CSE 587 Data intensive Computing Lab2

Aayush Kumar
aayushku@buffalo.edu

Sidharth Mishra
smishra9@buffalo.edu

Introduction

Data pipeline is a system that captures, organizes, and routes data so that it can be used to gain insights and is one of the most critical operations in today's data-driven enterprise. Raw data contains too many data points that may not be relevant. Data pipeline architecture organizes data events to make reporting, analysis, and using data easier.

For data Intensive computing Lab2 assignment, we have created a data pipeline using data aggregation from three data source Twitter [1], New York Times [2] and Common Crawl [3]. After data collection we do cleaning and use big data analytic methods of MapReduce to process the data. Once processed, data gets stored into the WORM infrastructure Hadoop. Finally, we built a visualization using Tableau.

Figure 1: A data pipeline model

We have chosen "economy for USA" as our major topic for this project. Again, the subtopics for this major topic are chosen to analyze the impact of these on Economy of USA and mentioned below.

- Technology
- Healthcare
- Stock Market
- Crime

The detailed explanation of the files, directory structure and steps to run the application is given in the README.txt file available inside the lab2 and a short video is also available at: https://buffalo.app.box.com/file/443466916555

The three modules Data Collection/Aggregation, Data Processing, Data Visualization is explained in details in the next section.

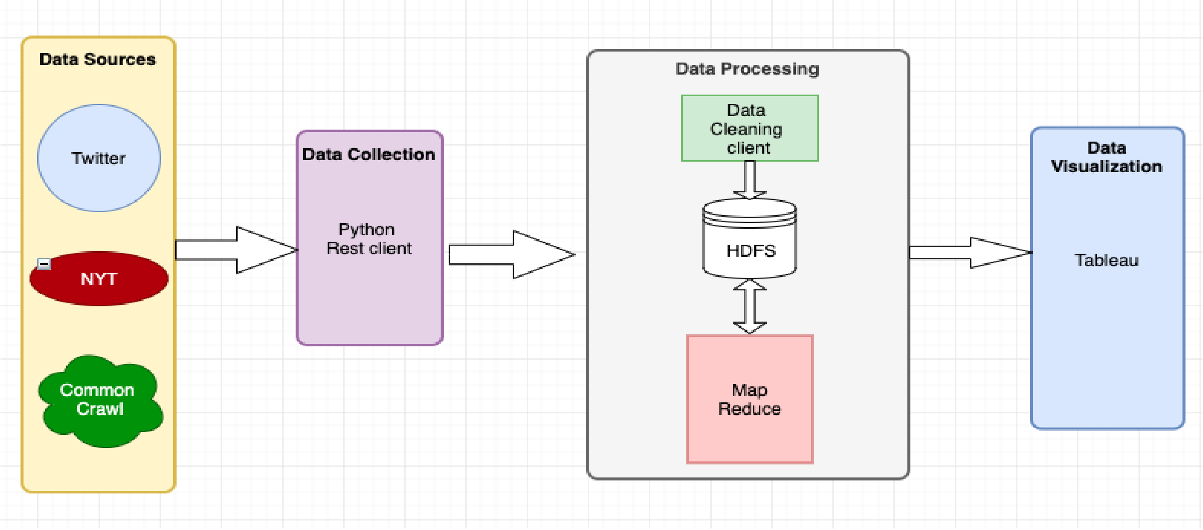


Figure 2 - **Architecture diagram of our system**

Data Collection:

As already mentioned after choosing main topic and subtopics some key phrases used to collect data are:

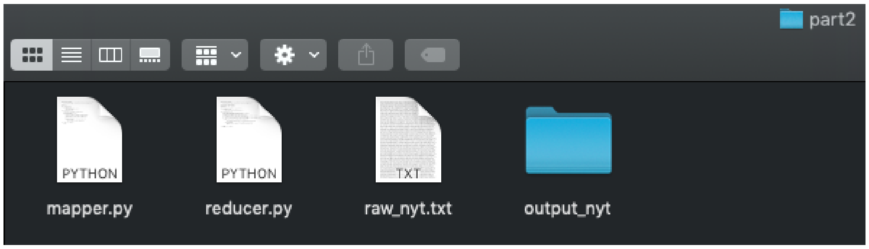
> *'nasdaq', 'exchange', 'trading', 'trader', 'share', 'stocks', 'retailers', 'drugs','money', 'commodity', 'diet', 'disease', 'hospital', 'patient','felony', 'authority',fitness', 'health', 'analysis', 'blockchain', 'technology', 'microsoft', 'judicial','digital', 'corruption', 'crime', 'alcohol', 'fraud', 'government', 'conspiracy'*

