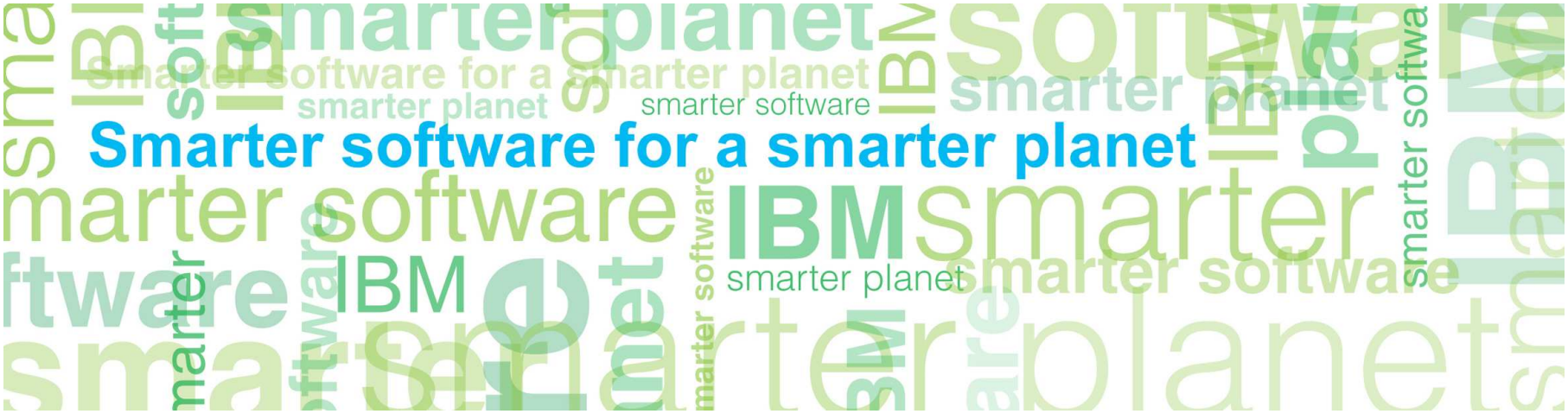
LanguageWare Resource Workbench 7.2

Analyze a collection of documents



Introduction

- **Module overview**
  - Analyzing collections
  - Saving and comparing annotations
  - Best practices
- **Target audience:**
  - All audiences
- **Prerequisites:**
  - Install LanguageWare® resource workbench
  - Create a LanguageWare project
  - Create a UIMA annotator
  - Create an IBM content analytics server connection file
- **Version release date:** LRW 7.2, ICA 2.2, released October, 2010

Module objectives

After this module you will be able to:

- Analyze a collection of documents using an annotator defined on the LRW
- Analyze a collection of documents using an annotator defined on IBM Content Analytics
- Save the annotations from a collection of documents
- Interpret how changes made to an annotator have affected the annotations generated by it. This is done by comparing annotations generated by the annotator with the saved annotations generated by the annotator before the change.

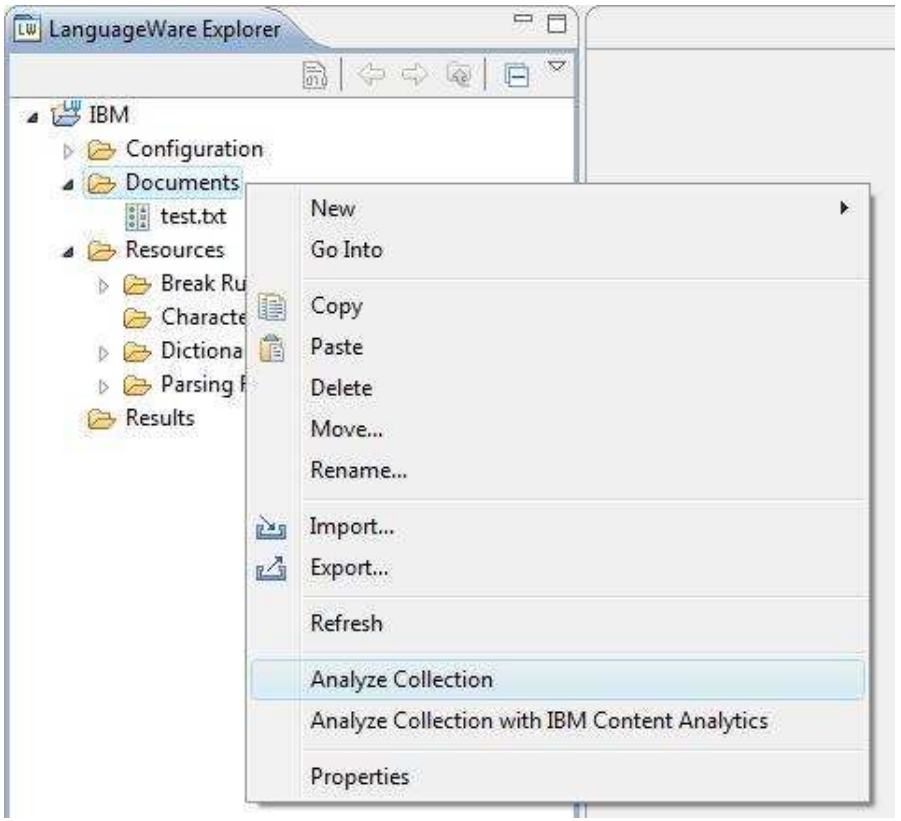
Module roadmap

- **Collections of documents**
  - Analyzing using the LRW
  - Analyzing using IBM Content Analytics
  - Saving annotations from a collection of documents
  - Comparing annotations generated by one annotator with the saved annotations generated by a previous annotator.
- **Summary and best practices**
- **Sample exercises**

