

PALS Userguide

This document contains instructions to perform the following:

- Use the data-pipeline to scrape data from Alberta Hansard, run it through the TextRank and output to the MySQL DB.
- Perform sentiment analysis on existing summaries.
- Use the web-application.

Data-pipeline

Setup:

Before using the data pipeline, the device in use must be connected to a minio, sql, and ssh client. This is done automatically through a config.ini script, which holds authentication values for each. Please contact an admin to obtain one of those files if you wish to use the pipeline.

Overview:

The data pipeline consists of many files, most of which don't need to be modified for use. If there is an exception to this, it will be mentioned in the specific use case. The two points of interest you may have are...

1. Scripts:

In general each use case is executed by calling one of more scripts. Most of the scripts don't take in arguments, and can be executed from the command shell by calling

```
python {script_name}.py
```

All scripts are found in the data-pipeline/etl directory. The additional folders contain code that the scripts call to perform their tasks.

2. Settings:

You may wish to adjust the settings of certain processes in the pipeline. All of the accepted modifications can be found in various files of the data-pipeline/etl/data, here:

mlas.csv: Contains mla data to be put in the SQL database. This can be modified if an MLA needs to be excluded or added.

parties.csv: Contains party data to be put in the SQL database. This can be modified if a party needs to be excluded or added.

sentences.txt: This file includes null sentences that text rank will ignore and not add to its list of sentences. If you wish to exclude a certain sentence that shouldn't be accounted for in the algorithm and stored in the database, place it here.

stopwords.txt: This file includes stop words that textrank will not include in its calculations. The sentences they belong to will still be included, however these words are ignored when the sentences get tokenized.

run_textrank_parties.py: The number of sentences used for the party rank calculation can be adjusted. Note that this increases the amount of time it will take to run the algorithm. The variable name is called number_of_sentences, and is found at line 21.

To load MLA and Party data:

1. Navigate to the data-pipeline/etl directory in your command shell.
2. Execute the following script by entering in the command line...

```
python load_mlas_from_csv.py
python load_parties_from_csv.py
```

To scrape sitting meta-data from Alberta Hansard:

1. Navigate to the data-pipeline/etl directory in your command shell.
2. Execute the following script by entering in the command line...

```
python load_documents.py
```

3. When executing the script, it will ask you if you would like to append to or overwrite existing data (by entering either a or o). If you select append, the existing documents in SQL and Minio will remain and new documents added to the stores. If overwrite is selected, then they will be removed and a fresh search will start. If neither are entered, the program will terminate.

To scrape all MLA portraits from Alberta Legislative website:

1. Navigate to the data-pipeline/etl directory in your command shell.
2. Execute the following script by entering in the command line...

```
python load_all_images.py
```

All images will then be uploaded to Minio.

To scrape and pre-process all sentences for all sittings from Alberta Hansard (REQ1, REQ2, REQ3):

1. Navigate to the data-pipeline/etl directory in your command shell.
2. Execute the following scripts by entering in the command line...

```
python extract_speeches.py
```

To perform TextRank summarization for each MLA (REQ5):

3. Navigate to the data-pipeline/etl directory in your command shell.
4. Execute the following script by entering in the command line...

```
python run_textrank_mlas.py
```

All textrank summarization will then automatically execute. Note that this task often takes around 30 minutes.

5. If this is the first time setting up the database, it is required that SQL views are created. If textrank is being run for the second time, this step is unnecessary. To generate views, the following scripts must be executed.

```
python create_summary_views_mlas.py  
python create_all_top_summaries_view.py
```

To perform TextRank summarization for Parties (REQ6):

1. Navigate to the data-pipeline/etl directory in your command shell.
2. Execute the following script by entering in the command line...

```
python run_textrank_parties.py
```

All party summarization will then automatically execute. The length of this task varies by the amount of sentences that are used for textrank by each MLA. To modify this, one can modify the `number_of_sentences` variable at the top of the `run_textrank_parties.py` file.

To perform obtain most similar/dissimilar MLAs (REQ11):

1. Navigate to the data-pipeline/etl/topic_analysis/jst/Debug directory in your terminal.

2. Compile the jst project by issuing a “make” command. To do so, you will need to install a C++ compiler. Please see the following for instructions on how to do so:

<https://www.cs.odu.edu/~zeil/cs250PreTest/latest/Public/installingACompiler/>

3. Return back to the data-pipeline/etl directory
4. Execute:
`python load_data_for_jst.py analyze`
5. Execute:
`./topic_analysis/jst/Debug/jst -inf -config topic_analysis/jst/analyze.properties`
6. Execute:
`python jst_analysis.py analyze`
7. Execute the following command:

`python jst_analysis.py analyze`

Web-application

To navigate to the web application, go to this link: <http://summarizer-ab.ca/>

To use the interactive map (REQ12):

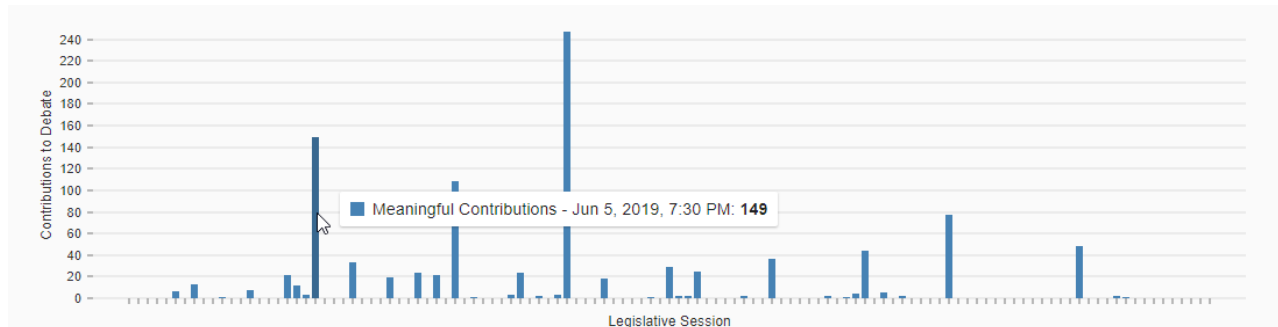
1. You can select MLAs using the map:
 - a. The interactive map is split into ridings.
 - b. Hover over and click the riding you want.
2. You can zoom into the map and move around:
 - a. To zoom in and out, use your scroll wheel in the map window.
 - b. To move around the map, click and drag around in the map window.

