ASSIGNMENT-2
REPORT

Name: Shrey Amin

Student ID: B00822245

Course: CSCI-5408

Cluster Setup

Following are steps for creating cloud cluster using Amazon Web Service (AWS):

- Create AWS account
- Create an EC2 Instance and download key pair.
- Generate private key from the key-pair downloaded.
- Connect EC2 instance from SSH using Putty and generate private key.
- Download spark and set JAVA_HOME, SPARK_HOME and PYSPARK_PYTHON path.
- Install Apache Spark and configure Master and Slave.
- Install MongoDB on the cloud cluster to store data and start MongoDB server by initiating mongod service.

Data Extraction & Transformation

First step it to setup the Twitter developer account and create an app for fetching the tweets. Search (Search Api) and Stream (Stream Api) programs are written in python using Tweepy library for downloading the tweets. The access tokens and customer keys are used in the programs to connect twitter app. Tweets are extracted and following cleaning and transformation steps are performed:

- The raw tweets along with their metadata are stored in JSON and Text files (search.json, search.txt, Tweets_stream.json and Tweets_stream.txt).
- Tweets are cleaned by removing URLs, special characters, images and emoticons by calling a function named tweet_cleaning.
- Retrieve attributes like created_at, text, ID, name, user_id, screen name and location from the tweet and store the clean data in csv files (search.csv and stream.csv).

created_at	id	name	text	user_id	screen_name	location
Wed Feb 20 01:23:07 +0000 2019	1.10E+18	Dani Evan	7 children	37981156	kikkicookie08	canada
Wed Feb 20 01:23:16 +0000 2019	1.10E+18	Gerard Ne	"7 children	1.05E+09	gerardnerr	Minot
Wed Feb 20 01:23:19 +0000 2019	1.10E+18	Andrew Kc	rt : less tha	8.56E+08	k0z1can	Halifax, Nova Scotia
Wed Feb 20 01:23:32 +0000 2019	1.10E+18	Crawford I	and so say	19042993	Crof	North Vancouver, BC

```
{'created_at': 'Tue Feb 19 23:43:02 +0000 2019', 'id': 1098005064894111747, 'id_str': '1098005064894111747', 'text
{'created_at': 'Tue Feb 19 23:42:58 +0000 2019', 'id': 1098005051627458560, 'id_str': '1098005051627458560', 'text
{'created_at': 'Tue Feb 19 23:42:58 +0000 2019', 'id': 1098005051455356928, 'id_str': '1098005051455356928', 'text
{'created_at': 'Tue Feb 19 23:42:37 +0000 2019', 'id': 1098004959914688514, 'id_str': '1098004959914688514', 'text
{'created_at': 'Tue Feb 19 23:42:35 +0000 2019', 'id': 1098004954571329537, 'id_str': '1098004954571329537', 'text
{'created_at': 'Tue Feb 19 23:45:04 +0000 2019', 'id': 1098005576909549568, 'id_str': '1098005576909549568', 'text
{'created_at': 'Tue Feb 19 23:45:00 +0000 2019', 'id': 1098005563613609986, 'id_str': '1098005563613609986', 'text
```

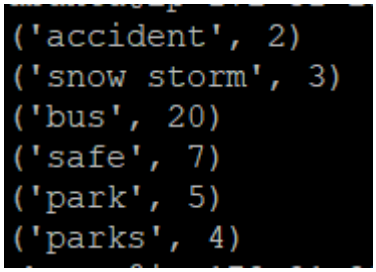


```
{
    "created_at": "Tue Feb 19 23:43:02 +0000 2019",
    "id": 1098005064894111747,
    "id_str": "1098005064894111747",
    "text": "RT @haligonia: Today is a sad and tragic d
    "truncated": false,
    "entities": {
        "hashtags": [],
        "symbols": [],
        "user_mentions": [
            {
                "screen_name": "haligonia",
                "name": "Halifax News & Info",
                "id": 16308920,
                "id_str": "16308920",
                "indices": [
                    3,
                    13
                ]
            }
        ],
        "urls": []
    },
    "metadata": {
        "iso_language_code": "en",
        "result_type": "recent"
```

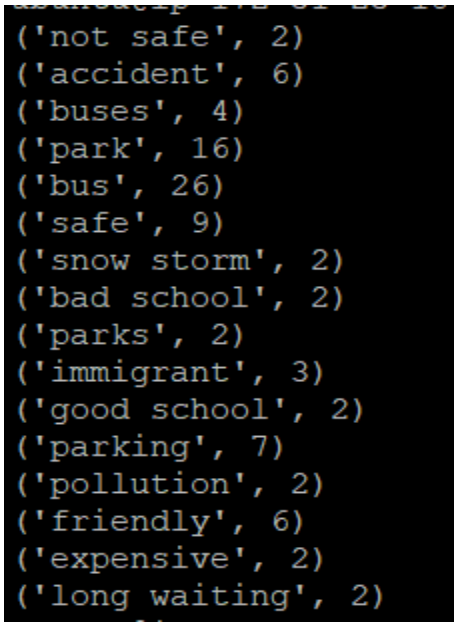

All the cleaned tweets are stored in MongoDB database on the cloud cluster. Tweets are imported into search and stream collections using “mongoimport” command.

Data Processing (MapReduce)

MapReduce program is executed in Apache Spark on “stream.csv” file. All the words to be found are mapped using the “Map” function. Then its output is reduced using “ReduceByKey” function to get final frequency count of each word. Initially word “bus” has the highest count of 20.



Again, the MapReduce program is executed on new tweets collected in “stream_2.csv” file. The frequency of every words has changed. The word “bus” still has highest count of 26. Further, the count of several words has increased.



Both the outputs of word count are stored in “word_frequency_count_1.txt” and “word_frequency_count_2.txt’ files.

References

- http://socialmedia-class.org/twittertutorial.html
- https://www.youtube.com/watch?v=jg7Z8ctKpEs&t=330s
- https://www.youtube.com/watch?v=wlnx-7cm4Gg
- https://stackoverflow.com/questions/33404752/removing-emojis-from-a-string-in-python
- https://tweepy.readthedocs.io/en/v3.5.0/
- https://www.youtube.com/watch?v=RxrMWh1lJ_k&t=649s
- Class lab work