Collections of documents

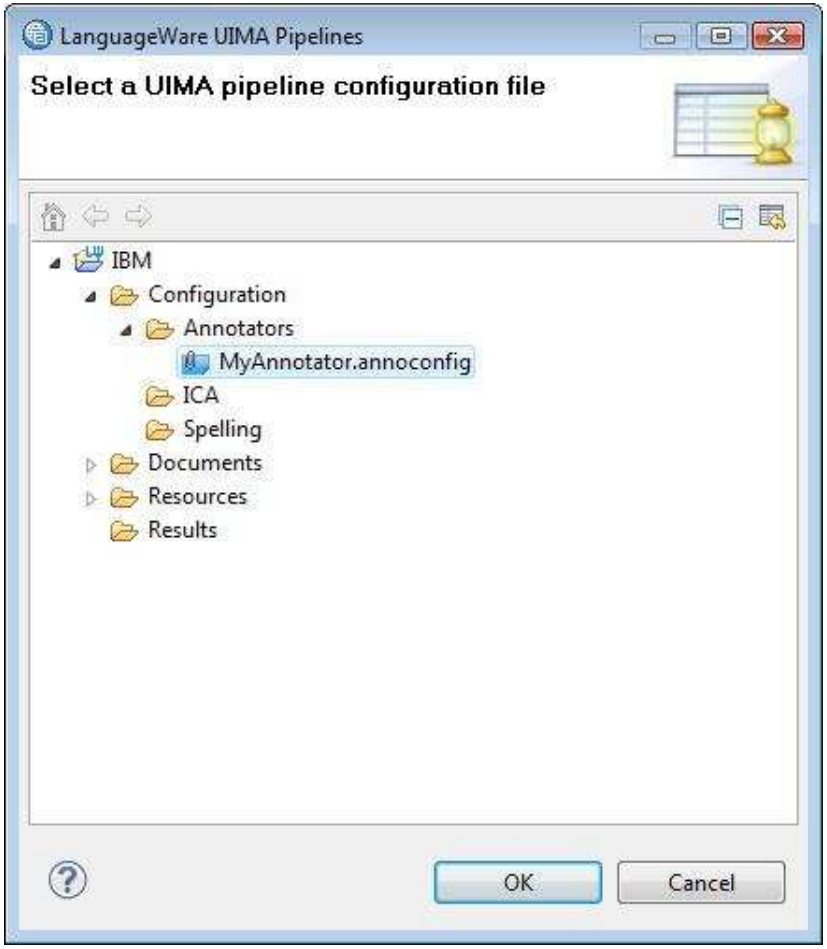
Analyzing using UIMA pipeline on the LRW

- Select one or more documents in the LanguageWare Explorer, or select a folder containing one or more documents
- Right click the selected documents, and select **Analyze Collection**.



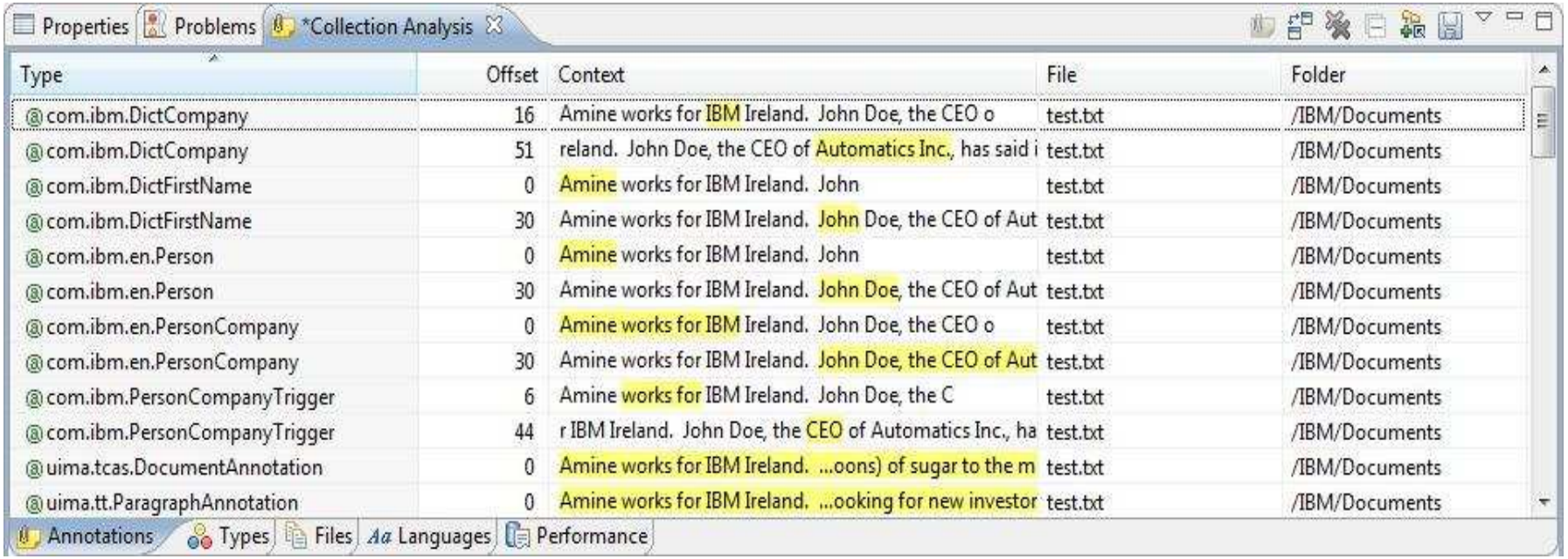
Collections of documents
Analyzing using UIMA Pipeline on the LRW

