Data Intensive Computing – Lab 2

Santosh Phani Vishnu Aita, UBID: saita

Harsha Kosta, UBID: harshako

Data Extraction:

Data extraction is the most important step in any computing or machine learning problem. Even though it is distributed in decent amounts by various organization for the developers to use. It is generally messy and unorganized. Every data has some kind of gibberish and noisy information. Noise is a hindrance to using data directly after extracting from the API or any kind of source.

For lab 2, Raw data is needed to feed to the Map Reduce application which removes all unrequired information like stopwords, punctuations, symbols etc.

Types of data:

For this lab, data from Twitter using RtweetAPI and NYTimes articles using nytimesarticleapi was extracted. This step is quite hideous. Tweets data comprised of location, dates, usernames and much more information. Only Tweet data had to be extracted. So tweet data was fetched and stored in separate files as it contains important information.

In case of NYTimes articles. The nytimes article API only provided JSON response which contained information like urls, dates, pages, location etc. Using the urls from the JSON response the article data was fetched.

Each url was passed to the BeautifulSoup (bs4), a package used in python for web scrapping and all the article data available in the paragraph tags were extracted and appended to form a larger data set.

We collected Tweets and NYTimes articles for Short-Span(one day) and Long-Span(one week). MR produced different results when visualized for weekly and daily data even though the key words were same.

The data from both Twitter and NYTimes was collected for three different keywords.

Hadoop Map Reduce:

As we know, Hadoop is a collection of open-source software utilities that facilitate using a network of many computers to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage and processing of big data using the MapReduce programming model.

We implemented Map Reduce functionality using Python3 as programming language.

Data collected from Twitter and NYTimes is separately fed to the Mapper and each word in the data set is selected if it is not in the stopwords, all the punctuations and unwanted symbols are removed and all the words are stemmed to eliminated similar meaning words using nltk lemmatize.

The list of stopwords was extracted from the nltk package available in python.

The output from the mapper is emitted as <Word, Count> for each word and this is fed as input to reducer after shuffle and sort.

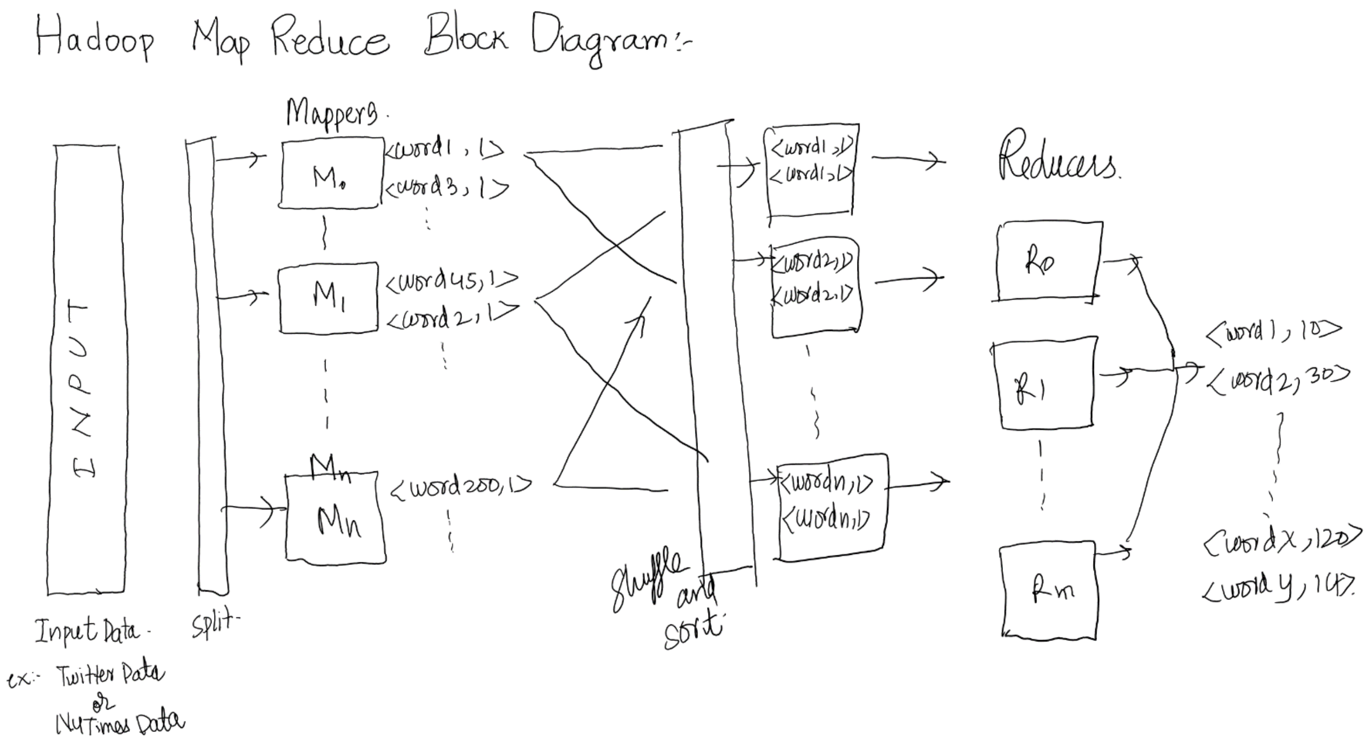
In shuffle and sort phase all the similar key words are paired together and sent to reduce as input.

