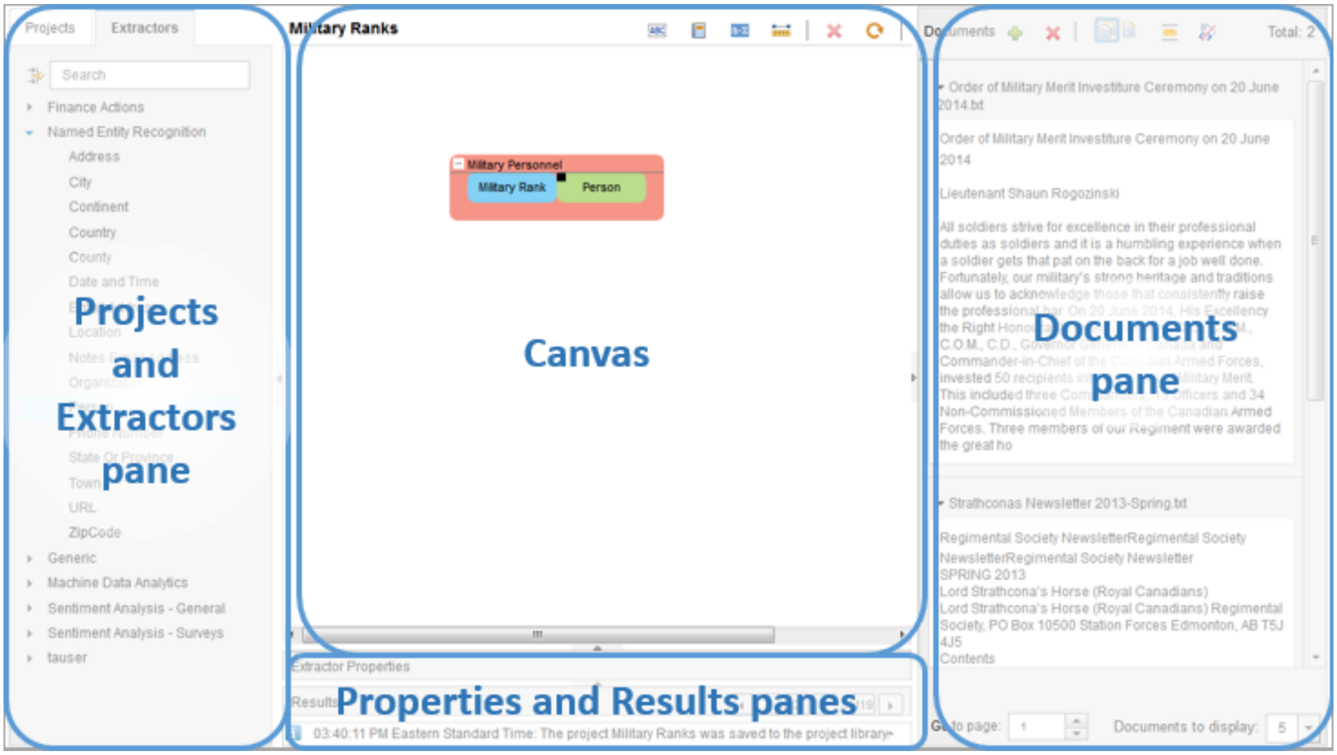
SystemT Advanced Rule Editor Lab

Part 1: Creating and Editing Extractors Visually

UI Components & Visual Constructs

In this lab you will explore learn about the core functionality of the SystemT Advanced Rule Editor.

These are the core UI components:



Order of Military Merit Investiture Ceremony on 20 June 2014.txt

Order of Military Merit Investiture Ceremony on 20 June 2014

Lieutenant Shaun Rogozinski

All soldiers strive for excellence in their professional duties as soldiers and it is a humbling experience when a soldier gets that pat on the back for a job well done. Fortunately, our military's strong heritage and traditions allow us to acknowledge those that consistently raise the professional bar. On 20 June 2014, His Excellency the Right Hono[illegible] M., C.O.M., C.D., Governor General of Canada and Commander-in-Chief of the Canadian Armed Forces, invested 50 recipients in[illegible] Military Merit. This included three Com[illegible] Officers and 34 Non-Commissioned Members of the Canadian Armed Forces. Three members of our Regiment were awarded the great ho

Strathconas Newsletter 2013-Spring.txt

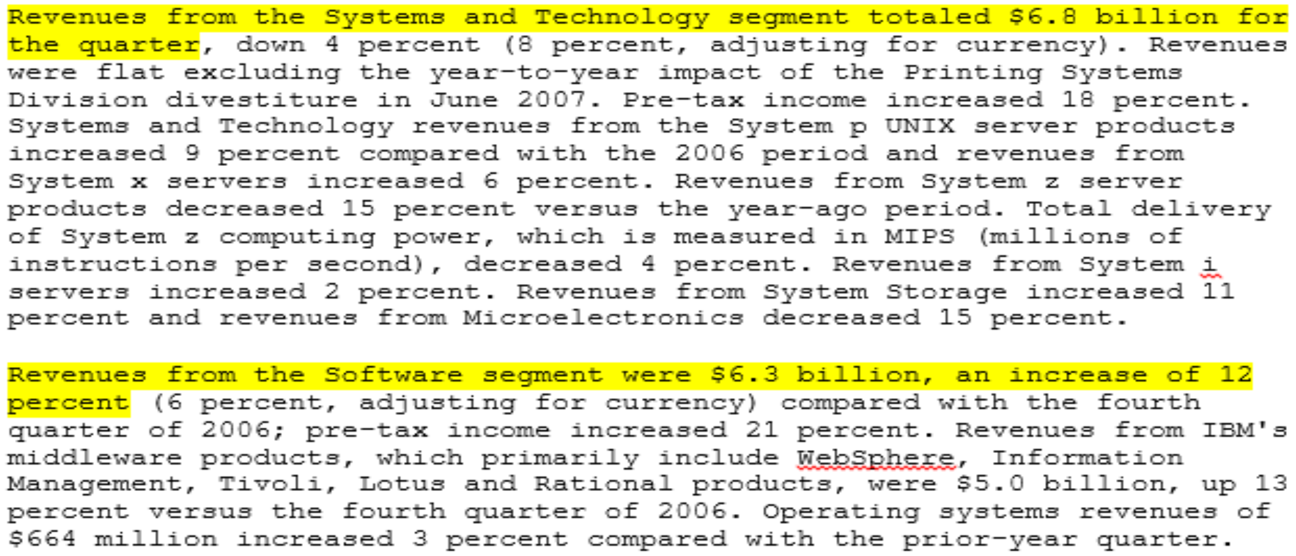
Regimental Society NewsletterRegimental Society NewsletterRegimental Society Newsletter
SPRING 2013
Lord Strathcona's Horse (Royal Canadians)
Lord Strathcona's Horse (Royal Canadians) Regimental Society, PO Box 10500 Station Forces Edmonton, AB T5J 4J5
Contents

03:40:11 PM Eastern Standard Time: The project Military Ranks was saved to the project library

At the center is your canvas, or workspace. To the right is the documents pane, showing the documents that you are analyzing. The bottom pane displays extractor output and properties that can be modified. To the left is the projects and extractors pane, contains available pre-built extractors as well as the extractors that you have developed and saved.

Lab Scenario 1: Analyzing company earnings reports

The goal of this lab is to have you creating your own extractors in a scenario which has been found to be useful for real application. In this scenario, a data analyst is trying to extract the revenue by division from a company's financial reports. The highlighted text below illustrates sample text of interest.



Revenues from the Systems and Technology segment totaled $6.8 billion for the quarter, down 4 percent (8 percent, adjusting for currency). Revenues were flat excluding the year-to-year impact of the Printing Systems Division divestiture in June 2007. Pre-tax income increased 18 percent. Systems and Technology revenues from the System p UNIX server products increased 9 percent compared with the 2006 period and revenues from System x servers increased 6 percent. Revenues from System z server products decreased 15 percent versus the year-ago period. Total delivery of System z computing power, which is measured in MIPS (millions of instructions per second), decreased 4 percent. Revenues from System i servers increased 2 percent. Revenues from System Storage increased 11 percent and revenues from Microelectronics decreased 15 percent.