- Select the UIMA pipeline configuration file to use to analyze the documents
- Click **OK**



Collections of documents
Analysis results

- The **Collection Analysis** view is displayed showing the annotations created in every document analyzed.
- Data can be sorted by clicking column headings.
- In addition to the list of annotations, the tabs at the bottom of the view give access to various statistics on the document annotations.

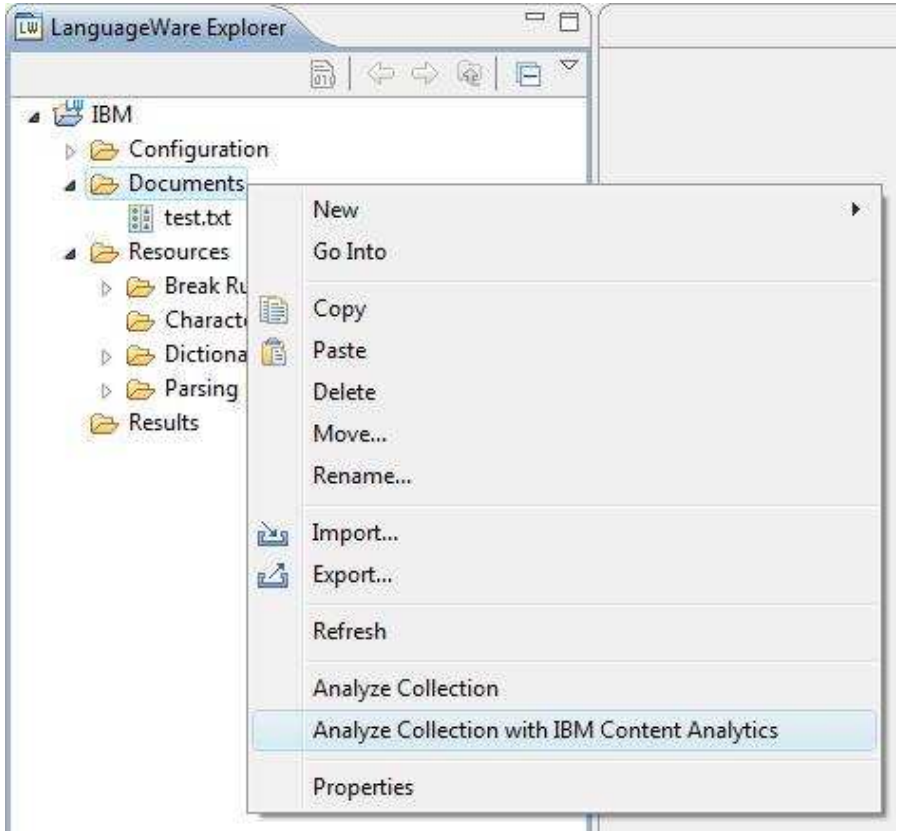


Type	Offset	Context	File	Folder
@com.ibm.DictCompany	16	Amine works for IBM Ireland. John Doe, the CEO o	test.txt	/IBM/Documents
@com.ibm.DictCompany	51	reland. John Doe, the CEO of Automatics Inc., has said i	test.txt	/IBM/Documents
@com.ibm.DictFirstName	0	Amine works for IBM Ireland. John	test.txt	/IBM/Documents
@com.ibm.DictFirstName	30	Amine works for IBM Ireland. John Doe, the CEO of Aut	test.txt	/IBM/Documents
@com.ibm.en.Person	0	Amine works for IBM Ireland. John	test.txt	/IBM/Documents
@com.ibm.en.Person	30	Amine works for IBM Ireland. John Doe, the CEO of Aut	test.txt	/IBM/Documents
@com.ibm.en.PersonCompany	0	Amine works for IBM Ireland. John Doe, the CEO o	test.txt	/IBM/Documents
@com.ibm.en.PersonCompany	30	Amine works for IBM Ireland. John Doe, the CEO of Aut	test.txt	/IBM/Documents
@com.ibm.PersonCompanyTrigger	6	Amine works for IBM Ireland. John Doe, the C	test.txt	/IBM/Documents
@com.ibm.PersonCompanyTrigger	44	r IBM Ireland. John Doe, the CEO of Automatics Inc., ha	test.txt	/IBM/Documents
@uima.tcas.DocumentAnnotation	0	Amine works for IBM Ireland. ...oons) of sugar to the m	test.txt	/IBM/Documents
@uima.tt.ParagraphAnnotation	0	Amine works for IBM Ireland. ...ooking for new investor	test.txt	/IBM/Documents

