University at Buffalo
The State University of New York

DATA AGGREGATION, BIG DATA ANALYSIS AND VISUALIZATION

Data Intensive Computing

CSE 587

SUBMITTED BY:

SAGAR POKALE (50288055)

SOUMITRA ALATE (50289133)

I. **Data Crawling**:

Crawling is performed on three data sources:

a) Twitter: Twitter data is crawled using TwitteR library in R. Oauth handler is used in the script which gives programmatic access to tweets. Oauth takes following parameters as input consumer key, consumer secret, access token, access secret. Preprocessing is performed on the tweets where retweets, hashtags, links stop words and special characters are removed using regular expressions. The preprocessed tweets are stored in text file and given as input to mapper.

b) New York Times: We have scraped articles from New York Times using articleAPI. We used nytimes articles library and NY Times key to access the data. Preprocessing is performed on the scrapped articles and results are saved in text file which is given as input to mapper.

c) Common Crawl: We have scraped common crawled data using warc and Beautiful Soup available in python and performed preprocessing on the extracted data.

II. **Map Reduce**:

Word Count: Hadoop Map-Reduce script is written which generates mapping for number of time the word has occurred in text.

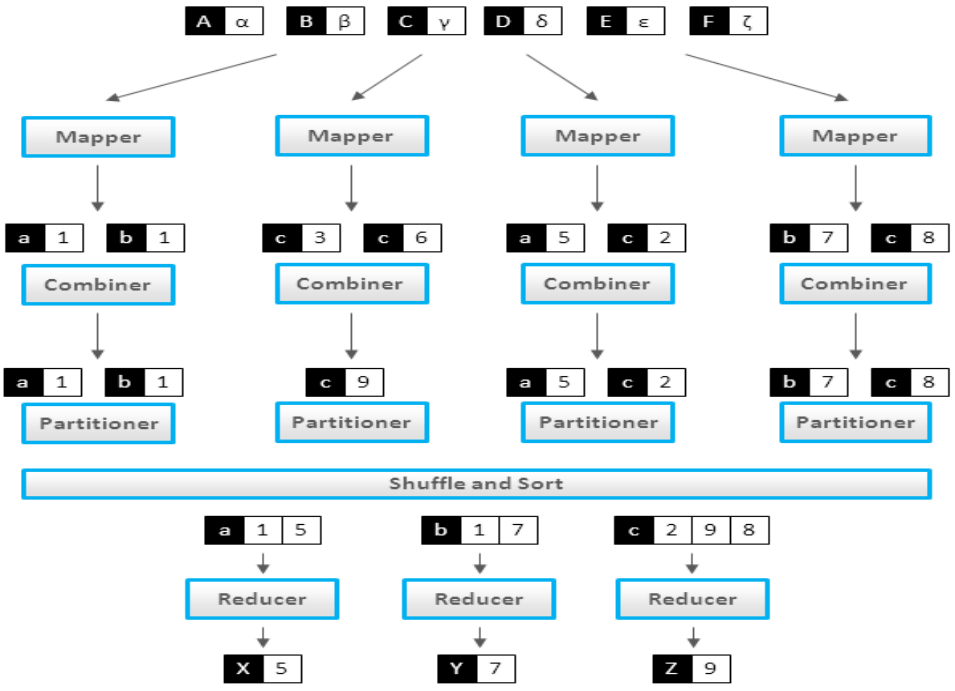


Fig : Map Reduce Architecture

III. Word co-occurrence:

In this the top 10 words from word count mappings are extracted and mappings for co-occurrence for those 10 words are generated.

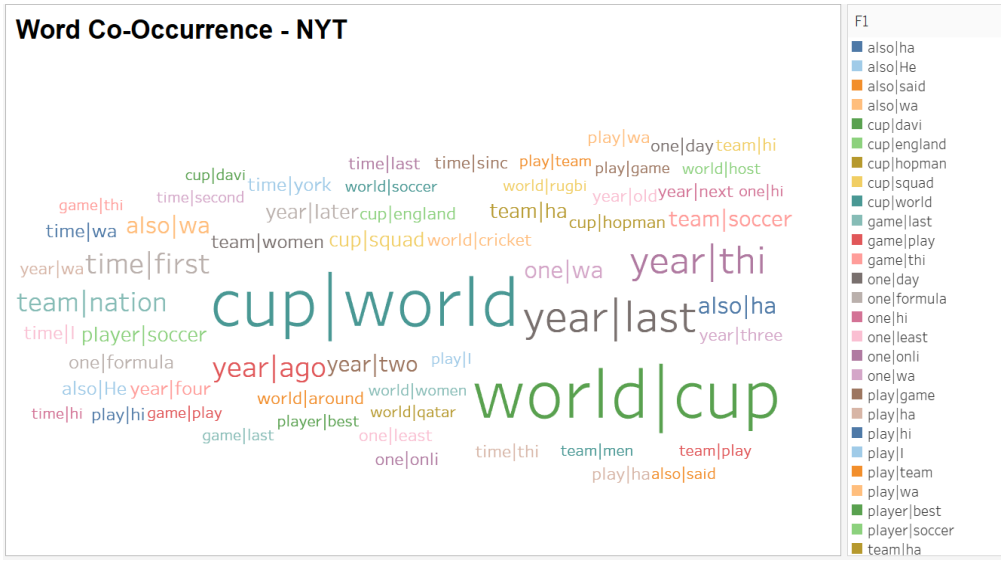
IV. Data Visualization:

For Data visualization we have used Tableau and created word cloud for word count and word co-occurrence.

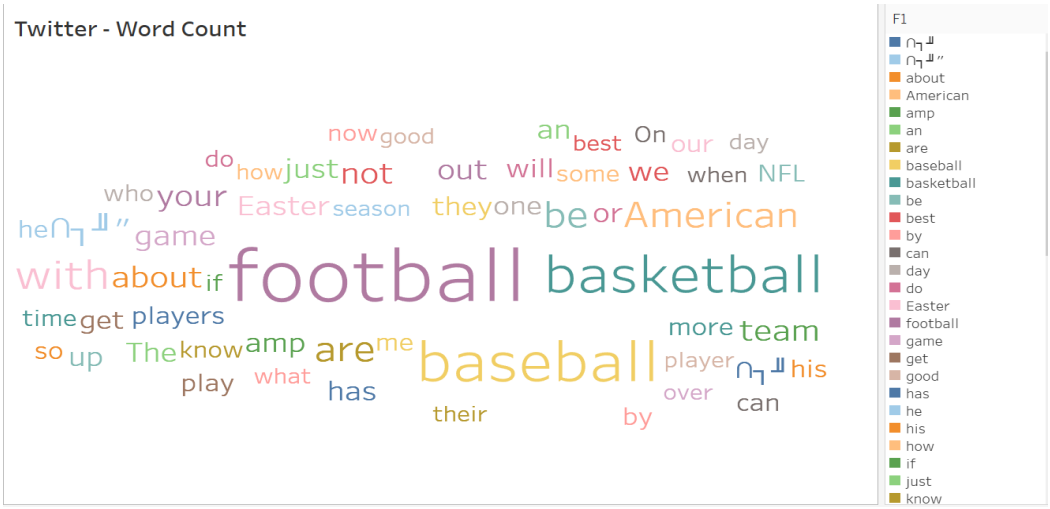
Screenshots of Results:

a) Word Count – NYT:

b) Word Co-Occurrence-NYT:

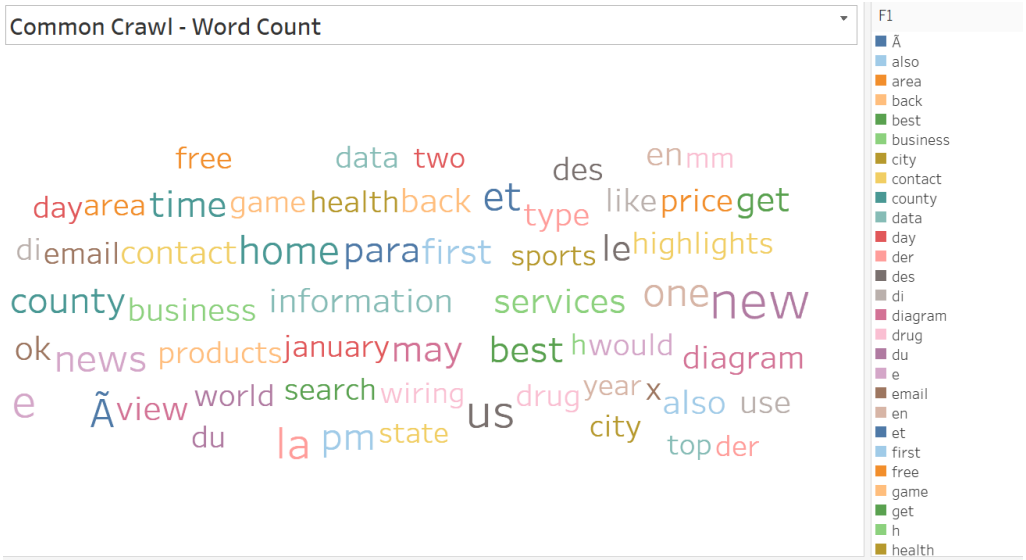


c) Word Count – Twitter:



d) Word Co-Occurrence – Twitter:

e) Word Count – Common Crawl:



f) Word Co-Occurrence – Common Crawl: