CSE 587 DATA INTENSIVE COMPUTING
DATA EXPLORATION, BIG DATA ANALYSIS USING HADOOP

Sravanthi Adibhatla(sadibhat,50288587)
Anupriya Goyal(anupriya, 50287108)

TOPIC of DATA:

Our topic for data collection is **"Politics"**.

DATA SETS:

We have collected large amount of data sets from 3 different sources:

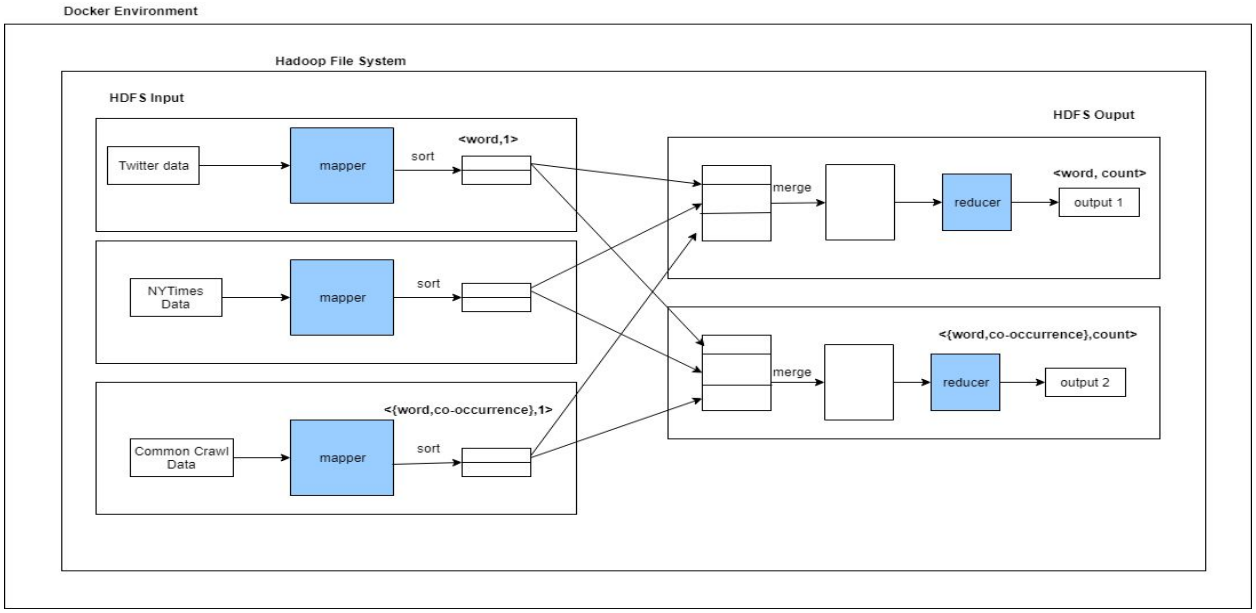
- Twitter
- New York Times
- Common Crawls

WORDS USED FOR DATA COLLECTION:

- Trump or Hilary
- Elections
- Government
- Vote
- Political party

We used the same set of words to collect data from all 3 data sources: Twitter, New York Times and Common Crawls.

BLOCK DIAGRAM FOR THE SYSTEM:



STEPS FOLLOWED :

- Initially, we took data from Twitter using Twitter API and 'rtweet' package in R using the script file- **'Tweets_collection.ipynb'.**
- Then we collected data from New York Times using New York Times API . We have used the following script file for that- '**NYT_collection.ipynb'.**
- Then we took data from Common Crawls using commoncrawl.org, where we downloaded selected latest march 2019 data, and then downloaded **.WET files**, crawled manually searching keywords and saved the relevant text and URLs.
- We collected the data for the period between Feb 1, 2019 to April 15, 2019.
- We cleaned the data, by performing lemmatization, then removed stop words and punctuations, then written to text file named according to the keyword and data source.
- That cleaned text file is input to the Mapper Class where we perform word count functionality and output the valid word and count as **< word,1>**.

- In the Reducer, we sum the value part of each word from all mappers and get the count of words and output **<word, count>.**
- We then use this output to generate word cloud to depict data using interactive visualization **Tableau.**
- Below are the screenshots for visualization of word count for all 3 data sources:

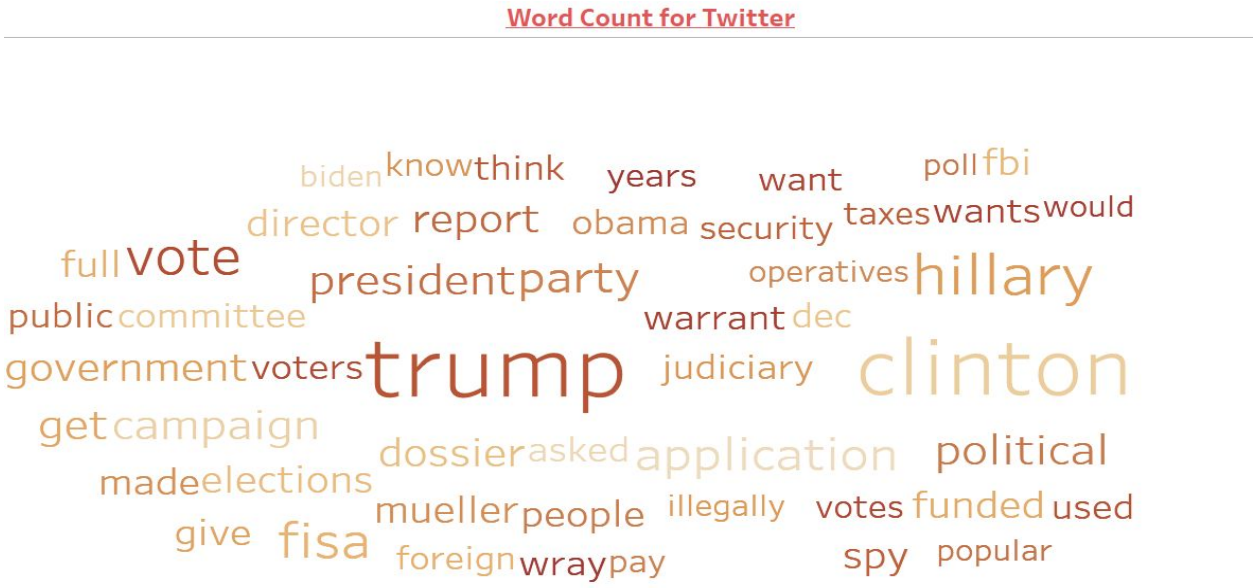


Fig 1: Word Count for Twitter

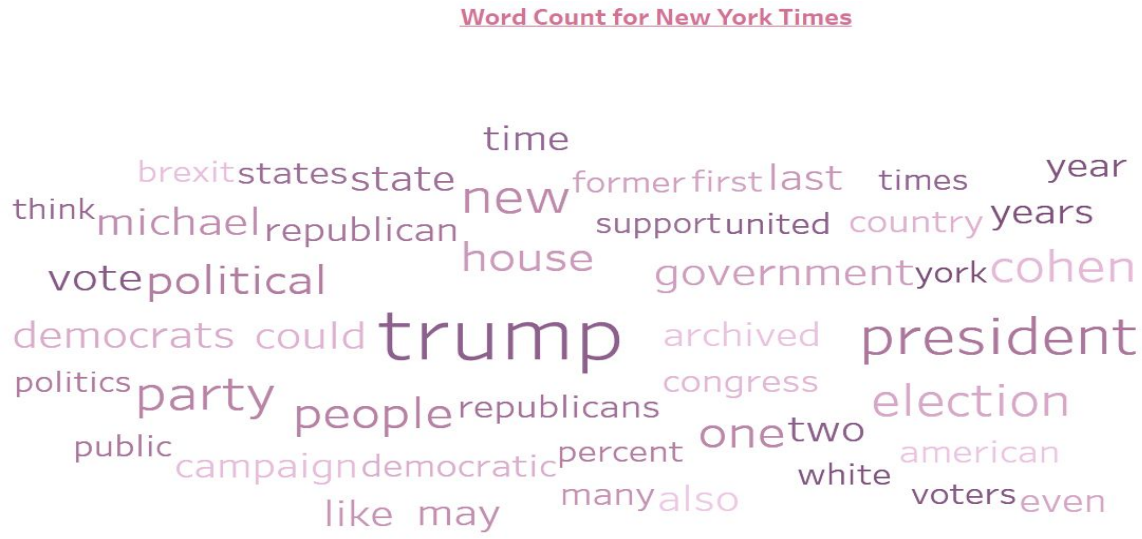


Fig 2: Word Count for New York Times

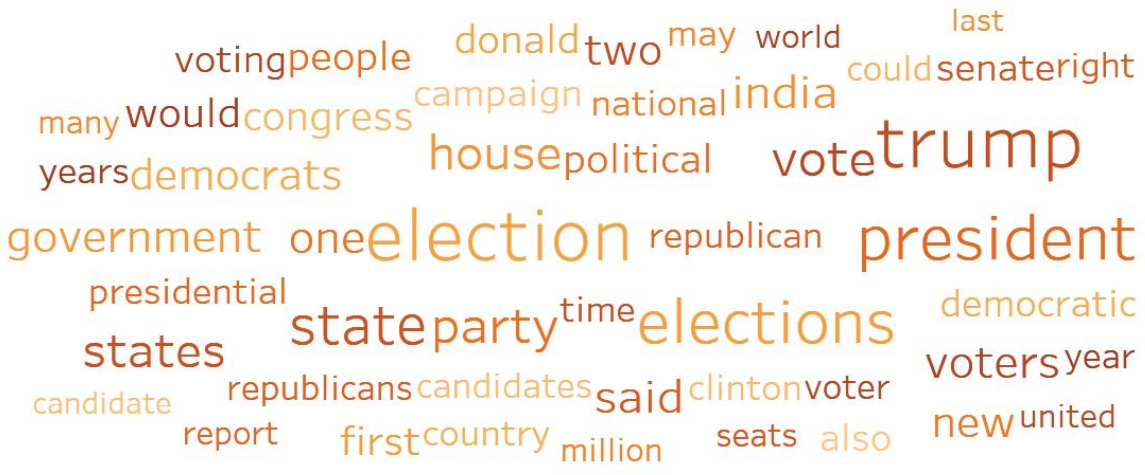


Fig 3: Word Count for Common Crawl

- From our visualization we found that **'Trump',** is the highest count word for both Twitter and New York Times but for Common Crawl it is **'Election'** with **count 384** and **'Trump'** with **count 380.**
- Then we are sorting the word-count pairs by value with **sorting.ipynb** and use the top 10 words from the sorted data to find the **co-occurrence words.**
- We found the co-occurrence words using a Map Reduce method for Twitter, New York Times and Common Crawl separately.
- In the Mapper, we used the Top 10 words from the sorted collection for Twitter, New York Times and Common Crawls data and we generate a key-value pair and output it to file.
- So, the output of the mapper will be of the form (**{word co-occurrence}, 1).**
- In the Reducer, we use the output from the Mapper and reduce it to get the count of the co-occurrence word.
- The output of the Reducer will be of the form (**{word co-occurrence}, count).**
- For visualizing the co-occurrence we used word cloud from **Tableau**.
  Below is the screenshot of word cloud:

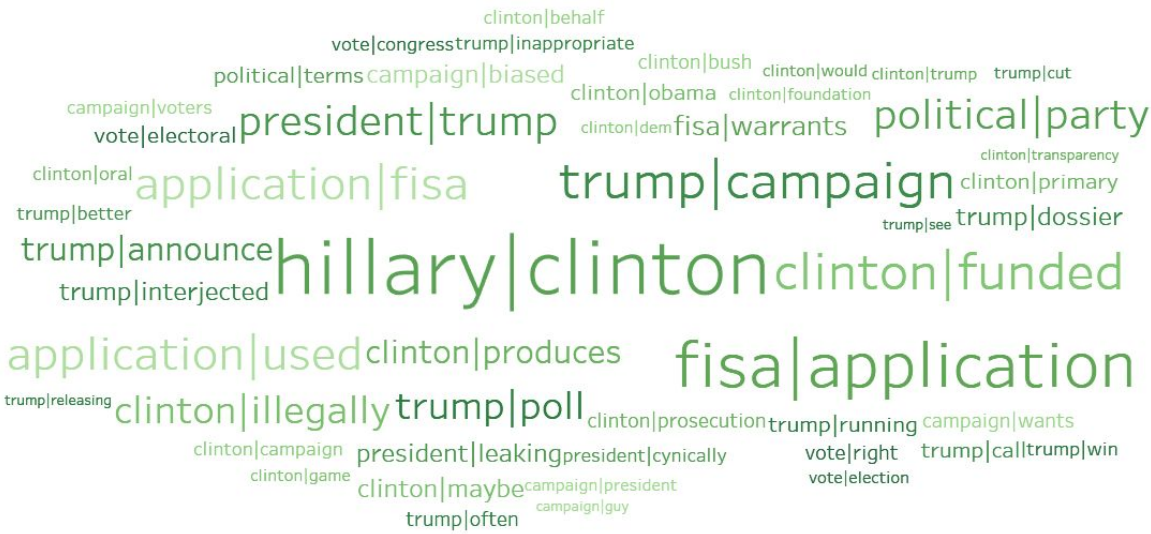


Fig 1: Twitter Word Co-occurrence for Top 10 words

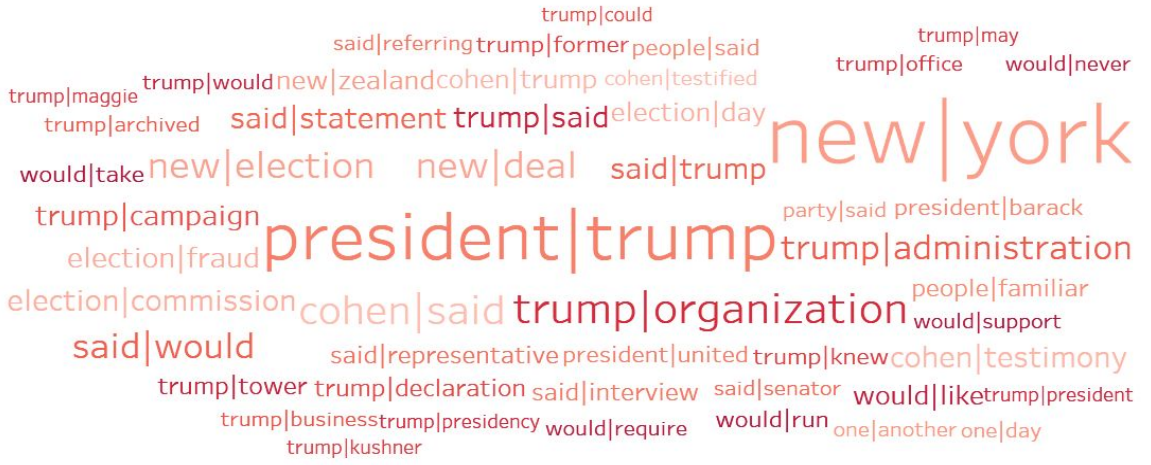


Fig 2: New York Times Word Co-occurrence pairs on Top 10 words

T

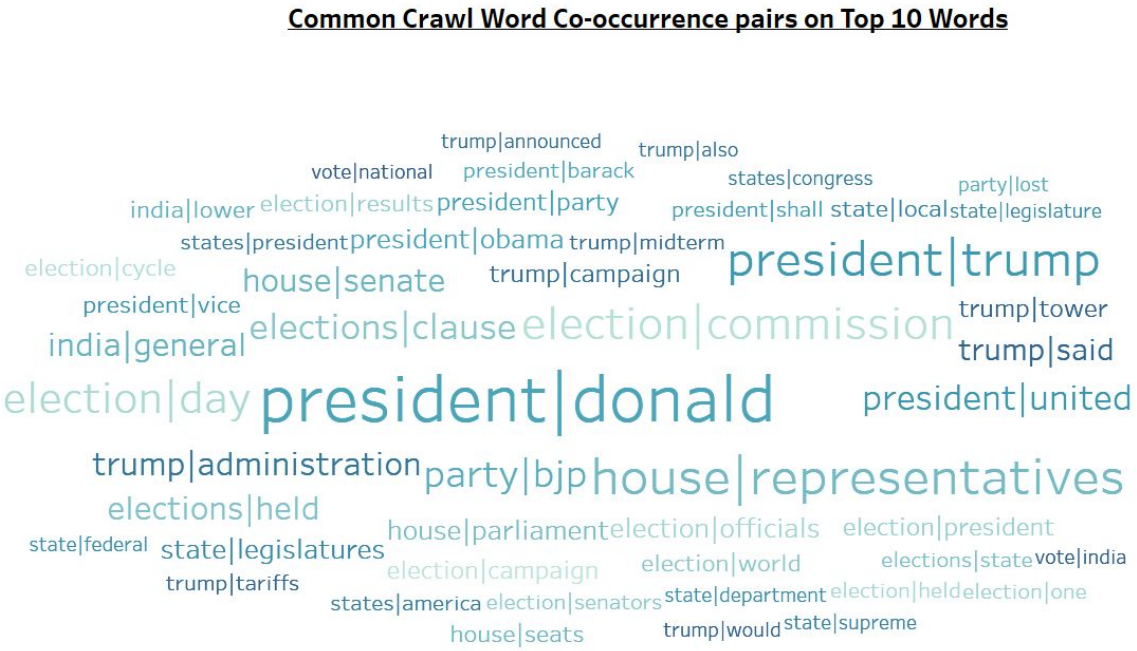


Fig 3: Common Crawl Word Co-occurrence pairs on Top 10 words

LEARNING:

- We learned Python Programming Language.
- We learned about data aggregation from API exposed by data sources - Twitter and New York Times, as well as about .WET, WARC files that is used by amazon aws to crawl and store data.
- We automated data collection from multiple sources using the API and python/R scripts.
- We gained the knowledge of how to import a docker image into Hadoop File Distributed System.
- We learnt how to use Mapper and Reducer, understood and implemented its functionality in Hadoop environment, and processed the data using big data algorithms.
- We used Tableau to use interactive visualization and depict our results.
- We got the knowledge to create web interface for visualizing the outcome of our analysis.

- We learned how to publish a Tableau workbook which can be accessed publicly to anyone.

WEBSITE/ PUBLISHED URL:

https://public.tableau.com/profile/anupriya.goyal?/vizhome/Workbook1_15558769608410/Twitter_all_word_co#!/