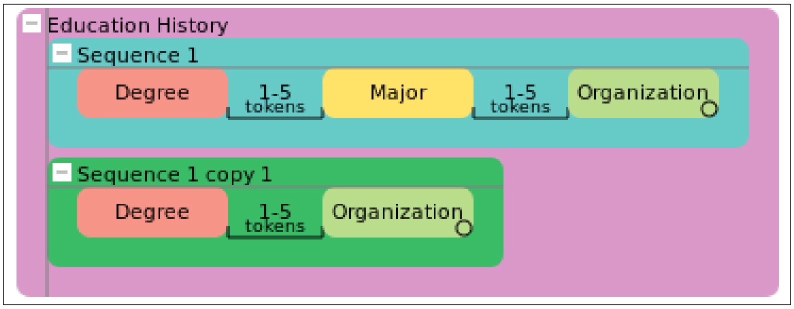
IBM Training

Demonstration 1

Extract education histories from biographies



Text Analytics Overview

Demonstration 1: Extract education histories from biographies

Demonstration 1: Extracting education histories from biographies

Purpose:
The purpose of this demonstration is to give you an end-to-end feel of how to use BigInsights Text Analytics to analyze text data. In subsequent units and demonstrations, you will get to work with the individual components to better understand how to use Text Analytics.

User ids / Passwords

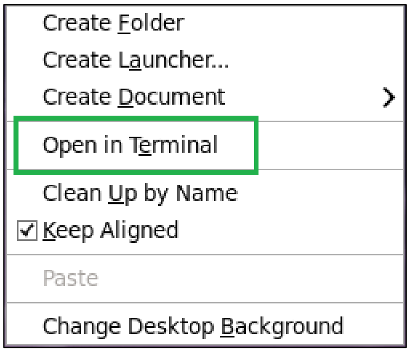
OS:	**biadmin/biadmin**
Root:	**root/dalvm3**
Ambari:	**admin/admin**
BigInsights Home:	**guest/guest-password**

Ambari Services Required:

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

Task 1. Starting up the required services via Ambari.

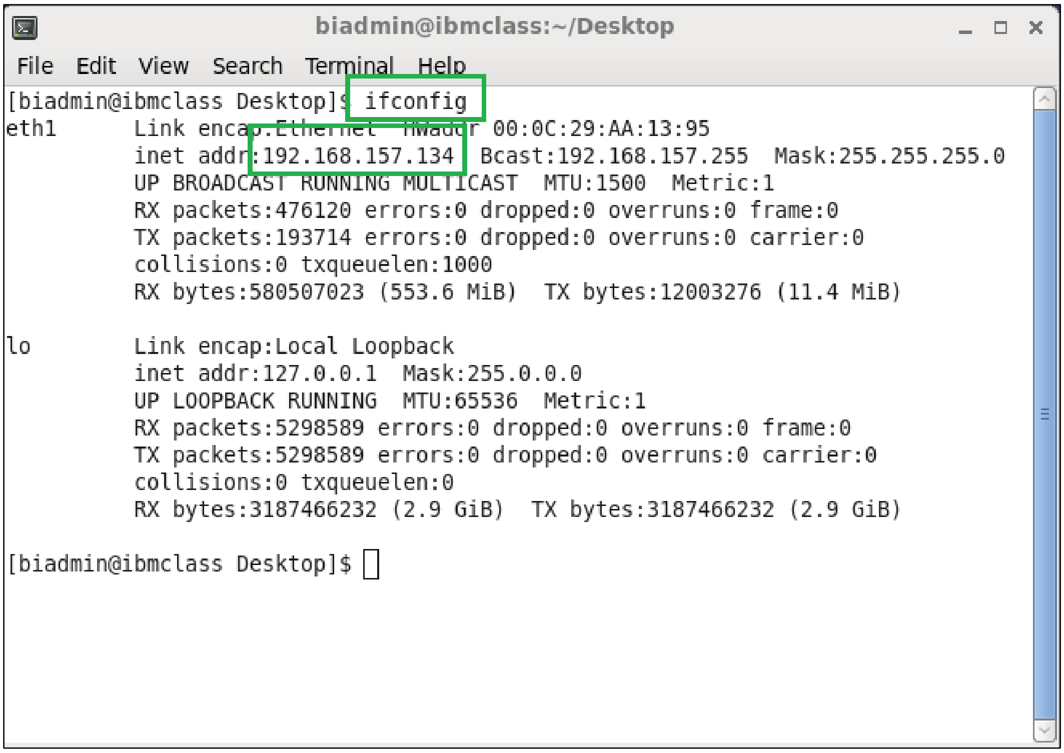
1. Login to the OS using **biadmin/biadmin**.
2. Once the OS has loaded, verify that the assigned IP address matches that in the /etc/hosts file. To do so, open up a new Terminal. Right-click somewhere on the desktop and select **Open in Terminal**.



3. In the terminal window that appears, type in:

```
ifconfig
```

4. Note the ip address that has been assigned. In the next few steps, you will update the /etc/hosts file if the ip address listed isn't the same as what is shown as a result of ifconfig.


biadmin@ibmclass:~/Desktop
File Edit View Search Terminal Help

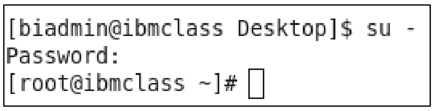
```
[biadmin@ibmclass Desktop]$ ifconfig
eth1      Link encap:Ethernet  HWaddr 00:0C:29:AA:13:95
          inet addr:192.168.157.134  Bcast:192.168.157.255  Mask:255.255.255.0
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:476120 errors:0 dropped:0 overruns:0 frame:0
          TX packets:193714 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:580507023 (553.6 MiB)  TX bytes:12003276 (11.4 MiB)

lo        Link encap:Local Loopback
          inet addr:127.0.0.1  Mask:255.0.0.0
          UP LOOPBACK RUNNING  MTU:65536  Metric:1
          RX packets:5298589 errors:0 dropped:0 overruns:0 frame:0
          TX packets:5298589 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:0
          RX bytes:3187466232 (2.9 GiB)  TX bytes:3187466232 (2.9 GiB)

[biadmin@ibmclass Desktop]$ 
```

5. Switch to the root user using the password **dalvm3**. Type in:

su -

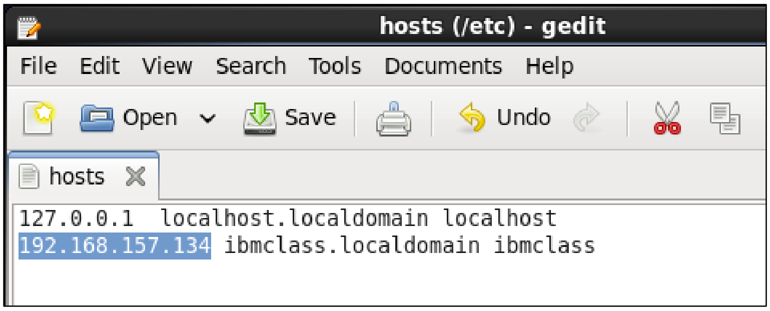


```
[biadmin@ibmclass Desktop]$ su -
Password:
[root@ibmclass ~]# 
```

6. Use your favorite text editor to open up the **/etc/hosts** file. I will be using gedit. Type in:

```
gedit /etc/hosts
```

7. Update the ip address to that of which was listed when you ran the *ifconfig* command.



```
127.0.0.1   localhost.localdomain localhost
192.168.157.134 ibmclass.localdomain ibmclass
```

8. Save and close that file.

9. Close the terminal window.

10. Open the **Firefox** web browser.
11. The login to **Ambari** is **admin/admin**. Go ahead and log in.
12. All the services on the left side should be showing a red triangle with an exclamation mark inside ⚠.

    This means that the services have not been started. If it is a yellow icon, it means that the IP address was not resolved since you updated the */etc/hosts* file, so you need to wait until it turns red before you can start up the services.

