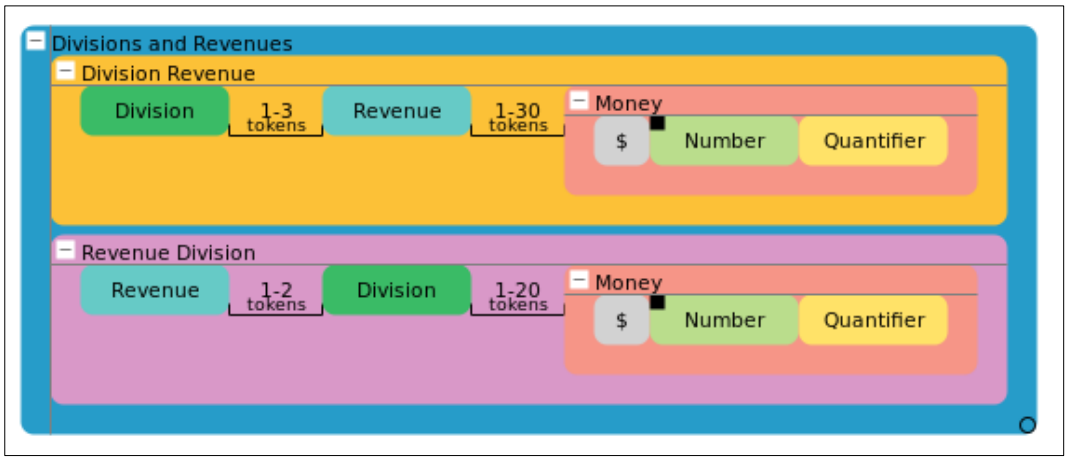
IBM Training

Demonstration 1

Analyzing quarterly reports using Text Analytics



Working with pre-built extractors

Demonstration 1: Analyzing quarterly reports using Text Analytics

Demonstration 1: Analyzing quarterly reports using Text Analytics

Purpose:
In this demo, you will work with some of the prebuilt extractors to analyze quarterly reports.

User ids / Passwords

OS:	**biadmin/biadmin**
Root:	**root/dalvm3**
Ambari:	**admin/admin**
BigInsights Home:	**guest/guest-password**

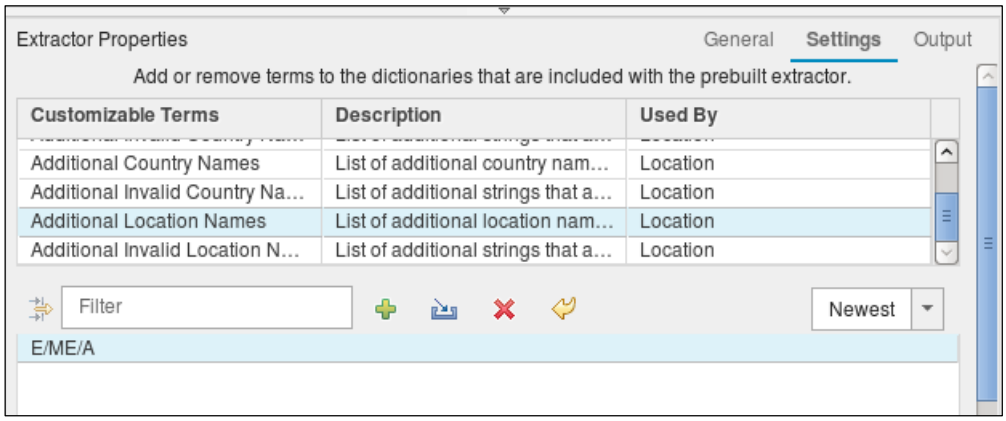
Ambari Services Required:

- HDFS
- MapReduce2
- YARN
- Knox (also start the Demo LDAP service)
- BigInsights - Text Analytics
- BigInsights - Home

Task 1. Setting up your project.

