Lab 2

Team:

Name	**UBIT Name**	**UBID**
Krithika Krishnan	kk242	50169047
Varad Tupe	varadsha	50249001

Topic used for analysis:
Gun violence

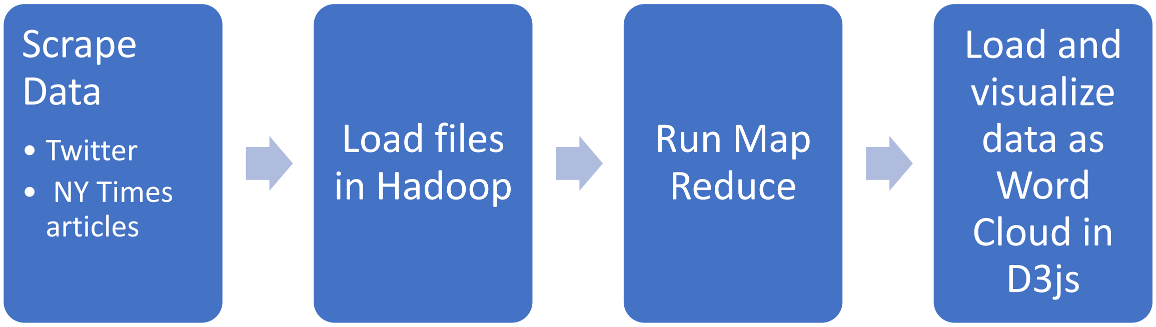
Data Source:

1. New York Times articles
   - All articles details are fetch Article Search API provided by NY Times.
   - Using request and Beautiful Soup packages each article is scraped using the article urls procured from the article search API.
   - Each article is written to a separate file.
2. Tweets from Twitter
   - All tweets are fetched using Tweepy Package in Python.
   - UTF-8 encoding is used while saving the tweets in file.
   - Tweets collected each day are stored in separate txt files.

Key words used in search:

gun violence, gun control, school shooting, parkland, gun, gun, mass shooting, gun laws

Data Flow



Algorithm

Top 100 words:

```
Mapper:
    1. Read input file.
    2. Remove unwanted symbols and punctuation
    3. Remove stop words.
    4. For all words in doc
       Emit (word, 1)
Reducer:
    1. Read input from Mapper.
    2. Group the data by key.
    3. For each key in data, put key and its count in dictionary.
    4. Sorted the dictionary based on count descending.
    5. For each key in dictionary
       Emit (key, count) [for first 100 keys]
```

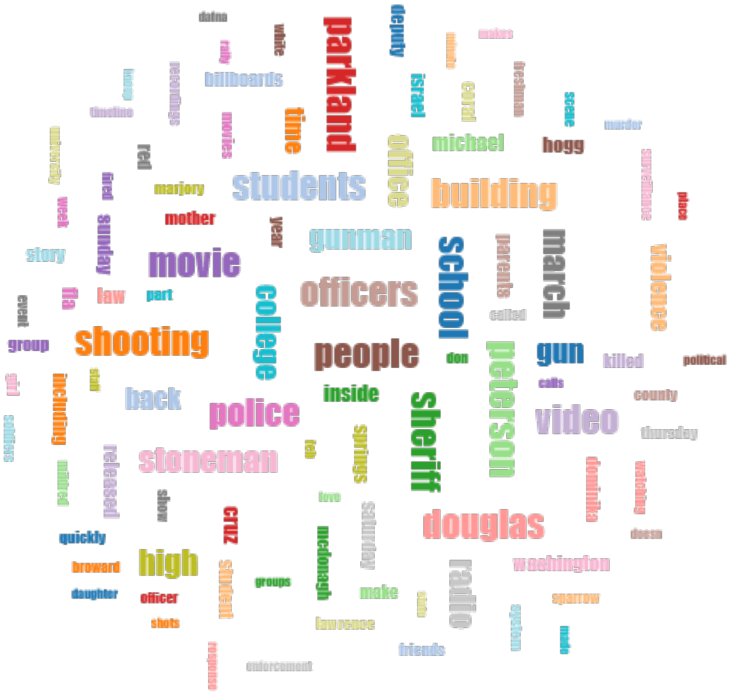
Top co-occurrence:

```
Mapper:
    1. Read input file.
    2. Remove unwanted symbols and punctuation
    3. Remove stop words.
    4. For all words w1 in doc
       For all words w2 in Neighbours(w1)
       Emit (w1-w2, 1)
Reducer:
    1. Read input from Mapper.
    2. Group the data by key.
    3. For each key in data, put key and its count in dictionary.
    4. Sorted the dictionary based on count descending.
    5. For each key in dictionary
       Emit (key, count) [for first 20 keys]
```