Collections of documents

Analyzing using UIMA pipeline on IBM Content Analytics

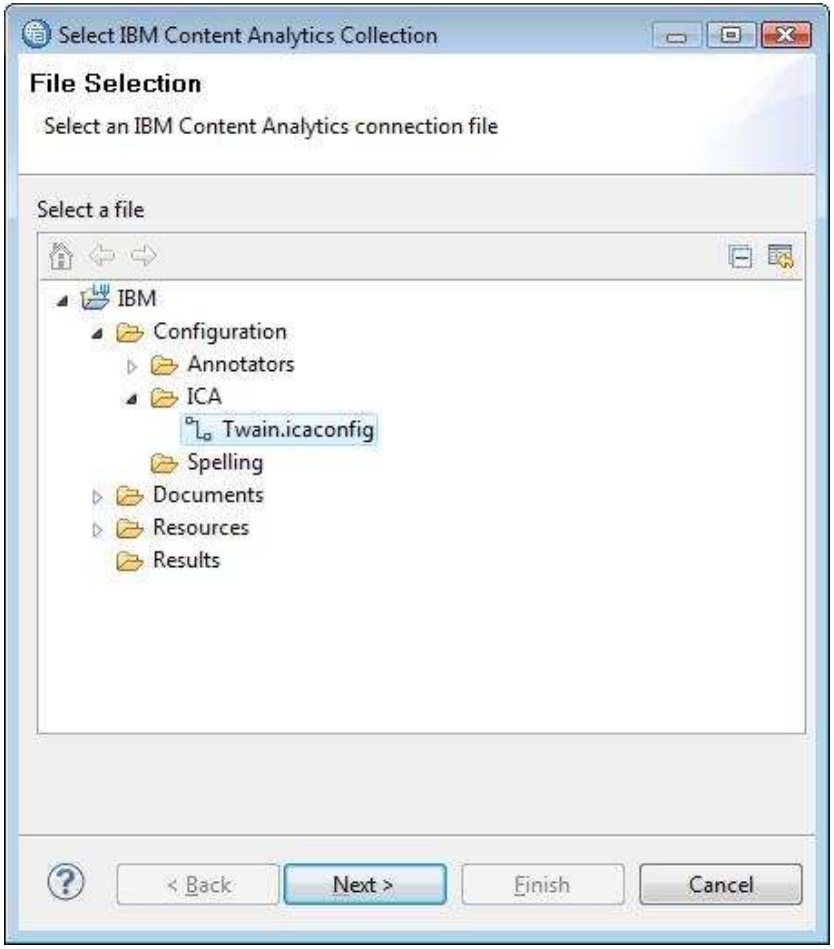
- Select one or more documents in the LanguageWare Explorer, or select a folder containing one or more documents
- Right click the selected documents, and select **Analyze Collection with IBM Content Analytics**.



Collections of documents

Analyzing using UIMA pipeline on IBM Content Analytics

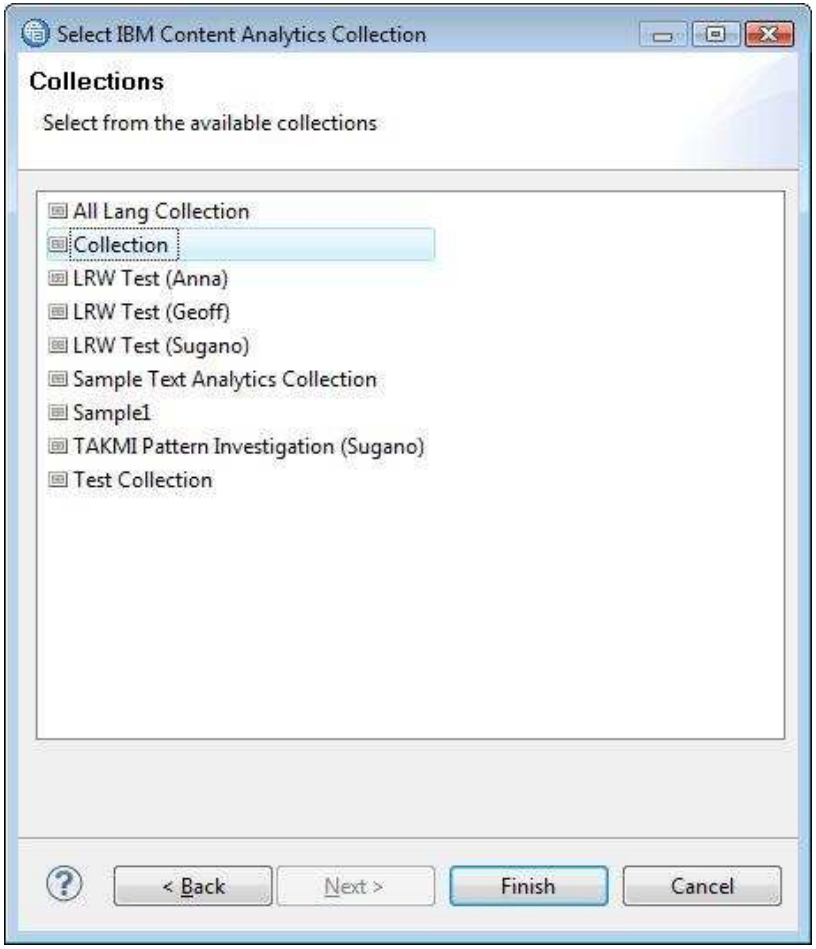
- Select the IBM Content Analytics Server Configuration file which defines the server on which the documents will be analyzed.
- Click **Next**