Number of records collected:

1. Tweets – 70,000 tweets approximately
2. NYT articles – 3500 articles approximately
3. Common Crawl – 2500 topic pages containing 6,80,000 lines approximately

Setup Big data Infrastructure:

Cloudera Hadoop docker has been used for processing dataset.



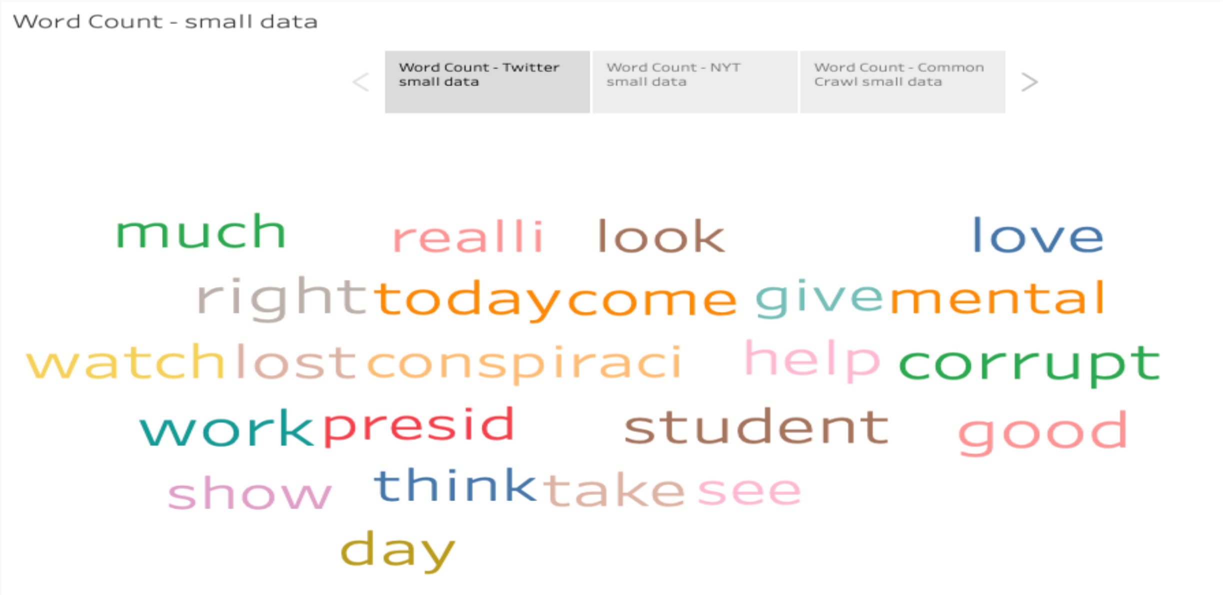
For the purpose of testing the setup, word count has been run of some of articles collected with small dataset. (available at LAB2 -> part2)