1. Create a new project called **Quarterly Reports**.
2. Add the five documents to the project from this location: **/home/biadmin/labfiles/TextAnalytics/IBMQuarterlyReports**
3. On the **Extractor** tab, expand the **Named Entity Recognition** category.
4. Drag and drop the **Location** extractor onto the canvas.
5. Run the extractor on the documents and you should get 63 rows returned.
6. Scroll through the results until you see the first result for the document listed for *4Q2007.txt*. Double click on that entry to see the results in the document pane. Notice that the acronym E/ME/A is not highlighted. This is an acronym that IBM uses for Europe/Middle East/Africa. You can modify the prebuilt extractor so that it accurately captures what you need.

7. On the **Extractor Properties** under the **Settings** tab, select the **Additional Location Names** and add the **E/ME/A** acronym to that list.



8. Run the extractor again and you see that there are now 64 matches.

Task 2. Analyzing documents and identifying examples.

For this process, you would typically enlist the help of someone who knows the document well to help you identify the examples of clues to search. Since you are interested in extracting revenue by division, you must read through to find spans of text that contain this information. Look for patterns and clues in the text to help improve the accuracy of the extractor.

An example that you might find is a phrase such as *Revenues from Software were $3.9 billion*. This has three important features:

"Software" is a division name

"$3.9 billion" is a revenue amount

"revenue"

You will use these features as context to identify instances of revenue by division.

It is a good idea to decompose the clues to the lowest level. This allows for flexibility and also it lets the extractor performs all the hard work of combining all the clues. Consider that *Money* has three basic features, a currency sign, followed by a number, followed by a quantifier such as million or billion.

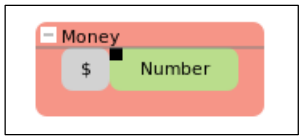
Two patterns that you may have picked up are:

Revenues for division were $x.x

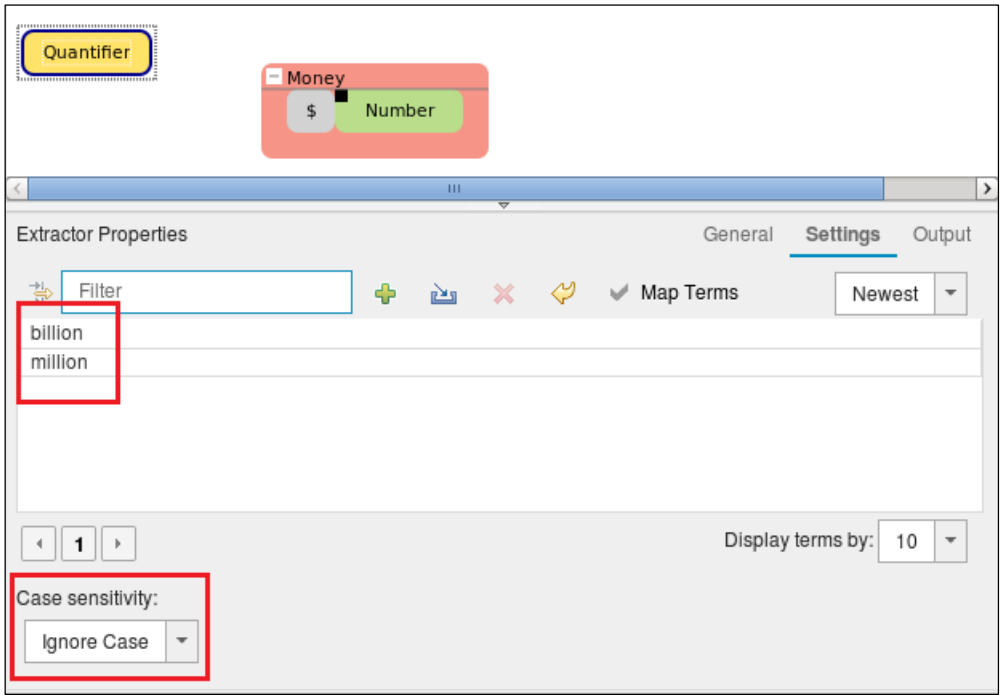
Division revenues were $x.x

Task 3. Creating and testing extractors.