The reducer aggregates or sums all the word count and emits the total number of occurrences of a word in the data set.

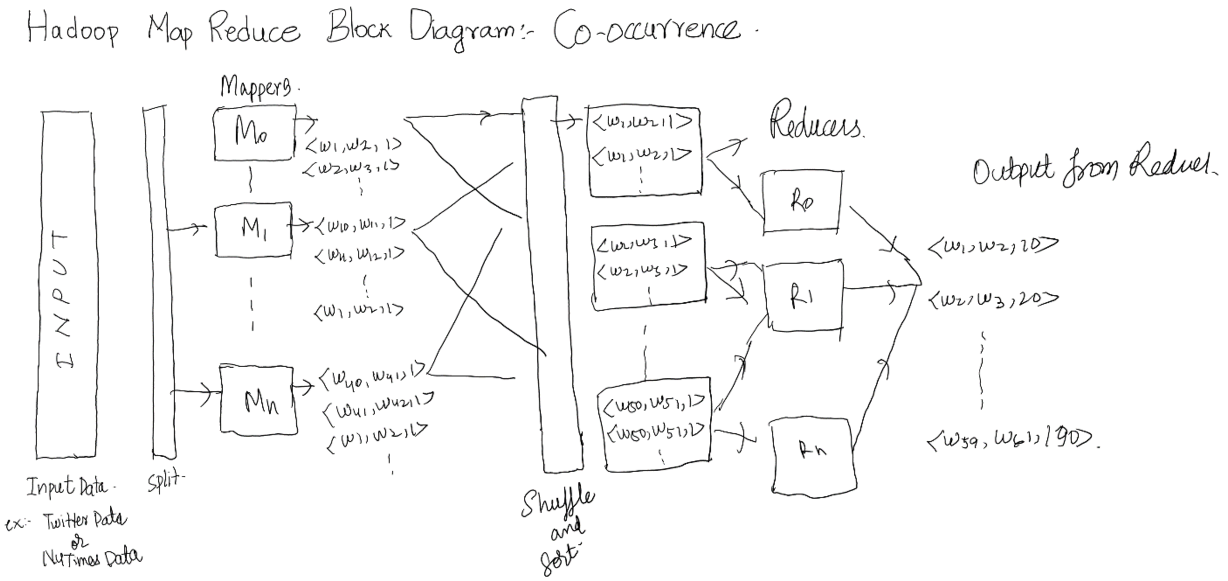
This is then fed to the d3.js to display as a word cloud on a webpage to visualize the data.

Block Diagrams

Word Count MR Problem:



Co-occurrences MR Problem:



Visualization using d3.js:

The output from Map Reduce for various key words and types of data like Twitter and NYTimes are plotted as word clouds using d3.js.

D3.js intakes <word_text, word_count> tuples stored in either a text file or a csv file as input. From this the word_text is the word which is supposed to be displayed in the word cloud and the word_count is used as the size of the word to be displayed on the web page.

The words having larger size mean that the particular word have comparatively more occurrence than other words. Lesser prevalent words are displayed in smaller sizes.