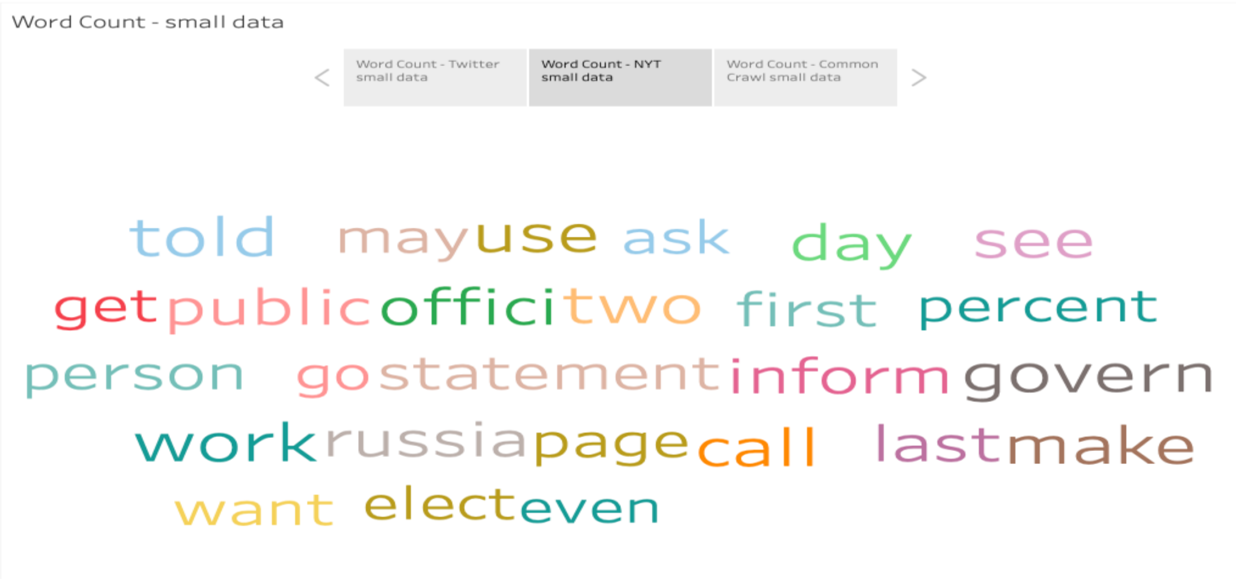
a. Following cleaning has been done before processing the data:

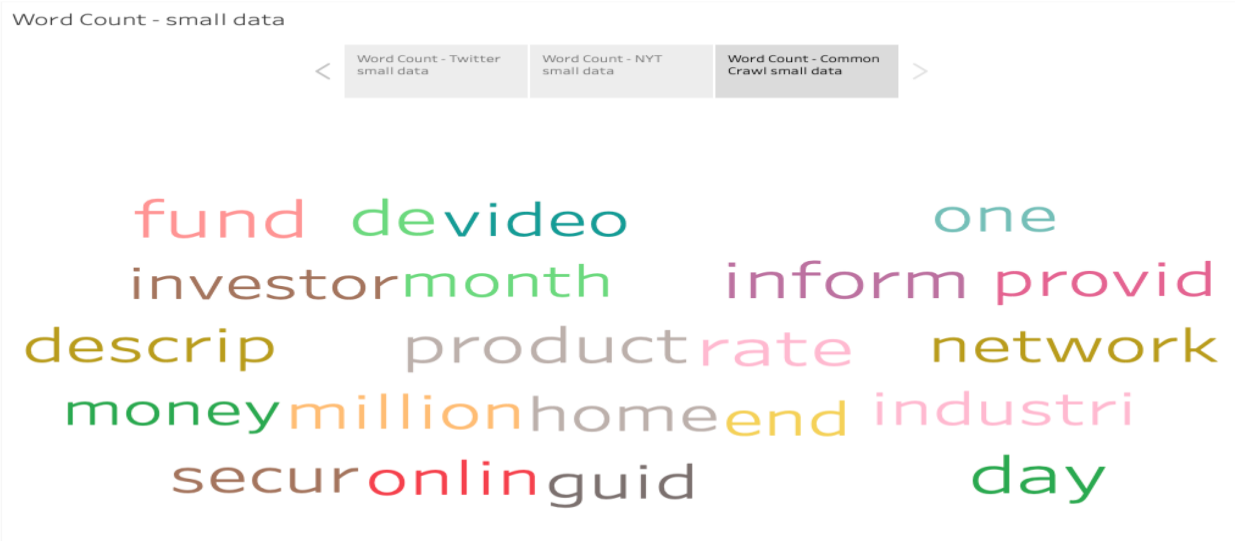
1. Removal of stop words
2. Removal of URLs
3. Removal of special characters
4. Stemming of complete dataset
5. Removal of alphanumeric and special characters.

