

DATA AGGREGATION, BIG DATA ANALYSIS AND VISUALIZATION

Introduction/Purpose

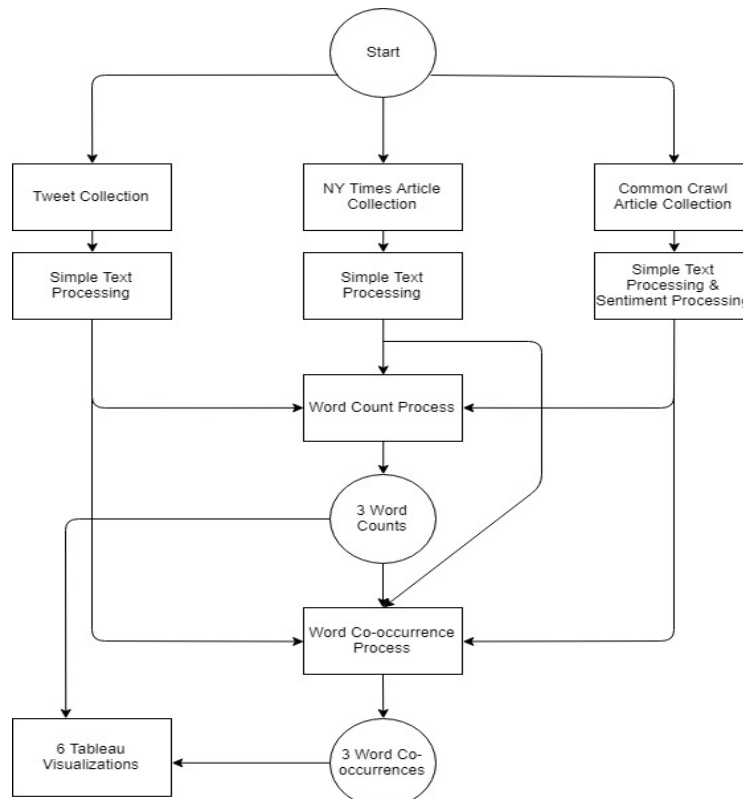
In this lab, we have expanded our skills in data exploration developed in Lab1 and enhanced them by adding big data analytics and visualization skills. This document describes the workflow for Lab2 which involves (i) data aggregation from more than one source using the APIs(Application Programming Interface) exposed by data sources, (ii) Applying classical big data analytics method of MapReduce to the unstructured data collected, (iii) store the data collected on WORM infrastructure Hadoop and (iii) building a visualization data product.

Implementation

We started by gathering the tweets from Twitter, articles from NY Times and articles from Common Crawl. Then clean the data to separate the meaningful data from the noise. Cleaned data is then processed to find the most used words and their co-occurrences with words in tweets. articles etc. Then we visualize the results to bring out the crux of the data from the articles and the tweets. To accomplish all that we collected data from following sources:

- Twitter
- New York Times
- Common Crawl

The workflow has been shown in the form of a flow chart below.



Twitter Data Collection

To gather the tweets on related topic, we used Rtweets API of Twitter. We used R Script to achieve this. The Script makes use of API Key and return tweets for given search query. We have used following subtopics strings to collected tweets from united states about our topic of Sports.

- Sports
- Basketball
- Baseball
- Hockey
- Football

The script retrieves the ands stores it into a csv file (for example, twitter_data.csv). We further extracted the tweets' text and stored them into the text file (for example, twitter_data.txt). Each tweet was stored in one line for future purpose.

New York Times Data Collection

To gather the articles, we used the API provided by NY Times. we used Python for achieving this. Using the API Key, search keywords, and date, we gathered the URL of the articles and stored it. The response of the API call is in JSON format. We parsed the response to extract only the useful information (URL of the articles) while discarding the rest of the information. We scraped the articles using the links. Again, we used Python's library BeautifulSoup to get the HTML Page of the articles. We then parsed the HTML page to get the body of the article. This information was stored in a file. Each paragraph was stored in one line of file for future purpose.

Common Crawl Data Collection

To gather the articles from common crawl we first chose the index of 2019-09 and 2019-13. This returned us a huge number of articles link roughly around 26 thousand. We then scraped each of the link and looked for words related to sports in it. If words were found we kept the articles otherwise we discarded them. When this was done we performed the same cleaning as we cleaned twitter and NY Times data.

Data Cleaning

We cleaned the data to keep only the words essential for our analysis. We removed stop words to make the data meaningful. Stop words are natural language words which have very little meaning, such as "and", "the", "a", "an" etc. These are commonly occurring words and will distort our analysis. We also removed website and emojis. We also removed hashtags. Once we only had the important words, We lemmatized the words to change words like 'following' to 'follow'. This way many redundant words were transformed to a base word.

Word Count Mapper/Reducer

Mapper: Emits (outputs) a key value pair for each word in the article or tweet. The key in this case is the word and value is 1. Later, in Reducer, we aggregated these 1 for every key (unique word) to

find the count of that word.