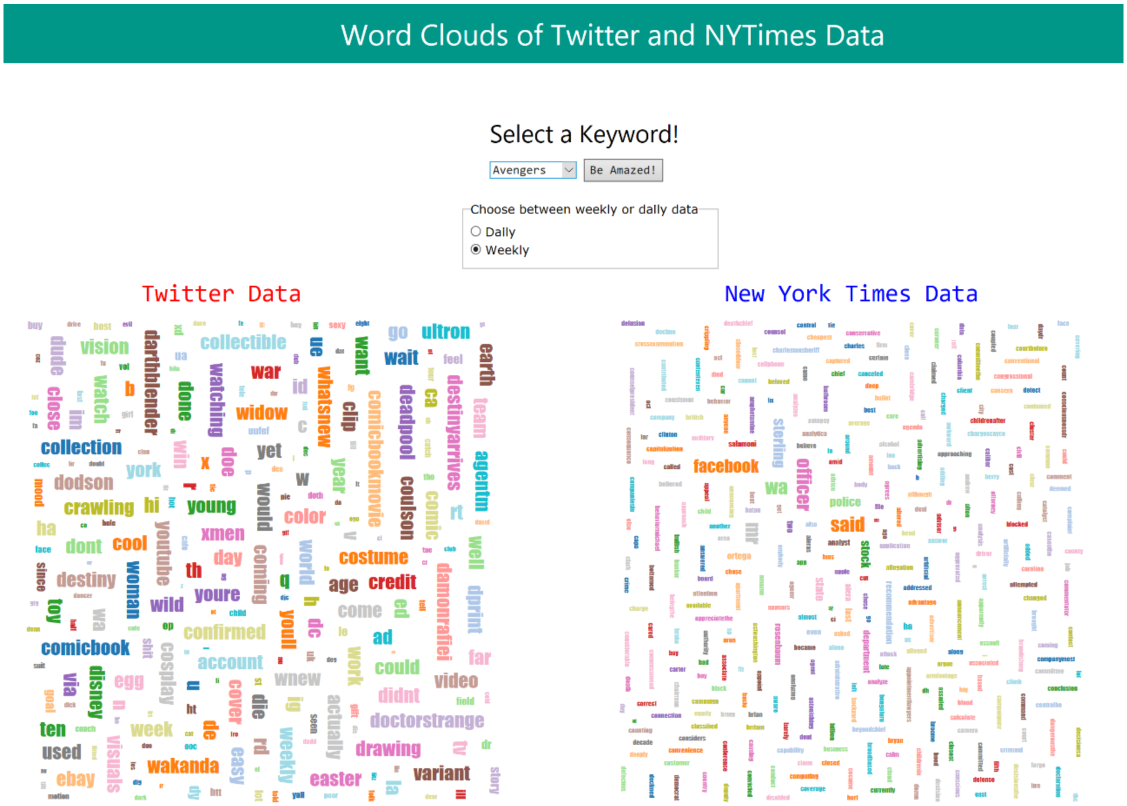
The two types of data NYTimes and Twitter resulted in different rampant words which is expected even though the key words were same because the are two different forms of media, one is social and other is professional new media.

There were many types of visualization mechanisms available in d3.js and are very widely used by various organization to visualize data.

We chose word cloud representation and it is really easy to understand the data by quantizing it effectively.

We chose the Avengers, Haunted and Psychopath as key words and extracted data from both Twitter and NYTimes articles and word counts, co-occurrences of words as pairs was evaluated using Hadoop map reduce environment. The output were fetched and stored as csv files and fed to d3.js draw method to produce word cloud.

The following is an example of comparison between Twitter and NYTimes article data for a keyword : Avengers and Span of data as week.



Word Clouds of Twitter and NYTimes Data

Select a Keyword!

Avengers ⌄ Be Amazed!

Choose between weekly or daily data

○ Daily

● Weekly

Twitter Data

New York Times Data

Exploration:

There are unlimited possibilities as the data available is enormous, the companies like facebook can make smart emojis/ trends using the top words or showcase the popular words used by various audiences.