Collections of documents

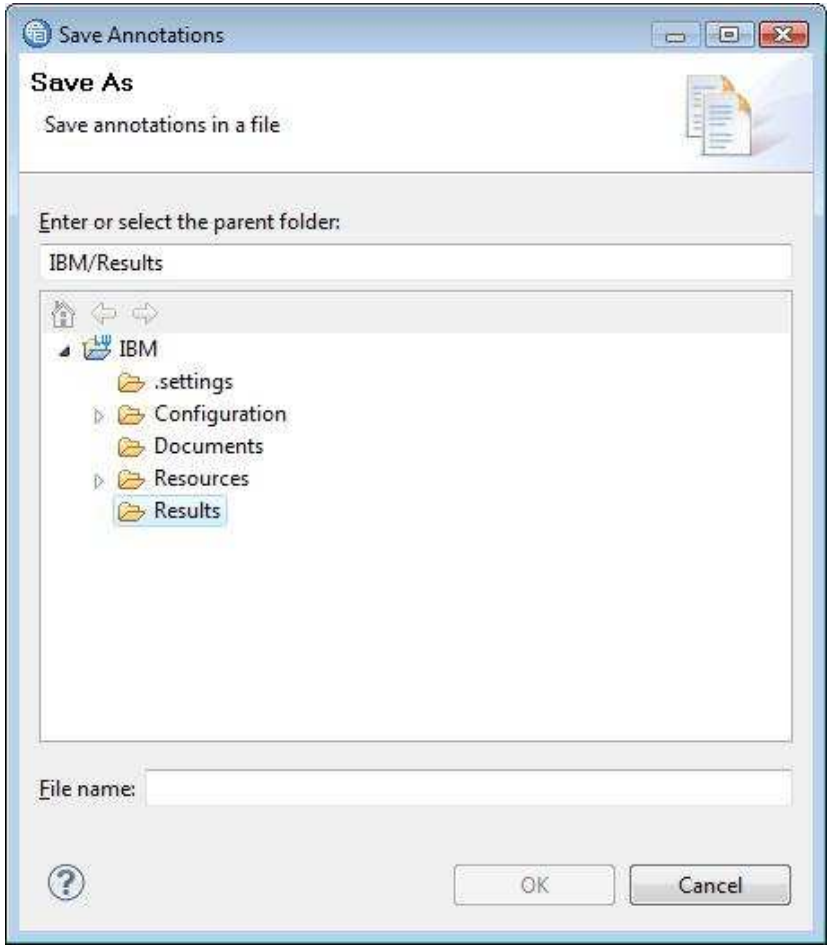
Analyzing using UIMA pipeline on IBM Content Analytics

- Select the collection on the IBM Content Analytics server on which the documents will be analyzed.
- Click **Finish**
- The documents will be sent to the IBM Content Analytics server for analysis, and the results returned to the LRW. These results will be displayed in exactly the same way as those annotated on the LRW.



Collections of documents
Saving annotations

- To save the annotations generated by annotating a collection of documents, click the Save button in the top right corner of the Collection Analysis view. This will display the Save Annotations dialog.
- Select the **Results** folder in your project as the location the annotations should be saved.
- Specify a meaningful name for the saved annotations (for example PersonAnnotator_20-01-2011).
- The extension .annotation will be added automatically to the file name.

