README

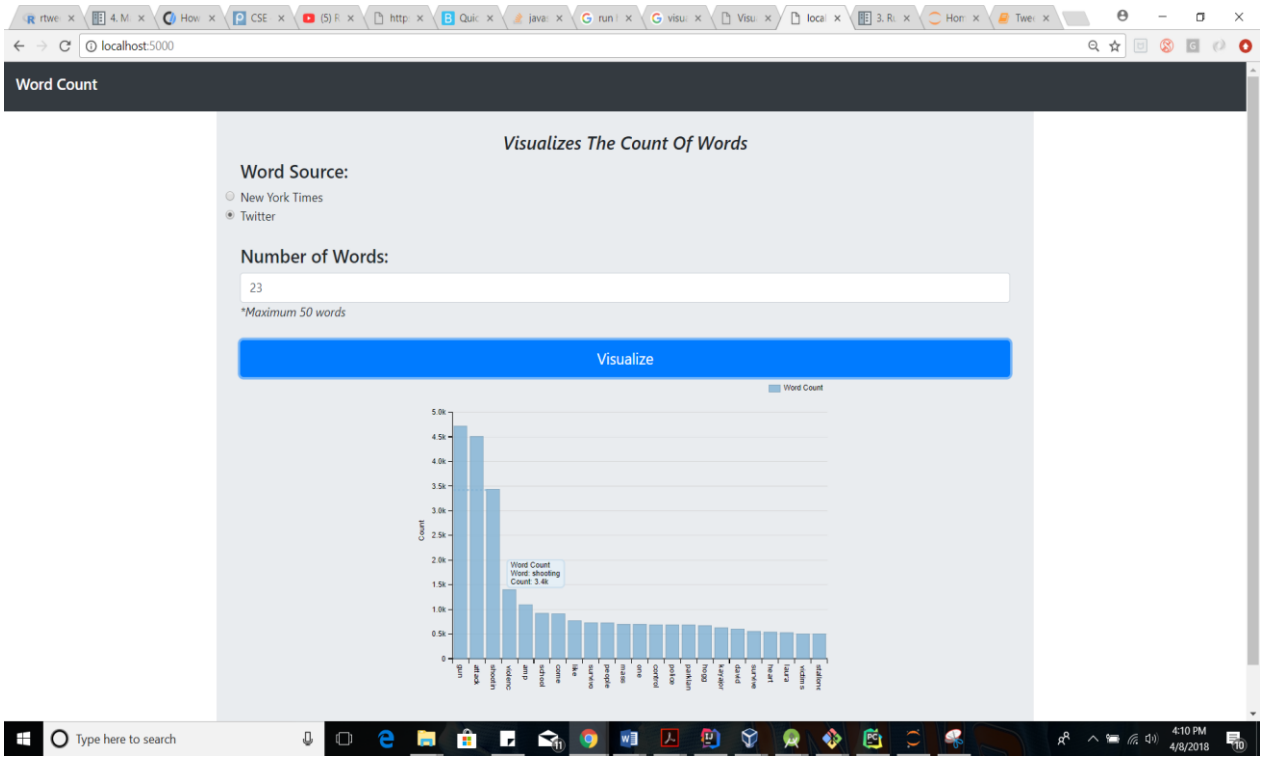
DATA AGGREGATION, BIG DATA ANALYSIS AND VISUALIZATION

Allen Daniel Yesa

Aditya Subramanian Muralidaran

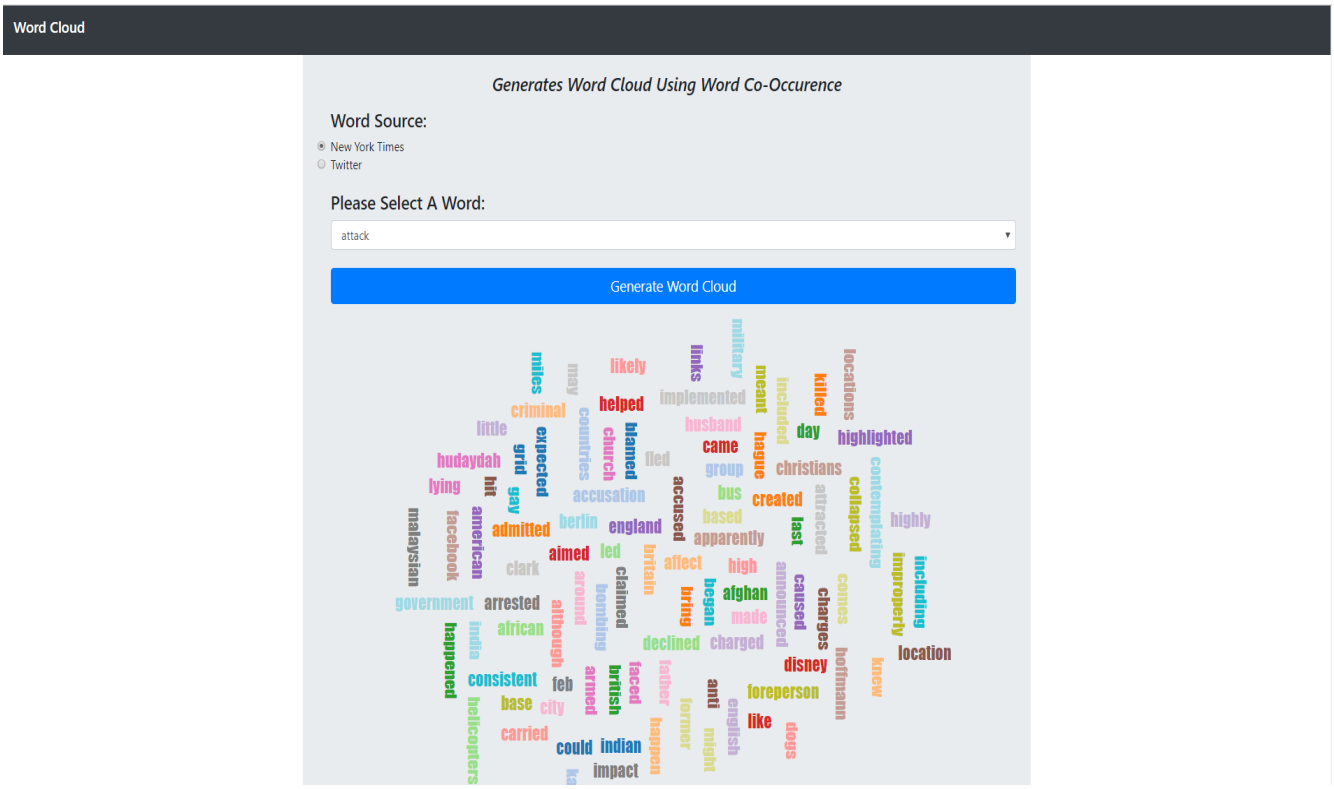
- Our topic for data collection was “Mass Shooting in US”.
- We used 2 data sources:
    - Twitter
    - New York Times
- The words which we used for collecting data was:
    - Shooting
    - Gun
    - Attack
- We used the same three words for collecting data from both the data sources - New York Times and Twitter.
- Steps Followed:
    - Initially we took data from Twitter using the Twitter API and ‘rtweet’ package in R using the script file - ‘TweetData_Program.ipynb’
    - Then we took the data from New York Times using New York Times API as using the script file – ‘GetNewsData.py’
    - We collected the data for the period between - 30-March-2018 and 06-April-2018
    - We cleaned the data, removed the stop words and removed the punctuations in the Mapper method and we just emit the valid word and count as (word \t 1).
    - In the Reducer, we sum the value part of each word from all the mappers and get the count of the words and emit the (word \t count)
    - We then use this output to generate the d3.js interactive visualization where we use a Bar Plot to depict the data.

