. 매크로 아이콘은 회색이 되고 매크로 자체는 편집할 수 없게 됩니다.

매크로 삭제

매크로를 제거하려면, 해당 매크로를 삭제할 수 있습니다. 이렇게 하면 해당 매크로를 계속 참조하는 규칙에서 오류가 발생할 수 있습니다. 매크로를 삭제하고 사용하지 않도록 설정할 때 주의하십시오.

1. 트리에서 매크로 이름을 클릭하십시오. 매크로는 오른쪽에 있는 편집기 분할창에서 열립니다.
2. 이름을 마우스 오른쪽 단추로 클릭하십시오.
3. 컨텍스트 메뉴로부터 **삭제**를 선택하십시오. 매크로가 목록에서 사라집니다.

오류 확인, 저장 및 취소

매크로 변경사항 적용

매크로 편집기 외부를 클릭하거나 **적용**을 클릭하는 경우, 매크로는 오류를 찾기 위해 자동으로 스캔됩니다. 오류가 발견되면, 애플리케이션의 다른 부분으로 이동하기 전에 수정해야 합니다.

그러나 덜 심각한 오류가 발견되면, 경고만 제공됩니다. 예를 들어, 매크로에 유형 또는 다른 매크로에 대한 완료되지 않거나 참조되지 않는 정의가 있는 경우, 경고 메시지가 표시됩니다. **적용**을 클릭하는 경우, 정정되지 않은 경고는 왼쪽 분할창에 있는 규칙 및 매크로 트리에서 매크로 이름의 왼쪽에 경고 아이콘이 나타나도록 합니다.

매크로를 적용해도 매크로가 영구적으로 저장됨을 의미하지는 않습니다. 적용하면 오류 및 경고에 대해 확인하기 위해 검증 프로세스가 발생합니다.

대화형 워크벤치 세션 내에서 자원 저장

1. 대화식 워크벤치 세션 동안 자원에 대해 작성한 변경사항을 저장하기 위해, 다음에 스트림을 실행할 때 변경사항을 가져올 수 있습니다.

- 다음에 스트림을 실행할 때 동일한 자원을 가져올 수 있도록 모델링 노드를 업데이트하십시오. 자세한 정보는 88 페이지의 『모델링 노드 업데이트 및 저장』의 내용을 참조하십시오. 그런 다음 스트림을 저장하십시오. 스트림을 저장하려면, 모델링 노드를 업데이트한 후 기본 IBM SPSS Modeler 창에서 저장을 수행하십시오.

2. 다른 스트림에서 사용할 수 있도록 대화식 워크벤치 세션 동안 자원에 대해 작성한 변경사항을 저장하기 위해, 다음을 수행할 수 있습니다.

- 사용한 템플리트를 업데이트하거나 새 템플리트를 작성하십시오. 자세한 정보는 173 페이지의 『템플리트 작성 및 업데이트』의 내용을 참조하십시오. 현재 노드에 대한 변경사항은 저장되지 않습니다(이전 단계 참조).
- 또는 사용한 TAP를 업데이트하십시오. 자세한 정보는 149 페이지의 『텍스트 분석 패키지 업데이트』의 내용을 참조하십시오.

템플리트 편집기 내에서 자원 저장

1. 먼저 라이브러리를 출판하십시오. 자세한 정보는 194 페이지의 『라이브러리 출판』의 내용을 참조하십시오.
2. 그런 다음, 메뉴에서 **파일 > 자원 템플리트 저장**을 통해 템플리트를 저장하십시오.

매크로 변경사항 취소

1. 변경사항을 삭제하려면 **취소**를 클릭하십시오.

특수 매크로: mTopic, mNonLingEntities, SEP

Opinions 템플리트(및 유사한 템플리트)와 기본 자원 템플리트는 mTopic 및 mNonLingEntities와 같은 두 개의 특수 매크로와 함께 제공됩니다.

mTopic

기본적으로, 매크로 mTopic은 유형이 의견(opinion) 유형(예: <Negative> 또는 <Positive>)이나 고급 자원에서 비언어 엔티티로 정의된 유형이 아닌 한, *코어* 라이브러리 유형 <Person>, <Organization>, <Location> 등과 같이 의견과 연결될 템플리트에서 제공되는 모든 유형을 그룹화합니다.

Opinions(또는 유사한) 템플리트에서 새 유형을 작성할 때마다, 제품은 이 유형이 고급 자원 탭의 비언어 엔티티 섹션이나 다른 매크로에서 지정되지 않는 한 매크로 mTopic에서 정의된 다른 유형과 동일한 방식으로 처리된다고 가정합니다.

Opinions 템플리트로부터 자원에서 새 유형 <Vegetables> 및 <Fruit>를 작성한다고 가정해 보십시오. 변경하지 않아도, 새 유형은 mTopic 유형으로 처리되므로, 새 유형에 대하여 긍정, 부정, 중립 및 컨텍스트 의견을 자동으로 노출할 수 있습니다. 추출 동안, 예를 들어 문장 "*I enjoy broccoli, but I hate grapefruit*"는 다음 두 가지의 출력 패턴을 생성합니다.

```
broccoli <Vegetables> + like <Positive>

grapefruit <Fruit> + dislike <Negative>
```

그러나 mTopic에서 다른 유형과 다르게 해당 유형을 처리하려면, mPos와 같은 유형 이름을 기존 매크로에 추가하거나(모든 긍정 의견 유형이 그룹화 됨) 나중에 하나 이상의 규칙에서 참조할 수 있는 새 매크로를 작성할 수 있습니다.

중요! <Vegetables>와 같은 새 유형을 작성하는 경우, 이 새 유형은 mTopic에서 유형으로 포함되지만, 이 유형 이름이 명시적으로 매크로 정의에 표시되지는 않습니다.

mNonLingEntities

마찬가지로, 새 비언어 엔티티를 고급 자원 탭의 **비언어 엔티티** 섹션에서 추가하면, 달리 지정하지 않는 한 자동으로 mNonLingEntities로 처리됩니다. 자세한 정보는 215 페이지의 『비언어 엔티티』의 내용을 참조하십시오.

SEP

또한 사전정의된 매크로 SEP를 사용할 수 있습니다. 이 매크로는 로컬 머신에서 정의된 전역 구분 문자(일반적으로 콤마(,))에 해당됩니다.

텍스트 링크 규칙에 대한 작업

텍스트 링크 분석 규칙은 문장에 매치를 수행하기 위해 사용되는 부울 쿼리입니다. 텍스트 링크 분석 규칙에는 유형, 매크로, 리터럴 문자열 또는 단어 간격 인수 중 하나 이상이 포함됩니다. TLA 결과를 추출하기 위해 하나 이상의 텍스트 링크 분석 규칙을 가지고 있어야 합니다.

다음 영역 및 필드가 텍스트 링크 규칙 탭, 규칙 편집기에 표시됩니다.

이름 필드. 텍스트 링크 규칙의 고유 이름.