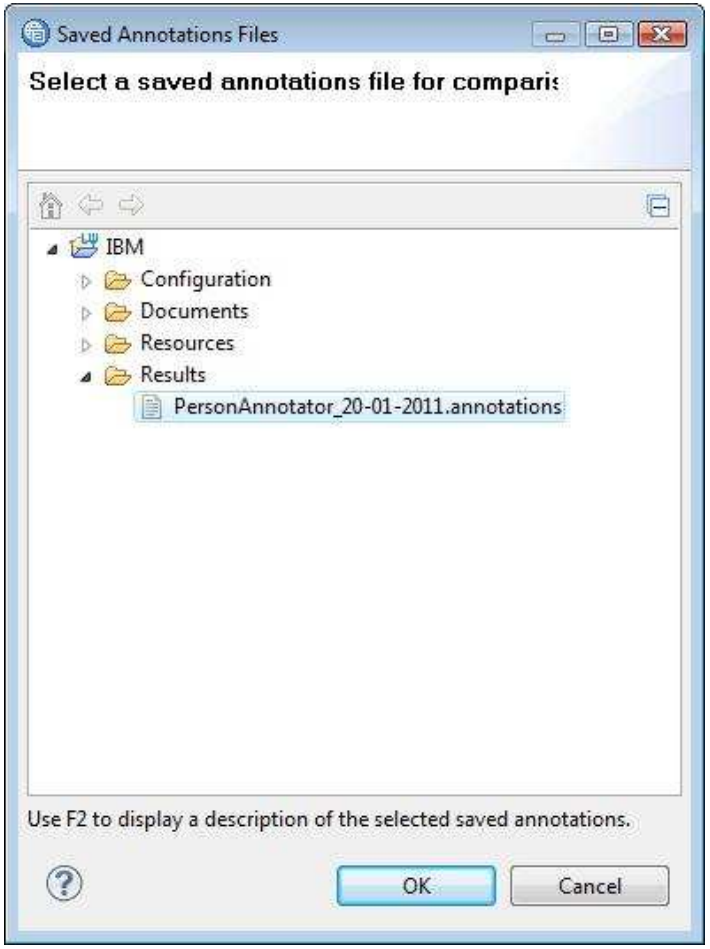


Collections of documents
Comparing annotations

- While developing a UIMA pipeline and its resources (dictionaries, and rules) it is useful to visualize how the changes you have made have affected the annotations generated by the pipeline.
- Comparing the annotations generated by a pipeline at various stages of its development will not only show the improvements that have been made to the quality of the pipeline. It may also show up unexpected and unwanted side effects of those changes.
- It is therefore good practice to regularly compare the results of your pipeline to ensure continued improvement in precision and recall.

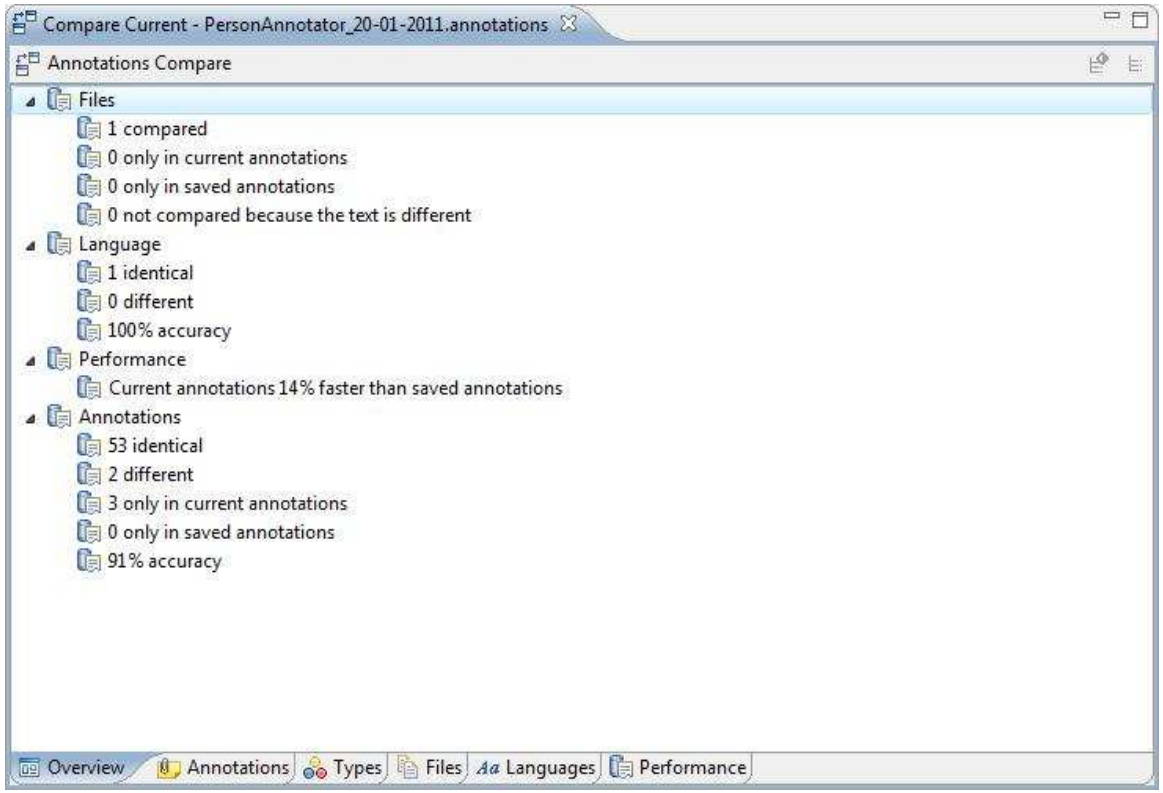
Collections of documents
Comparing annotations

- To compare annotations:
  - Annotate a collection with your UIMA Pipeline
  - Click the **Compare annotations** icon in the top right of the Collection Analysis view. This displays the Saved Annotations File dialog.
  - Select the file containing the annotations you saved earlier from a previous version of the UIMA Pipeline.
  - Click OK.



