4/8/2018

Lab 2 : DATA AGGREGATION, BIG DATA ANALYSIS AND VISUALIZATION

CSE 487/587

Palwinder Singh
50247454
palsingh@buffalo.edu

Gursimrat Singh
50212333
gsingh37@buffalo.edu

Work Description

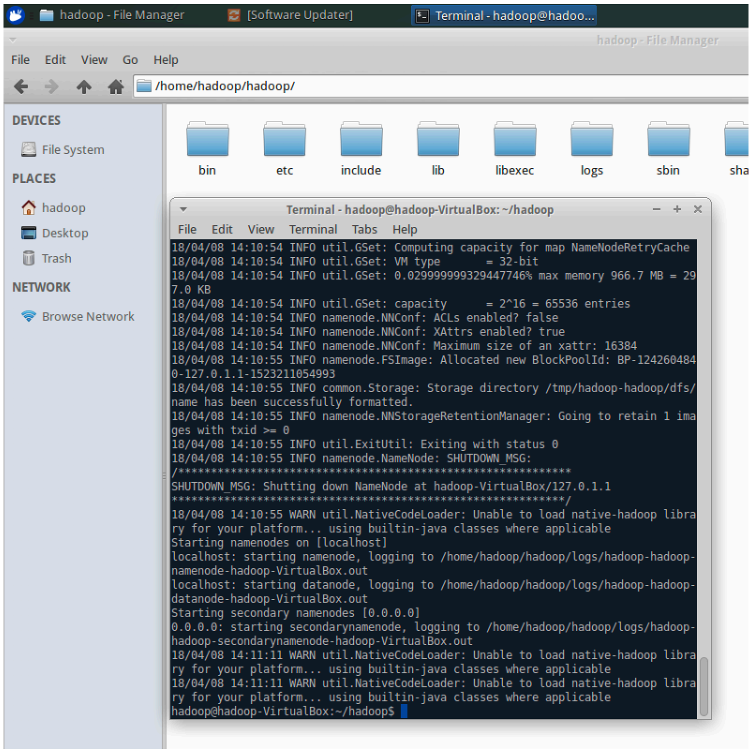
PART 1:

We ran all the code exercises present in the Chapters 3,4 & 5 from the book ‘**The Data Science Handbook-Field Cady’,** in Jupyter notebooks.

PART 2:

a. The topic we chose is the one in the news all the time,**’Gun Violence’**. We collected tweets using the twitter search API, and also fetched NY Times articles for the same topic by using the python library ‘Beautiful Soup4’.
   We got a lot of data from the above and we segregated the data in two different categories: TwitterData & NewsData, and stored the in different directories.

b. We installed the VM appliance for Hadoop infrastructure and test the basic commands with the sample data provided.



```
18/04/08 14:10:54 INFO util.GSet: Computing capacity for map NameNodeRetryCache
18/04/08 14:10:54 INFO util.GSet: VM type       = 32-bit
18/04/08 14:10:54 INFO util.GSet: 0.029999999329447746% max memory 966.7 MB = 29
7.0 KB
18/04/08 14:10:54 INFO util.GSet: capacity      = 2^16 = 65536 entries
18/04/08 14:10:54 INFO namenode.NNConf: ACLs enabled? false
18/04/08 14:10:54 INFO namenode.NNConf: XAttrs enabled? true
18/04/08 14:10:54 INFO namenode.NNConf: Maximum size of an xattr: 16384
18/04/08 14:10:55 INFO namenode.FSImage: Allocated new BlockPoolId: BP-124260484
0-127.0.1.1-1523211054993
18/04/08 14:10:55 INFO common.Storage: Storage directory /tmp/hadoop-hadoop/dfs/
name has been successfully formatted.
18/04/08 14:10:55 INFO namenode.NNStorageRetentionManager: Going to retain 1 ima
ges with txid >= 0
18/04/08 14:10:55 INFO util.ExitUtil: Exiting with status 0
18/04/08 14:10:55 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at hadoop-VirtualBox/127.0.1.1
************************************************************/
18/04/08 14:10:55 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /home/hadoop/hadoop/logs/hadoop-hadoop-
namenode-hadoop-VirtualBox.out
localhost: starting datanode, logging to /home/hadoop/hadoop/logs/hadoop-hadoop-
datanode-hadoop-VirtualBox.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/hadoop/hadoop/logs/hadoop-
hadoop-secondarynamenode-hadoop-VirtualBox.out
18/04/08 14:11:11 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
18/04/08 14:11:11 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
hadoop@hadoop-VirtualBox:~/hadoop$
```

c. We loaded the data as per instructions ,that is loaded the data aggregated in step (a) into the VM, two directories: TwitterData and NewsData. Each directory can have many files of data.

d. Coding language used: Python
We coded our Mapper and Reducer in python and cleaned and parsed the data sets into words, remove stop words, and reduce will count the useful words.
This is done in the Hadoop streaming jar.