    Ambari Services Required:

    HDFS

    MapReduce2

    YARN

    Knox (also start the Demo LDAP service)

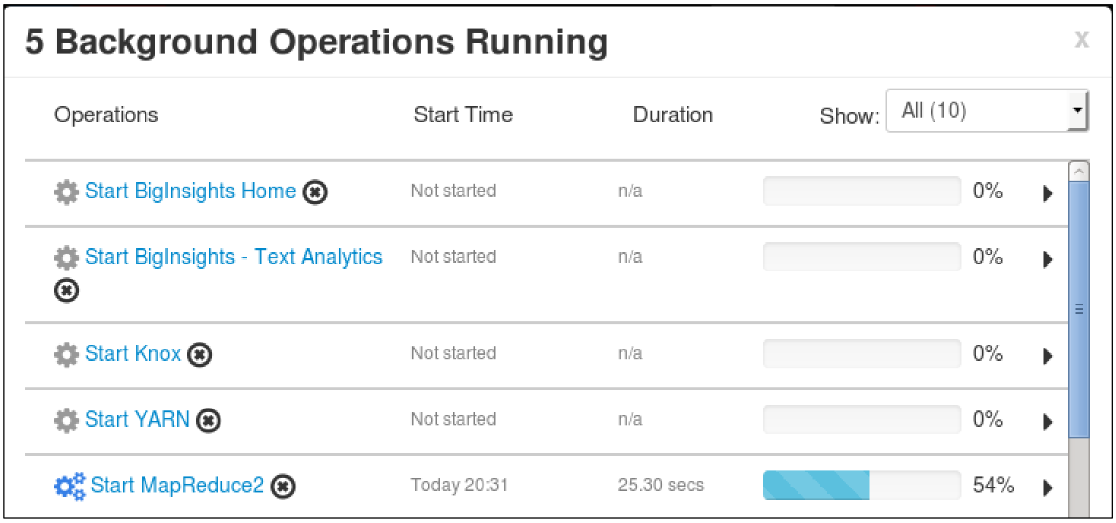
    BigInsights - Text Analytics

    BigInsights - Home

    Start up the services listed above by doing the following:

    - Select the service from the left side panel.
    - Click the **Service Actions** button from the right side.

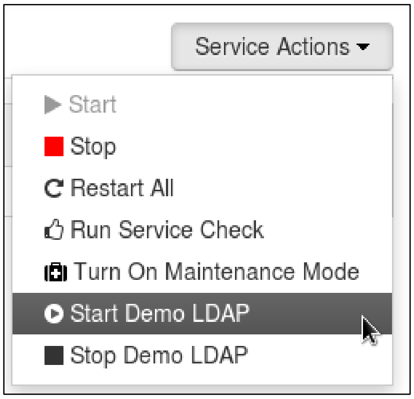
    - Depending on the service, you will have a number of options. Select the **Start** action to start up that service.
    - Confirm the start action in the popup.
    - Do the same thing for all the other services listed.
    - You can queue up each of these actions and view them in the background operations queue.



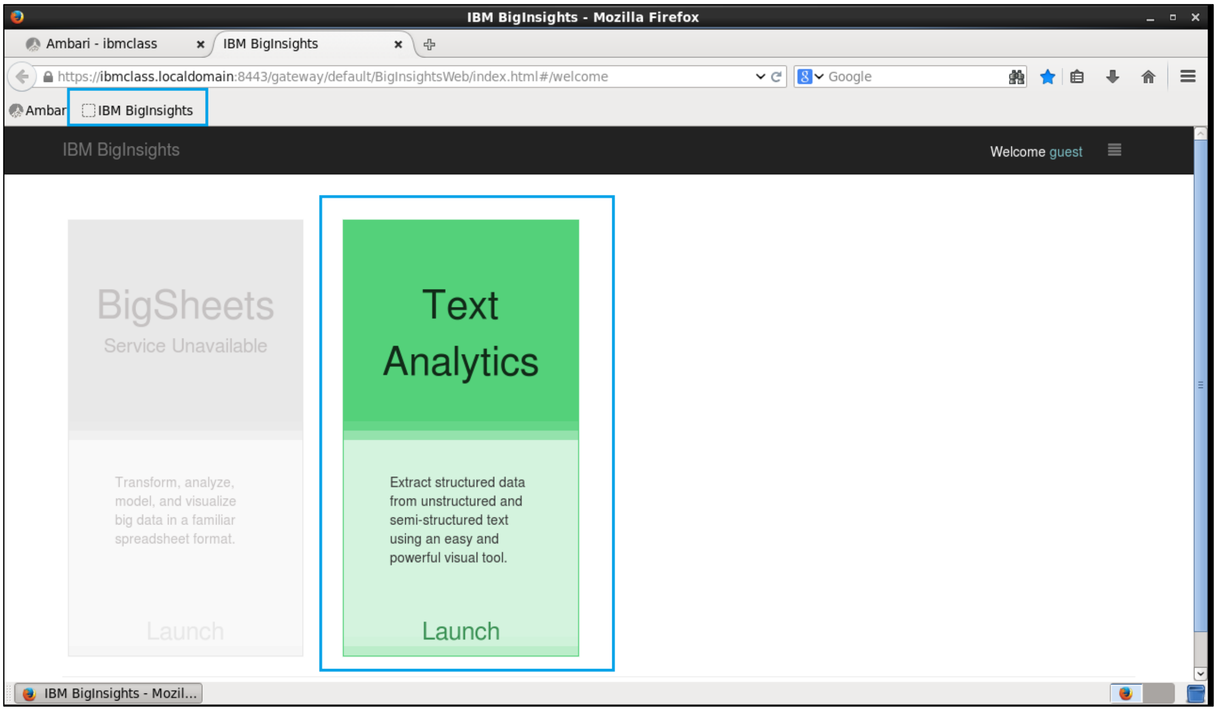
5 Background Operations Running

Operations	Start Time	Duration	Show: All (10)	
Start BigInsights Home	Not started	n/a	0%	
Start BigInsights - Text Analytics	Not started	n/a	0%	
Start Knox	Not started	n/a	0%	
Start YARN	Not started	n/a	0%	
Start MapReduce2	Today 20:31	25.30 secs	54%	

- When all of these services have started, click on the Knox service to start up start the **Demo LDAP** service under its **Service Action** menu. You need this service to authenticate into the BigInsights Home page.



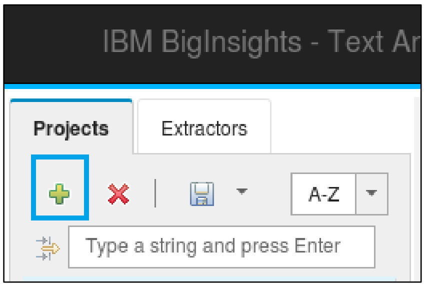
13. Once that has started, open up a new browser tab.
14. Navigate to the **BigInsights Home** page:
    https://ibmclass.localdomain:8443/gateway/default/BigInsightsWeb/index.html
    There is a bookmark already set up with this. The bookmark name is **IBM BigInsights**
15. Log in using **guest/guest-password**.


IBM BigInsights - Mozilla Firefox
Ambari - ibmclass
IBM BigInsights
https://ibmclass.localdomain:8443/gateway/default/BigInsightsWeb/index.html#/welcome
Google
Ambari
IBM BigInsights
IBM BigInsights
Welcome guest
BigSheets
Service Unavailable
Transform, analyze, model, and visualize big data in a familiar spreadsheet format.
Launch
Text Analytics
Extract structured data from unstructured and semi-structured text using an easy and powerful visual tool.
Launch
IBM BigInsights - Mozil...

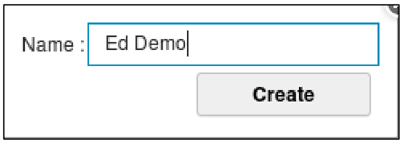
16. Click the **Text Analytics** link to bring up the Web UI.

Task 2. Creating a new Text Analytics project.