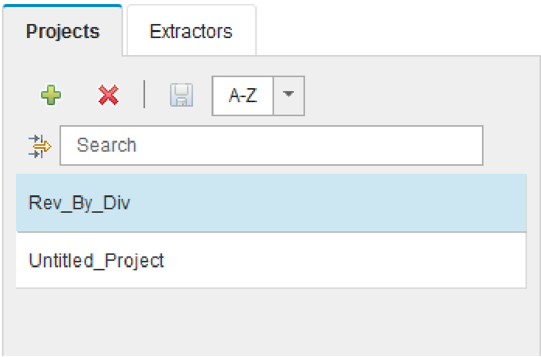
Revenues from the Software segment were $6.3 billion, an increase of 12 percent (6 percent, adjusting for currency) compared with the fourth quarter of 2006; pre-tax income increased 21 percent. Revenues from IBM's middleware products, which primarily include WebSphere, Information Management, Tivoli, Lotus and Rational products, were $5.0 billion, up 13 percent versus the fourth quarter of 2006. Operating systems revenues of $664 million increased 3 percent compared with the prior-year quarter.

Step 0: Launch the Advanced Rule Editor from your WKS instance on IBM cloud

If you have not yet provisioned a WKS instance on IBM cloud, follow the instructions in Appendix E (AppendixE-Provisioning-WKS-ARE.pdf) to get to your Advanced Rule Editor workspace.

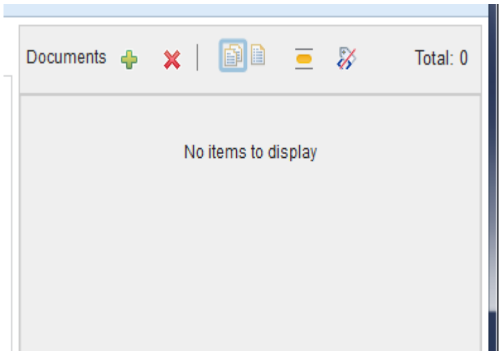
Step 1: Create a new project

We will begin by creating a new project. To do this, click the button ✚ in the projects pane and name the project **Rev_By_Div**.

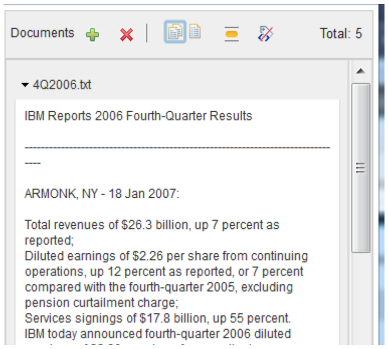


Step 2: Upload your documents.

Now you will import the documents for this project. Press the ✚ button in the documents pane on the right side of the screen, browse to the **financialStatements** folder, select all the .txt documents and confirm by clicking ‘Add’. Since we are importing text documents, we can use the default settings for importing.



No items to display



4Q2006.txt

IBM Reports 2006 Fourth-Quarter Results

ARMONK, NY - 18 Jan 2007:

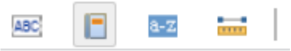
Total revenues of $26.3 billion, up 7 percent as reported;
Diluted earnings of $2.26 per share from continuing operations, up 12 percent as reported, or 7 percent compared with the fourth-quarter 2005, excluding pension curtailment charge;
Services signings of $17.8 billion, up 55 percent.
IBM today announced fourth-quarter 2006 diluted

From the example snippets earlier, we can see that in order to identify mention of revenue by division, we need to identify mentions of revenue and division first.

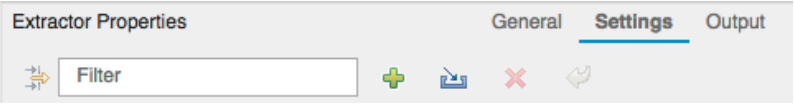
Step 3: Create an Extractor for Capturing Division

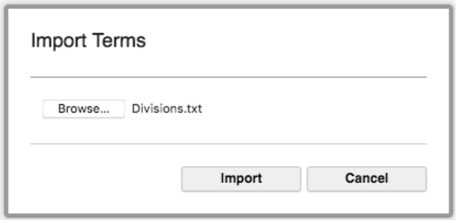
1. Create a dictionary titled *Division*.

   First click on the New Dictionary button present in the canvas toolbar. When the Dictionary Appears on the canvas simply type in the title “Division” and press Enter.



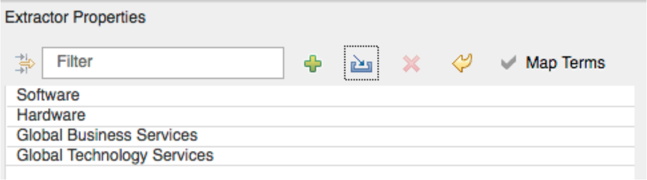
2. Import the terms for the division from a .txt file. Click the new blue “Division” object on your canvas, then Click the import button in the dictionary settings pane below the canvas. Browse to the file Divisions.txt and click ‘Import.





Divisions.txt

3. You should see a dictionary with four terms:



Software
Hardware
Global Business Services
Global Technology Services

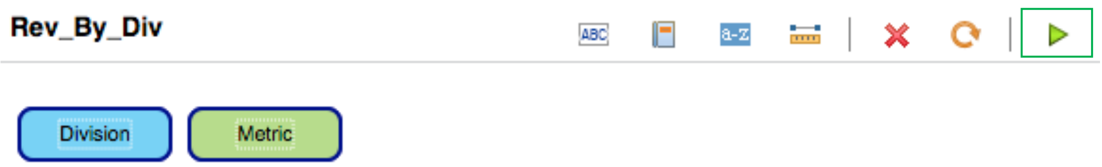
Step 4: Create an Extractor for Capturing Revenue

Similar to Step 3, create a Dictionary extractor called *Metric* on your own.

Add the term “Revenues” to ***Metric*** using the ➕ button on the dictionary settings pane.

Step 5: Run the Extractors Created in Step 3 and 4

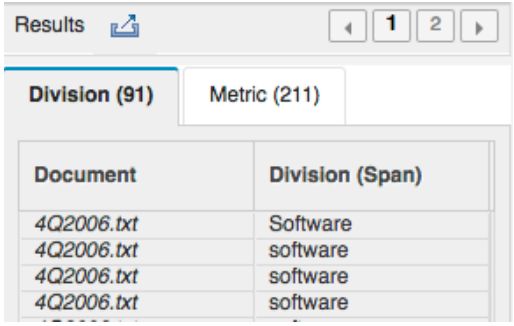
Drag to multi-select the two extractors and hit the ▶ button.



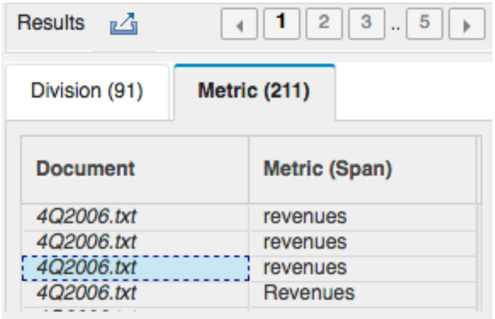
Matches found for the two dictionaries should be highlighted in the document pane and listed in the results grid.

Hint: View Individual Results

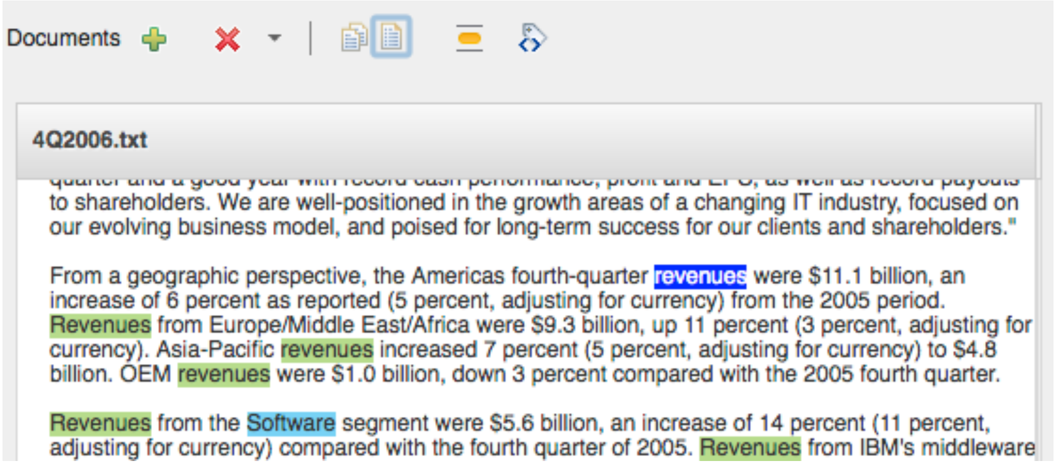
Click on any non-span attribute in a row of the result grid to see all the span attributes of the result highlighted in the document pane. Click on any span attribute in the row to see that span highlighted in the document pane.