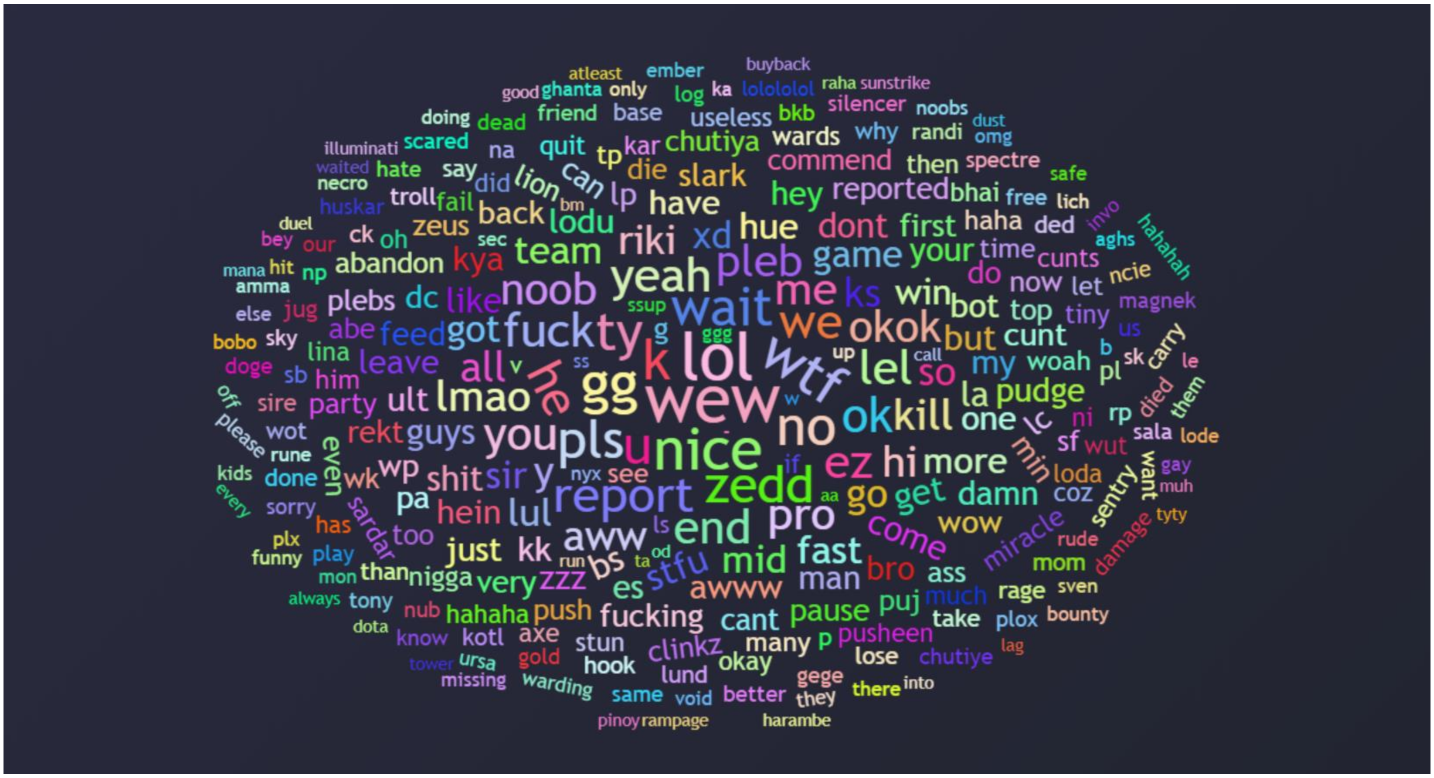
When a movie or a book is released the reviews data or posts on social media can be fetched and the word counts can be analyzed to check whether the movie has positive or negative ratings.

There are various possibilities in medical fields to determine commonly occurring molecules or gene data.

Usage in Organization:

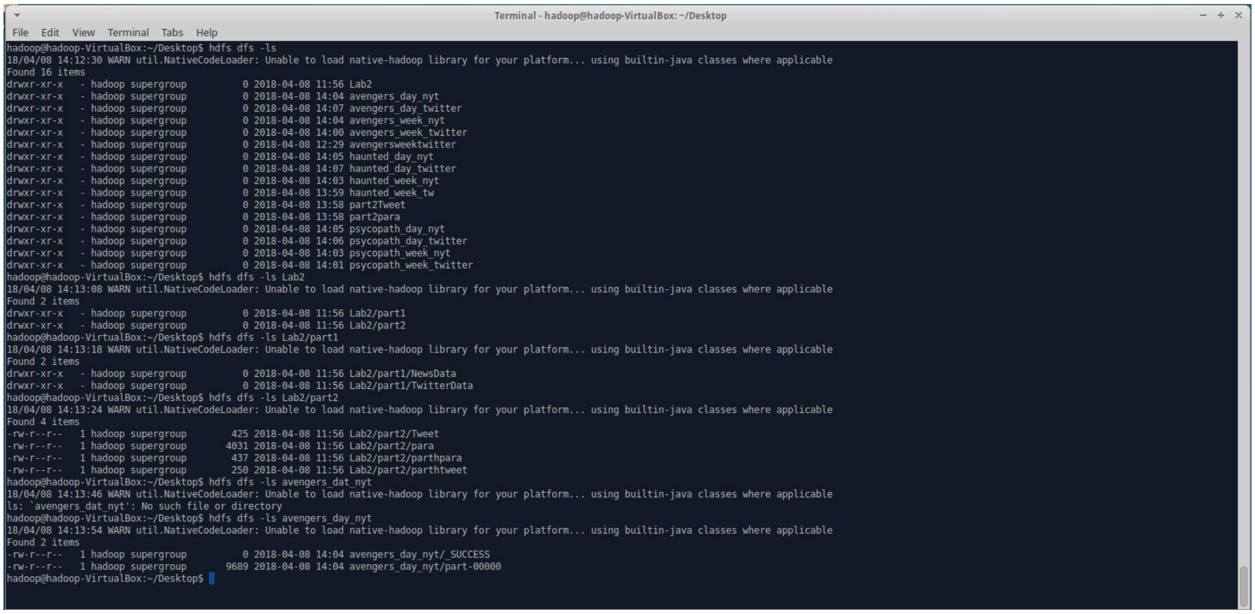
The Map Reduce is used by various organizations. A video game statistics company called Open Dota displays word clouds for each user using the data from the conversations he had while interacting with his team mates during all the video games.

The following is the word cloud generated by them for me:


atleast ember buyback
good ghanta only log ka lolololol raha sunstrike
doing dead friend base useless bkb silencer noobs dust
illuminati scared na quit tp kar chutiya wards why randi omg
waited hate say can die slark commend then spectre safe
necro did lion lp have hey reported bhai free lich
huskar troll fail bm back lodu hue dont first haha ded invo
duel ck zeus oh sec riki xd game your time aghs hahahah
bey our abandon kya team yeah pleb cunts ncie
mana hit np do now let
amma plebs dc like noob me ks win bot top tiny magnek
else jug abe feed got fuckty ssup wait we okok but cunt us
bobo sky lina g ggg up call b carry
doge leave all v ss k lol wtf lel so my woah sk le
sb him he gg pl
off sire party ult lmao wew w la pudge died
please wot no okkill one ic ni rp them
rune rekt guys you pls unice sf wut sala lode
kids done even wk wp shit sir y nyx see if ez hi more min loda gay
every sorry sardar pa hein lul report zedd aa go get damn coz sentry want muh
has too ls end pro come wow miracle tyty
plx funny play just kk aww od mid fast rude
mon than nigga very zzz run bs ta stfu bro ass mom damage
always tony es awww man much rage sven
dota nub hahaha push fucking cant pause puj take plox bounty
know kotl axe stun clinkz many p pusheen lag
tower ursa gold hook okay lose chutiye
missing warding same lund better gege there into
void they
pinoy rampage harambe

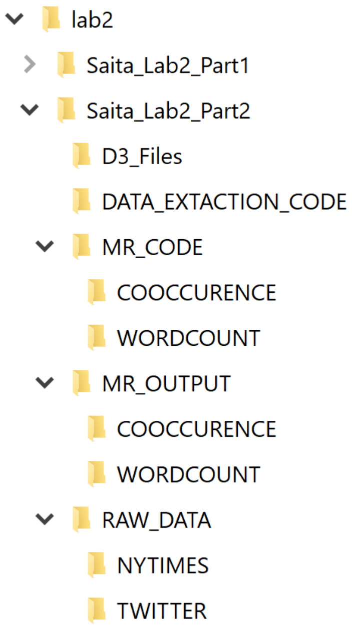
Directory Structure:

The following is the HDFS Structure for lab 2.