1. On the Projects tab on the left, create a new project. Click the **green plus** icon.

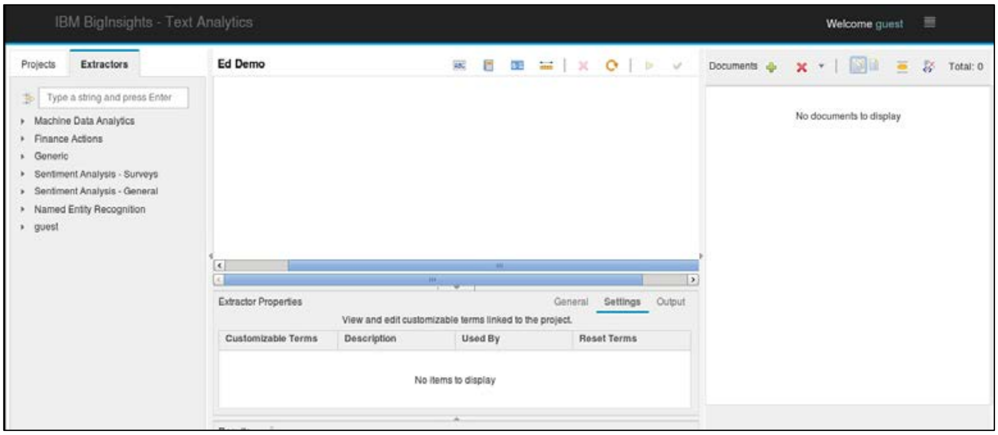


2. Name the project **Ed Demo.**



Name : Ed Demo

3. Click the **Create** button to create the project.



No documents to display

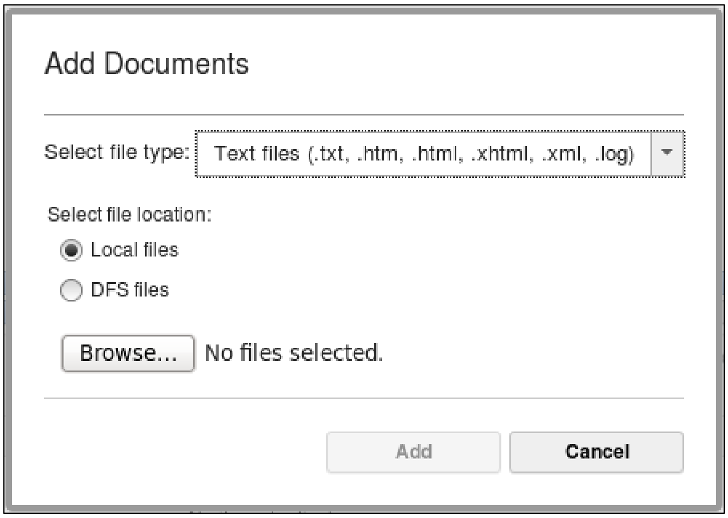
View and edit customizable terms linked to the project.

Customizable Terms	Description	Used By	Reset Terms

No items to display

Task 3. Importing documents into the Web UI.

1. On the **Documents** pane on the right. Click the **green plus** to add documents to the project.



Add Documents

Select file type: Text files (.txt, .htm, .html, .xhtml, .xml, .log)

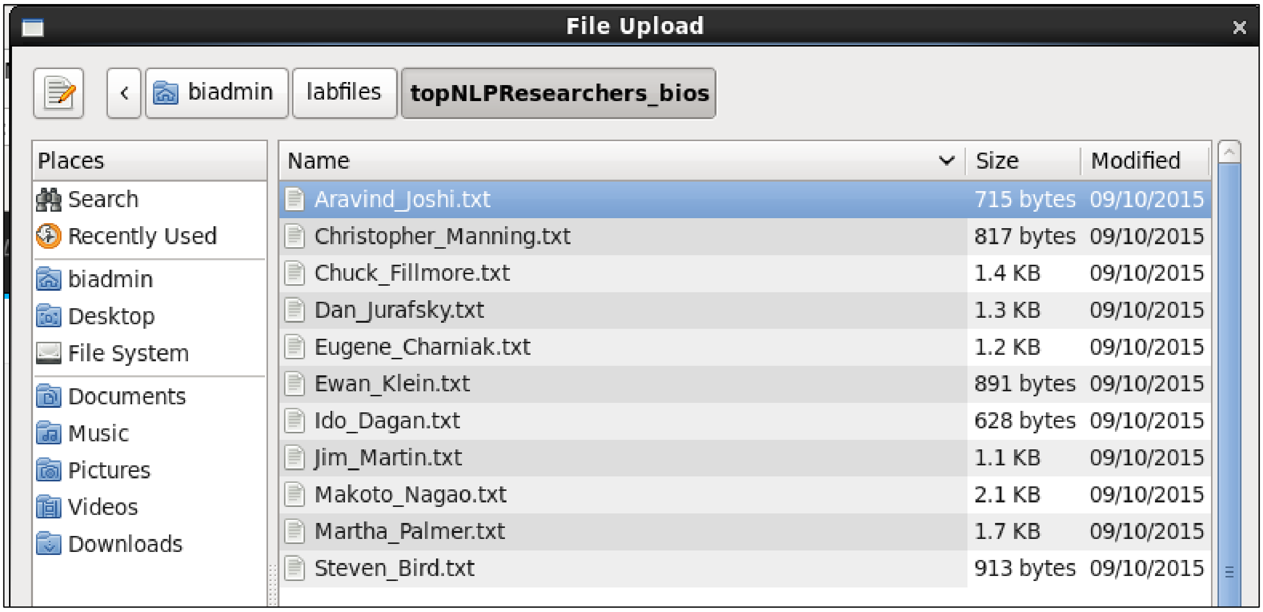
Select file location:

● Local files

○ DFS files

Browse... No files selected.

2. The file type is **Text files**. In our case, the files are located on the local filesystem. Alternatively, you can load files that are residing on your Distributed File System (DFS). Click **Browse** to select your files.
3. Navigate to */home/biadmin/labfiles/topNLPResearchers_bios/*

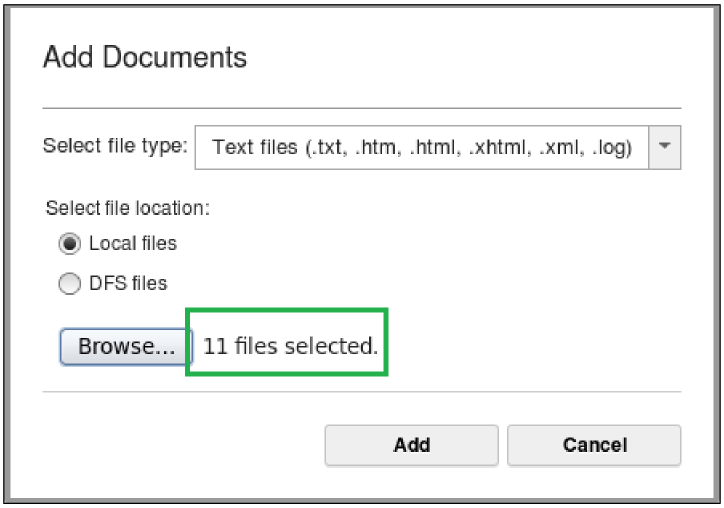


File Upload

biadmin / labfiles / **topNLPResearchers_bios**

Name	Size	Modified
Aravind_Joshi.txt	715 bytes	09/10/2015
Christopher_Manning.txt	817 bytes	09/10/2015
Chuck_Fillmore.txt	1.4 KB	09/10/2015
Dan_Jurafsky.txt	1.3 KB	09/10/2015
Eugene_Charniak.txt	1.2 KB	09/10/2015
Ewan_Klein.txt	891 bytes	09/10/2015
Ido_Dagan.txt	628 bytes	09/10/2015
Jim_Martin.txt	1.1 KB	09/10/2015
Makoto_Nagao.txt	2.1 KB	09/10/2015
Martha_Palmer.txt	1.7 KB	09/10/2015
Steven_Bird.txt	913 bytes	09/10/2015

4. Select all 11 files and click **Open.**



Add Documents

Select file type: Text files (.txt, .htm, .html, .xhtml, .xml, .log)

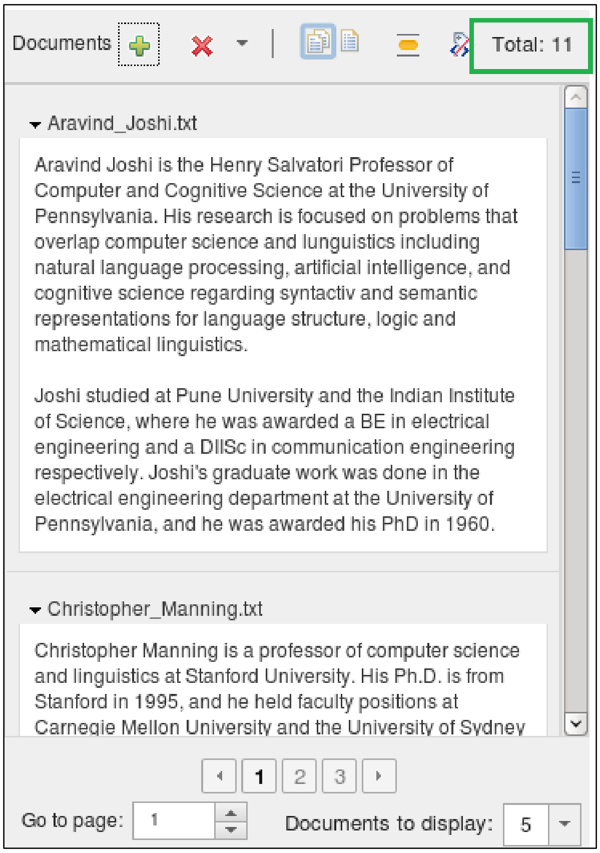
Select file location:

● Local files

○ DFS files

Browse... 11 files selected.

5. Finally, click **Add** to load the files into your project.
6. The files will show up on the **Documents** pane and are ready to be used.