예 필드. 선택적으로, 이 규칙에 의해 캡처되는 예 문장 또는 단어 시퀀스를 포함할 수 있습니다. 예를 사용하도록 하십시오. 이 편집기에서, 이 예 텍스트에서 토큰을 생성하여 텍스트가 규칙에 매치되는 방법과 출력될 방법을 볼 수 있습니다. **토큰**은 추출 프로세스 중에 식별된 단어 또는 단어 문구로 정의됩니다. 예를 들어, *My uncle lives in New York* 문장에서는 추출 중에 *my*, *uncle*, *lives*, *in* 및 *new york* 토큰을 발견할 수 있습니다. 또한 *uncle*을 개념으로 추출하고 <Unknown>으로 유형 지정하며, *new york*을 개념으로 추출하고 <Location>으로 유형 지정할 수 있습니다. 모든 개념은 토큰이지만 모든 토큰이 개념인 것은 아닙니다. 토큰은 다른 매크로, 리터럴 문자열, 단어 간격일 수도 있습니다. 유형이 지정된 해당 단어 또는 단어 문구만 개념이 될 수 있습니다.

규칙 값 테이블. 이 테이블에는 규칙을 문장에 매치하기 위해 사용되는 규칙 요소가 포함되어 있습니다. 오른쪽에 있는 단추를 사용하여 테이블에서 행을 추가하거나 제거할 수 있습니다. 테이블은 세 개의 열로 구성됩니다.

- **요소** 열. 유형, 리터럴 문자열, 단어 간격(<Any Token>) 또는 매크로 중 하나 또는 이들의 조합으로 값을 입력하십시오. 자세한 정보는 239 페이지의 『규칙 및 매크로에 대해 지원되는 요소』의 내용을 참조하십시오. 요소 셀을 두 번 클릭하여 정보를 직접 입력하십시오. 또는 공통 매크로, 유형 이름 및 비언어 종류 이

름의 목록을 제공하는 컨텍스트 메뉴를 표시하기 위해 셀에서 마우스 오른쪽 단추를 클릭하십시오. 정보를 입력하여 셀에 입력하는 경우, '$' 문자를 매크로 또는 유형 이름 앞에 붙이십시오(예: 매크로 mTopic의 경우 $mTopic). 요소 행을 작성하는 순서는 규칙이 텍스트에 매치되는 방법에 중요합니다. 인수를 조합할 때, 괄호 ( )를 사용하여 인수를 그룹화하고 문자 |를 사용하여 부울 OR을 표시해야 합니다. 값에서 대소문자가 구분됩니다.

- **양** 열. 이는 매치 발생을 위해 요소가 발견되어야 하는 최소 및 최대 횟수를 표시합니다. 예를 들어, 0 - 3개 단어의 다른 두 요소 사이에 간격 또는 단어 시리즈를 정의하려는 경우, 목록에서 **0 및 3 사이**를 선택하거나 직접 대화 상자에 숫자를 입력할 수 있습니다. 기본값은 '**정확히 1**'입니다. 일부 경우에는, 요소를 선택사항으로 만들려고 합니다. 이러한 경우, 최소 양이 0이고 최대 양이 0 보다 큽니다(즉, 0 또는 1, 0 및 2 사이). 규칙의 첫 번째 요소는 선택사항일 수 없으며, 이는 양이 0이 될 수 없음을 의미합니다.
- **예 토큰** 열. **토큰 가져오기**를 클릭하면, 프로그램은 **예** 텍스트를 토큰으로 분리하고 이 토큰들을 사용하여 이 열을 사용자가 정의한 요소와 매치되는 토큰으로 채웁니다. 출력 테이블에서 이 토큰을 볼 수도 있습니다(이와 같이 표시되도록 선택하는 경우).

규칙 출력 테이블. 이 테이블의 각 행은 TLA 패턴 출력이 결과에 나타나는 방식을 정의합니다. 규칙 출력은 최대 6개 개념/유형 열 쌍의 패턴을 생성할 수 있습니다. 각각은 슬롯을 나타냅니다. 예를 들어, 유형 패턴 <Location> + <Positive>은 두 개의 슬롯 패턴으로, 두 개의 개념/유형 열 쌍으로 구성됨을 의미합니다.

언어는 다양한 방식으로 동일한 기본 아이디어를 표현하기 위한 자유를 제공하므로, 동일한 기본 아이디어를 캡처하도록 정의된 여러 규칙이 있을 수 있습니다. 예를 들어, 텍스트 *"Paris is a place I love"* 및 텍스트 *"I really, really like Paris and Florence"*는 동일한 기본 아이디어(Paris is liked)를 나타내지만 다르게 표현되어 두 개의 다른 규칙 둘 다가 캡처되어야 합니다. 그러나, 유사한 아이디어가 함께 그룹화된 경우 패턴 결과에 대해 더 쉽게 작업할 수 있습니다. 이러한 이유로, 이 두 개의 구문을 캡처하기 위한 두 가지의 다른 규칙을 가지고 있어도, 두 규칙 모두에 대해 동일한 출력을 정의할 수 있습니다(예: 둘 다 텍스트를 나타내도록 유형 패턴 <Location> + <Positive>). 그리고 이 방식에서, 출력이 항상 원래 텍스트에서 발견된 단어 순서 또는 구조를 모방하지 않음을 볼 수 있습니다. 게다가, 이러한 유형 패턴은 다른 구문과 매치될 수 있고, paris + like 및 tokyo + like와 같은 개념 패턴을 생성할 수 있습니다.

오류를 적게 하면서 신속하게 출력을 정의하려면, 컨텍스트 메뉴를 사용하여 출력에서 보려고 하는 요소를 선택할 수 있습니다. 또는 규칙 값 테이블에서 출력으로 요소를 끌어다 놓을 수 있습니다. 예를 들어, 규칙 값 테이블의 행 2에서 mTopic 매크로에 대한 참조를 포함하는 규칙을 가지고 있고 해당 값이 출력되도록 하려면, 단지 mTopic에 대한 요소를 규칙 출력 테이블의 첫 번째 열 쌍으로 끌어다 놓으면 됩니다. 이와 같이 하면 선택한 쌍에 대한 개념 및 유형 둘 다 자동으로 채워집니다. 또는 규칙 값 테이블의 세 번째 요소(행 3)에 의해 정의된 유형으로 출력이 시작되도록 하려면, 해당 유형을 규칙 값 테이블에서 출력 테이블의 **유형 1** 셀로 끄십시오. 테이블은 괄호로 행 참조를 표시하도록 업데이트됩니다(3).

또는 출력할 각 **개념** 열에서 셀을 두 번 클릭하고 $와 행 번호를 차례로 입력하여(예: 규칙 값 테이블의 행 2에 정의된 용어를 참조할 경우 $2) 참조를 수동으로 테이블에 입력할 수도 있습니다. 수동으로 정보를 입력할 때, **유형** 열도 정의해야 하고, #와 행 번호를 차례로 입력해야 합니다(예: 규칙 값 테이블의 2 행에 정의된 요소를 참조하는 경우 #2).

