Sports Data Analysis Using Hadoop, Python and Tableau Visualization from sources Common Crawl, New York Times and Twitter

By Siddiq Syed

Data Collection: Part 1

1. Twitter:

A python script was used to extract data from twitter API using a package called tweepy.

The list of key words that are used to extract data are.

['sports', '#sports','NFL','#NFL','MLB','#MLB','NBA','#NBA','NHL','#NHL','NCAAB','#NCAAB','NCAAF','#NCAAF','GOLF','#GOLF','NASCAR','#NASCAR','INDYCAR','#INDYCAR','MMA','#MMA','FOOTBALL' ,'#FOOTBALL','BASKETBALL','#BASKETBALL','HOCKEY','#HOCKEY','#RACING','RACING' ,'TEAM',' #TEAM']

Initially data collection we started in the first month of April. About 20k has been collected as one data and about 7k as another data.

2. NYT:

From python package articleAPI (nytimesarticle) data is collect from NYT site. Topics used for the same are FOOTBALL, BASEBALL, BASKETBALL, GOLF, SOCCER.A total of about 1000 articles have been collected from NYT. All together the data is stored in one file and the same has been processed for mapreduce wordcount and word co-occurrence.

3. Common Crawl:

We have downloaded WET file from common crawl website which was of about 400 MB. In python we have used “warc” package to extract WET file and record.payload.read() function is used to extract data/blocks from WET file. Articles are being extracted from these blocks of using the key words sports, athlete, football. We have collected about 600 articles from these blocks.

Data Processing Code: Part 2

Oracle VM is used to get Unix Virtual machine in which Hadoop-3.1.2 was already configured. To process the data for word count and word co-occurrence, mapper and reducer programming is done in python. Two sets of mapper and reducer programs were written each for word count and word co-occurrence.

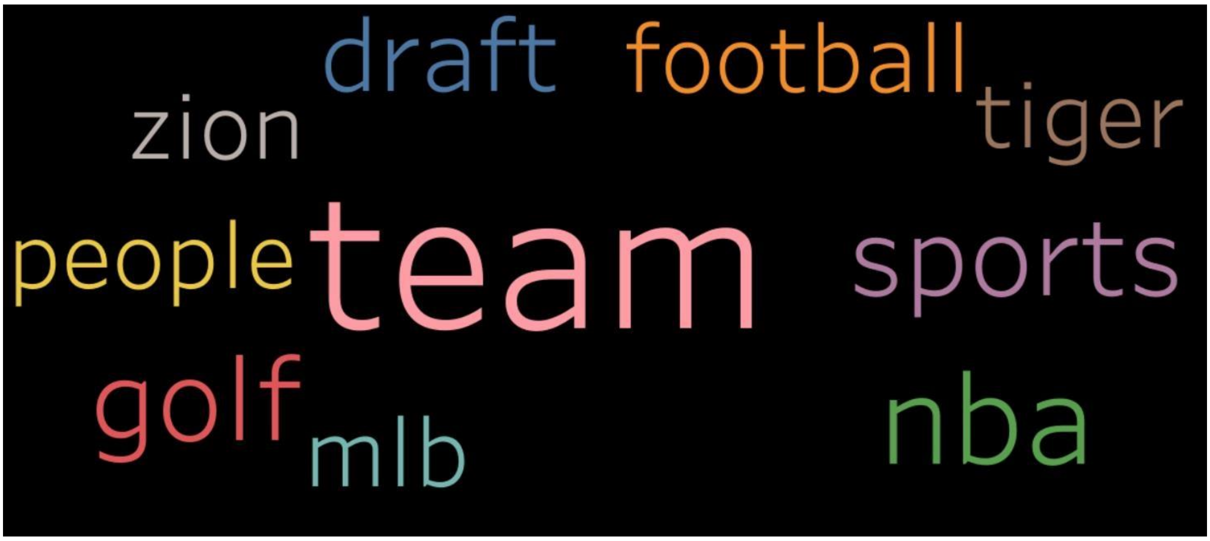
Before feeding the data to mapreduce the twitter, NYT and Common Crawl the data processed i.e.
cleaned to have only required words. The code for the same is provided in the folder.

Processing and Visualization: Part 3

1. Twitter:

For twitter data the tweets have many emoticons, links, @'s, hash tags, punctuations to clean them all "nltk.tokenize", "string", "nltk.corpus" package in python is used. The code for the same is provided in the folder.