Aravind_Joshi.txt

Aravind Joshi is the Henry Salvatori Professor of Computer and Cognitive Science at the University of Pennsylvania. His research is focused on problems that overlap computer science and lunguistics including natural language processing, artificial intelligence, and cognitive science regarding syntactiv and semantic representations for language structure, logic and mathematical linguistics.

Joshi studied at Pune University and the Indian Institute of Science, where he was awarded a BE in electrical engineering and a DIISc in communication engineering respectively. Joshi's graduate work was done in the electrical engineering department at the University of Pennsylvania, and he was awarded his PhD in 1960.

Christopher_Manning.txt

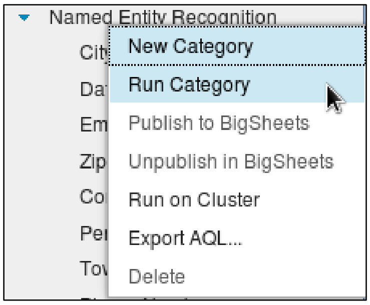
Christopher Manning is a professor of computer science and linguistics at Stanford University. His Ph.D. is from Stanford in 1995, and he held faculty positions at Carnegie Mellon University and the University of Sydney

Go to page: 1

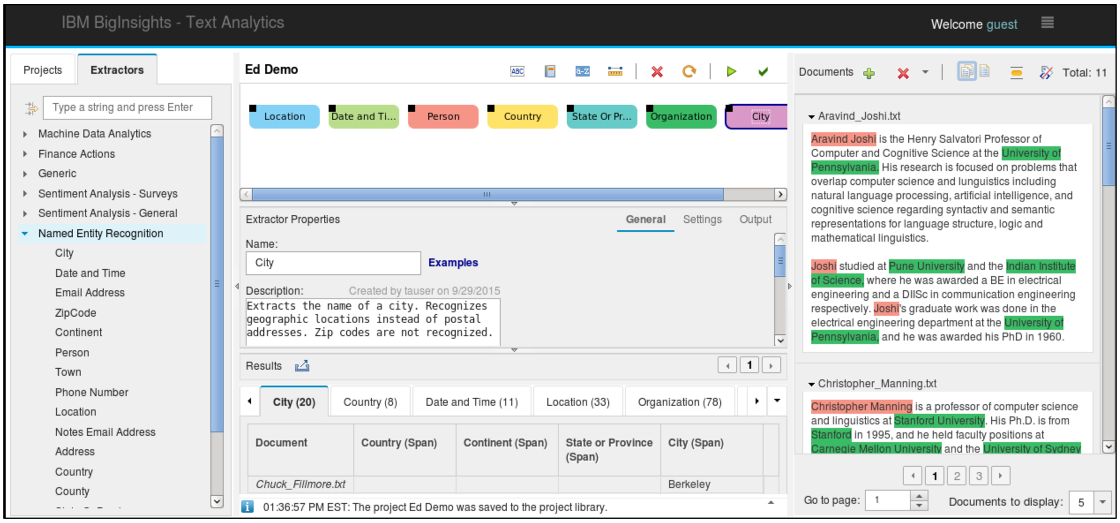
Documents to display: 5

Task 4. Running your first extractor.

1. Under the **Extractors** tab, expand the **Named Entity Recognition** category to see all the pre-built extractors underneath.
2. Run all of the extractors underneath the **Named Entity Recognition** category. Right-click the category and select **Run Category**.



3. When the run finishes, you will see something similar to this:



Name: City

Description: Created by tauser on 9/29/2015

```
Extracts the name of a city. Recognizes
geographic locations instead of postal
addresses. Zip codes are not recognized.
```

Document	Country (Span)	Continent (Span)	State or Province (Span)	City (Span)
Chuck_Fillmore.txt				Berkeley

01:36:57 PM EST: The project Ed Demo was saved to the project library.

Aravind_Joshi.txt

Aravind Joshi is the Henry Salvatori Professor of Computer and Cognitive Science at the University of Pennsylvania. His research is focused on problems that overlap computer science and lunguistics including natural language processing, artificial intelligence, and cognitive science regarding syntactiv and semantic representations for language structure, logic and mathematical linguistics.

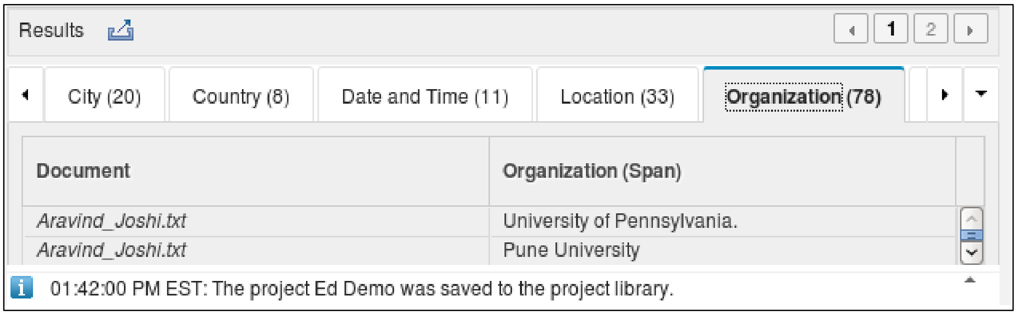
Joshi studied at Pune University and the Indian Institute of Science, where he was awarded a BE in electrical engineering and a DIISc in communication engineering respectively. Joshi's graduate work was done in the electrical engineering department at the University of Pennsylvania, and he was awarded his PhD in 1960.

Christopher_Manning.txt

Christopher Manning is a professor of computer science and linguistics at Stanford University. His Ph.D. is from Stanford in 1995, and he held faculty positions at Carnegie Mellon University and the University of Sydney

Task 5. Examining the results of the run.

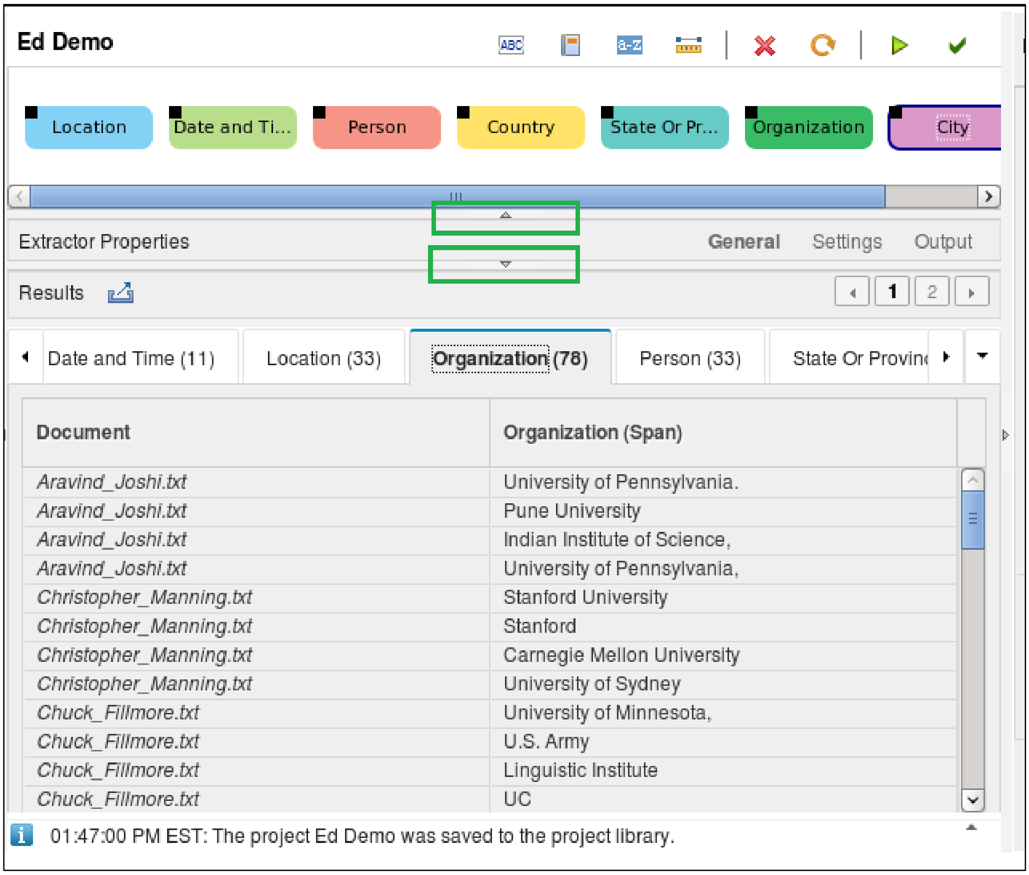
1. Each of those items on the canvas are the various extractors. The one of interest to us now is the **Organization**. In my output, it is the green rectangle. Yours may be different. On the **Results** pane at the bottom of the canvas, click the **Organization** tab to bring up its results.



Document	Organization (Span)
Aravind_Joshi.txt	University of Pennsylvania.
Aravind_Joshi.txt	Pune University

01:42:00 PM EST: The project Ed Demo was saved to the project library.

2. Go ahead and collapse the **Extractor Properties** pane and resize the **Results** pane to make more room. Click on the horizontal bar with the upside-down triangle to toggle the expand/collapse function. The same bar can be used to resize if you bring your mouse cursor over it and then click and drag to resize.



Document	Organization (Span)
Aravind_Joshi.txt	University of Pennsylvania.
Aravind_Joshi.txt	Pune University
Aravind_Joshi.txt	Indian Institute of Science,
Aravind_Joshi.txt	University of Pennsylvania,
Christopher_Manning.txt	Stanford University
Christopher_Manning.txt	Stanford
Christopher_Manning.txt	Carnegie Mellon University
Christopher_Manning.txt	University of Sydney
Chuck_Fillmore.txt	University of Minnesota,
Chuck_Fillmore.txt	U.S. Army
Chuck_Fillmore.txt	Linguistic Institute
Chuck_Fillmore.txt	UC

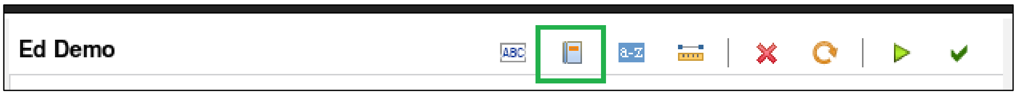
01:47:00 PM EST: The project Ed Demo was saved to the project library.

3. In the canvas, go ahead and delete all the other extractors. Keep the **Organization** one. Select the ones you wish to delete and press the red x. Alternatively, select and press the **Delete** key on your keyboard.

   Note that you can drag and drop extractors along the canvas.

Task 6. Creating a dictionary of clues to search.

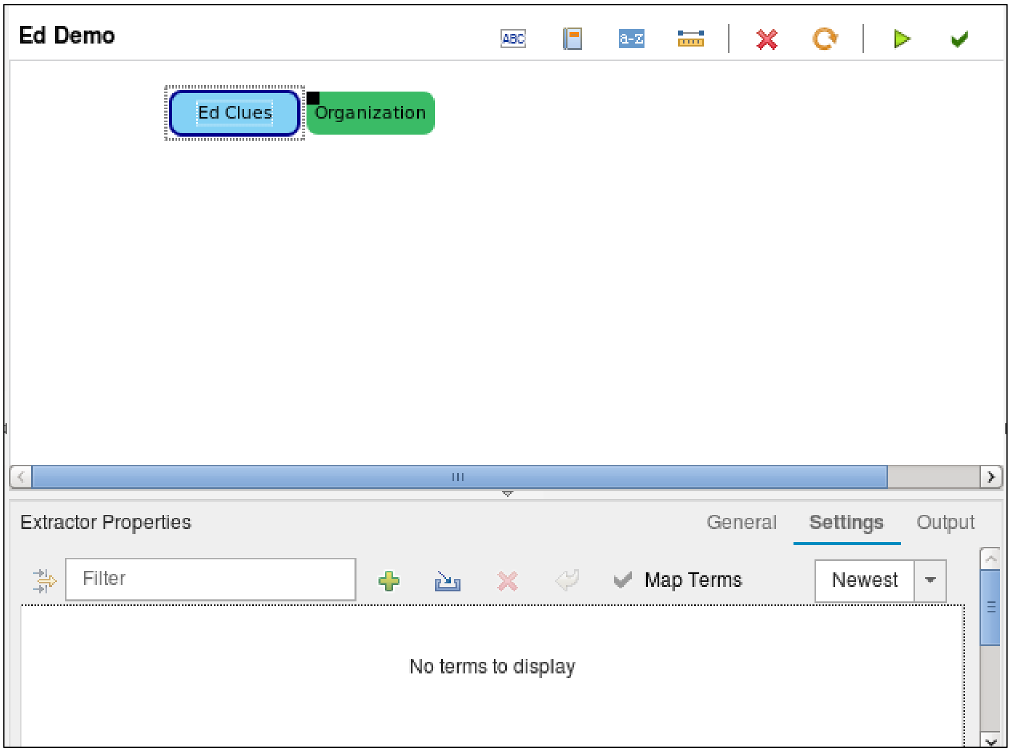
1. At the toolbar above the canvas, click the **New Dictionary** button.



2. On the canvas, the dictionary shows up. Name the dictionary, **Ed Clues.**

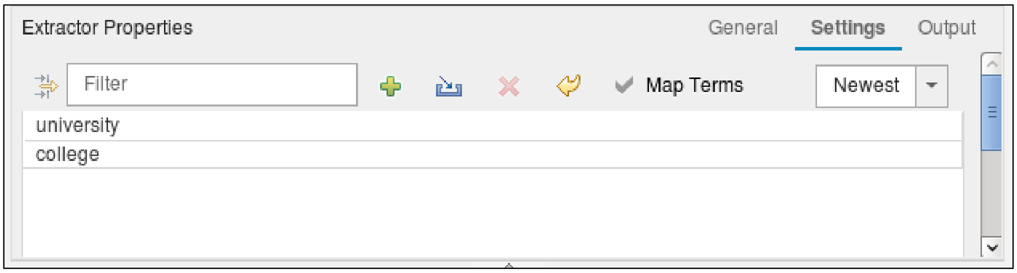
3. Add the following clues into the **Extractor Properties** pane:
   - college
   - university

   Go ahead and collapse the Results pane and expand the Extractor properties pane (if you haven't done so already).



No terms to display

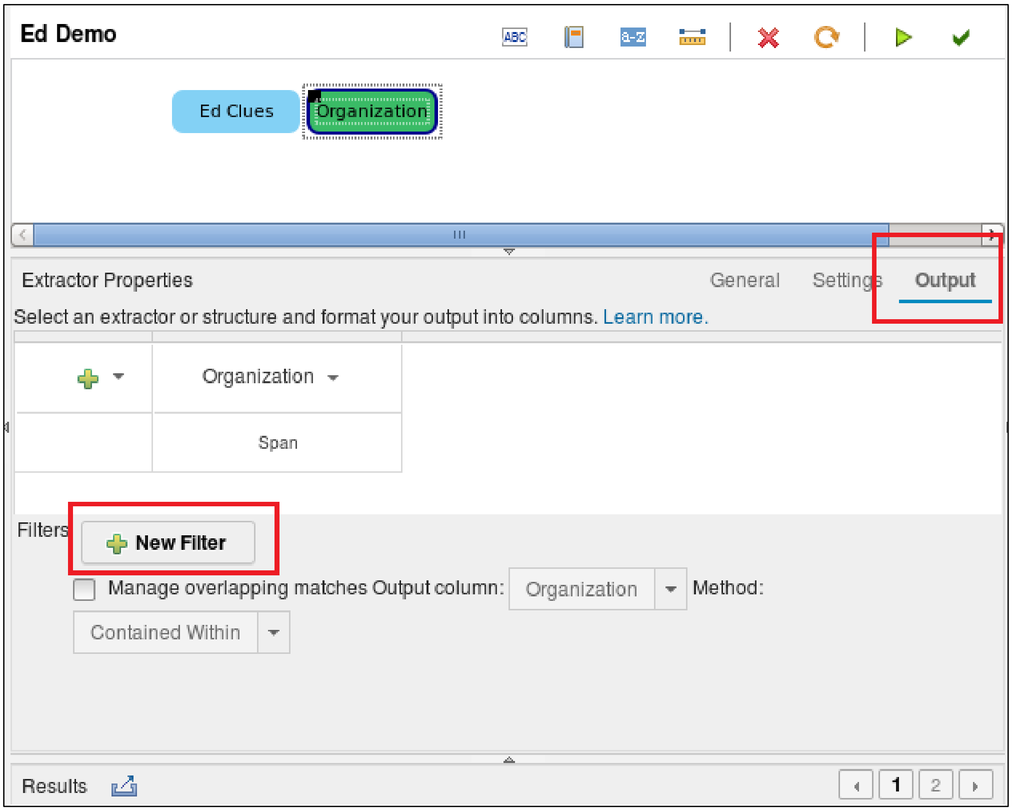
4. To add a new term. Click the **green plus** and add the clues.



university
college

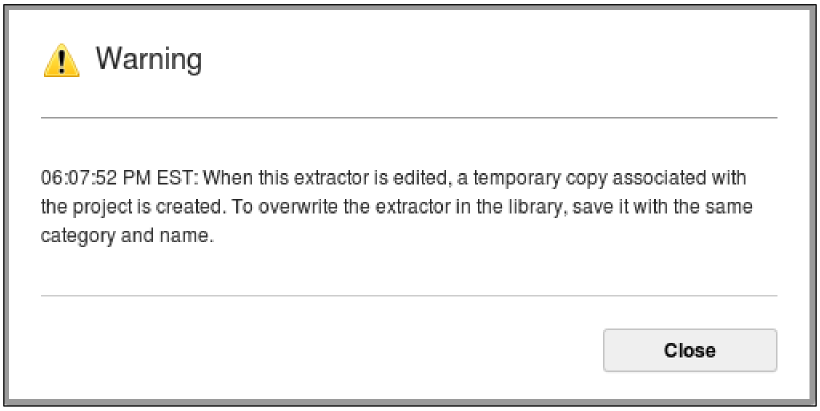
Task 7. Filtering the results and running the updated extractor.

1. On the canvas, click on the **Organization** extractor.
2. Click the **Output** tab.



Note a couple of things. To see all of the properties, I kept the **Results** pane collapsed and resized the **Extractor Properties** pane.

3. As you may have guessed, you are going to filter out rows that do not contain the clues in the dictionary. Click on the **New Filter** button.
4. Click **Close** to get out of that *Warning* dialog.

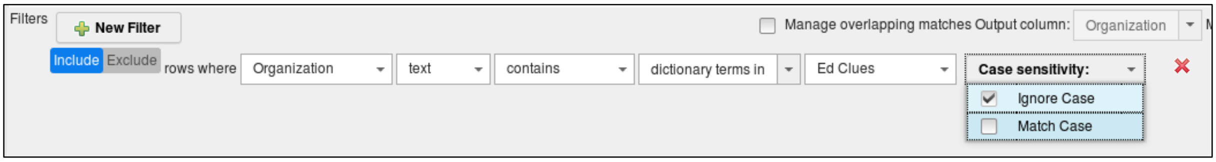


Warning

06:07:52 PM EST: When this extractor is edited, a temporary copy associated with the project is created. To overwrite the extractor in the library, save it with the same category and name.

5. Edit the filter to: **Include** rows where **Organization text contains dictionary terms in Ed Clues (case sensitivity:Ignore Case**).

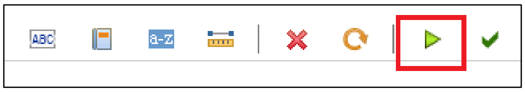
   For the purpose of showing the screenshot: I collapsed the panes to the left and right of it. You can do the same if you need more room to edit the filter.



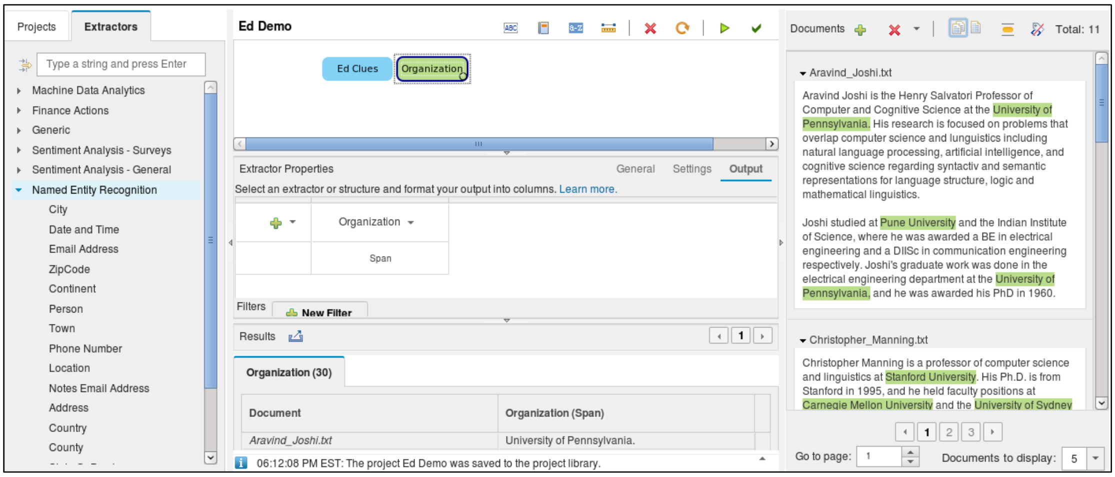
Filters | New Filter | □ Manage overlapping matches Output column: Organization

Include Exclude rows where Organization | text | contains | dictionary terms in | Ed Clues | Case sensitivity: ☑ Ignore Case □ Match Case

6. Before you run the extractor. Restore (expand) the Results pane to see that there are currently 78 rows where it originally matched. Run the extractor by clicking on the **green play arrow.**



7. Now note that the results went from 78 to 30 rows because you are only matching on the terms in the **Ed Clues** dictionary. The rest were filtered out.



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

Organization (30)

Document	Organization (Span)
Aravind_Joshi.txt	University of Pennsylvania.

06:12:08 PM EST: The project Ed Demo was saved to the project library.

Aravind_Joshi.txt

Aravind Joshi is the Henry Salvatori Professor of Computer and Cognitive Science at the University of Pennsylvania. His research is focused on problems that overlap computer science and lunguistics including natural language processing, artificial intelligence, and cognitive science regarding syntactiv and semantic representations for language structure, logic and mathematical linguistics.

Joshi studied at Pune University and the Indian Institute of Science, where he was awarded a BE in electrical engineering and a DIISc in communication engineering respectively. Joshi's graduate work was done in the electrical engineering department at the University of Pennsylvania, and he was awarded his PhD in 1960.

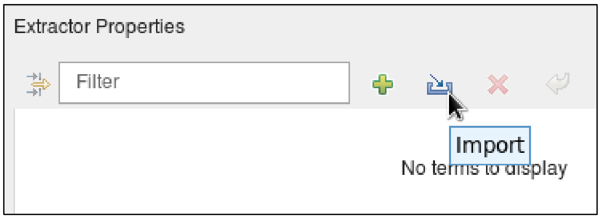
Christopher_Manning.txt

Christopher Manning is a professor of computer science and linguistics at Stanford University. His Ph.D. is from Stanford in 1995, and he held faculty positions at Carnegie Mellon University and the University of Sydney

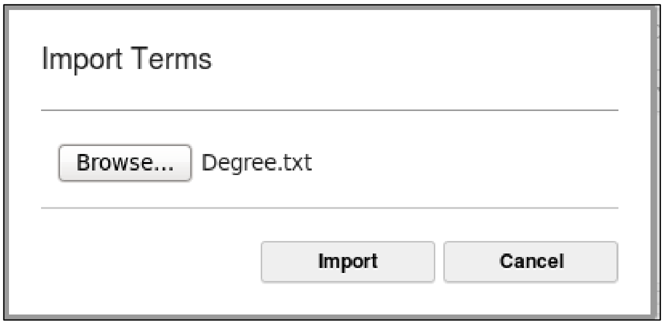
Go to page: 1 Documents to display: 5

Task 8. Creating a dictionary by importing a file.

1. Create a new dictionary by clicking the **New Dictionary** button.
2. Name the dictionary: **Degree**
3. With the **Degree** dictionary selected, click the **Import** button from the **Extractor Properties** pane:



4. Select the **Degree.txt** file from under */home/biadmin/labfiles* and import it.

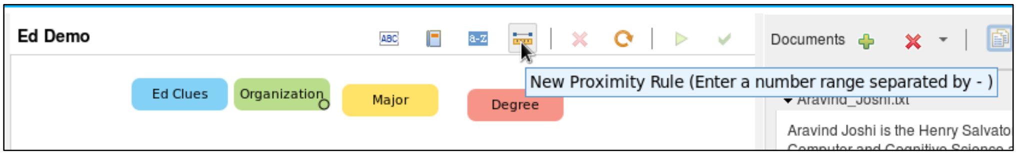


Import Terms

5. Create another dictionary called **Major** and import the **Majors.txt** file.
6. You should now have two new dictionaries created and loaded with terms: **Degree** and **Major**.

Task 9. Creating proximity rules.

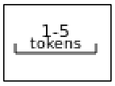
1. Create a Proximity rule. Click the **New Proximity Rule** button.



New Proximity Rule (Enter a number range separated by -)

Ed Demo
Documents
Ed Clues
Organization
Major
Degree
Aravind_Joshi.txt
Aravind Joshi is the Henry Salvato
Computer and Cognitive Science

2. Give the range of the proximity rule of **1 - 5** tokens.



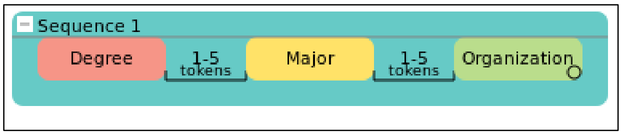
3. Create the same proximity rule again so that you have two proximity rules total.

