DATA AGGREGATION , BIG DATA ANALYSIS AND VISUALIZATION

SAI RAM PRASANTH TATIRAJU (50290579)
SANDEEP KOTHAPALLY (50289072)

Our topic for data collection is “Sports”.
We used 3 data sources:

- New York Times
- Twitter
- Common Crawl

The keywords which we used for collecting data is:

- Baseball
- Basketball
- Golf
- Tennis
- American Football

Using the keywords above we collected data between March 05 to April 19th from Twitter , Commoncrawl , New York Times. We collected

- 25000 tweets selecting 5000 from each of the above topics
- 500 common crawl URLS, data selecting 100 for each of the above topics
- 550 articles from New York Times selecting more than 100 for each of the above topics

Data Collection procedure

New York Times data collection:

- We used New York Times API in our script file NYT_collection.ipynb to collect articles

Twitter data collection:

- We used rtweet package in our Twitter_collection.ipynb file to collect tweets

Common Crawl data collection:

- First, we downloaded the March 2019 Index of wet paths and then downloaded few warc.wet files using the relative paths.
- We then manually searched for our topic keywords and copied the urls in a csv.
- We then crawled the urls and collected the data.

Data Cleaning:

From the collected data we converted all the text to lowercase , removed Stopwords using NLTK library and NLTK stopword corpus, removed punctuations, urls , emojis ,hashtags etc.

We then performed lemmatization on the text data to return base form of words.

Word Count:

In our mapper method we emitted the word and count as <word , 1>

In our reducer method we calculated the total count of each word from all the mappers and emitted the <word , count>

Word Co-occurrence:

From the word count output obtained in the above step we picked the top 10 high frequent words and used those words as our top words to obtain word co-occurrence count for each of the three data sources

Visualization :

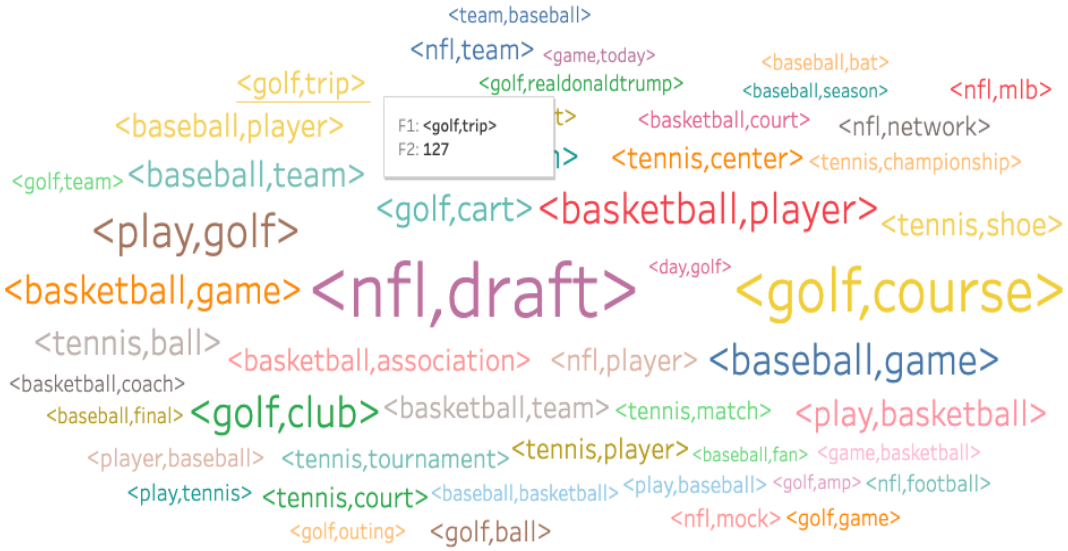
We used Tableau software to visualize the word count, word co-occurrence of the data from each of the above three sources using word cloud

Contd..

Word Clouds for Twitter Sports Data Word Count & Word Co-Occurrence

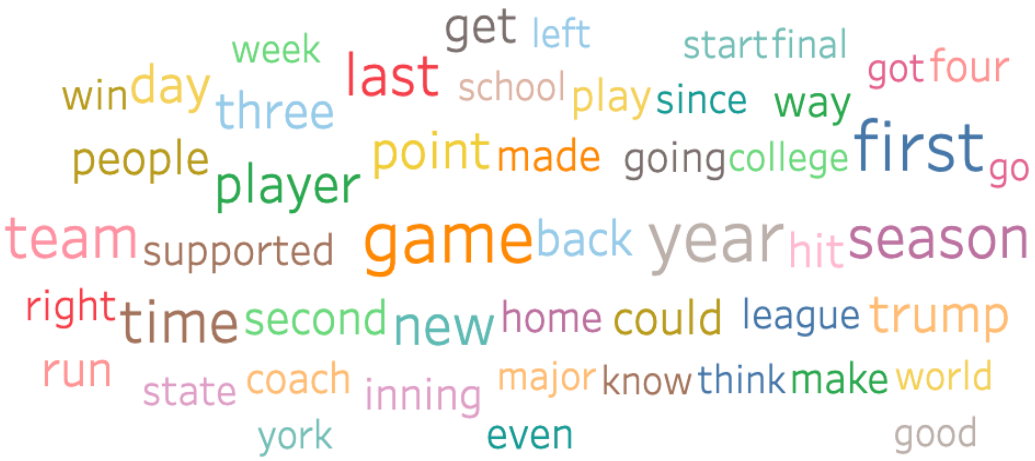
Twitter_WordCount

Twitter_WordCo-occurrence



Word Clouds for NYT Sports Data Word Count & Word Co-Occurrence

NYT_WordCount



NYT_WordCo-occurrence

Word Clouds for Common Crawl Sports Data Word Count & Word Co-Occurrence

CC_WordCount

CC_WordCo-occurence