After performing above operations, a cleaned dataset consisting of only text is used for further analysis.

b. Below is the world cloud for sample word count with small dataset.

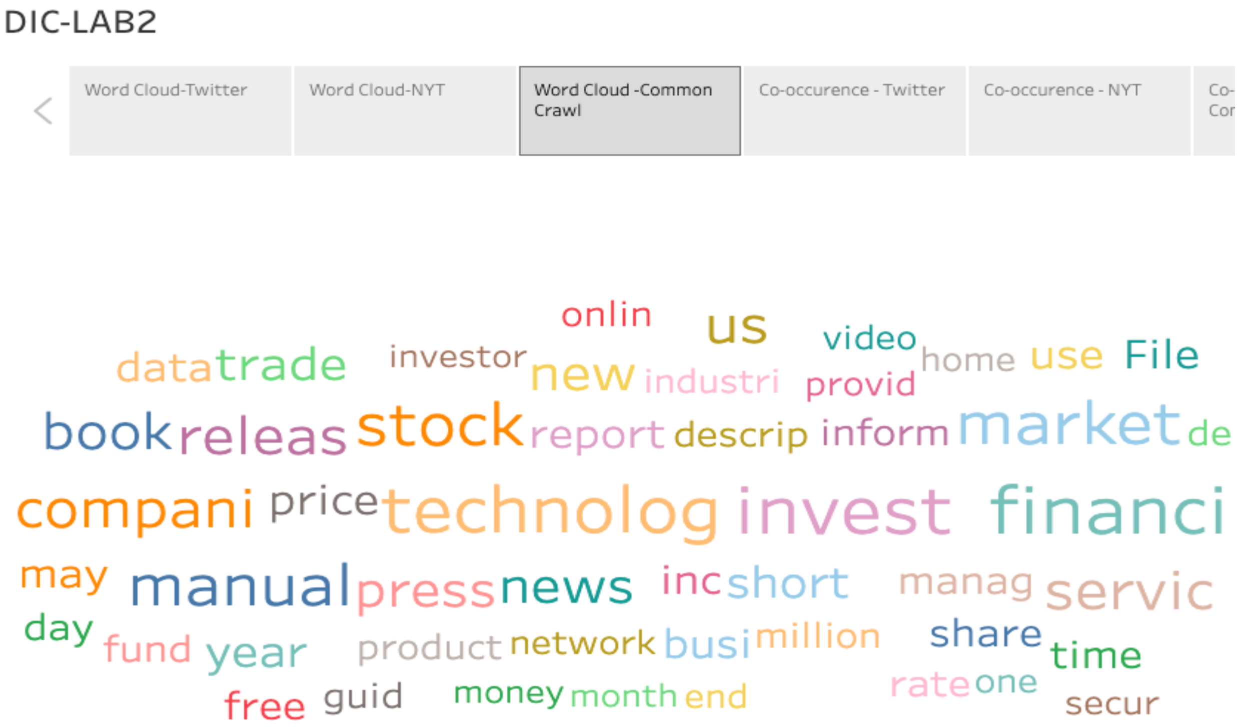


