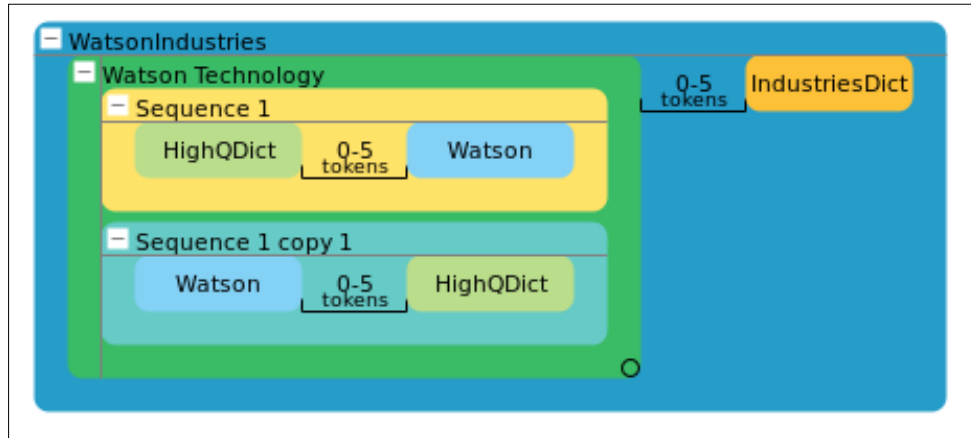


Demonstration 1

Filtering and consolidating



Filter and consolidation

© Copyright IBM Corporation 2015

Demonstration 1: Filtering and consolidating

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Demonstration 1: Filtering and consolidating

Purpose:

In this demonstration, you will filter and consolidate the results of the extractors

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
- BigSheets (optional)

Task 1. Removing duplicate results.

1. We are going to resume with the Watson project. Open it up from the **Projects** pane.
2. Run the **Union 1** extractor again. You should get 178 rows returned.
3. Now, look at the results more closely and you see some duplicates in the first file, SM001.txt. The word Watson was selected twice. Once for the IBM clue and again for the technology clue. We only need one occurrence of this.

Document	Sequence 1 (Span)	HighQDict (Span)	WatsonDict (Span)
SM001.txt	IBM Watson	IBM	Watson
SM001.txt	Watson technology	technology	Watson

4. Take a look at the **Extractor Properties**.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

5. Rename the **Sequence 1** column. On the **Output** tab, click the **Sequence 1 dropdown** and select **Rename**.
6. Name it **WatsonSpan**.
7. Click the Manage overlapping matches on the **WatsonSpan** column using the Method **Left to Right**.

Extractor Properties General Settings **Output**

Select an extractor or structure and format your output into columns. [Learn more.](#)

	WatsonSpan	HighQDict	WatsonDict
	Span	Span	Span

Filters New Filter

☒ Manage overlapping matches Output column: WatsonSpan Method: Left to Right

8. Run the extractor and note that there are now 124 rows returned. More importantly, the duplicates have been removed from the results.

HighQDict (1183)	Sequence 1 (99)	Sequence 1 copy 1 (79)	Union 1 (124)	WatsonDict (298)
Document	WatsonSpan (Span)	HighQDict (Span)	WatsonDict (Span)	
SM001.txt	IBM Watson	IBM	Watson	
SM002.txt	IBM Watson	IBM	Watson	
SM002.txt	Watson computer	computer	Watson	
SM002.txt	IBM Watson	IBM	Watson	
SM002.txt	system. Watson	system	Watson	

Task 2. Creating filters to remove instances of the Watson research center.

1. Under the same **Extractor Properties** on the **Output** tab, click the **New Filter** button.
2. Select **Exclude** rows where **WatsonSpan range occurs before** Extractor **ResearchDict** Column **ResearchDict** between **0** to **3** tokens.

Include **Exclude** rows where WatsonSpan range occurs before Extractor: ResearchDict Column: ResearchDict between 0 to 3 tokens

3. Run the extractor and see that the occurrence has been removed. 123 rows returned.

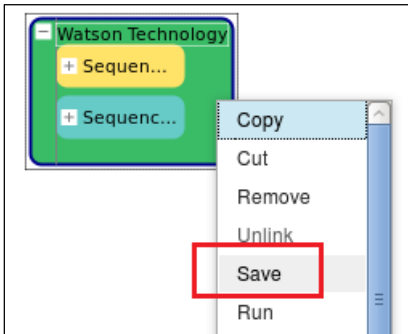
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Task 3. Using regular expression to filter out names.

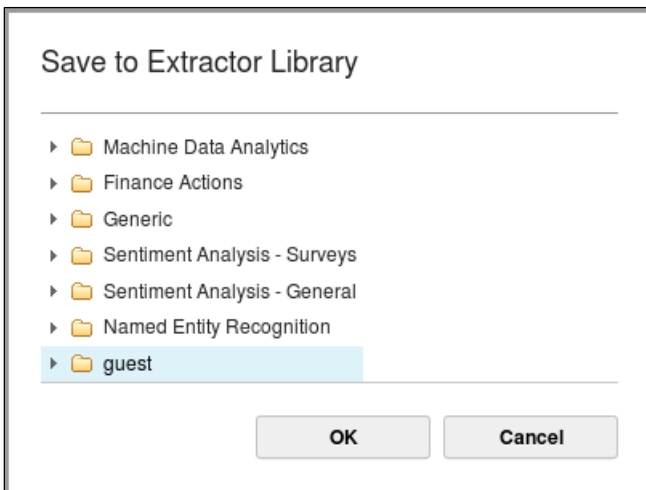
1. SM126.txt (on page 26 on the documents pane) has IBM Founder Thomas J Watson. We cannot eliminate this unless we come up with a regular expression for it.
2. I will leave this exercise up to you. Create a regular expression extractor which eliminates names such as the one located on the SM126.txt file.

