LAB2: DATA AGGREGATION, BIG DATA ANALYSIS AND VISUALIZATION

DEBOSHREE MAZUMDAR(deboshe), VANDANA PRASAD GAALLEE(vandanap)

April 21, 2019

1 Introduction

The project involves data collection from three major sources: Twitter. NewYork Times and Common Crawl. We have selected the main topic as Politics with subtopics: Trump, Democrats, Republicans, Election and President.

2 Data Collection

2.1 Twitter

We have collected the tweets using tweepy streaming API. Total tweets collected for 14 MB in size. The tweets have been cleaned and pre-processed using regular expressions and python nltk package.

2.2 NewYork Times

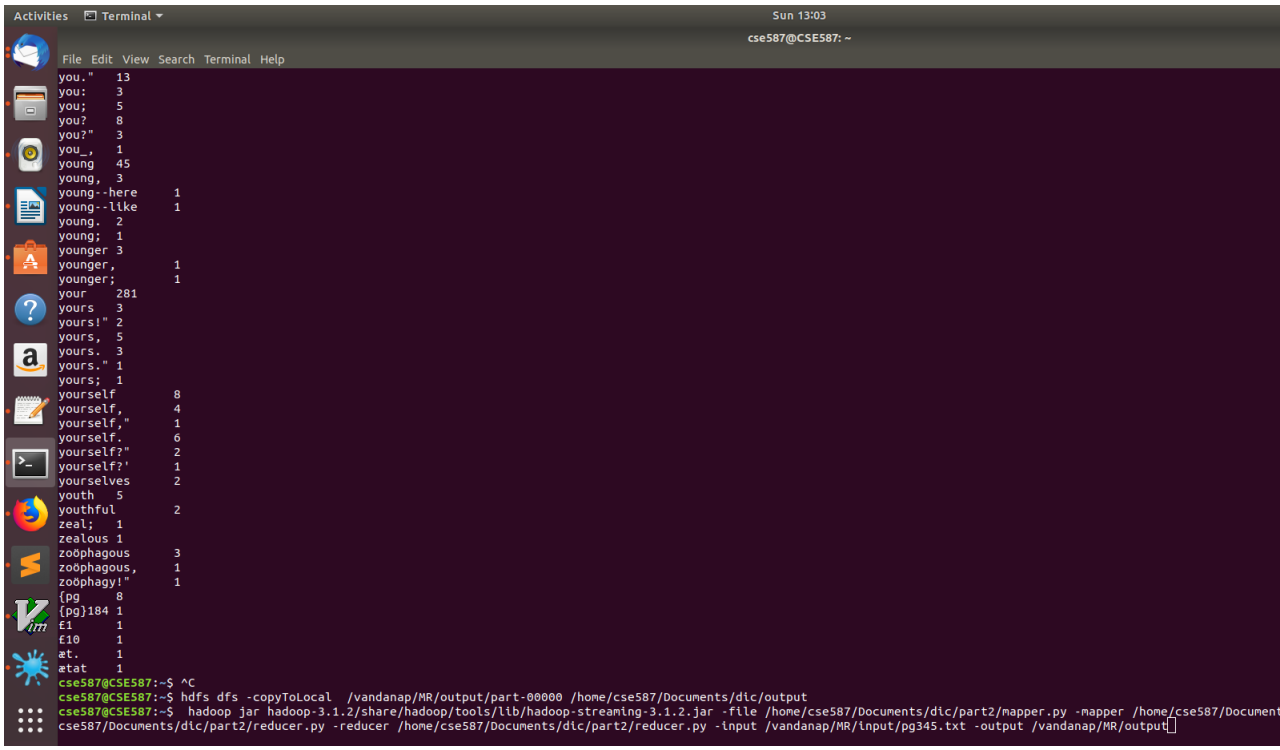
NewYork times has the article API which has been used to collect 500 articles.The data is approximately 19 MB in size. All the HTML tags have been parsed using beautiful soup.

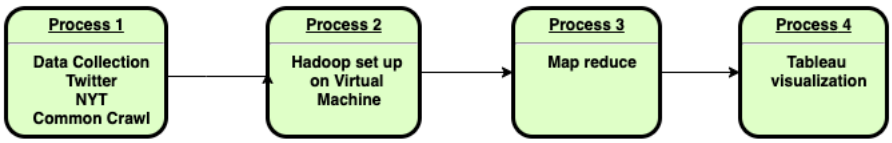
2.3 Common Crawl

We retrieved data from USA today filtering out with respect to the sub-topics we have. After collection, we processed the data using a python script. The data size is approximately 74 MB.