Document	Division (Span)
4Q2006.txt	Software
4Q2006.txt	software
4Q2006.txt	software
4Q2006.txt	software



Document	Metric (Span)
4Q2006.txt	revenues
4Q2006.txt	revenues
4Q2006.txt	revenues
4Q2006.txt	Revenues



quarter and a good year with record cash performance, profit and EPS, as well as record payouts to shareholders. We are well-positioned in the growth areas of a changing IT industry, focused on our evolving business model, and poised for long-term success for our clients and shareholders."

From a geographic perspective, the Americas fourth-quarter revenues were $11.1 billion, an increase of 6 percent as reported (5 percent, adjusting for currency) from the 2005 period. Revenues from Europe/Middle East/Africa were $9.3 billion, up 11 percent (3 percent, adjusting for currency). Asia-Pacific revenues increased 7 percent (5 percent, adjusting for currency) to $4.8 billion. OEM revenues were $1.0 billion, down 3 percent compared with the 2005 fourth quarter.

Revenues from the Software segment were $5.6 billion, an increase of 14 percent (11 percent, adjusting for currency) compared with the fourth quarter of 2005. Revenues from IBM's middleware

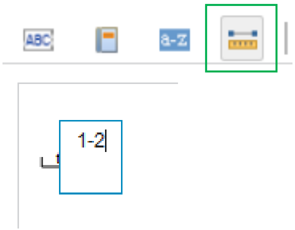
Step 6: Create a Pattern

We will now create a pattern to capture instances such as “Software segment revenue”

1. Create a proximity rule

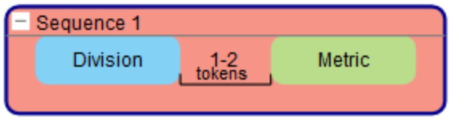
   Click on the *New Proximity rule* button present in the canvas toolbar. Enter the value “1-2” and hit Enter. Proximity Nodes allow for a number of Tokens to be allowed in a sequence between two matches.

   Tokens generally refer to words separated by spaces or other non-word characters (e.g. punctuation marks).



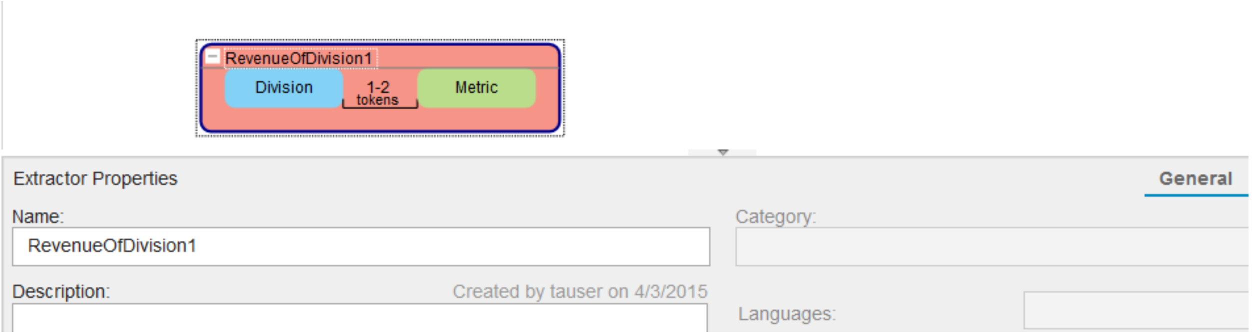
2. Create a Sequence

Create the following sequence by dragging the proximity after Division and before Metric.



3. Rename the Sequence

Rename the sequence to ‘RevenueOfDivision1’ by double-clicking on its title or by editing the “Name” field in the “General” tab, which contains Extractor Properties.



Extractor Properties

General

Name:

RevenueOfDivision1

Category:

Description:

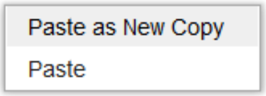
Created by tauser on 4/3/2015

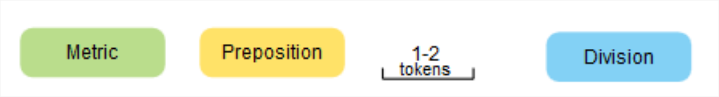
Languages:

Hint: How to Copy an Extractor

In the next step, you will want to copy and paste some extractors. Copying can be done by right-clicking a node and choosing copy, or by clicking on a node and pressing Ctrl-C (or ⌘-C). There are two ways to paste:
1) **Paste as New Copy**; 2) **Paste**
- To **Paste as New Copy**, right-click on the canvas and choose **Paste as New Copy** or press Ctrl-Shift-V. This will make a new copy of the node, which will function independently of the original node.
- To **Paste**, right-click on the canvas and choose **Paste** or press Ctrl-V. **Paste** will create a new node which points to the original node and is thus a linked copy. This is similar to the concepts of alias or symbolic link: all

the linked nodes refer to the same extractor. So adding a term to a linked dictionary will update all of the linked dictionaries with the new term.

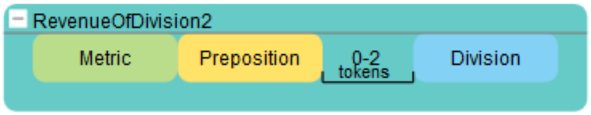




Step 7: Create a Second Sequence

We will now create a second sequence name ***RevenueOfDivision2*** to capture instances such as "Revenues from the Systems and Technology Group (S&TG)".

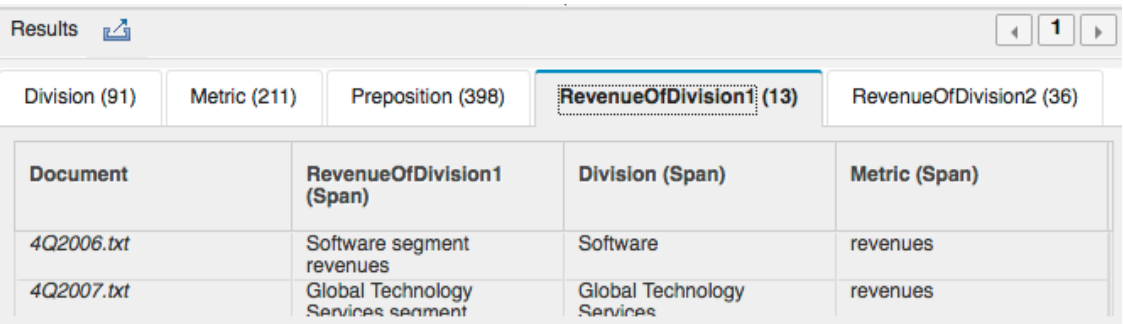
1. Copy ***Metric*** and ***Division*** from the first sequence to the canvas.
2. **Paste** (not **Paste as New Copy**) the sequence onto the canvas
3. Rename the pasted sequence "RevenueOfDivision2"
4. Create a dictionary named ***Preposition*** that contains the terms "from" and "for", and a 0-2 token proximity rule as shown below.
5. Finally, construct the following sequence, using similar steps as those described in Step 6:



This should be what your new sequence looks like.

Step 8: Run the two Sequences

Select both sequences on the canvas and run them. Check the matches using the Results pane:



Division (91) | Metric (211) | Preposition (398) | RevenueOfDivision1 (13) | RevenueOfDivision2 (36)

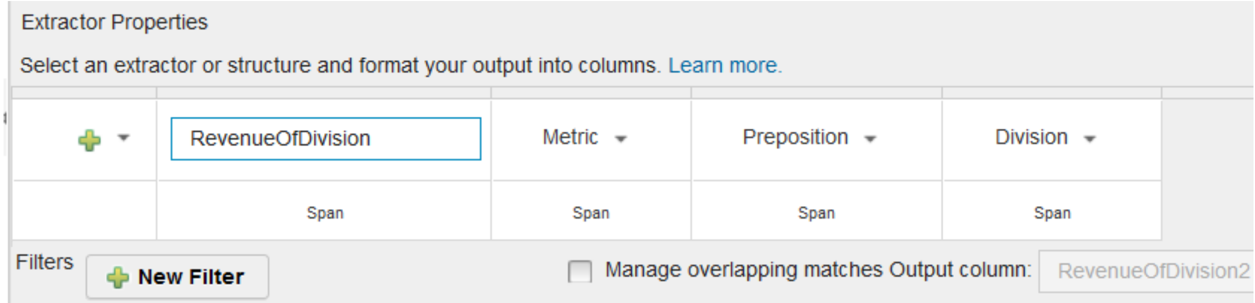
Document	RevenueOfDivision1 (Span)	Division (Span)	Metric (Span)
4Q2006.txt	Software segment revenues	Software	revenues
4Q2007.txt	Global Technology Services segment	Global Technology Services	revenues

Step 9: Union the two sequences

1. Rename the output columns of the two Sequences

To union the two sequences, we first need to make sure that their output columns are identical.

Click on **RevenueOfDivision1** and the select the Output tab in the properties pane. Rename the first output column to RevenueOfDivision by double-clicking on the column title and type in the new name.



Extractor Properties

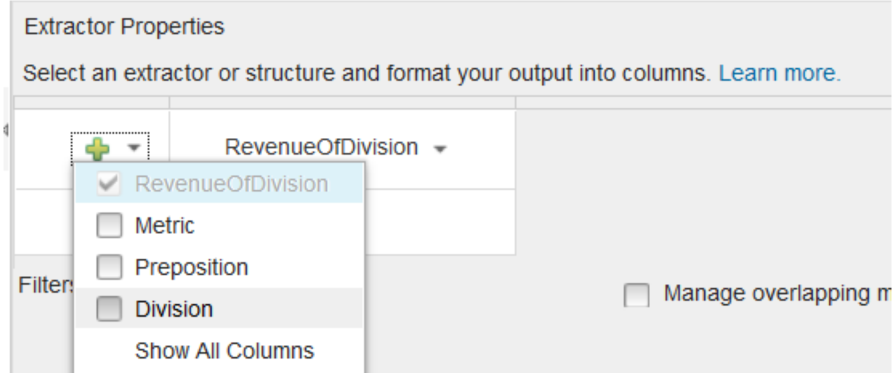
Select an extractor or structure and format your output into columns. Learn more.

➕ ▾	RevenueOfDivision	Metric ▾	Preposition ▾	Division ▾
	Span	Span	Span	Span

Filters ➕ New Filter

□ Manage overlapping matches Output column: RevenueOfDivision2

Then hide all other output columns by clicking on ➕ ▾ and choosing Hide All Columns. The only output column that should remain is RevenueOfDivision.



Extractor Properties

Select an extractor or structure and format your output into columns. Learn more.

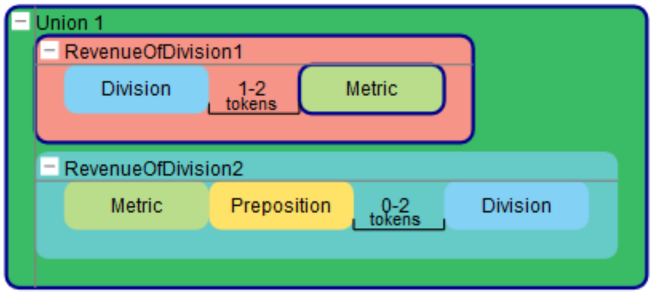
Repeat this step for **RevenueOfDivision2**. Click on **RevenueOfDivision2** and the select the Output tab in the properties pane. Rename the first output column to RevenueOfDivision. Then hide all other output columns.

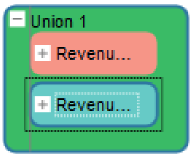
Hint: Union of Sequences

A union of sequences includes matches from all sequences in the union. The sequences in a union must have the same output schema (i.e., output column names and number of output columns). To create a union, simply drag one sequence under another.

2. Create a Union

Drag **RevenueOfDivision2** under **RevenueOfDivision1** to union the two sequences.



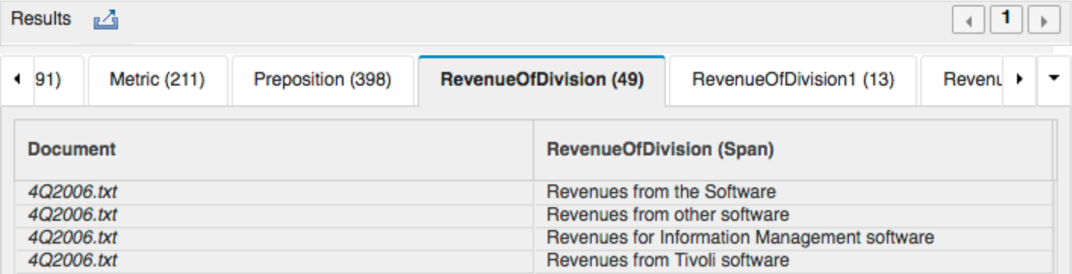


3. Rename the Union

Rename the Union to ***RevenueOfDivision*** the same way you did for the sequences.

Step 10: Run the Union

Run **RevenueOfDivision** and you will see the union combines the results of both sequences.



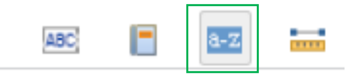
Document	RevenueOfDivision (Span)
4Q2006.txt	Revenues from the Software
4Q2006.txt	Revenues from other software
4Q2006.txt	Revenues for Information Management software
4Q2006.txt	Revenues from Tivoli software

Step 11: Create a Pattern to Capture Monetary Amount of Revenue for Each Division

You are now ready to create a pattern to capture cases such as "Revenues from the Software segment were $6.3 billion"

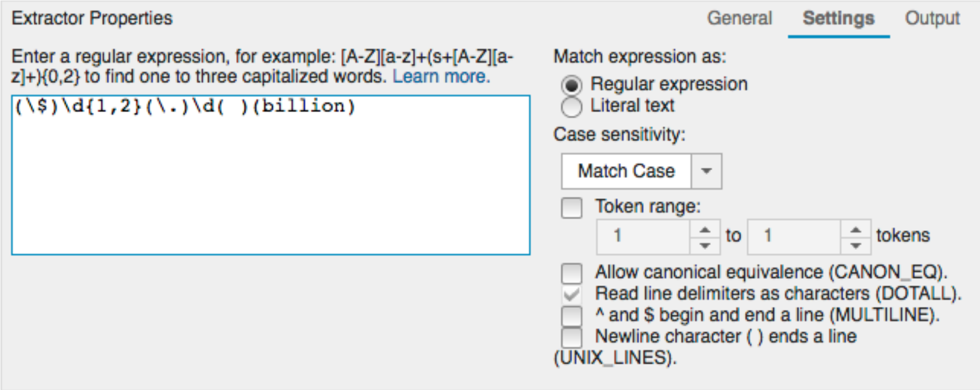
1. Create a Regular Expression

   Click on the *New Regular Expression* button present in the canvas toolbar. Name the Regular Expression ***AmountWithUnit***.



2. Edit the Regular Expression in the Settings for the Extractor

In the settings pane enter the expression: `(\$)\d{1,2}(\.)\d( )(billion)`
Note that there is a space inside the bracket before (billion).



Enter a regular expression, for example: [A-Z][a-z]+(s+[A-Z][a-z]+){0,2} to find one to three capitalized words. Learn more.

```
(\$)\d{1,2}(\.)\d( )(billion)
```

Match expression as:

● Regular expression
○ Literal text

Case sensitivity:

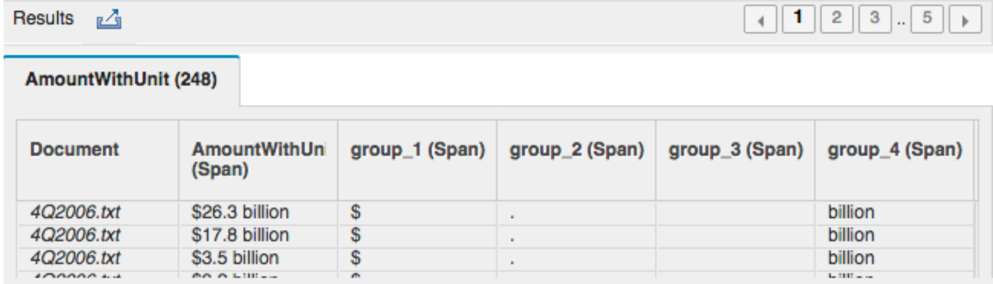
Match Case

□ Token range:
1 to 1 tokens

□ Allow canonical equivalence (CANON_EQ).
✓ Read line delimiters as characters (DOTALL).
□ ^ and $ begin and end a line (MULTILINE).
□ Newline character ( ) ends a line (UNIX_LINES).