게다가, 방법을 조합할 수도 있습니다. 규칙 값 테이블의 행 4에 유형 <Positive>가 있다고 가정해 보십시오. Type 2 열로 끌어온 후 Concept 2 열에서 셀을 두 번 클릭하여 수동으로 앞에 단어 '*not*'을 입력할 수 있습니다. 그러면 출력 열은 테이블에서 not (4)를 읽거나, 편집 모드 또는 소스 모드에서 not $4를 읽게 됩니다. 그러면 Type 1 열에서 마우스 오른쪽 단추를 클릭하고 예를 들어 mTopic이라고 하는 매크로를 선택할 수 있습니다. 그러면, 이 출력은 car + bad와 같은 개념 패턴이 될 수 있습니다.

대부분의 규칙에는 단 하나의 행이 있지만 두 개 이상의 출력이 가능하고 바람직한 경우가 있습니다. 이러한 경우, 규칙 출력 테이블에서 행마다 하나의 출력을 정의하십시오.

중요! 다른 언어적 처리 조작은 TLA 패턴의 추출 동안 수행됩니다. 따라서 출력이 t$3\t#3을 읽을 때, 이는 패턴이 궁극적으로 세 번째 요소에 대한 최종 개념을 표시할 것이고 모든 언어적 처리 후 세 번째 요소의 최종 유형이 적용됨을 의미합니다(동의어 및 기타 그룹).

- **지정된 대로 출력 표시.** 기본적으로 **규칙 값 테이블의 행에 대한 참조** 옵션이 선택되고 출력은 규칙 값 탭에 정의된 대로 행에 대한 숫자 참조를 사용하여 표시됩니다. 이전에 토큰 가져오기을 클릭하고 규칙 값 테이블의 예 토큰 열에 토큰이 있는 경우, 옵션을 선택하여 이러한 특정 토큰에 대한 출력을 볼 것을 선택할 수 있습니다.

참고: 출력 테이블에 충분한 개념/유형 출력 쌍이 표시되지 않은 경우, 편집기 도구 모음에서 추가 단추를 클릭하여 다른 쌍을 추가할 수 있습니다. 세 개의 쌍이 현재 표시되고 추가를 클릭하는 경우, 두 개 이상의 열(개념 4 및 유형 4)이 테이블에 추가됩니다. 이는 이제 모든 규칙에 대한 출력 테이블에서 네 개의 쌍이 표시됨을 의미합니다. 또한 이 라이브러리의 규칙 세트에 있는 다른 규칙이 해당 쌍을 사용하지 않는 한 사용되지 않는 쌍을 제거할 수도 있습니다.

예 규칙

자원에 다음 텍스트 링크 분석 규칙이 포함되고 TLA 결과 추출을 사용하도록 설정하였다고 가정해 보십시오.

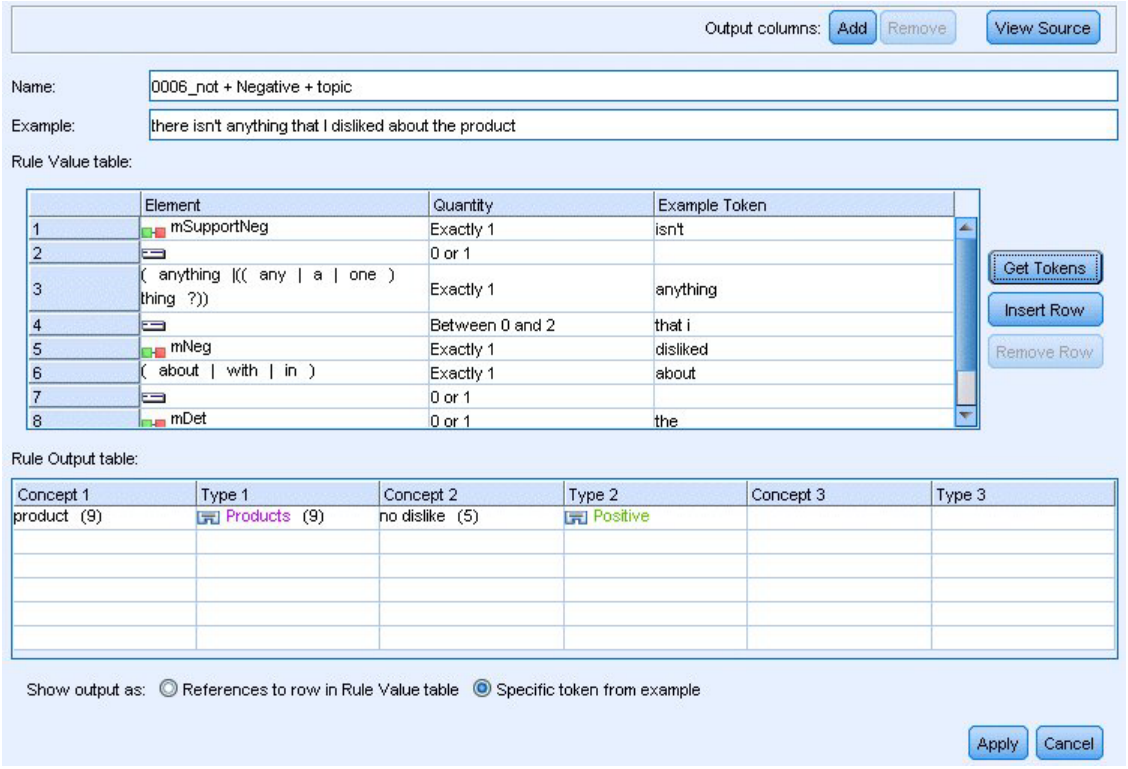


그림 43. 텍스트 링크 규칙 탭: 규칙 편집기

추출할 때마다, 추출 엔진은 각 문장을 읽고 다음 시퀀스와 매치하려고 합니다.

표 44. 추출 시퀀스 예

요소(행)	인수 설명
1	매크로 mPos 또는 mNeg에 의해 표시되는 유형 중 하나, 또는 유형 <Uncertain>의 개념.
2	매크로 mTopic에 의해 표시되는 유형 중 하나로 입력된 개념.
3	매크로 mBe에 의해 표시되는 단어 중 하나.
4	단어 간격 또는 <Any Token>으로도 언급되는 선택적 요소(0 또는 1개 단어).
5	매크로 mTopic에 의해 표시되는 유형 중 하나로 입력된 개념.

이 규칙에서 원하는 모든 것이 규칙 값 테이블의 행 5에 정의된 mTopic 매크로 + 규칙 값 테이블의 행 1에 정의된 대로 mPos, mNeg 또는 <Uncertain>에 해당하는 개념 또는 유형인 패턴임을 보여줍니다. 이는 sausage + like 또는 <Unknown> + <Positive>가 될 수 있습니다.

규칙 작성 및 편집

새 규칙을 작성하거나 기존 규칙을 편집할 수 있습니다. 규칙 편집기에 대한 지침과 설명을 따르십시오. 자세한 정보는 232 페이지의 『텍스트 링크 규칙에 대한 작업』의 내용을 참조하십시오.

새 규칙 작성

1. 메뉴에서 **도구 > 새 규칙**을 선택하십시오. 또는 트리 도구 모음에서 새 규칙 아이콘을 클릭하여 편집기에서 새 규칙을 여십시오.