Collections of documents
Comparison overview

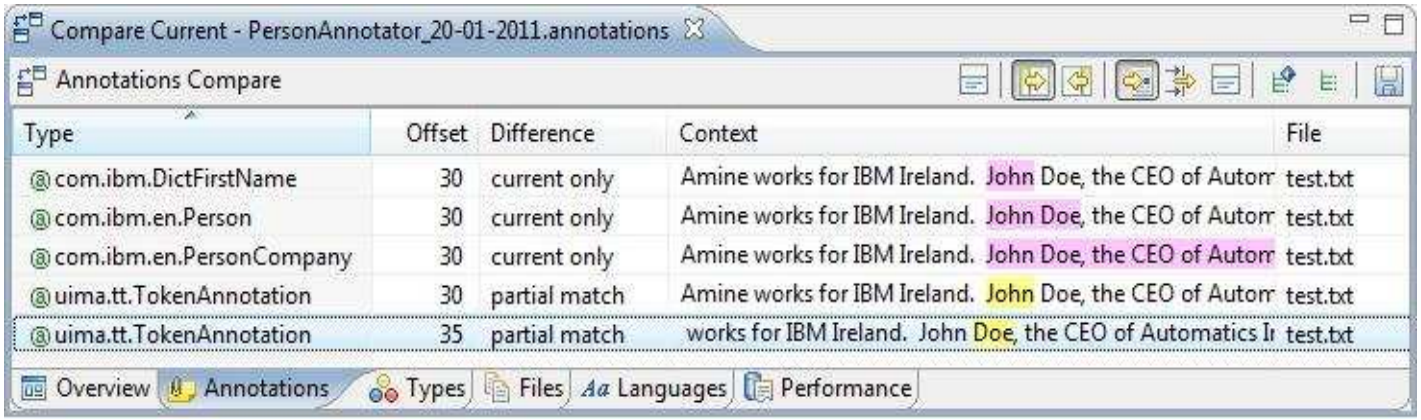
- The Comparison viewer initially displays an overview of the results
- Notice in the annotations there are two differences and three new annotations
- Switch to the Annotations tab to see more detail.



Collections of documents
Annotation differences

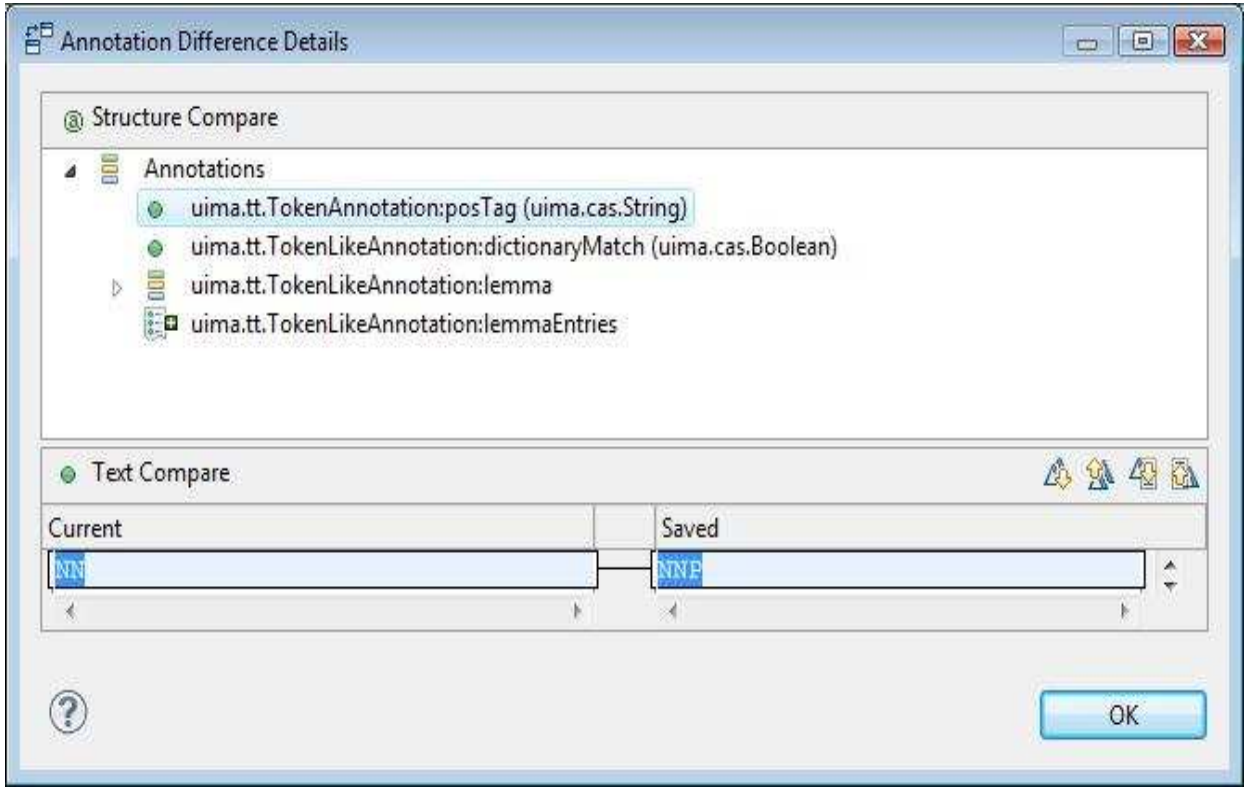
The columns in this view have the following meaning:

- **Type** – The UIMA Type that is different between the two sets of annotations, or exists in only one of the sets
- **Offset** – the offset (the position of the first character) of the annotation into the file
- **Difference** – The type of difference found:
  - Current only – the annotation is only in the current set of annotations
  - Saved only - the annotation is only in the saved set of annotations
  - Partial match – Both sets of annotation have an annotation over this span of text, but the value of one or more features is different.
- **Context** – text covered by the annotation is highlighted, and shown with its surrounding text.



Collections of documents
Annotation difference detail

- Right click any annotation marked as **partial match**, and select **Show Details** . This displays the Annotation Difference Details dialog, showing which features are different between the two sets of annotations.



Module roadmap

- **Collections of documents**
  - Analyzing using the LRW
  - Analyzing using IBM Content Analytics
  - Saving annotations from a collection of documents
  - Comparing annotations generated by one annotator with the saved annotations generated by a previous annotator.
- **Summary and best practices**
- **Sample exercises**

Module summary

You have completed this module and can:

- Analyze a collection of documents on the LRW
- Analyze a collection of documents on IBM Content Analytics
- Save the results of collection analysis
- Compare two sets of collection analysis results

See the LanguageWare help for more tips and advanced use cases.

Best practices

- Annotating an entire collection of documents can be useful to provide a better feeling for the precision and recall of an annotator.
- Annotations can be done either using a local UIMA pipeline, or a pipeline associated with an IBM Content Analytics collection.
- Annotations can be saved. These can later be used to compare against the annotations generated by an updated UIMA pipeline to help visualize the improvements in precision and recall that have been achieved by the recent changes to the pipeline.
- It is good practice to regularly compare annotations with previous results in order to identify any unexpected side effects of any changes that have been made to the pipeline.

Module roadmap

- **Collections of documents**
  - Analyzing using the LRW
  - Analyzing using IBM Content Analytics
  - Saving annotations from a collection of documents
  - Comparing annotations generated by one annotator with the saved annotations generated by a previous annotator.
- **Summary and best practices**
- **Sample exercises**

Practice exercises

- Analyze all documents in the Documents folder of the "ConfectionaryCompanyHelpline" project with the "AnalyseHelpline" annotator.
- Analyze the same documents on the "Sample Text Analytics Collection" on your IBM Content Analytics server.

Contacts

- If you have any questions, comments or suggestions, contact us using the LanguageWare email address *EMEALAN@ie.ibm.com* or on the developerWorks® forum.

Trademarks, copyrights, and disclaimers

IBM, the IBM logo, ibm.com, developerWorks, and LanguageWare are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of other IBM trademarks is available on the web at "Copyright and trademark information" at http://www.ibm.com/legal/copytrade.shtml

Other company, product, or service names may be trademarks or service marks of others.

THE INFORMATION CONTAINED IN THIS PRESENTATION IS PROVIDED FOR INFORMATIONAL PURPOSES ONLY. WHILE EFFORTS WERE MADE TO VERIFY THE COMPLETENESS AND ACCURACY OF THE INFORMATION CONTAINED IN THIS PRESENTATION, IT IS PROVIDED "AS IS" WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED. IN ADDITION, THIS INFORMATION IS BASED ON IBM'S CURRENT PRODUCT PLANS AND STRATEGY, WHICH ARE SUBJECT TO CHANGE BY IBM WITHOUT NOTICE. IBM SHALL NOT BE RESPONSIBLE FOR ANY DAMAGES ARISING OUT OF THE USE OF, OR OTHERWISE RELATED TO, THIS PRESENTATION OR ANY OTHER DOCUMENTATION. NOTHING CONTAINED IN THIS PRESENTATION IS INTENDED TO, NOR SHALL HAVE THE EFFECT OF, CREATING ANY WARRANTIES OR REPRESENTATIONS FROM IBM (OR ITS SUPPLIERS OR LICENSORS), OR ALTERING THE TERMS AND CONDITIONS OF ANY AGREEMENT OR LICENSE GOVERNING THE USE OF IBM PRODUCTS OR SOFTWARE.