3 Hadoop setup

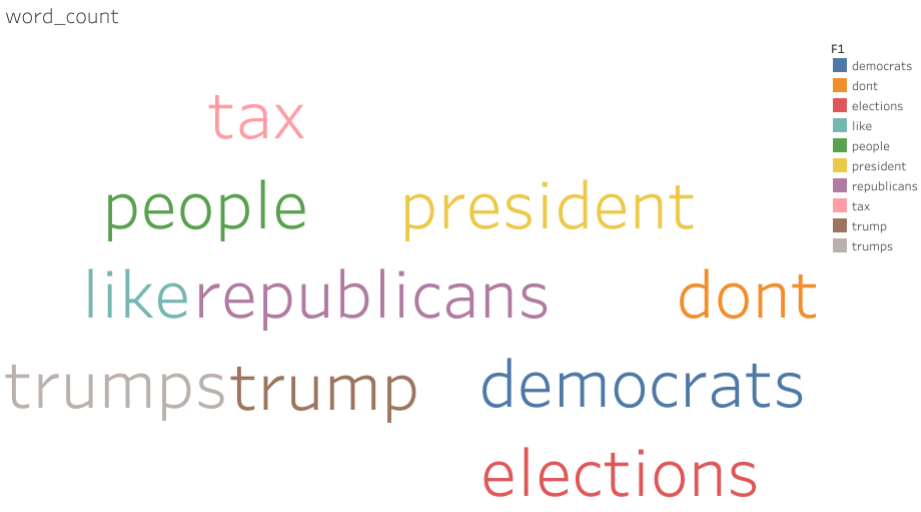
we installed oracle VM and added the image provided, created all folders in Hdfs file system and transferred the data and lastly, executed the map reduce program.