2. 고유한 이름을 입력하고 규칙 값 요소를 정의하십시오.
3. 오류 확인을 완료하면 **적용**을 클릭하십시오.

규칙 편집

1. 트리에서 규칙 이름을 클릭하십시오. 규칙은 오른쪽에 있는 편집기 분할창에서 열립니다.
2. 변경사항을 작성하십시오.
3. 오류 확인을 완료하면 **적용**을 클릭하십시오.

규칙 사용 안함 및 삭제

규칙 사용 안함

처리 중에 규칙이 무시되도록 하려면 이 규칙을 사용하지 않도록 설정할 수 있습니다. 규칙을 삭제하거나 사용하지 않도록 설정할 때 주의하십시오.

1. 트리에서 규칙 이름을 클릭하십시오. 규칙은 오른쪽에 있는 편집기 분할창에서 열립니다.
2. 이름을 마우스 오른쪽 단추로 클릭하십시오.
3. 컨텍스트 메뉴에서 **사용할 수 없음**을 선택하십시오. 규칙 아이콘은 회색이 되고 규칙 자체는 편집할 수 없게 됩니다.

규칙 삭제

규칙을 제거하기 위해 규칙을 삭제할 수 있습니다. 규칙을 삭제하거나 사용하지 않도록 설정할 때 주의하십시오.

1. 트리에서 규칙 이름을 클릭하십시오. 규칙은 오른쪽에 있는 편집기 분할창에서 열립니다.
2. 이름을 마우스 오른쪽 단추로 클릭하십시오.
3. 컨텍스트 메뉴로부터 **삭제**를 선택하십시오. 규칙이 목록에서 사라집니다.

오류 확인, 저장 및 취소

규칙 변경사항 적용

규칙 편집기 외부를 클릭하거나 **적용**을 클릭하는 경우, 규칙은 오류를 찾기 위해 자동으로 스캔됩니다. 오류가 발견되면, 애플리케이션의 다른 부분으로 이동하기 전에 수정해야 합니다.

그러나 덜 심각한 오류가 발견되면, 경고만 제공됩니다. 예를 들어, 규칙에 유형 또는 매크로에 대한 완료되지 않거나 참조되지 않는 정의가 있는 경우, 경고 메시지가 표시됩니다. **적용**을 클릭하는 경우, 정정되지 않은 경고는 왼쪽 분할창에 있는 트리에서 규칙 이름의 왼쪽에 경고 아이콘이 나타나도록 합니다.

규칙을 적용해도 규칙이 영구적으로 저장됨을 의미하지는 않습니다. 적용하면 오류 및 경고에 대해 확인하기 위해 검증 프로세스가 발생합니다.

대화형 워크벤치 세션 내에서 자원 저장

1. 대화식 워크벤치 세션 동안 자원에 대해 작성한 변경사항을 저장하기 위해, 다음에 스트림을 실행할 때 변경사항을 가져올 수 있습니다.
   - 다음에 스트림을 실행할 때 동일한 자원을 가져올 수 있도록 모델링 노드를 업데이트하십시오. 자세한 정보는 88 페이지의 『모델링 노드 업데이트 및 저장』의 내용을 참조하십시오. 그런 다음 스트림을 저장하십시오. 스트림을 저장하려면, 모델링 노드를 업데이트한 후 기본 IBM SPSS Modeler 창에서 저장을 수행하십시오.
2. 다른 스트림에서 사용할 수 있도록 대화식 워크벤치 세션 동안 자원에 대해 작성한 변경사항을 저장하기 위해, 다음을 수행할 수 있습니다.
   - 사용한 템플리트를 업데이트하거나 새 템플리트를 작성하십시오. 자세한 정보는 173 페이지의 『템플리트 작성 및 업데이트』의 내용을 참조하십시오. 현재 노드에 대한 변경사항은 저장되지 않습니다(이전 단계 참조).
   - 또는 사용한 TAP를 업데이트하십시오. 자세한 정보는 149 페이지의 『텍스트 분석 패키지 업데이트』의 내용을 참조하십시오.

템플리트 편집기 내에서 자원 저장

1. 먼저 라이브러리를 출판하십시오. 자세한 정보는 194 페이지의 『라이브러리 출판』의 내용을 참조하십시오.
2. 그런 다음, 메뉴에서 **파일 > 자원 템플리트 저장**을 통해 템플리트를 저장하십시오.

규칙 변경사항 취소

1. 변경사항을 삭제하려면, 편집기 창에서 **취소**를 클릭하십시오.

규칙 순서 처리

텍스트 링크 분석이 추출 동안 수행될 때, 매치가 발견되거나 모든 규칙이 소모될 때까지 각 규칙에 대해 차례로 "문장"(절, 단어, 구문)이 매치됩니다. 트리에서 위치는 규칙이 시도되는 순서를 알려줍니다. 우수 사례는 가장 특정한 규칙에서 가장 일반적인 규칙으로 순서를 지정하는 것입니다. 가장 특정한 규칙은 트리의 맨 위에 있어야 합니다. 특정 규칙 또는 규칙 세트의 순서를 변경하려면, 도구 모음의 위 및 아래 화살표나 규칙 및 매크로 트리 컨텍스트 메뉴를 통해 **위로 이동** 또는 **아래로 이동**을 선택하십시오.

소스 보기에 있는 경우, 편집기에서 이동하여 규칙의 순서를 변경할 수 없습니다. 소스 보기에서 규칙이 더 위에 표시될수록 더 빨리 처리됩니다. 복사/붙여넣기 문제를 방지하려면 트리에서만 규칙을 다시 정렬하도록 하십시오.

중요! IBM SPSS Modeler Text Analytics의 이전 버전에서는, 고유한 숫자 규칙 ID를 가지고 있어야 했습니다. 17.1 버전부터는 트리에서 규칙을 위 또는 아래로 이동하거나 소스 보기에서 해당 위치에 의해서만 처리 순서를 표시할 수 있습니다.

예를 들어, 텍스트에 다음 두 개의 문장이 포함되어 있다고 가정하십시오.

I love anchovies

I love anchovies and green peppers

또한 다음 값을 가지고 있는 두 개의 텍스트 링크 분석 규칙이 존재한다고 가정해 보십시오.

A

	Element	Quantity	Example Token
1	Positive	Exactly 1	
2	mDet	0 or 1	
3	mTopic	Exactly 1	
4			
5			
6			
7			

B

	Element	Quantity	Example Token
1	Positive	Exactly 1	
2	mDet	0 or 1	
3	mTopic	Exactly 1	
4	(SEP | and | or)	1 or 2	
5	mDet	0 or 1	
6	mTopic	Exactly 1	
7			

그림 44. 두 가지의 예 규칙

소스 보기에서, 규칙 값은 다음과 유사할 수 있습니다.

A: `value = $Positive $mDet? $mTopic`

B: `value = $Positive $mDet? $mTopic ($SEP|and|or){1,2} $mDet? $mTopic`

규칙 **A**가 규칙 **B**보다 트리에서 위(맨 위 가까이)에 있으면, 규칙 **A**가 먼저 처리되고 문장 *I love anchovies and green peppers*는 `$Positive $mDet? $mTopic`에 의해 먼저 매치되어, 불완전한 패턴 출력(`anchovies + like`)을 생성합니다. 두 개의 `$mTopic` 매치에 대해 찾지 못한 규칙에 의해 매치되었기 때문입니다.