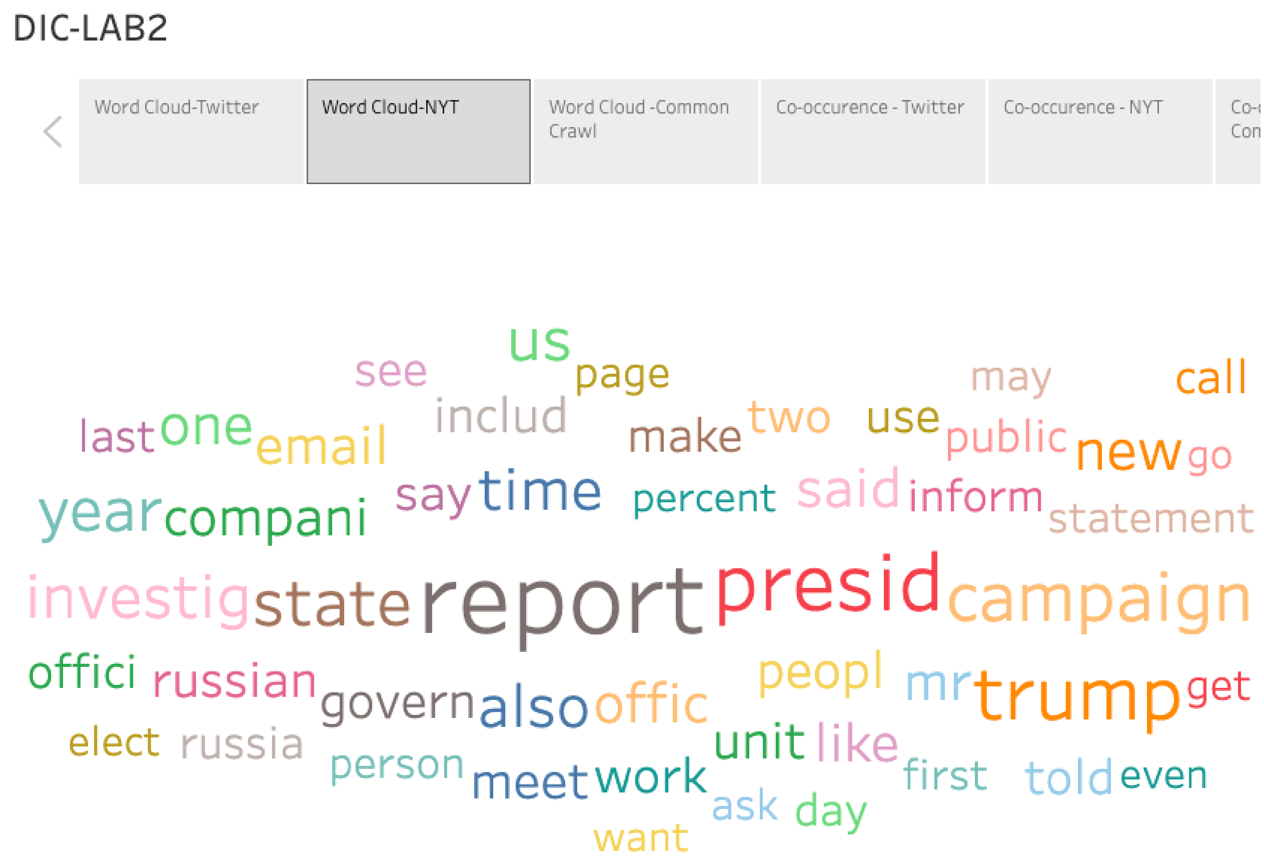


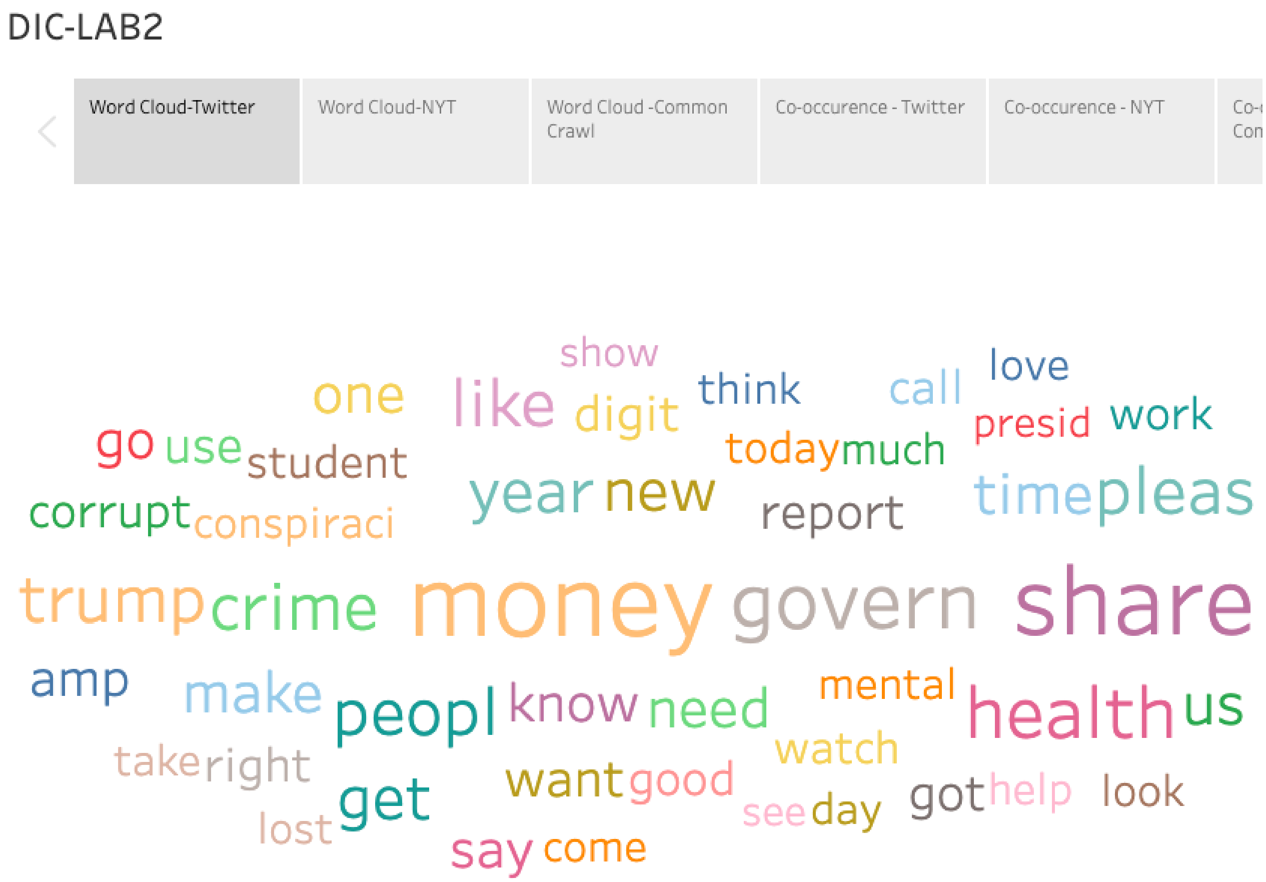


The visualization can be viewed in tableau from the link given below: https://public.tableau.com/profile/aayush2816#!/vizhome/Lab2-DICsmalldata/Story1

c. and d. The above task is repeated for big data and below is the word cloud for all the sources: We found some convergence in the text as the amount of data increased.