Reducer: Here We found the actual counts. Output of the mapper is fed as an input to the Reducer. The reducer collects <key,value> pairs which have the same key. The <key, value> pair are of the type <word, 1>. It then sums over all the values received for this key. This generates the count for that key. The reducer then emit (output) this value.

Top words for each source

Twitter:

- hockey
- football
- baseball
- sport
- basketball
- game
- team
- player
- playoff
- like

NY Times:

- player
- said
- baseball
- team
- league
- season
- game
- year
- hockey
- cuban

Common Crawl:

- sport
- football
- woman

- game
- life
- world
- year
- book
- day
- time

Co-occurrence using map-reduce frame work

We find out the co-occurring words in each article/ tweet. The aim is to find the pair of words which occur together most frequently. For this we used the top ten most frequently occurring words captured after running the Word Count on tweets and articles. The context for co-occurrence is chosen as a tweet or paragraph in case of NY Times or whole article in case of Common-crawl. The approach followed is similar to the one used in word count using Map Reduce. The Mapper and Reducer are implemented as follows.

Mapper: We select pairs of words from the tweet/article, such that at least one of those words lies in the top ten words which we have identified. This word-pair, consisting of two words, is emitted by the Mapper in <key, value> format as <word-pair, 1>

Reducer: In the Reducer, We collect the all the <key, value> pairs that are emitted by the Mapper. All the values (1) associated with the same key (word-pair) are aggregated to get the count of the word-pair. This is emitted by the Reducer as the output. This procedure is applied to every unique word-pair sent as the key. The resulting output gives us the frequency of co-occurrence of the word-pairs so that we can identify the co-occurring words with the highest frequencies.

Visualizing the output using Tableau

To show the comparison of top occurring words in Twitter data vs NYTimes articles, we have used Tableau visualization. This can be seen in the images below.

Tweets_Counts

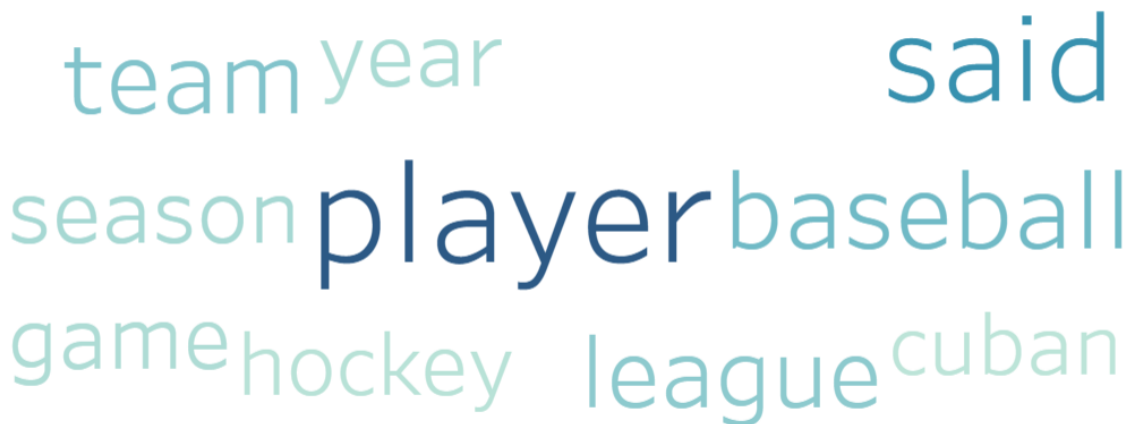


A word cloud for the 'Tweets_Counts' dataset. The words are arranged in a roughly circular pattern. The most prominent words are 'hockey', 'baseball', 'football', 'game', 'sport', 'team', 'player', 'basketball', 'playoff', and 'like'. The words are in various shades of blue and green, with 'hockey' and 'baseball' being the largest.

Word	Count (approximate)
hockey	15
baseball	12
football	10
game	8
sport	7
team	6
player	5
basketball	4
playoff	3
like	2

Figure 1: Top 10 Word Count for Twitter Data

NYTimes_Counts



A word cloud for the 'NYTimes_Counts' dataset. The words are arranged in a roughly circular pattern. The most prominent words are 'player', 'baseball', 'said', 'team', 'year', 'season', 'game', 'hockey', 'league', and 'cuban'. The words are in various shades of blue and green, with 'player' and 'baseball' being the largest.