따라서, 텍스트의 본질을 캡처하려면 가장 특정한 규칙(이 경우 **B**)이 더 일반적인 규칙(이 경우 규칙 **A**)보다 트리에서 위에 위치되어야 합니다.

규칙 세트에 대한 작업(다중 전달)

규칙 세트는 여러 전달 프로세스를 수행하도록 규칙 및 매크로 트리에서 함께 관련 규칙 세트를 그룹화하는 유용한 방법입니다. 규칙 세트에는 이름 이외의 어떤 정의 자체도 없으며 규칙을 의미있는 그룹으로 구성하기 위해 사용됩니다. 일부 컨텍스트에서, 텍스트는 너무 서식이 많고 단일 전달로 처리되기에는 다양합니다. 예를 들어, 보안 정보 데이터에 대해 작업할 때, 텍스트는 접속 방법(*x called y*), 가족 관계(*y's brother-in-law x*), 화폐 교환(*x wired $100 to y*) 등을 통해 노출되는 개별값 사이의 링크를 포함할 수 있습니다. 이러한 경우, 특수화된 텍스트 링크 분석 규칙 세트를 작성하는 것이 유용합니다. 이 세트는 노출되는 접속에 대한 하나의 관계, 노출되는 가족 구성원에 대한 다른 관계 등 특정 종류의 관계에 초점을 맞춥니다.

규칙 세트를 작성하려면, 규칙 및 매크로 트리 컨텍스트 메뉴나 도구 모음에서 "규칙 세트 작성"을 선택하십시오. 그러면 트리의 규칙 세트 노드 아래에서 직접 새 규칙을 작성하거나 규칙 세트에 기존 규칙을 이동할 수 있습니다.

규칙이 규칙 세트로 그룹화되는 자원을 사용하여 추출을 실행할 때, 추출 엔진은 각각의 전달에서 서로 다른 종류의 패턴과 매치하기 위해 텍스트를 통하여 여러 전달을 작성하도록 강요할 수 있습니다. 이러한 경우, "문장"은 각 규칙 세트에서 규칙에 매치될 수 있는 반면, 규칙 세트 없이는 단일 규칙에만 매치될 수 있습니다.

참고: 규칙 세트마다 최대 512개의 규칙을 추가할 수 있습니다.

새 규칙 세트 작성

1. 메뉴에서 **도구 > 새 규칙 세트**를 선택하십시오. 또는 트리 도구 모음에서 새 규칙 세트 아이콘을 클릭하십시오. 규칙 세트는 규칙 트리에 나타납니다.
2. 이 규칙 세트에 새 규칙을 추가하거나 기존 규칙을 세트로 이동하십시오.

규칙 세트 사용 안함

1. 트리에서 규칙 세트 이름을 마우스 오른쪽 단추로 클릭하십시오.
2. 컨텍스트 메뉴에서 **사용할 수 없음**을 선택하십시오. 규칙 세트 아이콘은 회색이 되고 해당 규칙 세트 내에 포함된 모든 규칙 역시 처리 동안 사용할 수 없도록 되어 무시됩니다.

규칙 세트 삭제

1. 트리에서 규칙 세트 이름을 마우스 오른쪽 단추로 클릭하십시오.
2. 컨텍스트 메뉴로부터 **삭제**를 선택하십시오. 포함하는 규칙 세트와 모든 규칙이 자원에서 삭제됩니다.

규칙 및 매크로에 대해 지원되는 요소

다음 인수는 텍스트 링크 분석 규칙 및 매크로의 값 매개변수에 대해 승인됩니다.

매크로

텍스트 링크 분석 규칙이나 다른 매크로에서 직접 매크로를 사용할 수 있습니다. 직접 매크로 이름을 입력하거나 소스 보기에서 입력하는 경우(컨텍스트 메뉴에서 매크로 이름을 선택하는 것과 반대로), 달러 부호 문자($)가 이름 앞에 붙여 있는지 확인하십시오(예: `$mTopic`). 매크로 이름에서는 대소문자가 구분됩니다. 컨텍스트 메뉴를 통해 매크로를 선택할 때 현재 텍스트 링크 규칙 탭에서 정의된 매크로에서 선택할 수 있습니다.

유형

텍스트 링크 분석 규칙이나 매크로에서 직접 유형을 사용할 수 있습니다. 직접 유형 이름을 입력하거나 소스 보기에서 입력하는 경우(컨텍스트 메뉴에서 유형을 선택하는 것과 반대로), 달러 부호 문자($)가 유형 이름 앞에 붙여 있는지 확인하십시오(예: `$Person`). 유형 이름에서는 대소문자가 구분됩니다. 컨텍스트 메뉴를 사용하는 경우, 사용되는 현재 자원 세트의 유형에서 선택할 수 있습니다.

인식되지 않은 유형을 참조하는 경우, 경고 메시지를 수신하고 규칙에는 사용자가 작성할 때까지 규칙 및 매크로 트리에 경로 아이콘이 있습니다.

리터럴 문자열

추출되지 않은 정보를 포함하기 위해, 추출 엔진이 검색할 리터럴 문자열을 정의할 수 있습니다. 추출된 모든 단어 또는 구문이 유형에 지정되었고 이러한 이유로 리터럴 문자열에서 사용할 수 없습니다. 추출된 단어를 사용하는 경우, 해당 유형이 <Unknown>인 경우에도 무시됩니다.

리터럴 문자열은 하나 이상의 단어가 될 수 있습니다. 다음 규칙은 리터럴 문자열을 정의할 때 적용됩니다.

- 문자열 목록을 (his)와 같이 괄호로 묶으십시오. 리터럴 문자열 선택사항이 있는 경우 각 문자열은 OR 연산자에 의해 구분됩니다(예: (a|an|the) 또는 (his|hers|its)).
- 단일 또는 복합 단어를 사용하십시오.
- 목록의 각 단어를 | 문자(부울 OR)로 구분하십시오.
- 단수 및 복수 양식 모두를 매치하려면 두 양식을 입력하십시오. 굴절은 자동으로 생성되지 않습니다.
- 소문자만 사용하십시오.
- 리터럴 문자열을 재사용하여면, 이 리터럴 문자열을 매크로로 정의한 후 다른 매크로 및 텍스트 링크 분석 규칙에서 해당 매크로를 사용하십시오.
- 문자열에 마침표(전체 중지)이나 하이픈이 있는 경우, 이들도 포함해야 합니다. 예를 들어, 텍스트에서 a.k.a를 매치하려면 리터럴 문자열로 문자 a.k.a와 함께 마침표를 입력하십시오.