Twitterdata->TwitterWords and NewsData ->NewsWords

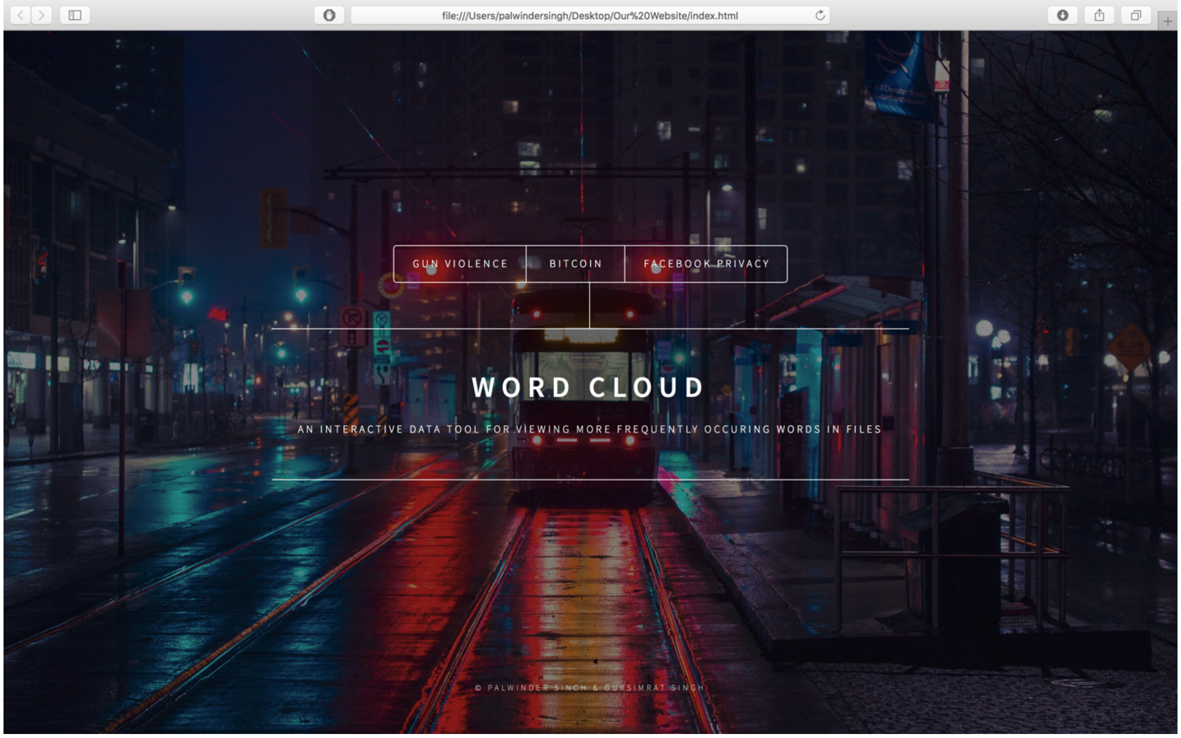
Now we also choose two more topics for our ananlysis : Bitcoin & Facebook Privacy.

e. UseNow for the visualization we used the wordcloud script from d3js.org .
The data we get from our reducer is a raw file which we will use as a csv or a tsv.
   - The script ref : Word cloud layout by Jason Davies, http://www.jasondavies.com/word-cloud/
   - Algorithm due to Jonathan Feinberg, http://static.mrfeinberg.com/bv_ch03.pdf

We got good results for the visualization,using the script.

f. Next we repeated the steps c) to e) for larger data set collected over week.We inferred that the word occurring the most in the smaller data set , is also the most occurring word in the larger data set collected .Eg : gun.

g. We made a website which gives us the word cloud for our chosen three topics,the webpage is interactive and when you click on the topic the wordcloud is displayed using the d3js script.



file:///Users/palwindersingh/Desktop/Our%20Website/index.html

WORD CLOUD

AN INTERACTIVE DATA TOOL FOR VIEWING MORE FREQUENTLY OCCURING WORDS IN FILES

© PALWINDER SINGH & GURSIMRAT SINGH

h. Now using the matrix algorithm we found out the co-occurrence and then we picked up the top ten co-occuring pairs which were non repeating .
Our “map” function emits <word$word,value> and your “reduce” function should collate the co-occurrences for the top ten words and output them in a suitable format.

Format: parkland$shooting 655

i. Documentation done.
j. A video explaining the project analysis and the visulizations has been put up in the google drive .

Link:

https://drive.google.com/file/d/1rUkLen56pIxCp0L4gsd2F_HXTHXQQZpe/view?usp=sharing

References & resources:

1. D3js.org -Word cloud script by Jason Davies.
2. Web crawling library code ref : https://stackoverflow.com/questions/1936466/beautifulsoup-grab-visible-webpage-text
3. NYtimes article API documentation and code: Stack Overflow help.
4. iMovie for editing, quicktime for video recording.
5. http://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/
6. Website template: https://onepagelove.com/templates/free-templates.