Task 10. Creating a sequence of extractors.

1. On the canvas, arrange the extractors into a sequence using drag and drop. Arrange the extractors in this order:

   Degree, 1-5 tokens, Major, 1-5 tokens, Organization

   When you drag an extractor or a rule next to another, a blue bar will appear on indicating that it will attach to that side when you let go of the mouse button.
2. When you have done it correctly, you would have created a new sequence:



3. With that sequence selected, run the sequence by clicking the **green play arrow** on the toolbar.

4. Once again, I resized the **Results** pane so that I can view them. Do so if you need to for your environment.

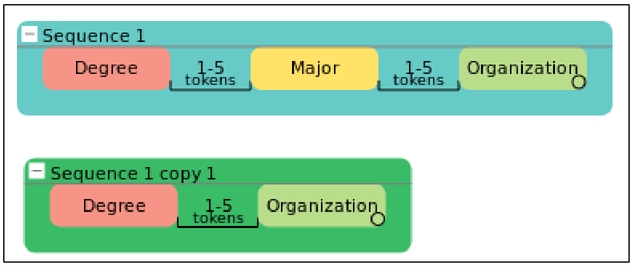
   Notice in the Results pane, there are four tabs. Each of the tabs represent the results from each individual extractors, plus the fourth tab for the sequence of the three extractors with the proximity rules. Because I know the data well, and this is a made up demonstration scenario, I know that in the sequence tab, you are missing the sequence from the University of Michigan "Ph.D in 1961 from the University of Michigan"

5. Right-click on **Sequence 1.** Select **Copy.**

6. Right-click on the canvas and select **Paste as New Copy.**

   Note: **Paste as New Copy** is essentially cloning the original extractor. If you did the normal paste. any changes made to the source will affect the copy as well.

7. In **Sequence 1 copy 1,** remove **Major** and one of the proximity rules by dragging it out of the sequence. You can delete those two items. This is what it should look like now.

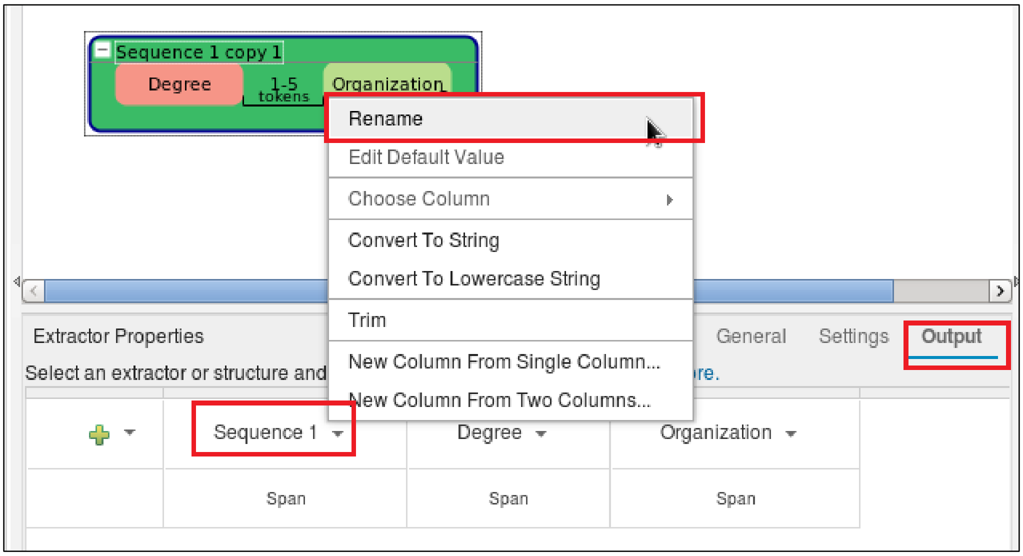


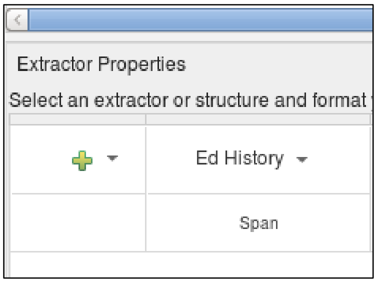
8. Go ahead and run **Sequence 1 copy 1.**

9. Note that the missing entry is now present.

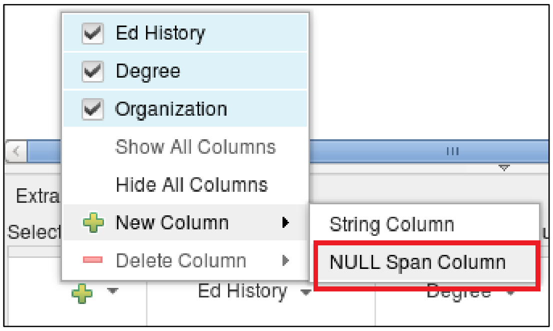
Task 11. Creating a union of extractors.

1. Under the **Extractor Properties**, on the **Output** tab, change the name of the **Sequence 1** column to **Ed History.** Select the **Sequence 1** column options and click **Rename**.



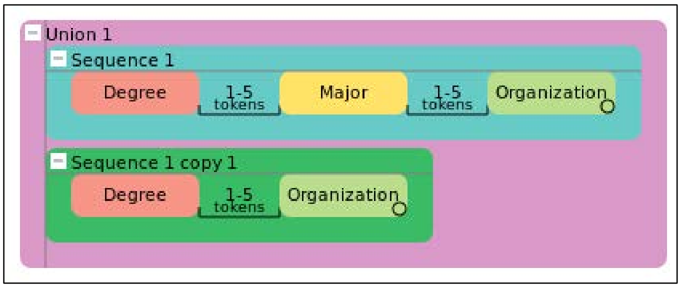


2. Add a new NULL Span Column. Click the **green plus** button → select **New Column → NULL Span Column**.

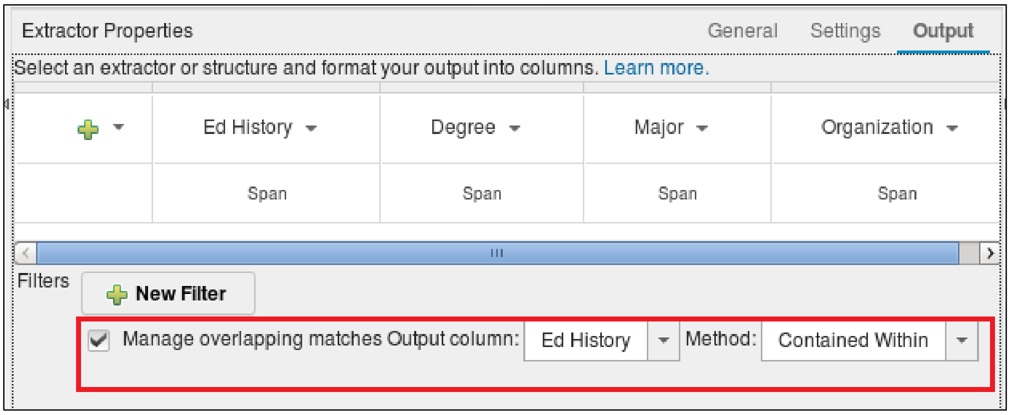


Name the new column Major.

3. At this point, you should have four columns for **Sequence 1 copy 1**.
4. Back on the canvas, click on the **Sequence 1** extractor and rename the **Sequence 1** column to Ed History (just as you had done for the Sequence 1 copy 1 extractor).
5. Now that both extractors have the same schema, drag and drop **Sequence 1** to align vertically with **Sequence 1 copy 1** to create Union 1. The blue bar should be at either the top or the bottom to indicate a union action.



6. Rename **Union 1** to **Education History**. Double-click the text directly on the canvas to rename it.
7. Run **Education History**.
8. Examine the results. There are 11 rows. Notice that there are duplicate values.
9. Back on the **Extractor Properties**, on the **Output tab**, check the box for **Manage overlapping matches.** You may need to resize the properties pane to see it.