제외 연산자

특정 슬롯 차지에서 부정 표현식을 중지하려면 제외 연산자로 !를 사용하십시오. 인라인 셀 편집을 통해 직접(규칙 값 테이블 또는 매크로 값 테이블에서 셀을 두 번 클릭) 추가하거나 소스 보기에서 추가할 수 있습니다. 예를 들어, $mTopic @{0,2} !($Positive) $Budget을 텍스트 링크 분석 규칙에 추가하는 경우, (1) mTopic 매크로에서 유형에 지정된 용어를 포함하고, (2) 0 - 2개 단어 길이의 단어 간격을 포함하며, (3) <Positive> 유형에 지정된 용어의 인스턴스는 전혀 없고, (4) <Budget> 유형에 지정된 용어를 포함하는 텍스트를 찾습니다. 이는 "*cars have an inflated price tag*"를 캡처할 수 있지만 "*store offers amazing discounts*"는 무시됩니다.

이 연산자를 사용하려면, 셀을 두 번 클릭하여 요소 셀에 수동으로 느낌표와 괄호를 입력해야 합니다.

단어 간격(<Any Token>)

<Any Token>이라고도 하는 단어 간격은 두 요소 사이에 있을 수 있는 토큰의 숫자 범위를 정의합니다. 단어 간격은 추가 한정사, 위치 문구, 형용사 또는 다른 이와 같은 단어의 존재로, 약간만 다를 수 있는 매우 유사한 구문과 매치할 때 아주 유용합니다.

표 45. 단어 간격 없이 규칙 값 테이블에서 요소의 예

#	요소
1	 Unknown
2	 mBeHave

표 45. 단어 간격 없이 규칙 값 테이블에서 요소의 예 (계속)

3	 양수(Positive)

참고: 소스 보기에서 이 값은 $Unknown $mBeHave $Positive로 정의됩니다.

이 값은 "*the hotel staff was nice*"와 같은 문장을 매치합니다. 여기서 *hotel staff*는 유형 <Unknown>에 속하며, *was*는 매크로 mBeHave 아래에 있고 *nice*는 <Positive> 아래에 있습니다. 그러나 "*the hotel staff was very nice*"는 매치하지 않습니다.

표 46. <Any Token> 단어 간격의 규칙 값 테이블의 요소 예

#	요소
1	 Unknown
2	 mBeHave
3	
4	 양수(Positive)

참고: 소스 보기에서 이 값은 $Unknown $mBeHave @{0,1} $Positive로 정의됩니다.

단어 간격을 규칙 값에 추가하면, "*the hotel staff was nice*" 및 "*the hotel staff was very nice*" 둘 다에 매치됩니다.

소스 보기에서, 또는 인라인 편집 사용 시 단어 간격의 명령문은 @{#,#}입니다. 여기서 @은 단어 간격을 나타내고 {#,#}은 이전 요소와 다음 요소 사이에 승인된 최소 및 최대 단어 수를 정의합니다. 예를 들어, @{1,3}은 최소 하나의 단어가 존재하지만 세 개 이하의 단어가 두 요소 사이에 나타나는 경우 정의된 두 요소 사이에 매치가 작성될 수 있음을 의미합니다. @{0,3}은 0, 1, 2 또는 3개 단어가 존재하지만 세 개 이하의 단어가 두 요소 사이에 나타나는 경우 정의된 두 요소 사이에 매치가 작성될 수 있음을 의미합니다.

소스 모드에서 보기 및 작업

각 규칙 및 매크로에 대해, TLA 편집기는 TLA 출력 매치 및 생성을 위해 추출기에서 사용되는 기본적인 소스 코드를 생성합니다. 코드 자체에 대해 작업하려는 경우, 이 소스 코드를 보고 편집기의 맨 위에 있는 "소스 보기" 단추를 클릭하여 직접 편집할 수 있습니다. 소스 보기는 현재 선택된 규칙 또는 매크로로 점프하여 강조표시합니다. 그러나 오류 가능성을 줄이기 위해 편집기 분할창을 사용할 것을 권장합니다.

소스 보기 및 편집을 완료하면 **소스 종료**를 클릭하십시오. 규칙에 대한 유효하지 구문이 생성되면, 먼저 이 구문을 수정하고 소스 코드를 종료해야 합니다.

중요사항: 소스 보기에서 편집하는 경우, 규칙과 매크로를 한 번에 하나씩 편집하도록 합니다. 매크로를 편집한 후, 추출한 결과의 유효성을 검증하십시오. 결과에 만족하면, 다른 변경을 작성하기 전에 템플리트를 저장하도록 하십시오. 결과에 만족하지 않거나 오류가 발생하면 저장된 자원으로 되돌리십시오.

소스 보기의 매크로

```
[macro]
name = macro_name
value = ([type_name|macro_name|literal_string|word_gap])
```

표 47. 매크로 항목

[macro]	각 매크로는 매크로의 시작을 표시하기 위해 [macro] 표시가 있는 행으로 시작해야 합니다.
name	매크로 정의의 이름. 각 이름은 고유해야 합니다.
value	하나 이상의 유형, 리터럴 문자열, 단어 간격 또는 매크로의 조합. 자세한 정보는 239 페이지의 『규칙 및 매크로에 대해 지원되는 요소』 주제를 참조하십시오. 인수를 조합할 때, 소괄호 ( )를 사용하여 인수를 그룹화하고 문자 |를 사용하여 부울 OR을 표시해야 합니다.

매크로의 섹션에 수록된 지침 및 구문 외에, 편집기 보기에서 작업할 때 필요하지 않은 몇 가지의 추가 지침이 소스 보기에 있습니다. 매크로는 또한 소스 모드에서 작업할 때 다음 사항을 준수해야 합니다.

- 각 매크로는 매크로 시작을 표시하기 위해 [macro] 표시가 있는 행으로 시작해야 합니다.
- 요소를 사용하지 않도록 설정하려면 각 행 앞에 주석 표시기(#)를 입력하십시오.

예. 다음 예는 mTopic이라고 하는 매크로를 정의합니다. mTopic의 값은 <Product>, <Person>, <Location>, <Organization>, <Budget> 또는 <Unknown> 유형 중 *하나*와 매치되는 항의 존재입니다.

```
[macro]
name=mTopic
value=($Unknown|$Product|$Person|$Location|$Organization|$Budget|$Currency)
```

소스 보기의 규칙

```
[pattern(ID)]
name = pattern_name
value = [$type_name|macro_name|word_gaps|literal_strings]
output = $digit[\t]#digit[\t]$digit[\t]#digit[\t]$digit[\t]#digit[\t]
```

표 48. 규칙 항목

[pattern (<ID>)]	해당 텍스트 링크 분석의 시작을 표시하고 처리 순서를 판별하기 위해 고유한 숫자 ID 사용을 제공합니다.
name	이 텍스트 링크 분석 규칙에 대한 고유 이름을 제공합니다.
value	텍스트에 매치될 인수 및 구문을 제공합니다. 자세한 정보는 239 페이지의 『규칙 및 매크로에 대해 지원되는 요소』의 내용을 참조하십시오.