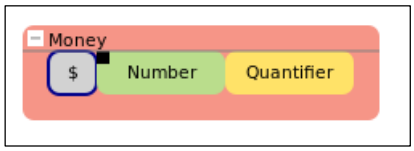
1. You will create the basic features on the canvas. Create one feature for each of the three basic features of money. Click the **New Literal** button. Type a dollar sign ($).
2. Run that extractor and you should see 389 results.
3. On the **Extractor**s catalog, expand the **Generic** category and drag the **Number** extractor to the canvas.
4. Run that and you will see 1722 results.
5. Create a sequence of these two by dragging the literal to the left of the Number extractor.
6. Rename it from **Sequence 1** to **Money**.



7. Run the Money extractor to test it out.
8. To add the Quantifier extractor, you will use a dictionary. Click the **New Dictionary** button. Name it **Quantifier.**
9. Enter in two terms for the dictionary: **million** and **billion**.
10. Ensure that the case sensitivity is set to **Ignore case** under the **Settings** tab.



11. Drag the **Quantifier** extractor into the **Money** extractor to complete the sequence.

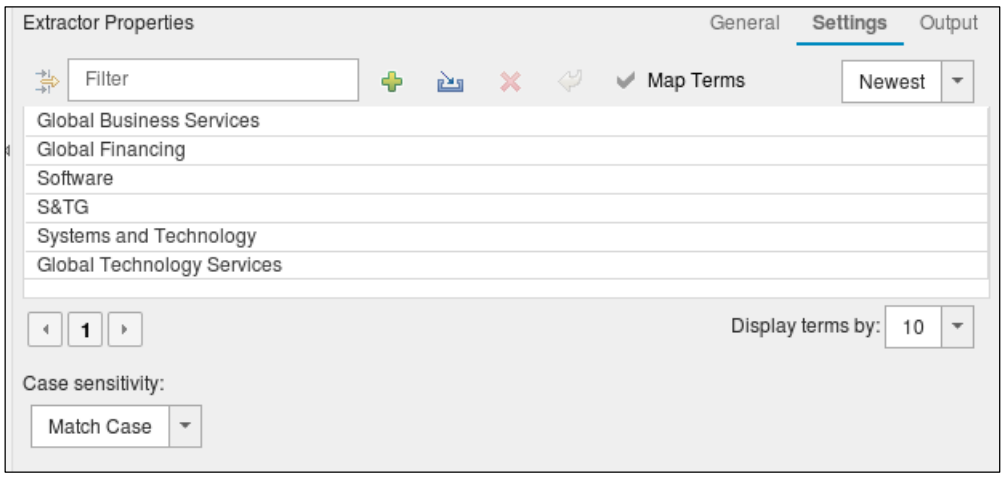


12. Run it and you should see 333 matches returned. You'll also see each tab for the rows specific to the individual extractors.

    You have now located all the instances of money. Next task is for you to combine with revenues and divisions.

Task 4. Writing and testing extractors for candidates.

1. Create a new dictionary. Name it **Revenue.**
2. Add two terms to it: **revenue** and **revenues**.
3. Run this extractor to test it out. There should be 238 matches.
4. Create a new dictionary for **Division** names.
5. Add the following terms to it: **Global Technology Services, Systems and Technology, S&TG, Software, Global Financing,** and **Global Business Services.**
6. Run the extractor. You should get 142 rows.
7. Notice that in the results, the terms software and global financing are picked up as division names. Because they are in lowercase, they are likely not division names. The problem can be fixed by choosing the **Match Case** option for the extractor.