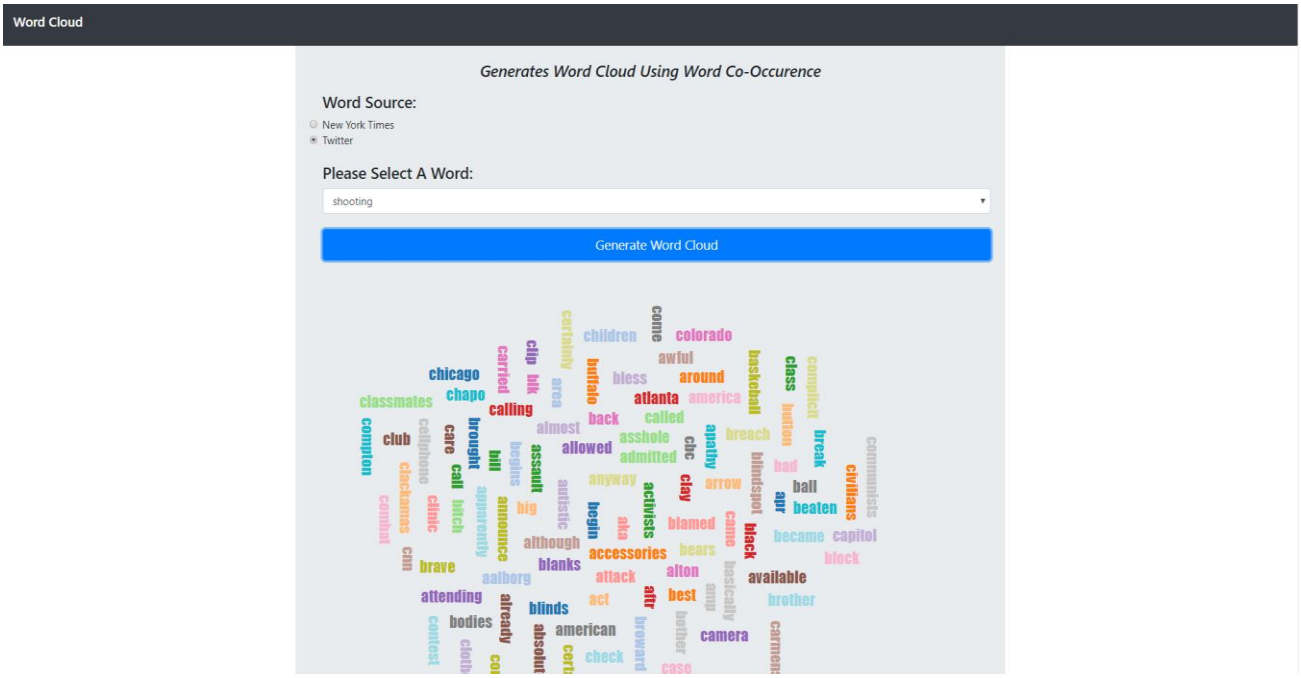
- Screenshot of word count virtualization:



- From our visualization we found that the count of words in the bar plot using Twitter data are similar to the count of words we found from the News data.

- Then we used the script ‘Sorting_WordCount.py’ to sort the data based on word count and used the top 10 words from the sorted data to find the co-occurrence words.
- We found the co-occurrence words by using a Map Reduce method for Twitter and New York Times data separately.
- In the Mapper, we used the top 10 words from the sorted collection for both the Twitter and New York Times data and we generated a pair as a key and a value (which will be 1) and emitted it.
- 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). We use this to visualize the co-occurrence.
- For visualizing the co-occurrence we use the word cloud from d3.js
- Screenshot of word cloud:





➢ Learning:

- We learned python programming language.
- We learned about data aggregation from more than one source using the APIs (Application programming interface) exposed by data sources – Twitter and New York Times
- We automated data collection from multiple sources using the APIs offered by the businesses and python/R scripts.
- We got the knowledge to install a virtual machine (VM) image for data storage in HDFS and Hadoop infrastructure.
- We learnt how to use Mapper and Reducer in Hadoop environment and learned about Hadoop 2.x, HDFS and process the data using big data algorithms.
- We used d3.js to learn modern visualization methods and disseminate results using the web/mobile interface
- We got the knowledge to create a responsive web interface (web tool) for visualizing the outcome of your analysis.