표 48. 규칙 항목 (계속)

결과	텍스트에서 발견되는, 매치 패턴 결과에 대한 출력 형식. 출력은 소스 텍스트에서 요소의 정확한 원래 위치와 항상 유사하지는 않습니다. 또한 각각의 출력을 별도의 행에 놓아서 제공된 텍스트 링크 분석 규칙에 대해 여러 출력을 수반할 수 있습니다. 출력 구문: • 출력을 탭 코드 \t로 구분하십시오(예: $1\t#1\t$3\t#3). • $ 및 숫자는 해당 위치에서 값 매개변수에 정의된 인수와 매치되는 발견된 항을 요청합니다. 따라서 $1은 값에 대해 정의된 첫 번째 인수와 매치되는 항을 의미합니다. • # 및 숫자는 해당 위치에서 요소의 유형 이름을 요청합니다. 항목이 리터럴 문자열 목록인 경우, 유형 <Unknown>이 지정됩니다. • Null\tNull 값은 어떤 출력도 작성하지 않습니다.

규칙의 섹션에 수록된 지침 및 구문 외에, 편집기 보기에서 작업할 때 필요하지 않은 몇 가지의 추가 지침이 소스 보기에 있습니다. 규칙은 또한 소스 모드에서 작업할 때 다음 사항을 준수해야 합니다.

- 두 개 이상의 요소가 정의될 때마다, 선택사항 여부에 관계없이 괄호로 묶어야 합니다(예: ($Negative|$Positive) 또는 ($mCoord|$SEP)?). $SEP는 콤마를 나타냅니다.
- 텍스트 링크 분석 규칙에서 첫 번째 요소는 선택적 요소가 될 수 없습니다. 예를 들어, value = $mTopic? 또는 value = @{0,1}로 시작할 수 없습니다.
- 양(또는 인스턴스 개수)을 토큰과 연관시킬 수 있습니다. 이는 각 케이스에 대해 별도의 규칙을 작성하는 대신 모든 케이스를 포함하는 단 하나의 규칙을 작성할 때 유용합니다. 예를 들어, ,(콤마) 또는 and와 매치하는 경우 리터럴 문자열 ($SEP|and)를 사용할 수 있습니다. 리터럴 문자열이 ($SEP|and){1,2}가 되도록 양을 추가하여 이를 확장하는 경우, 이제는 "," "and" ", and" 인스턴스와 매치됩니다.
- 공백은 텍스트 링크 분석 규칙 value의 $ 및 ? 문자와 매크로 이름 사이에서 지원되지 않습니다.
- 공백은 텍스트 링크 분석 규칙 output에서 지원되지 않습니다.
- 요소를 사용하지 않도록 설정하려면 각 행 앞에 주석 표시기(#)를 입력하십시오.

예. 자원에 다음 TLA 텍스트 링크 분석 규칙이 포함되고 TLA 결과 추출을 사용하도록 설정하였다고 가정해 보십시오.

```
## Jean Doe was the former HR director of IBM in France
[pattern(201)]
name= 1_201
value = $Person ($SEP|$mDet|$mSupport|as|then){1,2} @{0,1} $Function
 (of|with|for|in|to|at) @{0,1} $Organization @{0,2} $Location
output = $1\t#1\t$4\t#4\t$7\t#7\t$9\t#9
```

추출할 때마다, 추출 엔진은 각 문장을 읽고 다음 시퀀스와 매치하려고 합니다.

표 49. 추출 시퀀스 예

위치	인수 설명
1	사람의 이름($Person),
2	콤마($SEP), 한정사($mDet), 보조 동사($mSupport), 문자열 “then” 또는 “as” 중 하나 또는 둘.

표 49. 추출 시퀀스 예 (계속)

위치	인수 설명
3	0 또는 1개 단어(@{0,1})
4	함수($Function)
5	“of”, “with”, “for”, “in”, “to” 또는 “at”문자열 중 하나.
6	0 또는 1개 단어(@{0,1})
7	조직의 이름($Organization)
8	0, 1 또는 2개 단어(@{0,2})
9	위치의 이름($Location)

이 샘플 텍스트 링크 분석 규칙은 다음과 같이 문장 또는 구문과 매치합니다.

Jean Doe, the HR director of IBM in France

Jean Doe was the former HR director of IBM in France

IBM appointed Jean Doe as the HR director of IBM in France

이 샘플 텍스트 링크 분석 규칙은 다음 출력을 생성합니다.

```
jean doe <Person> hr director <Function> ibm <Organization> france <Location>
```

여기서,

- jean doe는 $1(텍스트 링크 분석 규칙에서 첫 번째 요소)에 대응하는 항이고 <Person>은 jean doe(#1)의 유형입니다.
- hr director는 $4(텍스트 링크 분석 규칙에서 네 번째 요소)에 해당되는 항이고 <Function>은 hr director(#4)의 유형입니다.
- ibm은 '$7(텍스트 링크 분석 규칙에서 7번째 요소)에 해당하는 항이고 <Organization>이 ibm(#7)의 유형입니다.
- france는 '$9(텍스트 링크 분석 규칙에서 9번째 요소)에 해당하는 항이고 <Location>이 france(#9)의 유형입니다.

소스 보기의 규칙 세트

```
[set(<ID>)]
```

여기서 [set (<ID>)]는 규칙 세트의 시작을 표시하고 세트의 처리 순서를 판별하기 위해 사용하는 고유 숫자 ID를 제공합니다.

예. 다음 문장에는 개인, 회사 내에서 개인의 기능, 해당 회사의 합병/인수 활동에 대한 정보가 포함됩니다.

```
Org1 Inc has entered into a definitive merger agreement with Org2 Ltd, said John Doe, CEO
of Org2 Ltd.
```

가능한 모든 출력을 처리하도록 몇 가지의 출력이 있는 하나의 규칙을 작성할 수 있습니다.

```
## 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
```

이는 다음의 두 가지 출력 패턴을 생성합니다.

- org1 inc<Organization> + merges with <ActiveVerb> + org2 ltd<Organization>
- john doe <Person> + ceo <Function> + org2 ltd<Organization>

중요! 다른 언어적 처리 조작은 TLA 패턴의 추출 동안 수행됩니다. 이러한 경우, merger은 추출 프로세스의 동의어 그룹화 단계 동안 merges with 아래에서 그룹화됩니다. 그리고 merges with는 <ActiveVerb> 유형에 속하므로, 이 유형 이름은 마지막 TLA 패턴 출력에 나타납니다. 따라서 출력이 t$3\t#3을 읽을 때, 이는 패턴이 궁극적으로 세 번째 요소에 대한 최종 개념을 표시할 것이고 모든 언어적 처리 후 세 번째 요소의 최종 유형이 적용됨을 의미합니다(동의어 및 기타 그룹).

이전과 같이 복합 규칙을 작성하는 대신, 두 개의 규칙에 대해 관리하고 작업하는 것이 더 쉬울 수 있습니다. 첫 번째는 회사 사이의 합병/인수를 찾을 때 특수화됩니다.

```
[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
```

이는 다음을 생성합니다. org1 inc<Organization> + merges with <ActiveVerb> + org2 ltd <Organization>

두 번째는 다음과 같이 개인/기능/회사에서 특수화됩니다.

```
[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
```

이는 다음을 생성합니다. which would produce john doe <Person> + ceo <Function> + org2 ltd <Organization>

주의사항

이 정보는 전 세계에 제공된 제품 및 서비스를 위해 개발되었습니다.