Word	Count (approximate)
player	15
baseball	12
said	10
team	8
year	7
season	6
game	5
hockey	4
league	3
cuban	2

Figure 2: Top 10 Word Count for NYTimes Data

CommonCrawl_Counts

A word cloud where the size of each word corresponds to its frequency. The words are arranged in a roughly circular pattern. The word 'sport' is the largest and is dark blue. Other words include 'game', 'woman', 'football', 'world', 'year', 'daytime', 'life', 'book', and 'world'.

book life game
world football woman
daytime sport year

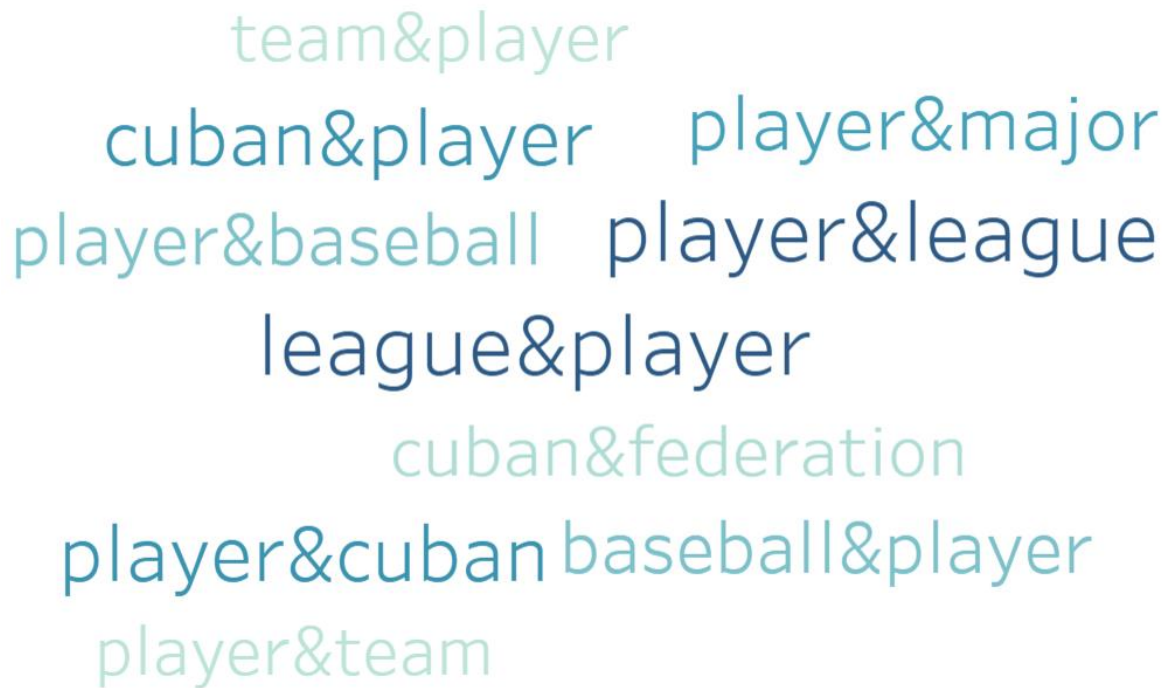
Figure 3: Top 10 Word Count for Common Crawl Data

Tweets_Co-occurrence

A word cloud where the size of each phrase corresponds to its frequency. The phrases are arranged in a roughly circular pattern. The phrase 'football&league' is the largest and is dark blue. Other phrases include 'baseball&game', 'basketball&coach', 'football&national', 'hockey&playoff', 'basketbal', 'football&news', 'playoff&hockey', 'basketball&association', and 'basketball&national'.

baseball&game
basketball&coach football&league
football&national
hockey&playoff basketbal
football&news playoff&hockey
basketball&association
basketball&national

Figure 4: Top 10 words Co-occurrence for Twitter Data



A word cloud showing the top 10 co-occurrences for NYTimes Data. The words are arranged in a roughly circular pattern, with 'team&player' at the top, 'cuban&player' and 'player&major' below it, 'player&baseball' and 'player&league' in the middle, 'league&player' below that, 'cuban&federation' at the bottom, and 'player&cuban', 'baseball&player', and 'player&team' at the very bottom. The colors range from light green to dark blue.

team&player
cuban&player player&major
player&baseball player&league
league&player
cuban&federation
player&cuban baseball&player
player&team

Figure 5: Top 10 words Co-occurrence for NYTimes Data



A word cloud showing the top 10 co-occurrences for Common Crawl Data. The words are arranged in a roughly circular pattern, with 'life&woman' and 'woman&life' at the top, 'woman&book' and 'woman&world' below it, 'woman&story' in the middle, 'day&woman' and 'world&woman' below that, 'woman&child' and 'book&woman' at the bottom, and 'woman&day' at the very bottom. The colors range from light green to dark blue.

life&woman woman&life
woman&book woman&world
woman&story
day&woman world&woman
woman&child book&woman
woman&day

Figure 6: Top 10 words Co-occurrence for Common Crawl Data

References

<https://stackoverflow.com/>

<https://stats.stackexchange.com>

Lab 2 Documentation

Tableau Documentation

<https://cse.buffalo.edu/~bina/cse487/spring2019/Lectures/Lab2/>