Task 4. Saving the extractor.

1. Rename the Union 1 extractor as **Watson Technology**.
2. Save the extractor. Right-click and select **Save**.



3. Save it as **Watson Technology** under the **guest** folder.

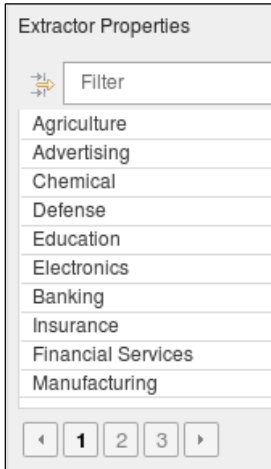


4. Now you can drag and drop this extractor from the catalog to use it in other projects.

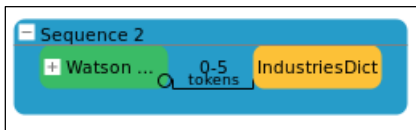
This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

Task 5. Working with Watson industries.

1. Continuing with the same project, create a new dictionary: **IndustriesDict**
2. This time, import the dictionary from the **/home/biadmin/labfiles/WatsonData/Dictionary/Industries.dict** file



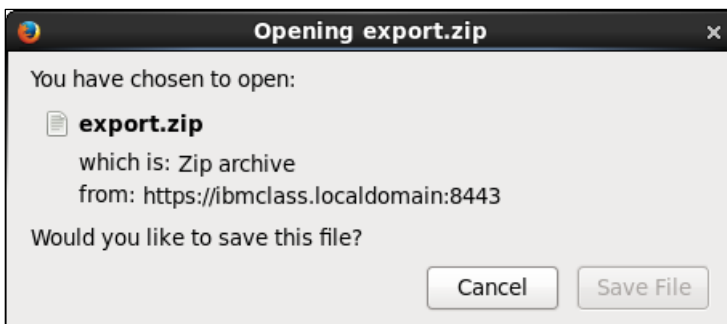
3. Create a proximity rule of **0-5 tokens**.
4. Join the **Watson Technology** extractor with the **IndustriesDict** using the proximity rule.



5. Run the extractor.
6. Likewise, you can create a union of these and put IndustriesDict preceding the proximity rule. I leave this optional exercise up to you.
7. Rename and Save this Sequence 2 extractor as **WatsonIndustries**

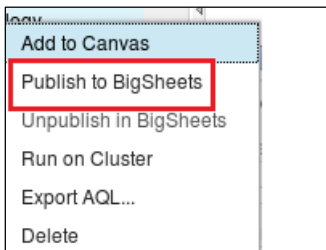
Task 6. Exporting the results.

1. On the Extractor catalog, right-click any of the extractors under the guest folder and select the Export AQL option to get the AQL code out as a zipped file.



This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

2. Click Cancel to the Opening export.zip dialog. We are not going to export.
3. You can export the extractor as a function within BigSheets. The BigSheets service has to be installed and started for the function to work. If you want to try this now, go ahead and start up the **BigInsights-BigSheets** service in Ambari.
4. Once the BigSheets service has started, restart the **BigInsights-Home** service.
5. Refresh the **BigInsights-Home** page in the **Firefox** Browser. You should see the BigSheets panel enabled (no longer greyed out).
6. Click the **Text Analytics** link to go back into your projects.
7. Back on the Extractor catalog on the left side, under the guest folder, right-click the **Watson Technology** extractor and select **Publish to BigSheets**.



8. Click **Next>**.

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

9. Rename the extractor. Double click the **Watson_Technology** name. Name it **WatsonTechExtractor**.


BigSheets Plugin

List of extensions:

Big Sheets Function Name	Description	Big Sheets Function Category	Replace
WatsonTechExtractor		textanalytics	None ▼

< Back OK Cancel

10. Click **OK** to publish it.
11. Click the **Close** button.

 Information

02:58:02 PM EST: Publish Results:

Function Name	Concept Name	Status
WatsonTechExtractor	Watson Technology	Success

Close

This material is meant for IBM Academic Initiative use only. NOT FOR RESALE

12. From the **Extractors** tab **guest>Watson Technology** right click and select **Run on Cluster**.
13. A final way to use your extractor outside of **Text Analytics** is to run it on the Cluster. Select the **Run on Cluster** option.

Run on Cluster

Data source:

No Directory Selected

Write results to output folder:

No Directory Selected

☐ **Save extractor artifacts:**

No Directory Selected

File name:

Select the extractors to run:

☐ Watson Technology

14. Specify the options that you wish and run the extractor. I leave this as an optional exercise for you.

Results:

You have learned to filter and consolidate the results of the extractors.