IBM은 다른 국가에서 이 책에 기술된 제품, 서비스 또는 기능을 제공하지 않을 수도 있습니다. 현재 사용할 수 있는 제품 및 서비스에 대한 정보는 한국 IBM 담당자에게 문의하십시오. 이 책에서 IBM 제품, 프로그램 또는 서비스를 언급했다고 해서 해당 IBM 제품, 프로그램 또는 서비스만을 사용할 수 있다는 것을 의미하지는 않습니다. IBM의 지적 재산권을 침해하지 않는 한, 기능상으로 동등한 제품, 프로그램 또는 서비스를 대신 사용할 수도 있습니다. 그러나 비IBM 제품, 프로그램 또는 서비스의 운영에 대한 평가 및 검증은 사용자의 책임입니다.

IBM은 이 책에서 다루고 있는 특정 내용에 대해 특허를 보유하고 있거나 현재 특허 출원 중일 수 있습니다. 이 책을 제공한다고 해서 특허에 대한 라이센스까지 부여하는 것은 아닙니다. 라이센스에 대한 의문사항은 다음으로 문의하십시오.

150-945
서울특별시 영등포구
국제금융로 10, 3IFC
한국 아이.비.엠 주식회사
대표전화서비스: 02-3781-7114

2바이트(DBCS) 정보에 관한 라이센스 문의는 한국 IBM에 문의하거나 다음 주소로 서면 문의하시기 바랍니다.

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502 Japan

다음 단락은 현지법과 상충하는 영국이나 기타 국가에서는 적용되지 않습니다. IBM은 타인의 권리 비침해, 상품성 및 특정 목적에의 적합성에 대한 묵시적 보증을 포함하여(단, 이에 한하지 않음) 명시적 또는 묵시적인 일체의 보증 없이 이 책을 "현상태대로" 제공합니다. 일부 국가에서는 특정 거래에서 명시적 또는 묵시적 보증의 면책사항을 허용하지 않으므로, 이 사항이 적용되지 않을 수도 있습니다.

이 정보에는 기술적으로 부정확한 내용이나 인쇄상의 오류가 있을 수 있습니다. 이 정보는 주기적으로 변경되며, 변경된 사항은 최신판에 통합됩니다. IBM은 이 책에서 설명한 제품 및/또는 프로그램을 사전 통지 없이 언제든지 개선 및/또는 변경할 수 있습니다.

이 정보에서 언급되는 비IBM의 웹 사이트는 단지 편의상 제공된 것으로, 어떤 방식으로든 이들 웹 사이트를 옹호하고자 하는 것은 아닙니다. 해당 웹 사이트의 자료는 본 IBM 제품 자료의 일부가 아니므로 해당 웹 사이트 사용으로 인한 위험은 사용자 본인이 감수해야 합니다.

IBM은 귀하의 권리를 침해하지 않는 범위 내에서 적절하다고 생각하는 방식으로 귀하가 제공한 정보를 사용하거나 배포할 수 있습니다.

(i) 독립적으로 작성된 프로그램과 기타 프로그램(본 프로그램 포함) 간의 정보 교환 및 (ii) 교환된 정보의 상호 이용을 목적으로 본 프로그램에 관한 정보를 얻고자 하는 라이센스 사용자는 다음 주소로 문의하십시오.

150-945
서울특별시 영등포구
국제금융로 10, 3IFC
한국 아이.비.엠 주식회사
대표전화서비스: 02-3781-7114

이러한 정보는 해당 조건(예를 들면, 사용료 지불 등)하에서 사용될 수 있습니다.

이 정보에 기술된 라이센스가 부여된 프로그램 및 프로그램에 대해 사용 가능한 모든 라이센스가 부여된 자료는 IBM이 IBM 기본 계약, IBM 프로그램 라이센스 계약(IPLA) 또는 이와 동등한 계약에 따라 제공한 것입니다.

본 문서에 포함된 모든 성능 데이터는 제한된 환경에서 산출된 것입니다. 따라서 다른 운영 환경에서 얻어진 결과는 상당히 다를 수 있습니다. 일부 성능은 개발 단계의 시스템에서 측정되었을 수 있으므로 이러한 측정치가 일반적으로 사용되고 있는 시스템에서도 동일하게 나타날 것이라고는 보증할 수 없습니다. 또한 일부 성능은 추정을 통해 추측되었을 수도 있으므로 실제 결과는 다를 수 있습니다. 이 책의 사용자는 해당 데이터를 본인의 특정 환경에서 검증해야 합니다.

비IBM 제품에 관한 정보는 해당 제품의 공급업체, 공개 자료 또는 기타 범용 소스로부터 얻은 것입니다. IBM에서는 이러한 제품들을 테스트하지 않았으므로, 비IBM 제품과 관련된 성능의 정확성, 호환성 또는 기타 청구에 대해서는 확신할 수 없습니다. 비IBM 제품의 성능에 대한 의문사항은 해당 제품의 공급업체에 문의하십시오.

IBM이 제시하는 방향 또는 의도에 관한 모든 언급은 특별한 통지 없이 변경될 수 있습니다.

이 정보에는 일상의 비즈니스 운영에서 사용되는 자료 및 보고서에 대한 예제가 들어 있습니다. 이들 예제에는 개념을 가능한 완벽하게 설명하기 위하여 개인, 회사, 상표 및 제품의 이름이 사용될 수 있습니다. 이들 이름은 모두 가공의 것이며 실제 기업의 이름 및 주소와 유사하더라도 이는 전적으로 우연입니다.

이 정보를 소프트카피로 확인하는 경우에는 사진과 컬러 삽화가 제대로 나타나지 않을 수도 있습니다.

상표

IBM, IBM 로고 및 ibm.com은 전세계 여러 국가에 등록된 International Business Machines Corp.의 상표 또는 등록상표입니다. 기타 제품 및 서비스 이름은 IBM 또는 타사의 상표입니다. 현재 IBM 상표 목록은 웹 (www.ibm.com/legal/copytrade.shtml)의 "저작권 및 상표 정보"에 있습니다.

Intel, Intel 로고, Intel Inside, Intel Inside 로고, Intel Centrino, Intel Centrino 로고, Celeron, Intel Xeon, Intel SpeedStep, Itanium 및 Pentium은 미국 또는 기타 국가에서 사용되는 Intel Corporation 또는 그 자회사의 상표 또는 등록상표입니다.

Linux는 미국 또는 기타 국가에서 사용되는 Linus Torvalds의 등록상표입니다.

Microsoft, Windows, Windows NT 및 Windows 로고는 미국 또는 기타 국가에서 사용되는 Microsoft Corporation의 상표입니다.

UNIX는 미국 및 기타 국가에서 사용되는 The Open Group의 등록상표입니다.

Java 및 모든 Java 기반 상표와 로고는 Oracle 및/또는 그 계열사의 상표 또는 등록상표입니다.

기타 제품 및 서비스 이름은 IBM 또는 타사의 상표입니다.

색인

[가]

[나]

[사]

[아]

[자]

[차]

[카]

[타]

[파]

A

B

F

H

I

L

M

N

O

P

T

U

W

[특수 문자]