**LAB2 – PART2 Description**

Below are the Steps that needs to be done for Part2 of this project

1. Download the Tweets and New York times article data using the API . I have used **Bitcoin** as the search term. The codes are in the folder “***codes for downloading Tweets and Articles***”.

* Use ***Tweets download*** code in R for tweet download. Output is in folder “***Twitter Data***”
* Use ***NYT download*** code in python for articles download. Output is in folder “***NewsData***”

1. Now we have the data it’s time to run the code. I have also made stop word list that has to be used to run the code. Code used is TwitterWordCount.jar

* Transfer the files to Hadoop system

hdfs dfs –put Twitterdata input1

hdfs dfs –put Newsdata input2

hdfs dfs –put stopwordlist

* Now run the command given below for Twitter Data

hadoop jar /home/hadoop/hadoop/share/hadoop/mapreduce/TwitterWordCount.jar TwitterWordCount -Dwordcount.case.sensitive=false input1 output1 -skip

stopwordslist.txt

* Now run the command given below for News Data

hadoop jar /home/hadoop/hadoop/share/hadoop/mapreduce/TwitterWordCount.jar TwitterWordCount -Dwordcount.case.sensitive=false input2 output2 -skip stopwordslist.txt

* Transfer the files back to system using

hdfs dfs –get output1

hdfs dfs –get output2

* Now copy the output to csv file as Twitter\_Output.csv and NYT\_Output.csv and save it in the ***“Twitter\_One\_Week “***and ***“New\_York\_Times\_Full\_Data”*** folder with header as “word” and “Count”
* Now time to run the html file in the folder given. It has html file , d3.layout.cloud.js and medley.js to get the output. ***NOTE: Code runs in Microsoft Edge , doesn’t seem to run in chrome in my system***
* Twitter Output



* NYT Output



***Word Cooccurence***

1. Now we have to find word co-occurrences using the data . I have also made stop word list that has to be used to run the code. This time code used to Word-Cooucurence.jar

* Now run the command given below for Twitter Data

hadoop jar /home/hadoop/hadoop/share/hadoop/mapreduce/WordCoOccurence.jar WordCoocurence -Dwordcount.case.sensitive=false input1 output3 -skip stopwordslist.txt

* Now run the command given below for News Data

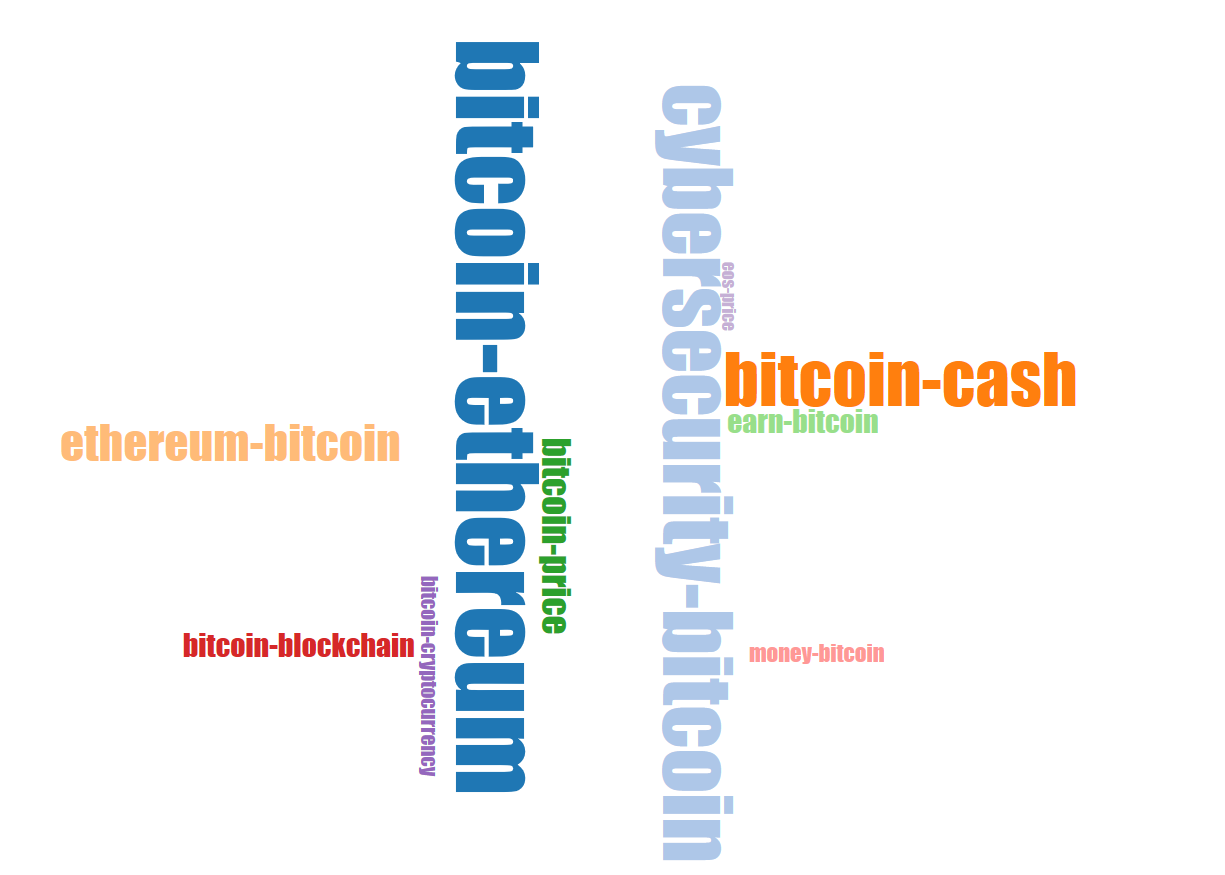
hadoop jar /home/hadoop/hadoop/share/hadoop/mapreduce/WordCoOccurence.jar WordCoocurence -Dwordcount.case.sensitive=false input2 output4 -skip stopwordslist.txt

* Transfer the files back to system using

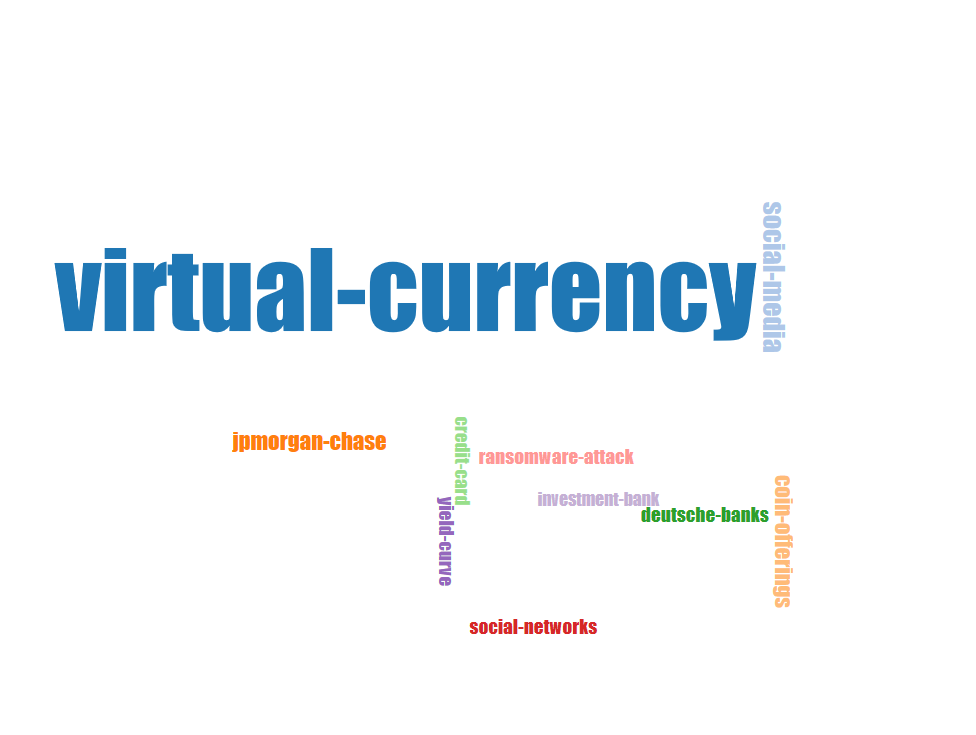
hdfs dfs –get output3

hdfs dfs –get output4

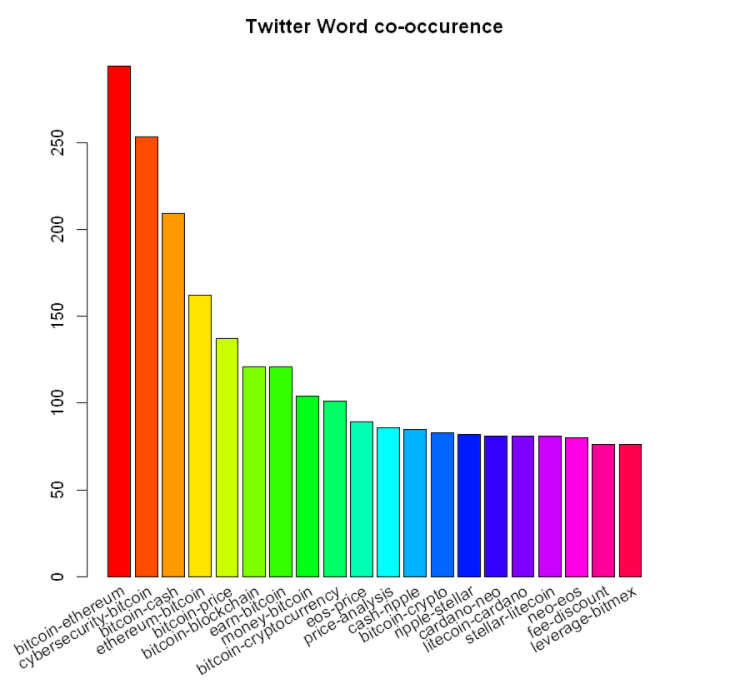
* Now copy the output to csv file as Twitter\_word\_coocurence.csv and NYT\_word\_coocurence.csv and save it in the ***“Twitter\_Word\_Co\_occurnece “***and ***“New\_York\_Times\_Word\_Co\_occurence”*** folder with header as “word” and “Count”
* Now time to run the html file in the folder given. It has html file , d3.layout.cloud.js and medley.js to get the output. ***NOTE: Code runs in Microsoft Edge , doesn’t seem to run in chrome in my system***
* Twitter Output

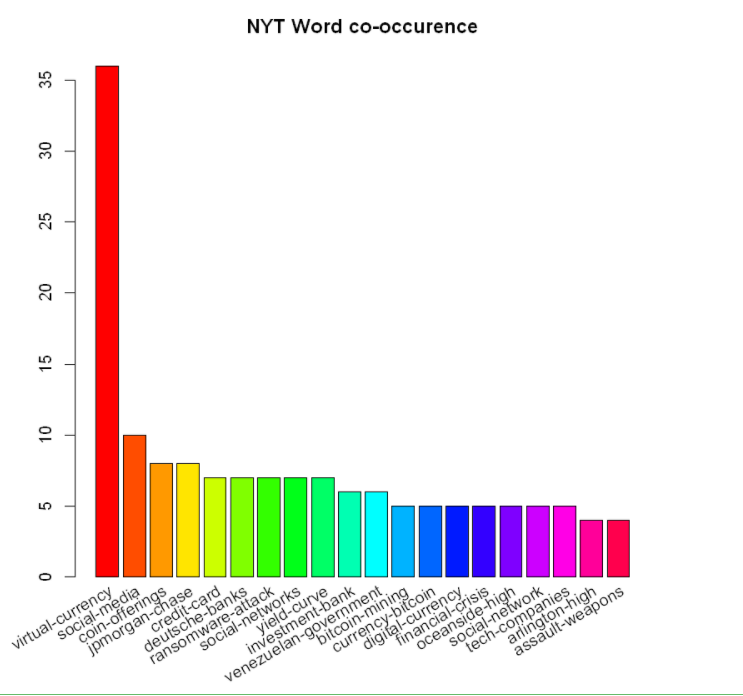


* NYT Output



* We can also visualize using by running the R code in the jupyter file given to visualize the output. Just choose appropriate file when asked.





* In this way we used API to download, then Hadoop to process and finally D3.js to visualize the output
* I have also done this processing for some other data also which is in the folder “***Other Examples***”
* A video describing the procedure is also included

AUTHOR and Project Done by :

Sachin Kumar Kuppayya

50248933

Refernce :

1. <https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html>
2. <https://github.com/d3/d3/wiki/Gallery>
3. <https://www.youtube.com/watch?v=1KEiTIu0k44> – Rapid D3.js