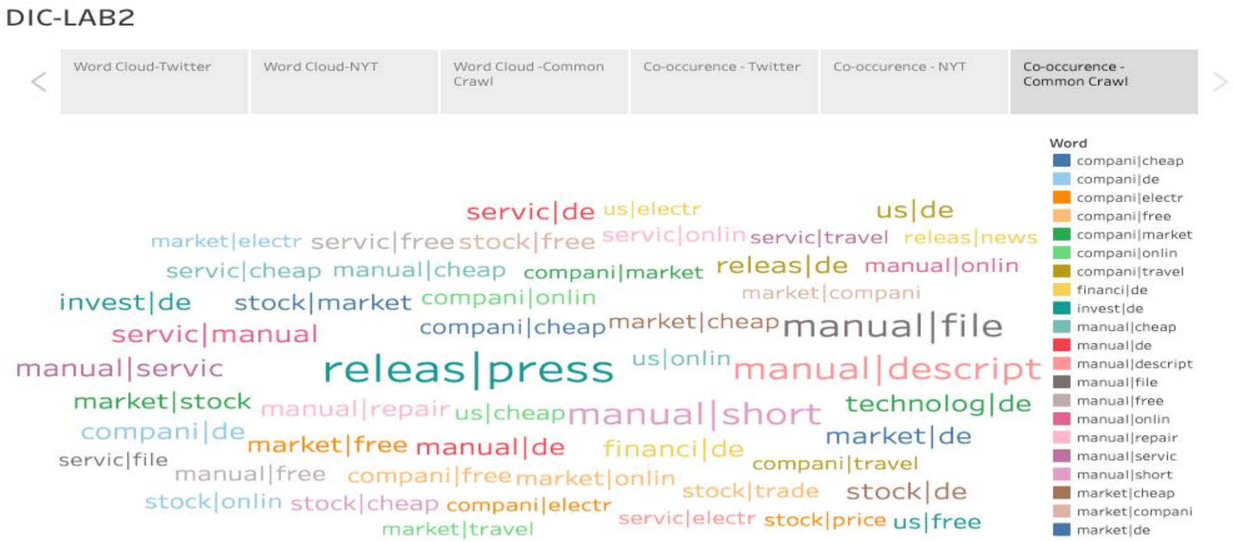

DIC-LAB2
Word Cloud-Twitter
Word Cloud-NYT
Word Cloud -Common Crawl
Co-occurence - Twitter
Co-occurence - NYT
Co-
Cor
onlin
us
video
data
trade
investor
new
industri
provid
home
use
File
book
releas
stock
report
descrip
inform
market
de
compani
price
technolog
invest
financi
may
manual
press
news
inc
short
manag
servic
day
fund
year
product
network
busi
million
share
time
rate
one
free
guid
money
month
end
secur

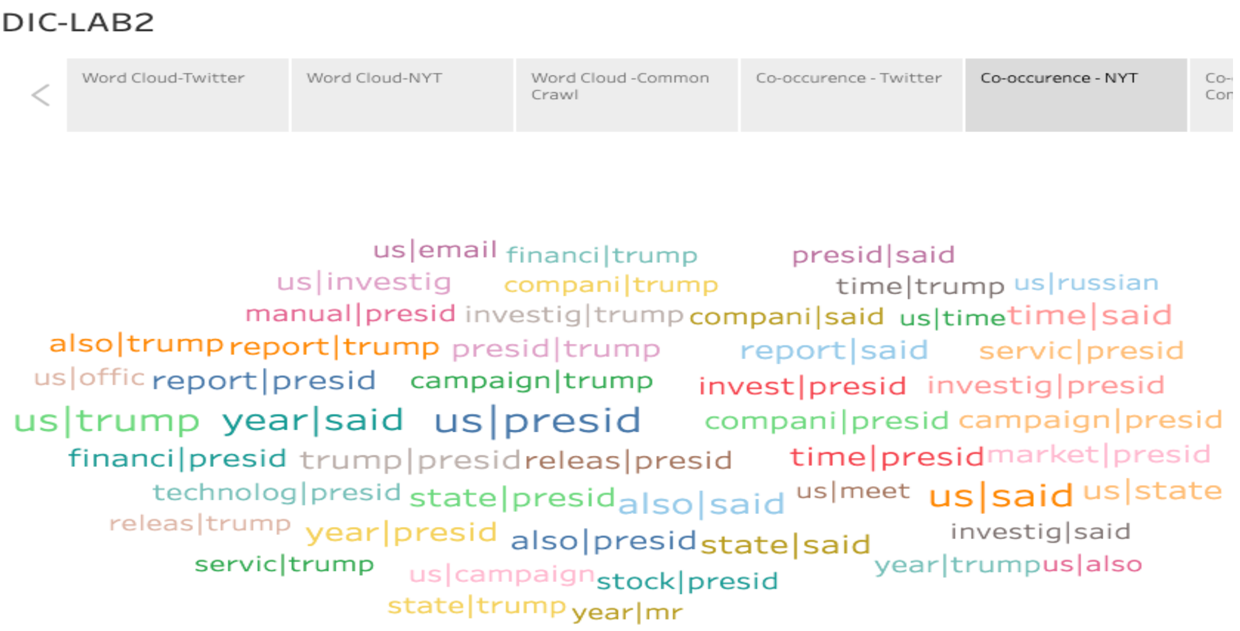

DIC-LAB2
Word Cloud-Twitter
Word Cloud-NYT
Word Cloud -Common Crawl
Co-occurence - Twitter
Co-occurence - NYT
Co-
Con
us
see
page
may
call
includ
last
one
email
make
two
use
public
new
go
year
compani
say
time
percent
said
inform
statement
investig
state
report
presid
campaign
offici
russian
govern
also
offic
peopl
mr
trump
get
elect
russia
person
meet
work
unit
like
first
told
even
ask
day
want