Terminal - hadoop@hadoop-VirtualBox: ~/Desktop
File Edit View Terminal Tabs Help

```
hadoop@hadoop-VirtualBox:~/Desktop$ hdfs dfs -ls
18/04/08 14:12:30 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 16 items
drwxr-xr-x   - hadoop supergroup          0 2018-04-08 11:56 Lab2
drwxr-xr-x   - hadoop supergroup          0 2018-04-08 14:04 avengers_day_nyt
drwxr-xr-x   - hadoop supergroup          0 2018-04-08 14:07 avengers_day_twitter
drwxr-xr-x   - hadoop supergroup          0 2018-04-08 14:04 avengers_week_nyt
drwxr-xr-x   - hadoop supergroup          0 2018-04-08 14:00 avengers_week_twitter
drwxr-xr-x   - hadoop supergroup          0 2018-04-08 12:29 avengersweektwitter
drwxr-xr-x   - hadoop supergroup          0 2018-04-08 14:05 haunted_day_nyt
drwxr-xr-x   - hadoop supergroup          0 2018-04-08 14:07 haunted_day_twitter
drwxr-xr-x   - hadoop supergroup          0 2018-04-08 14:03 haunted_week_nyt
drwxr-xr-x   - hadoop supergroup          0 2018-04-08 13:59 haunted_week_tw
drwxr-xr-x   - hadoop supergroup          0 2018-04-08 13:58 part2Tweet
drwxr-xr-x   - hadoop supergroup          0 2018-04-08 13:58 part2para
drwxr-xr-x   - hadoop supergroup          0 2018-04-08 14:05 psycopath_day_nyt
drwxr-xr-x   - hadoop supergroup          0 2018-04-08 14:06 psycopath_day_twitter
drwxr-xr-x   - hadoop supergroup          0 2018-04-08 14:03 psycopath_week_nyt
drwxr-xr-x   - hadoop supergroup          0 2018-04-08 14:01 psycopath_week_twitter
hadoop@hadoop-VirtualBox:~/Desktop$ hdfs dfs -ls Lab2
18/04/08 14:13:08 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
drwxr-xr-x   - hadoop supergroup          0 2018-04-08 11:56 Lab2/part1
drwxr-xr-x   - hadoop supergroup          0 2018-04-08 11:56 Lab2/part2
hadoop@hadoop-VirtualBox:~/Desktop$ hdfs dfs -ls Lab2/part1
18/04/08 14:13:18 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
drwxr-xr-x   - hadoop supergroup          0 2018-04-08 11:56 Lab2/part1/NewsData
drwxr-xr-x   - hadoop supergroup          0 2018-04-08 11:56 Lab2/part1/TwitterData
hadoop@hadoop-VirtualBox:~/Desktop$ hdfs dfs -ls Lab2/part2
18/04/08 14:13:24 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 4 items
-rw-r--r--   1 hadoop supergroup        425 2018-04-08 11:56 Lab2/part2/Tweet
-rw-r--r--   1 hadoop supergroup       4031 2018-04-08 11:56 Lab2/part2/para
-rw-r--r--   1 hadoop supergroup        437 2018-04-08 11:56 Lab2/part2/parthpara
-rw-r--r--   1 hadoop supergroup        250 2018-04-08 11:56 Lab2/part2/parthtweet
hadoop@hadoop-VirtualBox:~/Desktop$ hdfs dfs -ls avengers_dat_nyt
18/04/08 14:13:46 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
ls: `avengers_dat_nyt': No such file or directory
hadoop@hadoop-VirtualBox:~/Desktop$ hdfs dfs -ls avengers_day_nyt
18/04/08 14:13:54 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 hadoop supergroup          0 2018-04-08 14:04 avengers_day_nyt/_SUCCESS
-rw-r--r--   1 hadoop supergroup       9689 2018-04-08 14:04 avengers_day_nyt/part-00000
hadoop@hadoop-VirtualBox:~/Desktop$ 
```

The following is the directory structure of the submission for lab 2.



```
lab2
  Saita_Lab2_Part1
  Saita_Lab2_Part2
    D3_Files
    DATA_EXTACTION_CODE
    MR_CODE
      COOCCURENCE
      WORDCOUNT
    MR_OUTPUT
      COOCCURENCE
      WORDCOUNT
    RAW_DATA
      NYTIMES
      TWITTER
```