8. Run the extractor again and you should get 98 rows now.

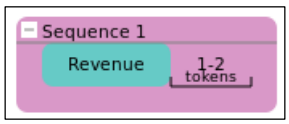
You have now extracted the three key basic features: money, revenue, and division. The next step is to extract candidates that match the two patterns that you identified earlier. To generate candidates, you combine extractors into sequences, building on the extractors that you created in the previous tasks.

If you remember, the two patterns are: revenues for division were $x.x and division revenues were $x.x

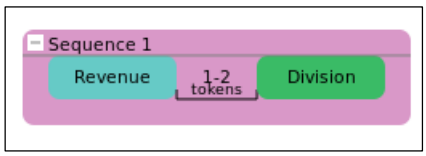
The first pattern looks for examples where the word revenue is followed by a division name and then a money amount, with some number of tokens in between. This is the conceptual view of the first pattern.

<Revenue><1 to 2 Tokens><Division><1 to 20 Tokens><Money>

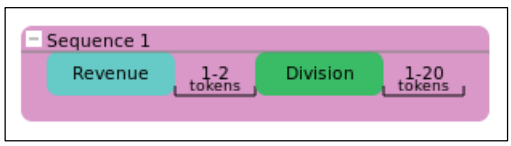
9. Add a new proximity rule with 1-2 tokens.
10. Drag the proximity rule to the right of the Revenue extractor to create a new sequence.



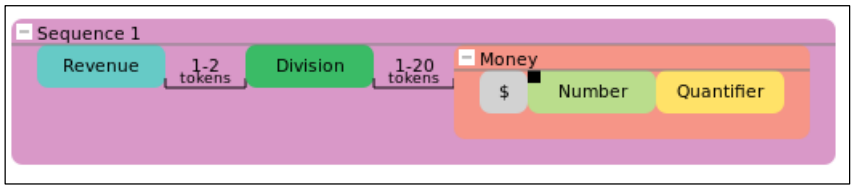
11. Drag the Division extractor to the right of the proximity rule to add to the sequence.



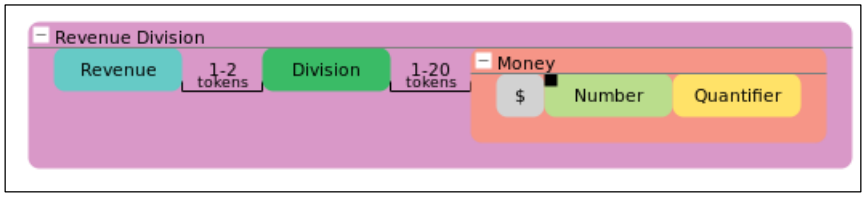
12. In order to create the next proximity rule, right-click on the **Division** extractor and choose **Add After** and then **Proximity Rule**. Fill in 1-20 in the text box.



13. Drag the **Money** extractor into the sequence.



14. Rename **Sequence 1** to **Revenue Division**.



15. Run it and you should see 22 results.
16. For the second pattern, it looks for examples where a division name is followed by the word revenue and a money amount. For example, *Global Financing segment revenues* increased *3 percent* (flat, adjusting for currency) in the fourth quarter to *$620 million*. When you look at these patterns in the document, you find that there are between 1 and 3 words between *Division* and *Revenue*, but perhaps as many as 30 between *Revenue* and *Money*.

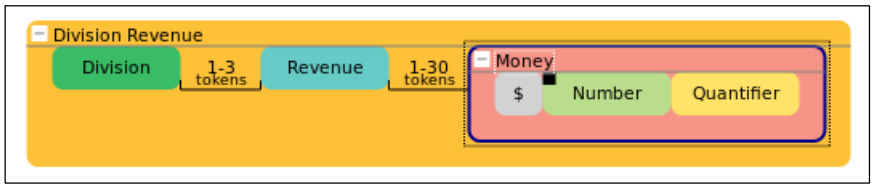
    Conceptually, this looks like:

    ```
    <Division><1 to 3 Tokens><Revenue><1 to 30 Tokens><Money>
    ```

17. Right-click on the **Revenue** extractor and choose **Copy.**
18. Right-click on the canvas and choose **Paste.**
19. Do the same for the **Division** and **Money** extractors.

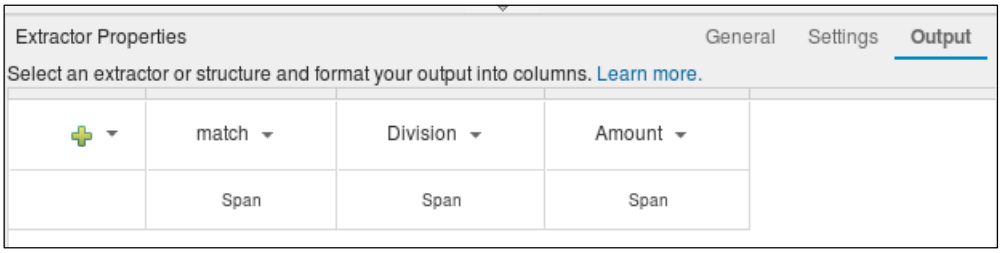
    You should now have linked copies of the three extractors. Remember, linked copies are affected when you change one. If you needed a new copy, you would select Paste as New Copy. Notice that the linked copies are the same color.

20. Create a new sequence with these three extractors and proximity rules to create an extractor for the second pattern. Name it **Division Revenue**.

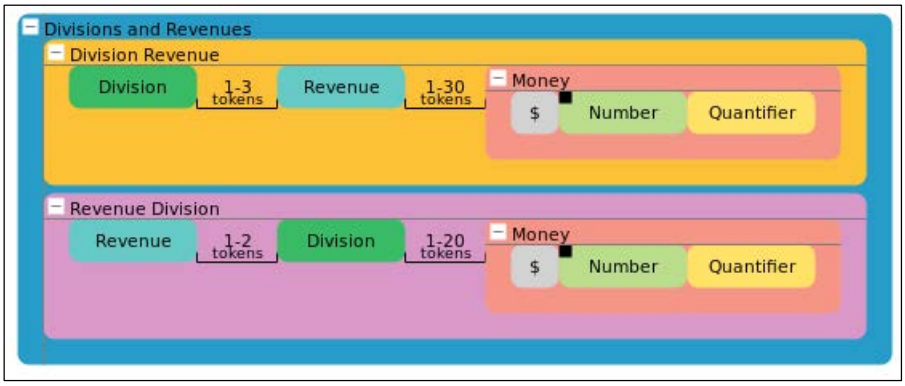


21. Run the extractor. You should get 40 matches.

22. Union these two extractors together to yield a full picture. First, the columns must match. Modify the output specification to make them match.
    a) Go to the Output tab of the Properties pane.
    b) Deselect the **Revenue** column. We don't need this column. Click on the dropdown next to the green plus.
    c) Rename the **Division Revenue** column to **match**.
    d) Rename the **Money** column to **Amount**.



23. Both extractor's output column must match.
24. Create the union by dragging and dropping one of them on top of the other.
25. Change the name of the union to **Divisions and Revenues.**



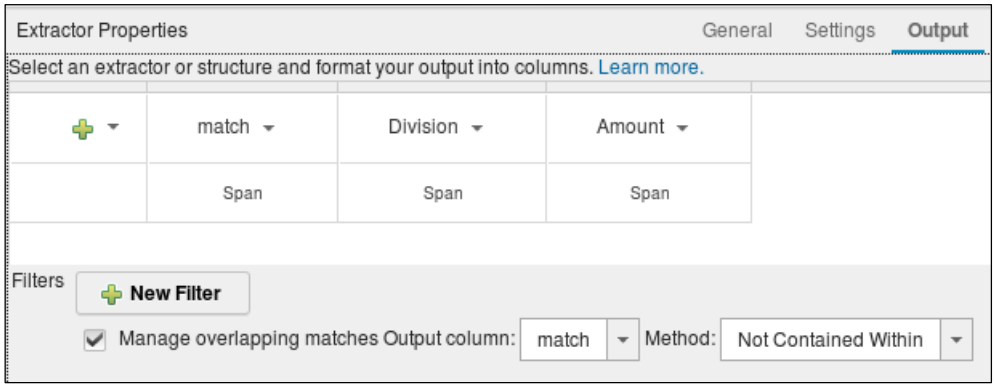
26. Run the extractor and you should get 62 matches.