To view an individual MLA's information, summaries and most similar/dissimilar MLAs (REQ7, REQ11, REQ15):

1. Select an MLA using the instructions in “To use the interactive map” or in “To select MLA from search page”.
2. On the top right, the MLA info and summaries are displayed in the MLA Profile card.
3. Click the “MLA Info” tab to see the MLA name, political party, riding, photo, phone # and email.
4. Click the “Summaries” tab to see the summaries for the MLA in tabular format ordered by their ranking from TextRank.
 - a. You will initially see the top most rated summary. You can view additional summaries by clicking on the “>” or “<” buttons on the bottom right of the MLA Profile card.
 - b. You can change the “Rows per page” by clicking on the drop down and selecting a new value. This will change the number of summaries displayed per page.

5. Click the “MLA Comparisons” tab to see the most similar and dissimilar MLAs in terms of sentiment for the chosen MLA.

Involvement over time graph



- The bar graph displays the number of summaries for the MLA at any given date.
- You can hover over each bar to display the exact number of summaries (meaningful contributions) for a given date time.


To use the interactive involvement over time graph (REQ8, REQ10):

1. Ensure you have an MLA selected.
2. Click on the “Summaries” tab to view the summaries for that MLA.
3. Now, you can select a session to look at from the graph by hovering over and clicking on a bar in the graph.
4. When clicked, the bar will turn orange to indicate that only summaries from the session will be displayed.
5. You can click multiple bars to display summaries from multiple sessions at a time.
6. If you no longer want to view summaries for a session, click the same bar again and it will turn blue to indicate that summaries from the session are no longer being displayed.

To navigate to the search page, go to this link: <http://summarizer-ab.ca/search>

1. You can also navigate to the search page by clicking on the magnifying glass icon on the navigation bar to the right.

Summaries are displayed in a tabular format with the following information:

Rows per page: 5 1-5 of 1000 < >				
Person	MLA Rank	Party Rank	Date ↓	Summary
 Whitney Issik	4	25	2/27/2020	<p>Given that communities like those surrounding the Industrial Heartland rely on the jobs and economic benefits that come from industrial innovation and continued growth from our energy sectors and given that Alberta's low natural gas prices have created an opportunity for natural gas products to add value to industries like the petrochemical sector, to the associate minister: what is this government doing to encourage corporations like those in the petrochemical sector to consider Alberta a viable jurisdiction for long-term investment? ></p>

- Person: Contains image of MLA and name.
- MLA Rank: The rank of the sentence relative to that MLA's other sentences. For example, an MLA Rank of 1 means that sentence best summarizes the ideas of that specific MLA.
- Party Rank: The rank of the sentence relative to that Party's total sentences. For example, a Party Rank of 1 means that sentence best summarizes the ideas of that specific party that the MLA is in.
- Date: The date of the sitting that this sentence was spoken.

To filter summaries by riding, MLA, token and date (REQ13, REQ14):

1. Navigate to the search page using the above instructions.
2. To filter by:
 - a. Riding - Click the combo box for "Please select a riding" and select a riding.
 - b. MLA - Click the combo box for "Please select an MLA" and select an MLA.
 - c. Token - Enter the text you want to find in the list of summaries. **Note:** Token search will only find summaries that **contain** the searched text.
 - d. Date - Search for summaries in a date range **from** a specific date **to** a specific date. To change the date, hover over the date field and click the triangle to display a Date picker box. Select the date using this Date picker.

Note: The web application initially loads 1000 summaries. If you search for a summary that is not within this initial list, the application will request additional summaries up to 11000 until it finds a matching summary.

To filter summaries by Party (REQ9, REQ16):

1. Navigate to the search page using the above instructions.
2. To filter by party, click on the party you want to filter by on the Party button toggle.
3. The results will update to only show summaries for MLAs in that party.
4. You can view the top ranked summaries for that party by ordering by Party Rank using the instructions below.

To order summaries by MLA Rank, Party Rank and date (REQ16):

1. Navigate to the search page and ensure you have some summaries displayed.
2. To order by:
 - a. MLA Rank: Hover over the MLA Rank cell in the search results table header, you should see an arrow pointing up appear. Click once to sort by ascending MLA Rank and again to sort by descending MLA Rank.
 - b. Party Rank: Similar to above, except hover over the Party Rank cell.
 - c. Date: Similar to above, except hover over the Date cell.

To view the original statement from a summary (REQ17):

1. Ensure you have some summaries displayed.
2. Hover over the text of the summary you want to view, the text should become underlined and dark blue.
3. Click the summary to open a tab for the Hansard document the summary was pulled from.

To search for an MLA and then view them on the front page (REQ10, REQ13):

1. Ensure you have some summaries displayed.
2. Hover over the portrait of the MLA you want to view, the picture should darken slightly to show you are hovering over it.
3. Click the portrait to view the MLA on the front page including their meta-data information, summaries and interactive involvement over time graph.