**Project done by SHUBHAM SHAILESH PANDEY, UBID - spandey4**

This document will clarify the steps you need to take in order to recreate the output I got –

I have collected data for 3 topics – Cambridge Analytica, Syria and Gun Violence.

This is the process for Cambridge Analytica. You can work on the other 2 topics similarly.

* In lab2->Part2 folder you will see two files - **'NYTimesData.ipynb'** and **'TwitterData.ipynb'**. Simply execute them on Jupyter to see 2 new files formed in the same directory. This is how I've collected data over the weeks.
* Collection of tweets files is in lab2->Part2->Cambridge Analytica-> **'TwitterData'** folder. Collection of News article files is in lab2->Part2->Cambridge Analytica **->'NewsData'** folder.
* Now, our aim is to get word count for either datasets. For this, I have **'mapper.py'** and **'reducer.py'** files ready. Locate them in lab2->Part2 folder.
* Now copy **'TwitterData'** folder, **'NewsData'** folder, **'mapper.py'** and **'reducer.py'** to Hadoop VM Desktop.
* Locate the file **'commands to execute for count.txt'** in the lab2->Part2->Cambridge Analytica folder. This txt file contains all commands to be sequentially executed in order to get the word count output.
* Open a terminal on the Desktop in VM. Execute each command listed in **'commands to execute for count.txt'** one by one.
* You will see two new folders on your Hadoop VM Desktop. The folders are **'TwitterWords'** and **'NewsWords'**. The same folders have been copied by me in the lab2->Part2->Cambridge Analyticafolder.
* Now we have our tab separated output files, we need to use them for displaying on a webpage. All the webpage related files can be found in the **'wordcloud'** folder. Simply, double clicking html files will NOT work. You need to setup a local server and copy the **'wordcloud'** folder to its directory to visualize the data.

**How to setup a local server -**

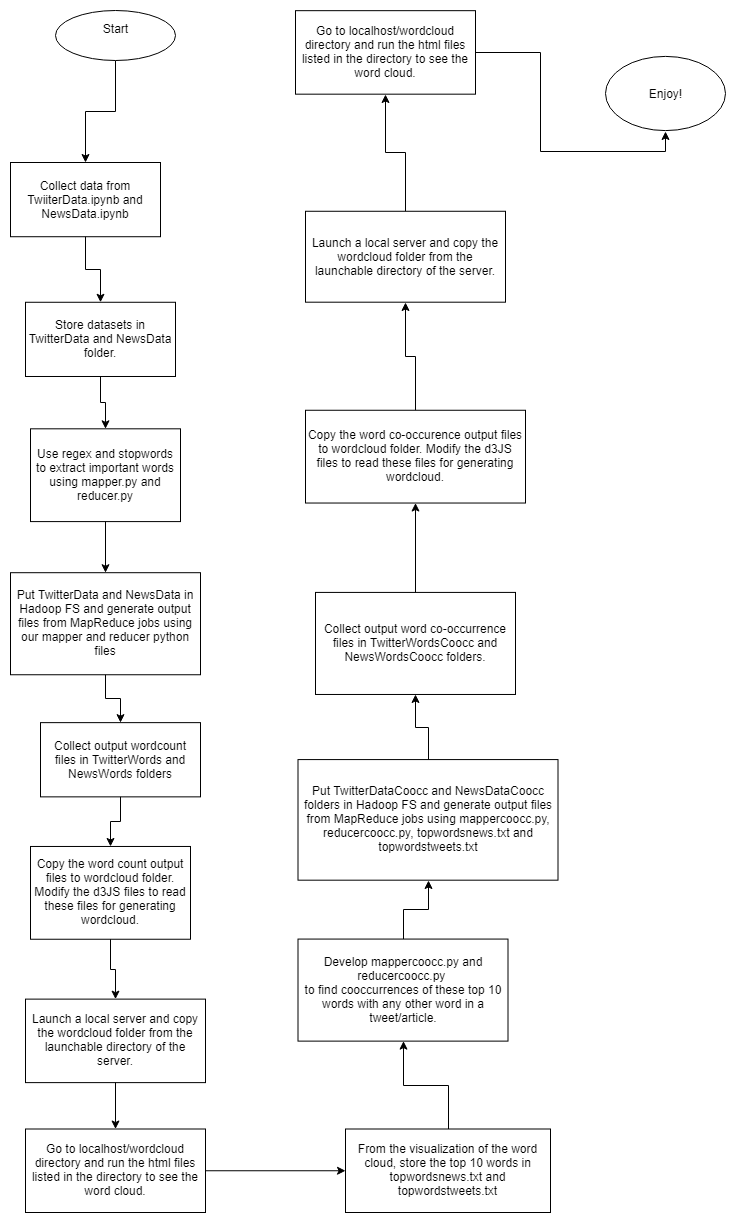
* Download XAMPP. Go to this URL [https://www.apachefriends.org/download.html](https://www.apachefriends.org/download.html%20) and download the version based on your OS.
* Open XAMPP Control Panel application after installation has completed. Click 'Start' for Apache service.
* Go to the location where XAMPP was installed. You will see a folder named **'htdocs'**. Copy the **'wordcloud'** folder inside the **'htdocs'** folder. Now, open your browser and go to "<http://localhost/wordcloud/index.html>”

**THIS CONCLUDES OUR WORD COUNT PART. INSTRUCTIONS FOR WORD CO-OCCURRENCE FOLLOW.**

* Tweet file is in lab2->Part2->Cambridge Analytica->**'TwitterDataCoocc'** folder. News article files is in lab2->Part2->Cambridge Analytica->**'NewsDataCoocc'** folder.
* Now, our aim is to get word co-occurrence for either datasets. For this, I have **'mappercoocc.py'** and **'reducercoocc.py'** files ready. They are in lab2->Part2 folder.
* We have to find co-occurrences for the 10 most frequent words. I thought it would be better to fetch these words from a file rather than hard-coding them. Hence, the files for the top 10 words for each dataset are **'topwordsnews.txt '** and **'topwordstweets.txt’**. They can be located in lab2->Part2->Cambridge Analyticafolder.
* Now copy **'TwitterDataCoocc'** folder, **'NewsDataCoocc'** folder, **'topwordsnews.txt'**, **'topwordstweets.txt'**, **'mappercoocc.py'** and **'reducercoocc.py'** to Hadoop VM Desktop.
* Locate the file **'commands to execute for cooccurence.txt'** in the lab2->Part2->Cambridge Analyticafolder. This txt file contains all commands to be sequentially executed in order to get the word co-occurrence output.
* Open a terminal on the Desktop in VM. Execute each command listed in **'commands to execute for cooccurence.txt'** one by one.
* You will see two new folders on your Hadoop VM Desktop. The folders are **'TwitterWordsCoocc'** and **'NewsWordsCoocc'**. The same folders have been copied by me in the lab2->Part2->Cambridge Analyticafolder.
* Now we have our tab separated output files, we need to use them for displaying on a webpage. All the webpage related files can be found in the **'wordcloud'** folder. Simply, double clicking html files will NOT work. You need to setup a local server and copy the **'wordcloud'** folder to its directory to visualize the data.
* Open XAMPP Control Panel application. Click 'Start' for Apache service.
* Go to the location where XAMPP was installed. You will see a folder named **'htdocs'**. Copy the **'wordcloud'** folder inside the **'htdocs'** folder. Now, open your browser and go to "<http://localhost/wordcloud/index.html>”

**PLEASE SEE THE VIDEO IF ANY CONFUSION STILL EXISTS.**

**BELOW IS THE WORKFLOW DIAGRAM I USED FOR THIS PROJECT.**

****