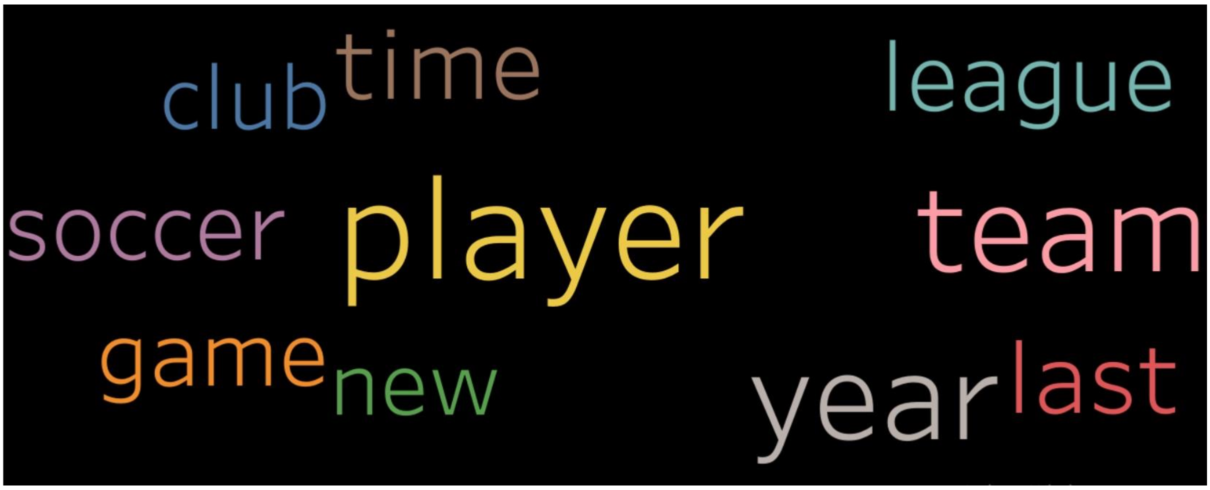
Now for the cleaned data in Hadoop, written code for mapper (word count and word cooccurrence) and reducer is executed. The outputs from HDFS file system are taken for visualization. Below is visualization plot from Tableau.



2. NYT:

Similarly for NYT data each article is extracted as a paragraph in a text file. As part of cleaning Beautifulsoup package was used to extract passages from articles and stemming operation is done on paragraphs along with removing pronouns, punctuations and junk words. Cleaned paragraphs are now appended to a list and downloaded as text files.

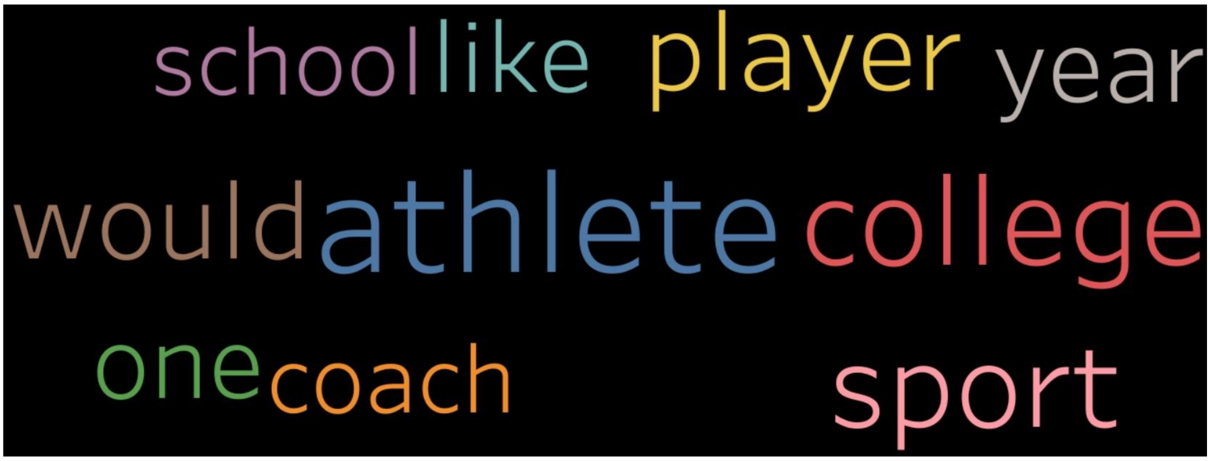
The above obtained text file is processed using mapreduce code in Hadoop for word count and word co-occurrences. The top 10 counts are visualized as below from tableau.