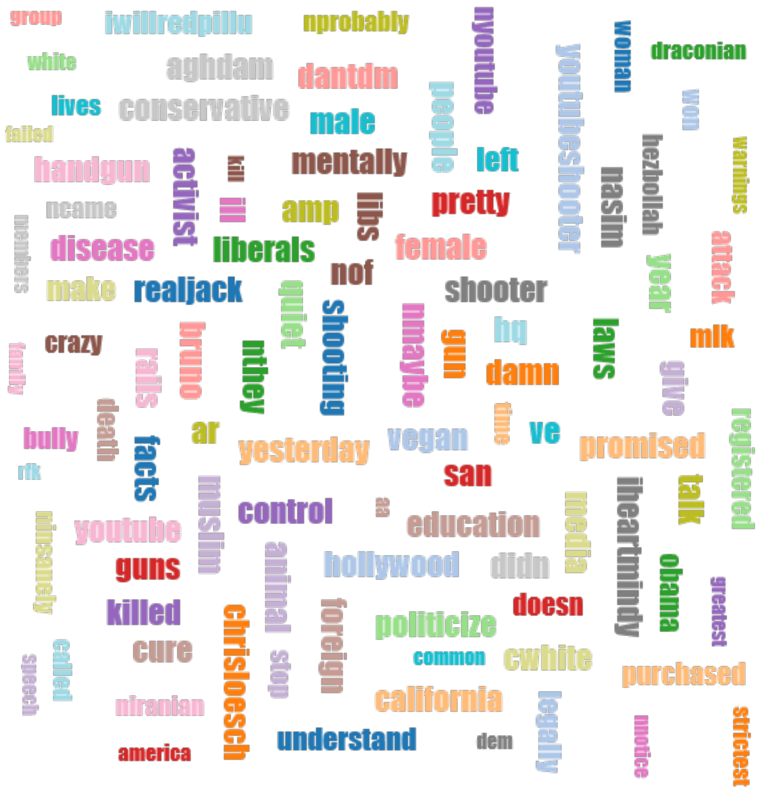
Results:
At first, data from a single day was fed into map reduce.
In the NY Times article, commonly occurring words included,
Students, shooting, gun, school, parkland,..
In the Twitter data, commonly occurring words included,
Students, shooting, sheriff, school, parkland,..
There were commonly occurring words between the two sets of data.

When the mapreduce was performed for the bigger dataset, collected over a week,
The commonly occurring words among the NY articles and Twitter data converged
with school, gun, shooting featuring among the top 20 words along with some interesting
words like vegan, democrats, female and education, along with the youtube.

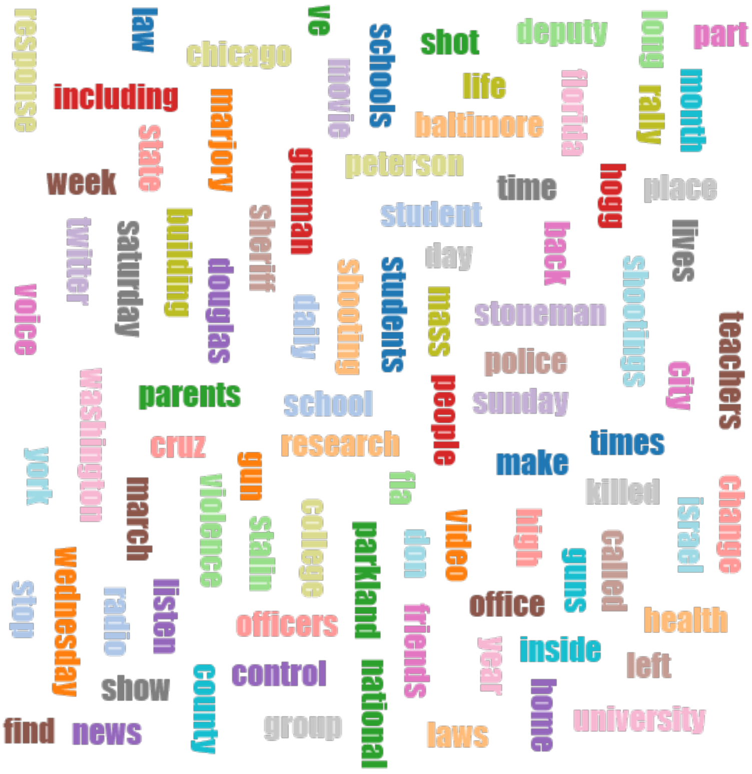
Prototype wordcloud from NY Times article