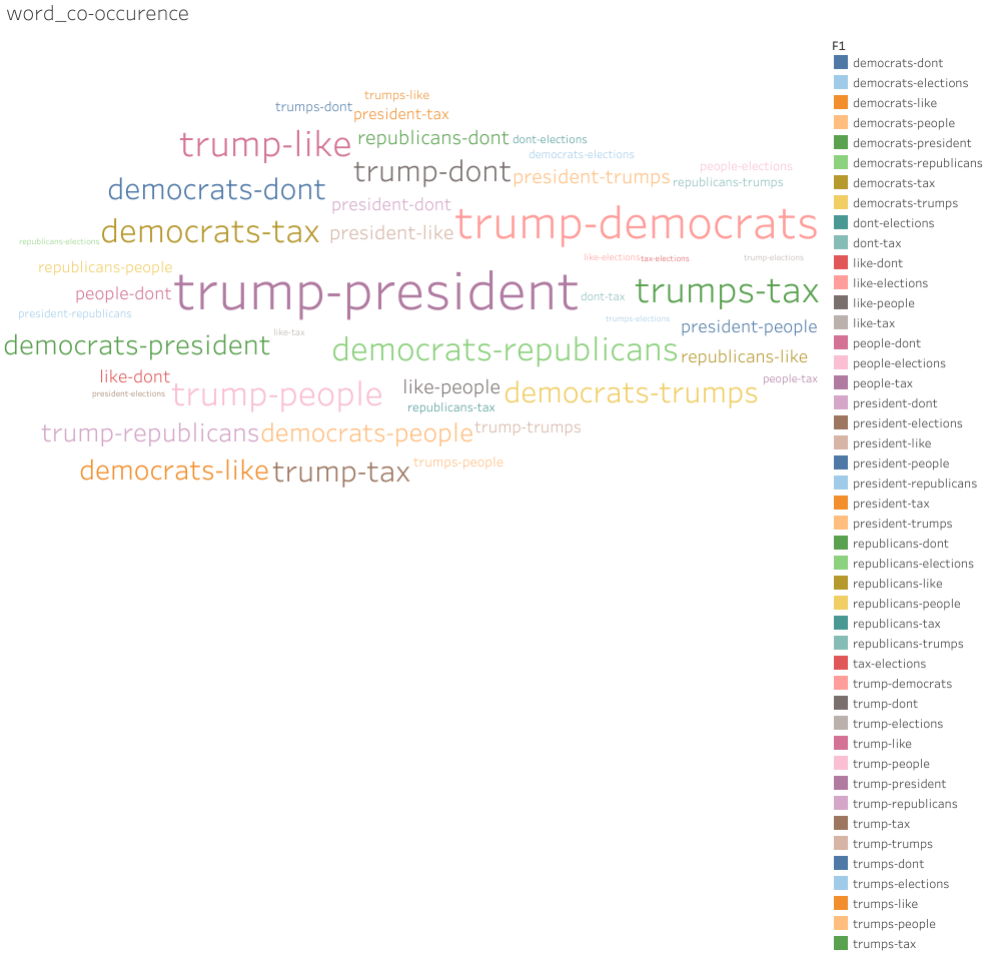


4 Tableau visualization

4.1 Twitter

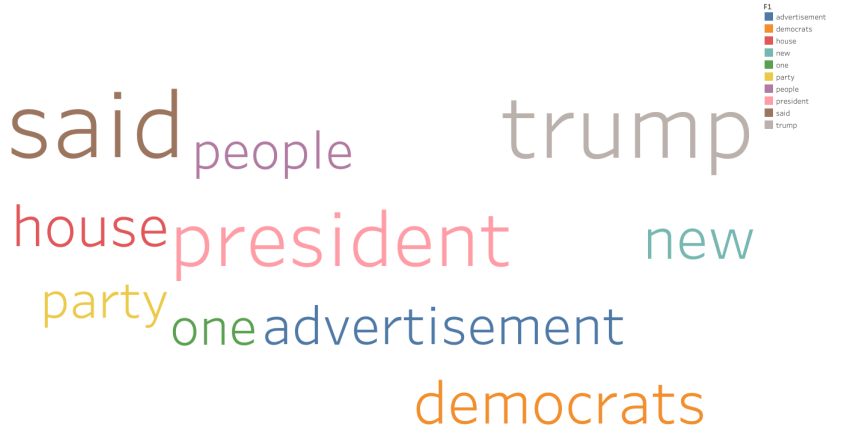


4.2 NewYork Times

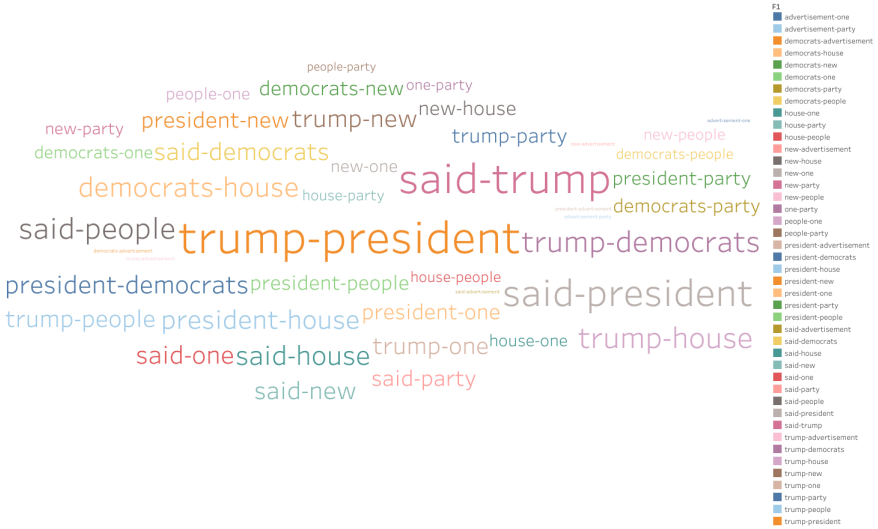


4.3 Common Crawl

NewYork Times word count



Newyork Times word co-occurence




word_count
word_co-occurence
Dashboard 1
Common Crawl Word count
F1
first
getty
lady
new
one
photo
today
trump
usa
white
white
one
usa
today
trump
getty
lady
first
photo
new


word_count
word_co-occurence
Dashboard 1
Common Crawl Word Co-occurence
F1
first-lady
first-one
first-photo
first-today
first-white
getty-first
getty-lady
getty-new
getty-one
getty-photo
getty-today
getty-white
lady-one
lady-photo
lady-today
lady-white
new-first
new-lady
new-one
new-photo
new-today
new-white
one-photo
one-today
one-white
today-photo
trump-first
trump-getty
trump-lady
trump-new
trump-one
trump-photo
trump-today
trump-white
usa-first
usa-getty
usa-lady
usa-new
usa-one
usa-photo
usa-today
trump-today
lady-photo
getty-one
usa-white
new-lady
trump-white
lady-today
white-today
usa-trump
trump-photo
today-photo
first-one
getty-new
getty-lady
getty-today
trump-first
usa-one
new-first
white-photo
one-photo
first-photo
first-lady
trump-lady
usa-photo
usa-new
new-photo
usa-today
usa-first
new-today
new-white
F1: new-white
F2: 1,581
new-one
getty-first
usa-getty
getty-white
lady-one
getty-photo
one-today
first-white
lady-white
trump-one
one-white
first-today
trump-new
usa-lady
trump-getty