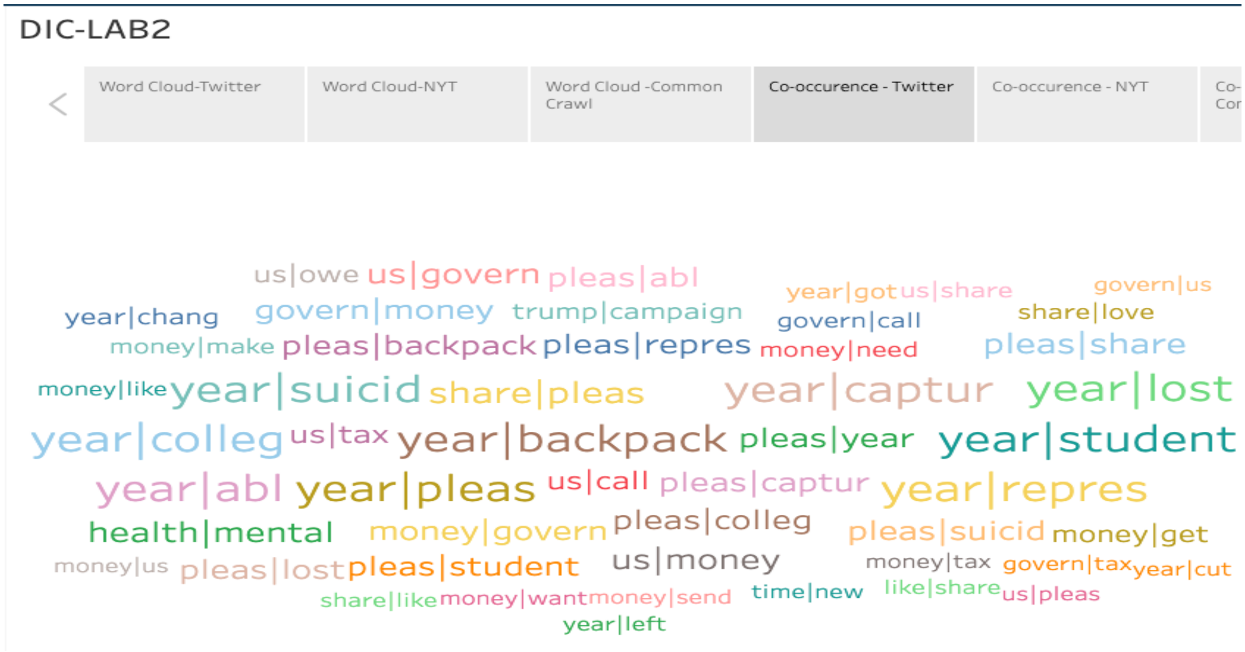


e. Deeper analysis using word co-occurrence:

The top ten words generated from the word cloud have been used to analyze the pair of words which co-occur in Common crawl pages, tweets and news articles. Pair co-occurrence algorithm is implemented to compute the same and below are the results:







All the above visualization can be viewed in tableau from the link given below: https://public.tableau.com/profile/aayush2816#!/vizhome/DIC-LAB2/Story2

Conclusion:

From the Tableau visualization, it's quite interesting to see some top words and how they co-occur with other words in all the data sources. For all the three data sources, we see similar outcome where the top results are pertaining to stock market, trading, crime, money, trump, market, finance, invest, suicide, etc. trending in past two months. The similar data pipeline approach can be applied for other data sources to get good insight of data.

References:

1. Twitter API. Twitter Developer https://dev.twitter.com/
2. https://developer.nytimes.com/docs/articlesearch-product/1/overview
3. http://commoncrawl.org/2015/04/announcing-the-common-crawl-index/
4. https://onlinehelp.tableau.com/current/pro/desktop/en-us/copy_b_wkbks.htm