3. Run the Regular Expression

Run your regular expression to verify it is finding matches in the documents

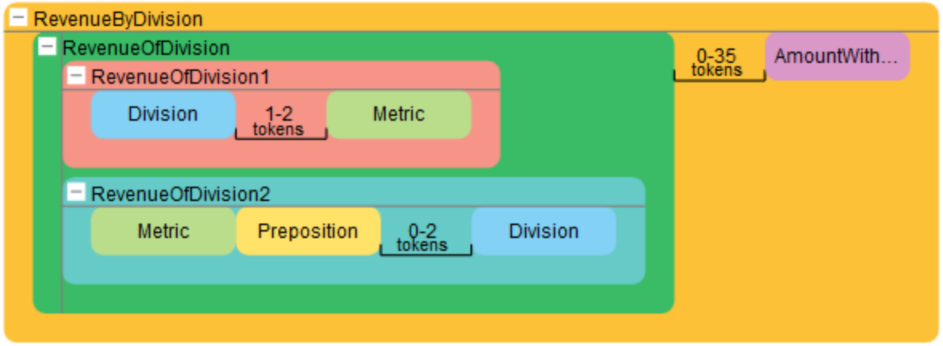


AmountWithUnit (248)

Document	AmountWithUn (Span)	group_1 (Span)	group_2 (Span)	group_3 (Span)	group_4 (Span)
4Q2006.txt	$26.3 billion	$	.		billion
4Q2006.txt	$17.8 billion	$	.		billion
4Q2006.txt	$3.5 billion	$	.		billion

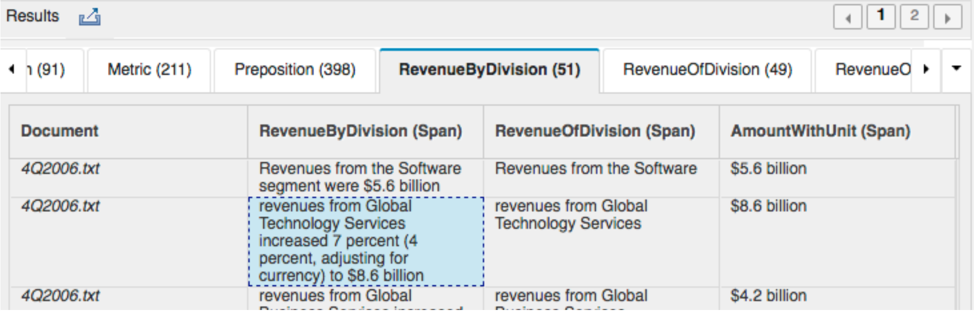
4. Create a Sequence with Union and Regular Expression

Create a new sequence consisting of ***RevenueOfDivision*** followed by *0-35 tokens* (using a new Proximity Rule) and then ***AmountWithUnit.*** Rename it to ***RevenueByDivision***.

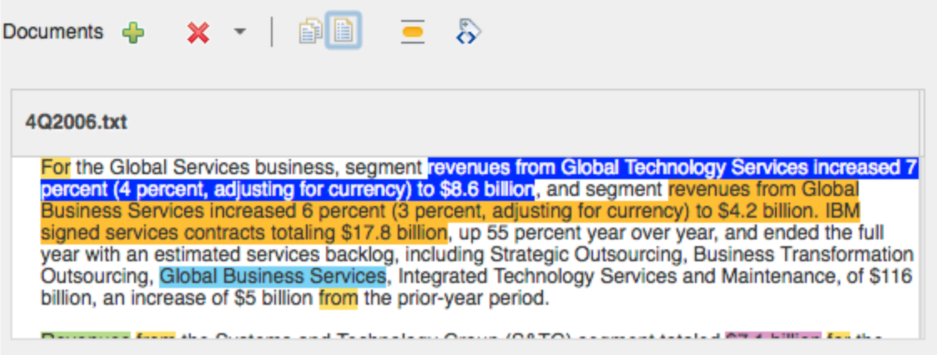


Step 12: Run the Sequence

Run the ***RevenueByDivision*** sequence to see the results:



Document	RevenueByDivision (Span)	RevenueOfDivision (Span)	AmountWithUnit (Span)
4Q2006.txt	Revenues from the Software segment were $5.6 billion	Revenues from the Software	$5.6 billion
4Q2006.txt	revenues from Global Technology Services increased 7 percent (4 percent, adjusting for currency) to $8.6 billion	revenues from Global Technology Services	$8.6 billion
4Q2006.txt	revenues from Global	revenues from Global	$4.2 billion

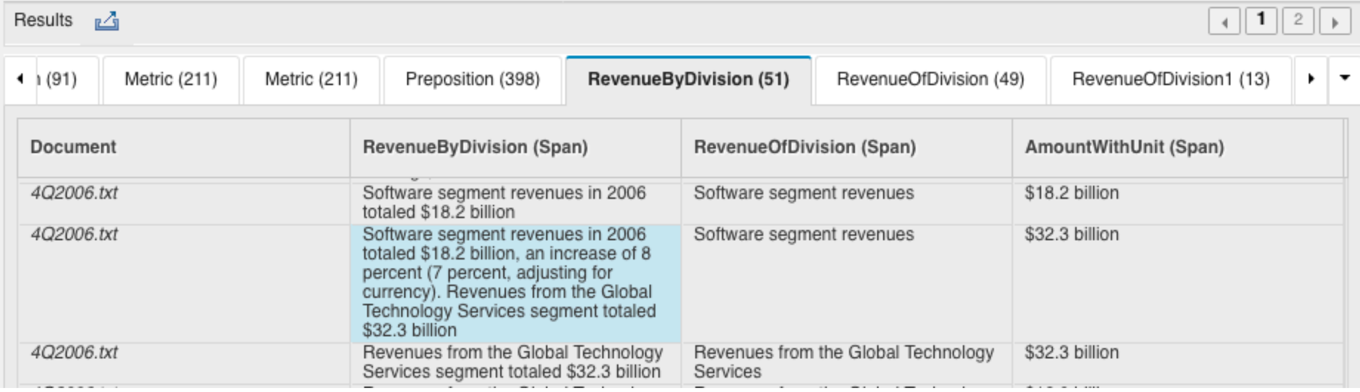


4Q2006.txt

For the Global Services business, segment revenues from Global Technology Services increased 7 percent (4 percent, adjusting for currency) to $8.6 billion, and segment revenues from Global Business Services increased 6 percent (3 percent, adjusting for currency) to $4.2 billion. IBM signed services contracts totaling $17.8 billion, up 55 percent year over year, and ended the full year with an estimated services backlog, including Strategic Outsourcing, Business Transformation Outsourcing, Global Business Services, Integrated Technology Services and Maintenance, of $116 billion, an increase of $5 billion from the prior-year period.

Step 13: Extractor refinement: Managing overlapping matches

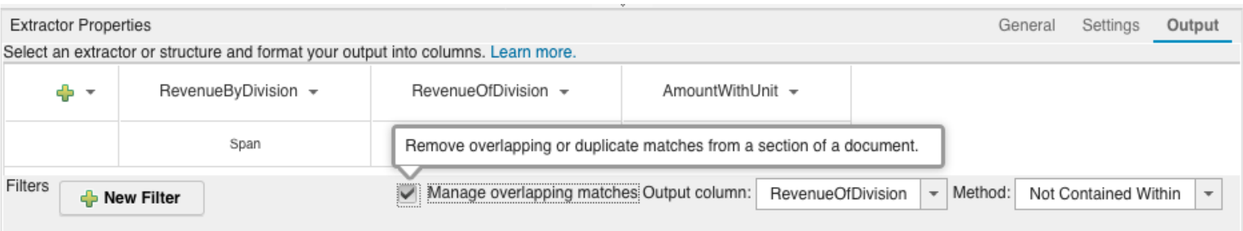
Look at the results of the ***RevenueByDivision*** sequence. Although they may look reasonable at first glance, you should quickly be able to find several examples where the extractor pairs the division and the amount incorrectly. For example:



Document	RevenueByDivision (Span)	RevenueOfDivision (Span)	AmountWithUnit (Span)
4Q2006.txt	Software segment revenues in 2006 totaled $18.2 billion	Software segment revenues	$18.2 billion
4Q2006.txt	Software segment revenues in 2006 totaled $18.2 billion, an increase of 8 percent (7 percent, adjusting for currency). Revenues from the Global Technology Services segment totaled $32.3 billion	Software segment revenues	$32.3 billion
4Q2006.txt	Revenues from the Global Technology Services segment totaled $32.3 billion	Revenues from the Global Technology Services	$32.3 billion

The highlighted example is incorrectly pairing “Software segment revenues” with “$32.3 billion.” This is because the consolidation policy for the extractor has not been specified.

1. In the ***RevenueByDivision*** extractor properties, go to the Output pane. Check the box for *Manage overlapping matches*. For Output column, select ***RevenueOfDivision***. For Method, select *NotContainedWithin*.

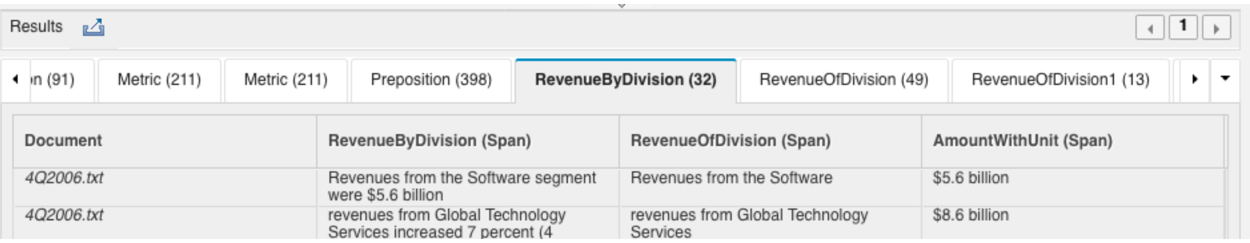


Select an extractor or structure and format your output into columns. Learn more.

Remove overlapping or duplicate matches from a section of a document.

☑ Manage overlapping matches Output column: RevenueOfDivision Method: Not Contained Within

2. Run the extractor again. Note that there are now significantly fewer results.

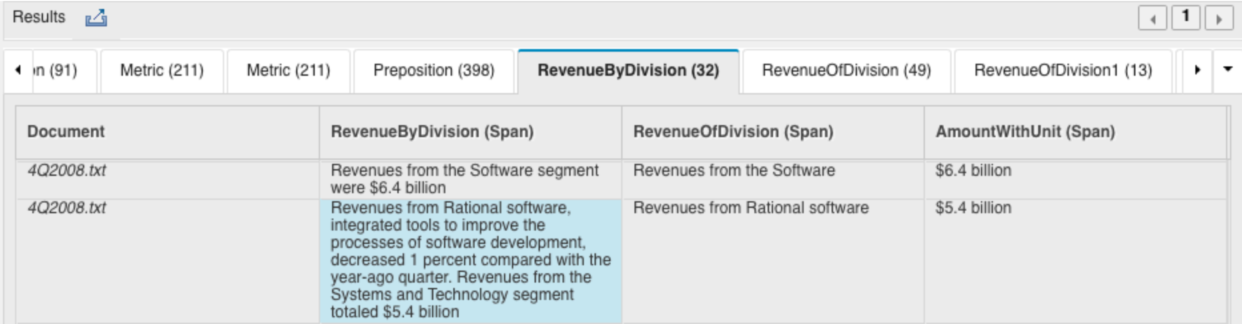


Document	RevenueByDivision (Span)	RevenueOfDivision (Span)	AmountWithUnit (Span)
4Q2006.txt	Revenues from the Software segment were $5.6 billion	Revenues from the Software	$5.6 billion
4Q2006.txt	revenues from Global Technology Services increased 7 percent (4	revenues from Global Technology Services	$8.6 billion

Step 14: Identifying errors in the extractor results

Look closely again at the results of the ***RevenueByDivision*** sequence.

The following example demonstrates that the extractor is still making some mistakes by pairing the division from one sentence with an amount from a subsequent sentence:

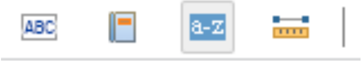


Document	RevenueByDivision (Span)	RevenueOfDivision (Span)	AmountWithUnit (Span)
4Q2008.txt	Revenues from the Software segment were $6.4 billion	Revenues from the Software	$6.4 billion
4Q2008.txt	Revenues from Rational software, integrated tools to improve the processes of software development, decreased 1 percent compared with the year-ago quarter. Revenues from the Systems and Technology segment totaled $5.4 billion	Revenues from Rational software	$5.4 billion

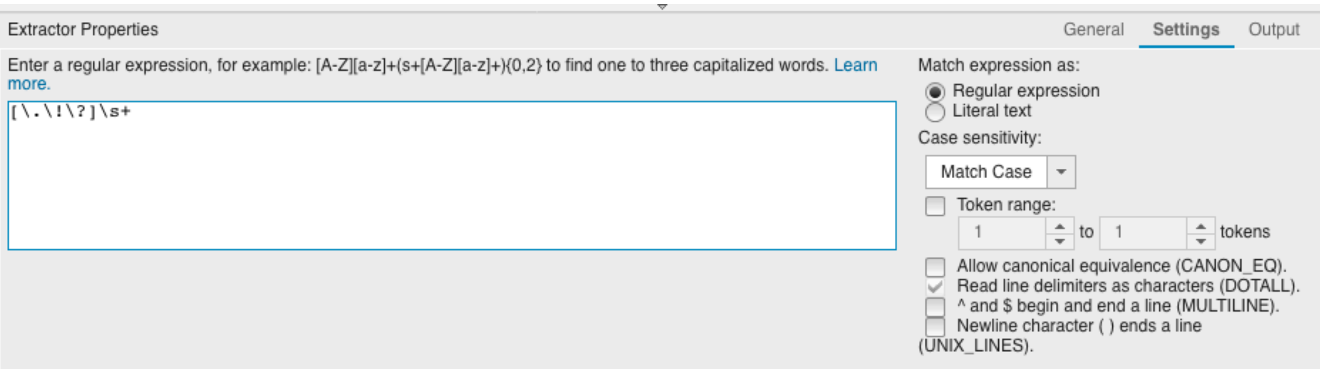
To fix this, you will need to create an extractor that detects sentence boundaries, and then use it to filter the ***RevenueByDivision*** extractor.

Step 15: Extractor refinement: Sentence boundary detection

1. Create a new regex extractor by clicking on the New Regular Expression button in the canvas toolbar. When the extractor appears, type ***SentenceBoundary*** and press Enter.



In the Extractor Properties Settings pane, copy and paste the following expression: `[\.\!\?]\s+`
Check that your extractor Settings are correct.



Enter a regular expression, for example: [A-Z][a-z]+(s+[A-Z][a-z]+){0,2} to find one to three capitalized words. Learn more.

```
[\.\!\?]\s+
```

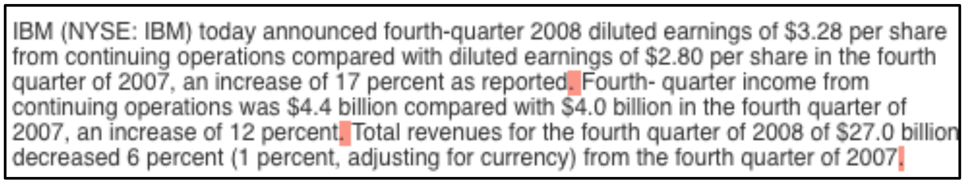
Match expression as:
● Regular expression
○ Literal text

Case sensitivity:
Match Case

□ Token range:
1 to 1 tokens

□ Allow canonical equivalence (CANON_EQ).
✓ Read line delimiters as characters (DOTALL).
□ ^ and $ begin and end a line (MULTILINE).
□ Newline character ( ) ends a line (UNIX_LINES).

2. Run the extractor and look at the text to see that it does in fact find sentence boundaries:



IBM (NYSE: IBM) today announced fourth-quarter 2008 diluted earnings of $3.28 per share from continuing operations compared with diluted earnings of $2.80 per share in the fourth quarter of 2007, an increase of 17 percent as reported. Fourth- quarter income from continuing operations was $4.4 billion compared with $4.0 billion in the fourth quarter of 2007, an increase of 12 percent. Total revenues for the fourth quarter of 2008 of $27.0 billion decreased 6 percent (1 percent, adjusting for currency) from the fourth quarter of 2007.

Step 16: Extractor refinement: Filter *SentenceBoundary* with Abbreviation detection

Our sentence boundary extractor is not quite sophisticated enough yet. If you review the results closely you will find that it incorrectly identifies spans like the period after the J in “Samuel J. Palmisano”, and the period after the No in “FASB Interpretation No. 47” (both of these examples can be seen in 4Q2006.txt).

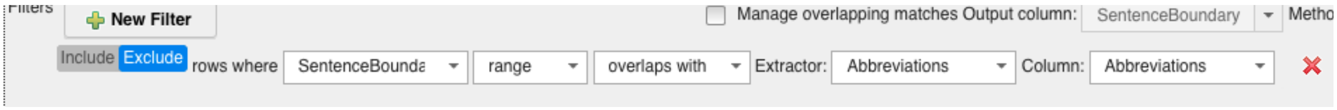
To fix this, we will need to filter our sentence boundary extractor to ignore abbreviations.

1. First, create a new dictionary extractor and import the terms from a .txt file, just like you did for the Division extractor earlier, by clicking on the New Dictionary button in the toolbar.

2. Name the extractor Abbreviations and use the file Abbreviations.txt

3. Select the ***SentenceBoundary*** extractor and go to the Extractor Properties Output pane. Click the button to add a new filter to the output of the extractor

Set the filter to exclude (by clicking on Exclude) rows where the ***SentenceBoundary*** extractor overlaps with the newly created Abbreviations extractor:



Filters   New Filter   ☐ Manage overlapping matches Output column: SentenceBoundary   Metho

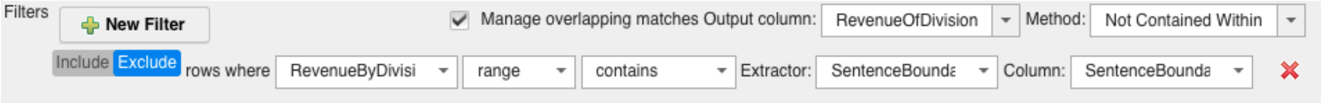
Include Exclude rows where SentenceBounda   range   overlaps with   Extractor: Abbreviations   Column: Abbreviations   ✖

4. Run the ***SentenceBoundary*** extractor again. You will see that the number of results has decreased, and abbreviations such as the ones we had noticed are no longer being picked up.

Step 17: Extractor refinement: Use *SentenceBoundary* to filter *RevenueByDivision*

1. Go back to ***RevenueByDivision*** and use ***SentenceBoundary*** as an output filter in the same way as you just used Abbreviations.

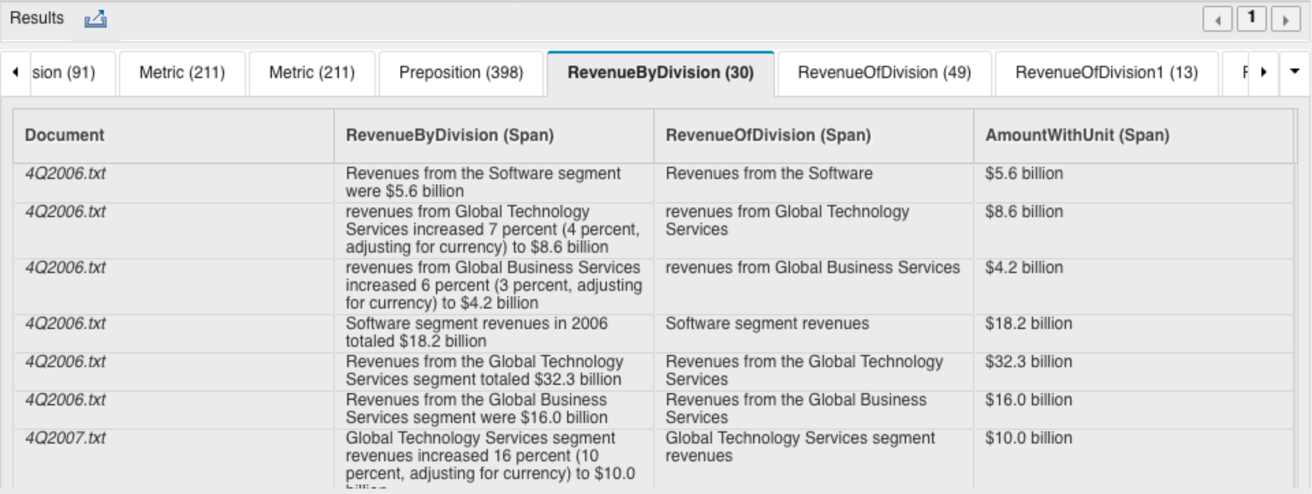
2. Go to the Extractor Properties Output pane, click the button to add a new filter to the output of the extractor, and set it to exclude rows that contain a sentence boundary:



Filters   New Filter   ☑ Manage overlapping matches Output column: RevenueOfDivision   Method: Not Contained Within

Include Exclude rows where RevenueByDivisi   range   contains   Extractor: SentenceBounda   Column: SentenceBounda   ✖

Run ***RevenueByDivision*** again and review the results:



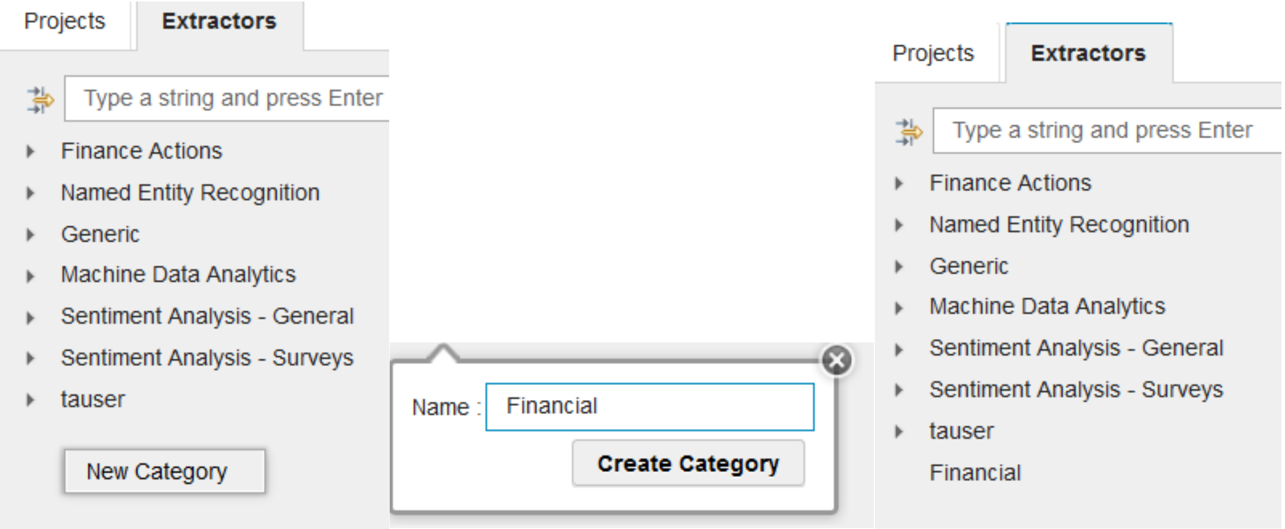
Document	RevenueByDivision (Span)	RevenueOfDivision (Span)	AmountWithUnit (Span)
4Q2006.txt	Revenues from the Software segment were $5.6 billion	Revenues from the Software	$5.6 billion
4Q2006.txt	revenues from Global Technology Services increased 7 percent (4 percent, adjusting for currency) to $8.6 billion	revenues from Global Technology Services	$8.6 billion
4Q2006.txt	revenues from Global Business Services increased 6 percent (3 percent, adjusting for currency) to $4.2 billion	revenues from Global Business Services	$4.2 billion
4Q2006.txt	Software segment revenues in 2006 totaled $18.2 billion	Software segment revenues	$18.2 billion
4Q2006.txt	Revenues from the Global Technology Services segment totaled $32.3 billion	Revenues from the Global Technology Services	$32.3 billion
4Q2006.txt	Revenues from the Global Business Services segment were $16.0 billion	Revenues from the Global Business Services	$16.0 billion
4Q2007.txt	Global Technology Services segment revenues increased 16 percent (10 percent, adjusting for currency) to $10.0	Global Technology Services segment revenues	$10.0 billion

Step 18: Save Your Extractors

1. Create a custom category

Right click on the Extractors pane on the left hand side of the screen and click on *New Category*.