Extractor Properties

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

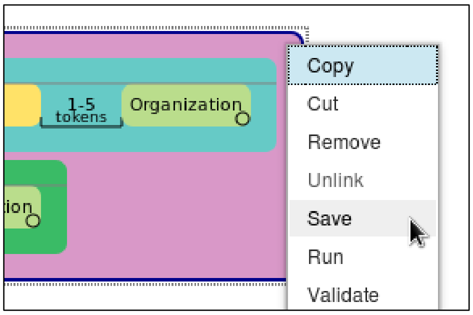
	Ed History	Degree	Major	Organization
	Span	Span	Span	Span

☑ Manage overlapping matches Output column: Ed History Method: Contained Within

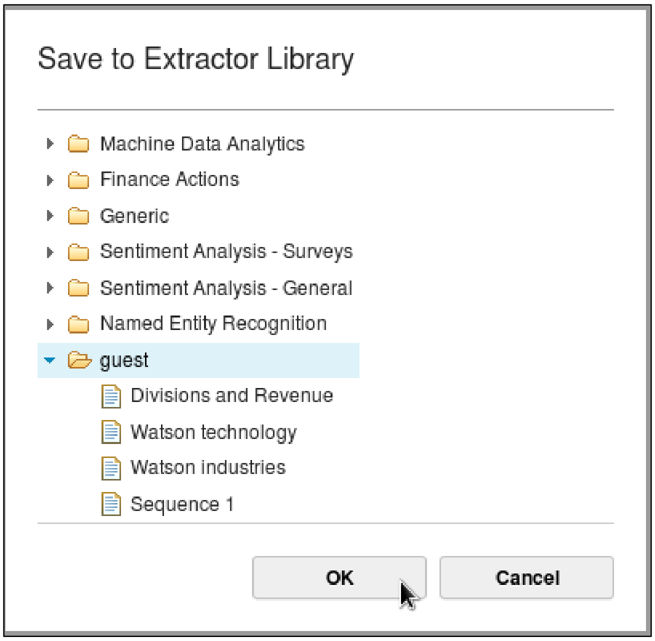
10. Run the **Education History** extractor again. Note that the number of returned rows went from 11 to 7.

Task 12. Saving extractors and exporting results - Optional.

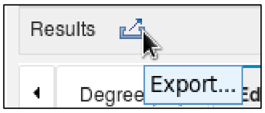
1. Right-click on the **Education History** extractor and select **Save**.



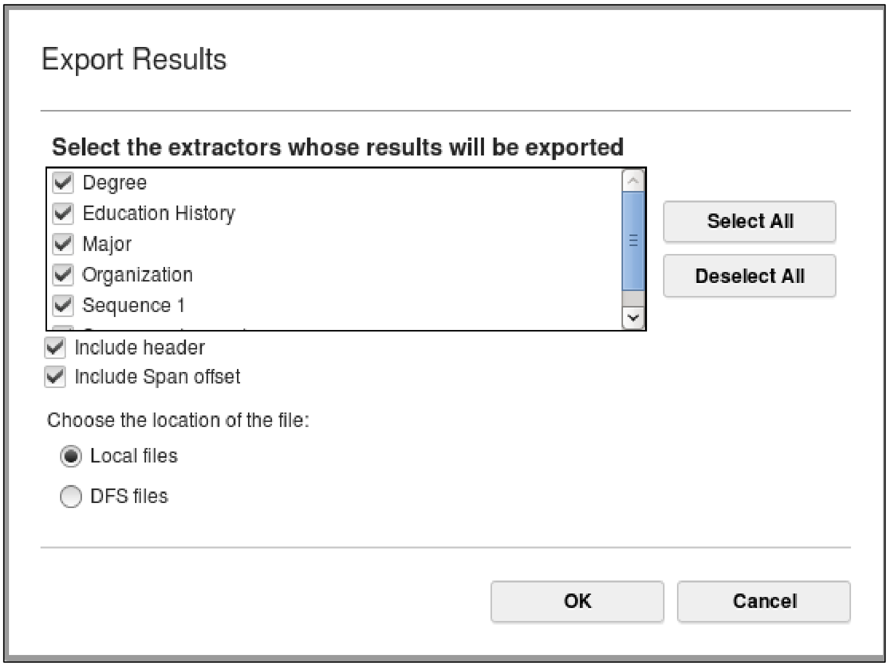
2. Select the **guest** directory and click **OK**.



3. In the Results view, click the export button to export the results to csv format.



4. Select the results, specify your options, choose your location, and click **OK**.



Export Results

Select the extractors whose results will be exported

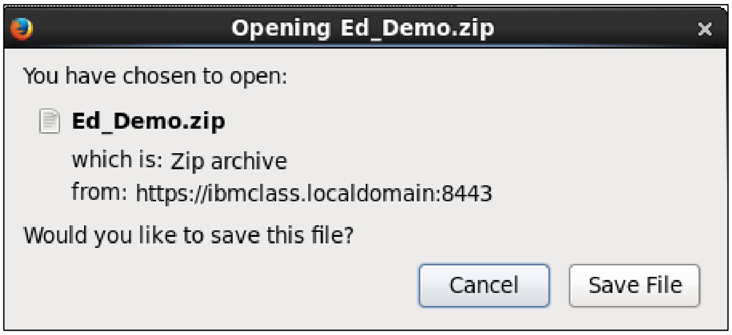
☑ Degree
☑ Education History
☑ Major
☑ Organization
☑ Sequence 1

☑ Include header
☑ Include Span offset

Choose the location of the file:

● Local files
○ DFS files

5. You will be prompted to download the file by your web browser:



Opening Ed_Demo.zip

You have chosen to open:

Ed_Demo.zip

which is: Zip archive

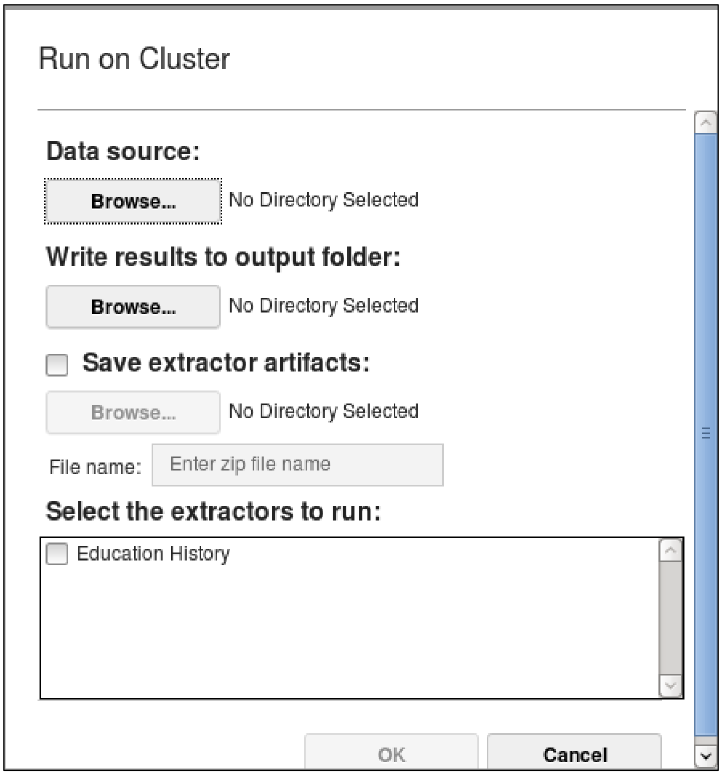
from: https://ibmclass.localdomain:8443

Would you like to save this file?

Task 13. Running on a MapReduce cluster - Optional.

1. From the **Extractor** catalog, expand **guest**, right-click **Education History** and choose **Run on Cluster**

2. Specify the data source, output folder, and the extractor to run.



Run on Cluster

Data source:

Browse... No Directory Selected

Write results to output folder:

Browse... No Directory Selected

□ Save extractor artifacts:

Browse... No Directory Selected

File name: Enter zip file name

Select the extractors to run:

□ Education History

OK Cancel

Task 14. Publishing to BigSheets - Optional.

1. Similar process as pushing the job onto a MapReduce cluster. Make sure that your BigSheets service has started. Since this task is optional, the BigSheets service wasn't required to be started earlier. You can explore this on your own.
2. Note on publishing NULL spans to BigSheets. There is a bug with this release where you will not be able to publish NULL spans to BigSheets. Our Education History extractor contains a NULL span column, if you recall. We added this NULL span column in order to ensure that both of the extractors in the union had the same number of columns. You can still publish the Sequence 1 extractor as a BigSheets function since that one didn't have any NULL spans within.
3. Once published to BigSheets, the extractor will be a function from which you can use to create child worksheets to extract data. From there, you can use BigSheets visualization to paint a picture of your data.

Results:
You created a Text Analytics project which analyzed text data to find the education histories. You created extractors to locate the degree, major, and the organization within the biography files. Then you consolidated and finalized those extractors to create the final Education History extractor. In subsequent demos, you will see how to use each of the individual text analytics steps in more detail.