DATA AGGREGATION , BIG DATA ANALYSIS AND VISUALIZATION REPORT

Data Intensive Computing- CSE-587

Jayant Solanki

UBIT Name: jayantso
Person Number: 50246821
SPRING 2018 CSE

Anant Gupta

UBIT Name: anantram
Person Number: 50249127
SPRING 2018 CSE

INTRODUCTION

This report contains the implementation details for the LAB 2 of DIC-CSE 587. We have successfully completed the Part 1 and Part 2 of the Lab 2.

OBJECTIVES ACHIEVED

We used the APIs from NyTimes and Twitter to collect data based on certain topics. We then used MapReduce on the unstructured data to build a visualizing data product with the help of javascripts and D3js. The product displayed the word cloud of most trending topics.

IMPLEMENTATION

Part 1

Part 1 has been done as per described in the prescribed Text book for the Data Analysis.

Part 2

The overall workflow for the part 2 implementation has been shown in the Figure 1 below.

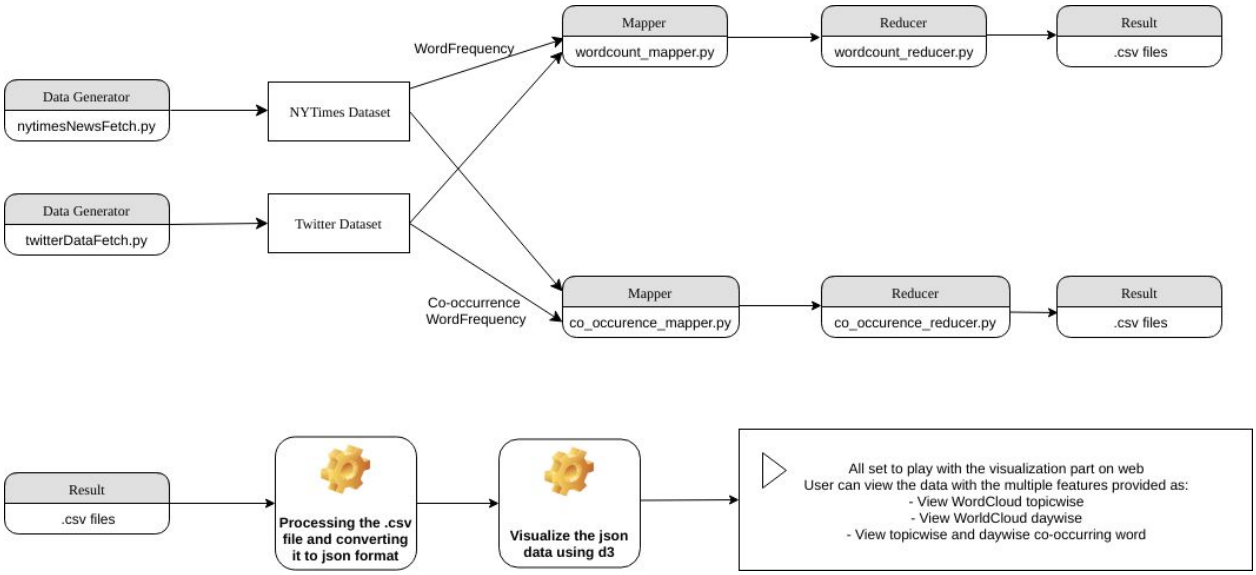

Figure 1

We created two scripts for collecting the data from NY Times and Twitter for five search topics. Data for whole week has been collected ranging from 1st April to 7th April, 2018.

a. We chose 5 topics, trade war, donald trump, south china sea, cambridge analytica, and gun crime.
b. We then collected the data in the textcorpus folder and applied mapreduce to find the word count and word-co-occurrence count.
c. The csv files then generated were used in the d3js visualization.

DIRECTORY LAYOUT

There are 2 folders under **Assignment-2** directory:

- **code :**
    1. **Part1:** Contains the data and the script files for the Lab1 chapter-wise (CH3,CH4,CH5 subfolders in it)
    2. **Part2**: Contains the data and the script files for the Lab2
       Nytimes and twitter folder contains the final csv files which are being used for the web visualization.
       input,output,output_co folders are used to store the text files which can be used further to perform the map-reduce operation.

textcorpus will have the files which are being generated by the data generator python scripts (twitterDataFetch.py and nytimesDataFetch.py)

- **report:** Contains the ipynb files for Lab 2.

VIDEO:

https://buffalo.box.com/s/ltdluqv33nk36y9ojz4wk6bviyh8i03q

INSTALLATION

1. Please copy the whole folder structure in the **/var/www/html/foldername of ubuntu.**
2. The website needs to be there to be accessed.
3. For any difficulty in the installation please do contact us on our email, jayantso@buffalo.edu or anantram@buffalo.edu

SOFTWARE/HARDWARE USED

- Anaconda
- Jupyter Notebook
- Python 3.6 Environment based on Anaconda
- Ubuntu 17, Intel core i7 processor

REFERENCES

1. Stackoverflow.com
2. The Data Science Handbook