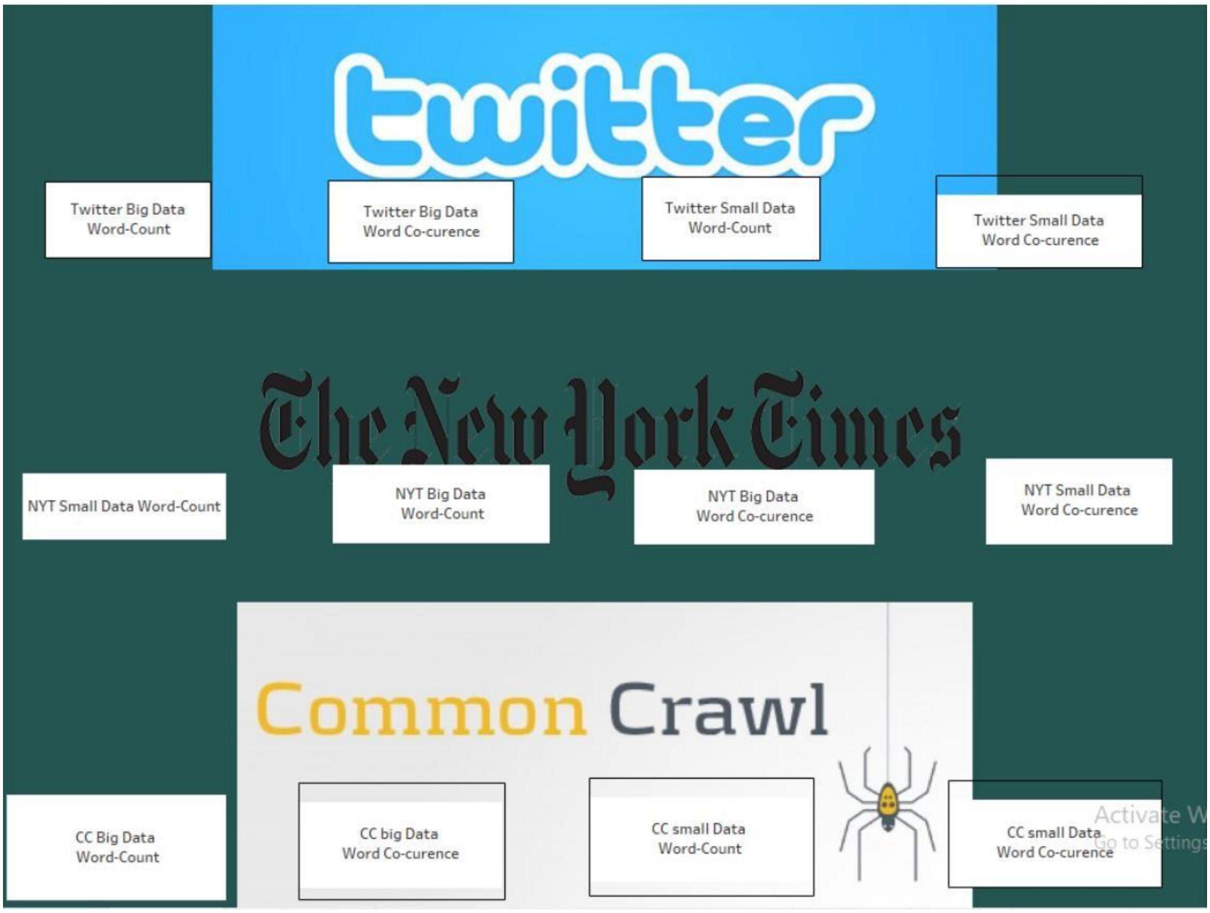
3. Common Crawl:

In common crawl WET files it is first searched for English language and then for key words like sports, athlete and football and the web pages with these content are being fetch. Further data cleaning for junk words and junk content is removed from paragraphs

The final text data is then processed with mapreduce for word count and word cooccurrence and top 10 counts are used to form a word cloud in Tableau.



Using Tableau an Interactive dashboard has been made and below is the screenshot for the same.


twitter
Twitter Big Data Word-Count
Twitter Big Data Word Co-curence
Twitter Small Data Word-Count
Twitter Small Data Word Co-curence
The New York Times
NYT Small Data Word-Count
NYT Big Data Word-Count
NYT Big Data Word Co-curence
NYT Small Data Word Co-curence
Common Crawl
CC Big Data Word-Count
CC big Data Word Co-curence
CC small Data Word-Count
CC small Data Word Co-curence
Activate W
Go to Settings