Name the category ***Financial*** and click on *Create Category*.

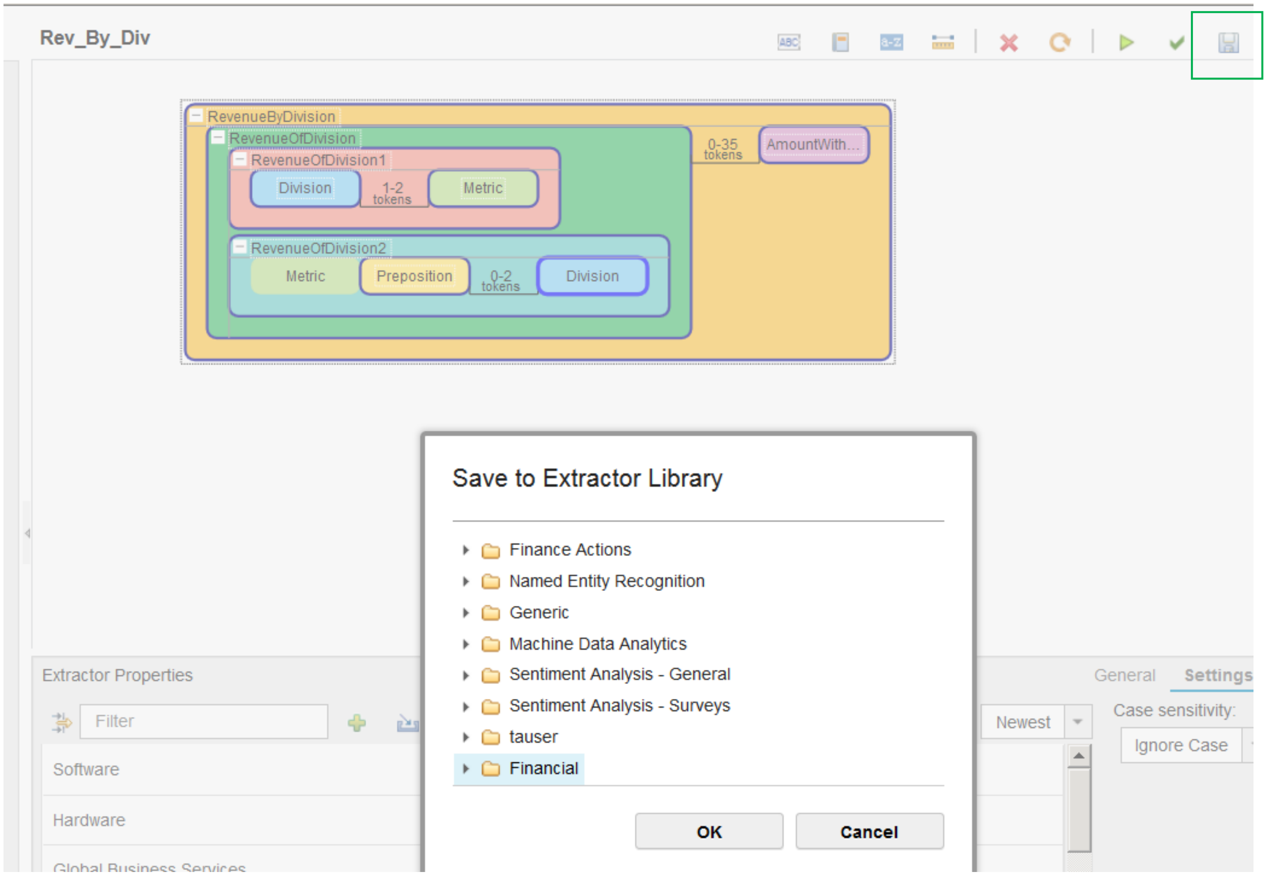


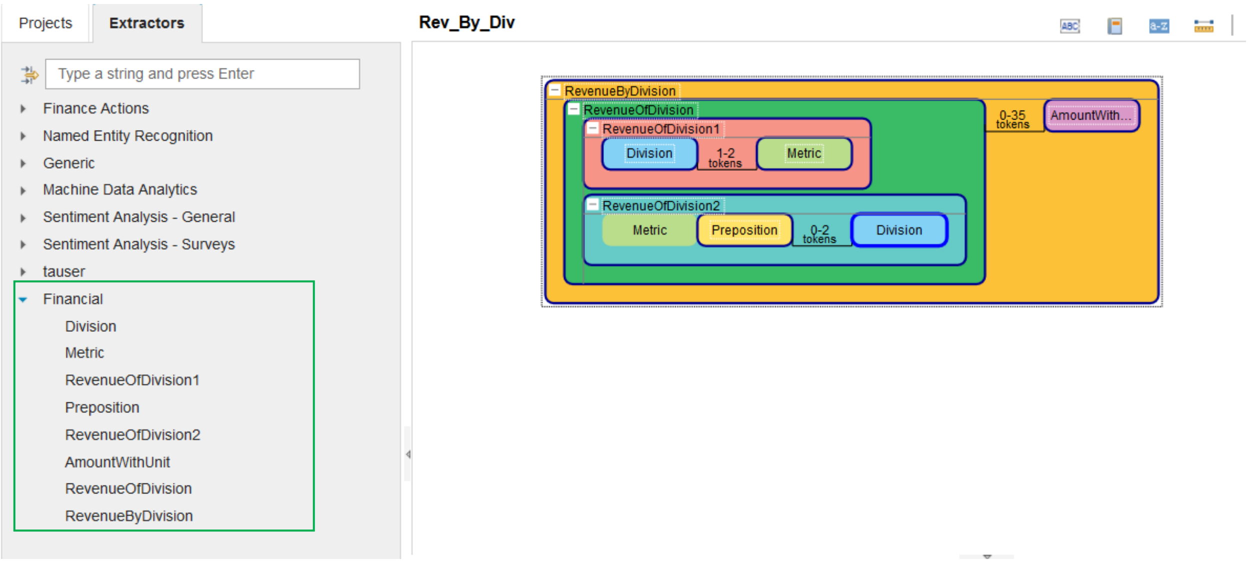
Name : Financial

2. Save the extractors to the Catalog under the newly created category

Save the sequence **RevenueByDivision** to **Financial** category in the extractor catalog.

Multi-select all the extractors by clicking and dragging your mouse to select them all, then click on the Save Extractor button in the canvas toolbar. Then select the Financial category and click OK.


Rev_By_Div
RevenueByDivision
RevenueOfDivision
RevenueOfDivision1
Division
1-2 tokens
Metric
0-35 tokens
AmountWith...
RevenueOfDivision2
Metric
Preposition
0-2 tokens
Division
Save to Extractor Library
Finance Actions
Named Entity Recognition
Generic
Machine Data Analytics
Sentiment Analysis - General
Sentiment Analysis - Surveys
tauser
Financial
OK
Cancel
Extractor Properties
Filter
Software
Hardware
Global Business Services
General
Settings
Newest
Case sensitivity:
Ignore Case


Projects
Extractors
Type a string and press Enter
Finance Actions
Named Entity Recognition
Generic
Machine Data Analytics
Sentiment Analysis - General
Sentiment Analysis - Surveys
tauser
Financial
Division
Metric
RevenueOfDivision1
Preposition
RevenueOfDivision2
AmountWithUnit
RevenueOfDivision
RevenueByDivision
Rev_By_Div
RevenueByDivision
RevenueOfDivision
RevenueOfDivision1
Division
1-2 tokens
Metric
0-35 tokens
AmountWith...
RevenueOfDivision2
Metric
Preposition
0-2 tokens
Division