Task 5. Creating and testing final extractors.

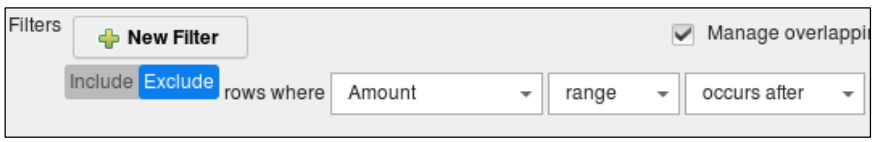
The first part of consolidating is to remove duplicate information. As you scroll through the 62 results, notice the last two entries for *4Q2006.txt* that one of the results is contained within the other results, causing duplicated matches.

1. Right-click the union extractor and select **Edit Output**. This will bring you to the Output tab of the Extractor Properties pane.
2. Click **Manage overlapping matches**.

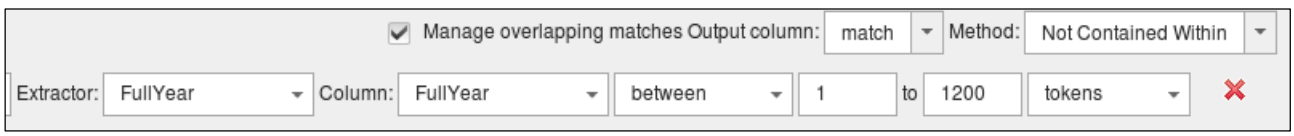
3. Choose output column **match** and Method **Not Contained Within**. This specifies that we only want matches which are not contained within another match.



4. Run the extractor again and verify that none of the 49 matches are contained within another.
5. Look at the results again. Notice that there are two values for the Software division in *the 4Q2006.txt* file. Looking more closely, one of these results was for 4Q, and the other for the full year.
6. On examining the document, we see that the unwanted results have their Money amount within a proximity of 1200 tokens from a phrase like `Full-Year 2006 Results`. To match multiple years, we can create a regular expression to match this clue for unwanted results. Click the **New Regular Expression** button.
7. Type in **FullYear**
8. Type in **Full-Year \d{4} Results** as the regular expression.
9. Run the regular expression to test it. You should see five results, one per document.
10. Select the **Divisions and Revenues** extractor. Under the Output tab, click the **New Filter** button.
11. Click **Exclude,** because we want to exclude some rows.
12. Choose the **Amount** column, the **range** type, and the **occurs after** option.



13. Select the **FullYear** extractor and **FullYear** Column choose **between** and fill in **1** and **1200** for the **tokens**.



14. Run the extractor and verify that only 25 results should be shown now.

    This view contains exactly the information that you need for further analysis. When you apply text analytics to more complex documents, and when you are extracting more sophisticated information, you would expect to spend time improving the precision and recall of your extractor. You can also profile your extractor to understand and improve its performance characteristics.

Task 6. Finalizing and saving the extractor.

1. Click on the extractor and click the **Save** icon.
2. Select the **guest** category to save the extractor.
3. You can choose to **Export AQL**, **Publish to BigSheets** or **Run on the Cluster**.

Results:
In this demonstration you have learned to use some of the pre-built extractors to analyze IBM quarterly reports to figure out the revenues from each division.