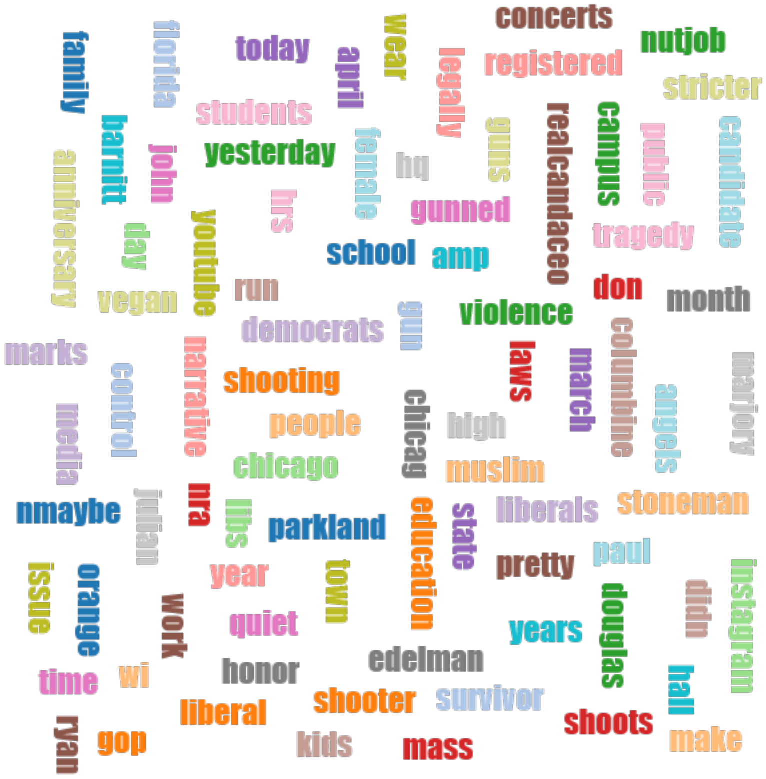
Prototype wordcloud from Twitter



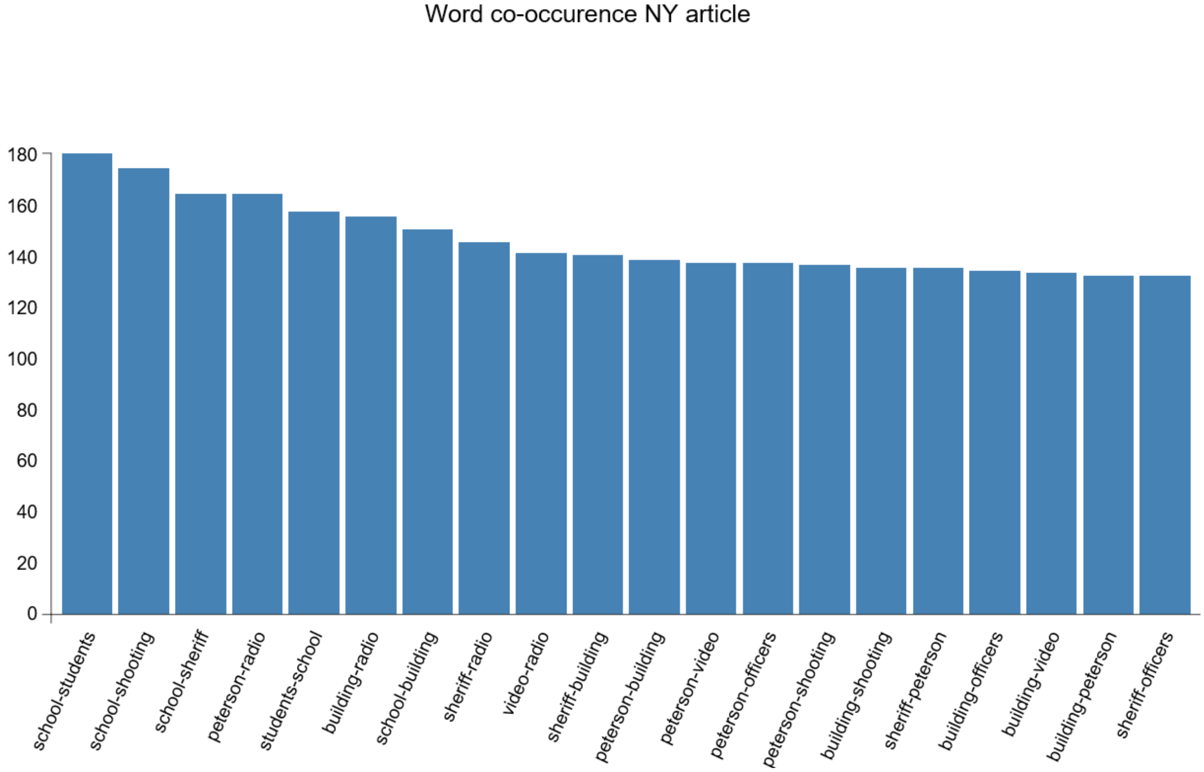
NY Times article collected over a week



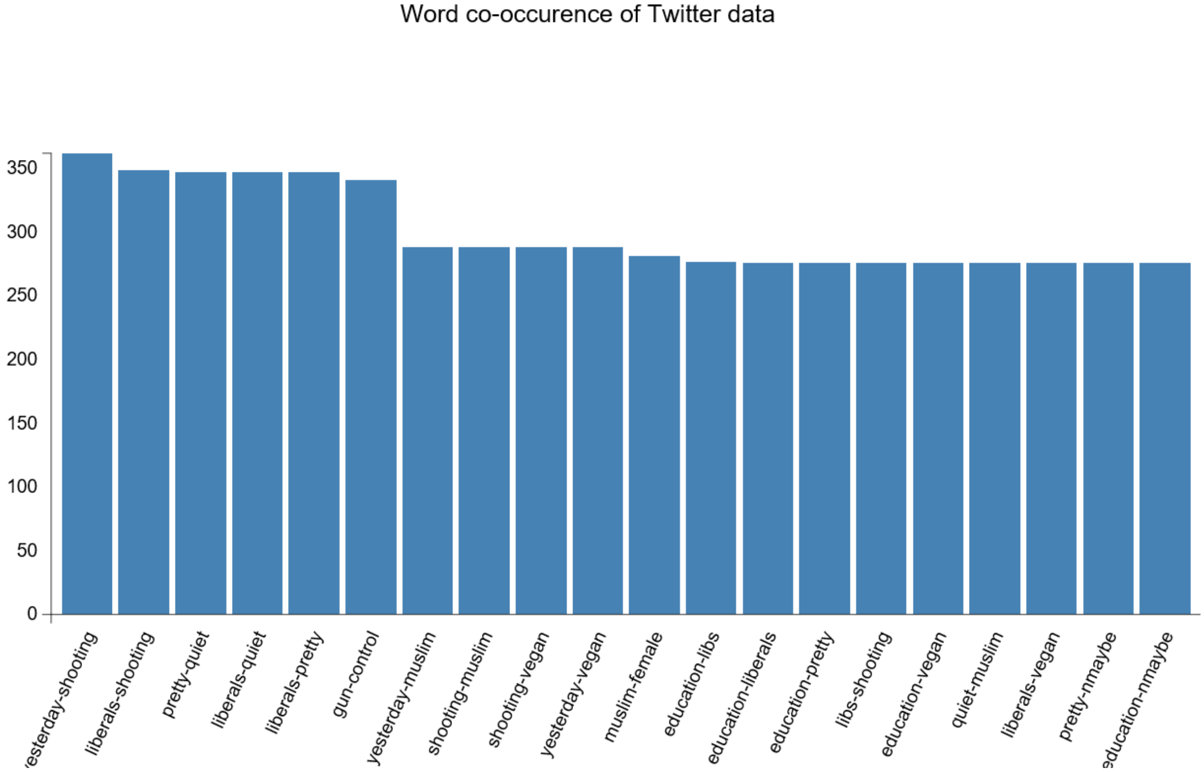
Twitter data collected over a week



Word co-occurrence:
NY Data


Word co-occurence NY article

Twitter Data


Word co-occurence of Twitter data

Discussion of Word pair results:

In the graphs above, the actual count of the word co-occuring are plotted as bar plots with the count on the y axis and the corresponding word pairs plotted in the x axis.

- There seems to be repeated word pairing in the NY and twitter data.
- School-students, school-shooting, students-school featured in among the top 20 word co-occurences for the NY Times articles.
- Yesterday-shooting, Liberals-shooting, gun-control, are among the featured in among the top 20 word co-occurences for the Twitter data.

There seems to be a co-occurrence of shooting, students on the most co-occuring word pairs.

Folder Structure:

- lab2
    - dataFetch
        - Python notebooks for fetching NY & Twitter data
    - Hadoop
        - Input
            - TwitterData
            - protoTwitterData
            - NewsData
            - protoNewsData
        - Output
            - Twitter & NY wordcount for top 100
            - Twitter & NY wordpair for top 20 co-occurences
        - README.txt
    - results
        - visualization
            - index.html (to view wordcloud)
            - barchart
                - barchartTwitter.html
                - barchartNY.html

Citation:
Michael Noll for mapReduce
And d3.js from Jason Davies d3.js wordcloud template