# ASSIGNMENT-2

# REPORT

**Name:** Kushal Tushar Reshamdalal

**Student ID:** B00824760

**Course:** CSCI-5408

## 1. Cloud Setup process

- The below steps are followed to setup the cloud using Amazon's AWS EC2 instance.
- I created free account on Amazon' aws for accessing instance and storage.
- I created EC2 instance and generated public and private keys pair.
- In the third step I connected EC2 instance using SSH from PUTTY software.
- Then I set up ubuntu authentication.
- Started downloading python, exporting JAVA_HOME, SPARK_HOME.
- In the following step, I downloaded Apache sparks and configured master and slave.
- Finally, I downloaded MongoDB and initiated mongod service.

## 2. Data Extraction

- In the first step I created Twitter account and then I set up Twitter developer account by answering some security questions.
- In the following step I created one Twitter application which generated consumer API keys and access token required python script to fetch tweets from Twitter.
- In this assignment part there are two types tweets fetched
  I. Static tweets (fetched using keywords)
  II. Live tweets (fetched using keywords-streaming)
- Without access tokens(credentials) python code can not connect twitter API and data.
- For this assignment I have used python for tweets extraction and also for articles extraction.

## 3. Tweet Cleaning

- This is the most important part for getting accurate data.
- There were so many characters and emojis which required to be cleaned from the tweets.
- I considered following entities which needed to be removed for getting cleaned tweets using cleaning_tweet() function.
  I. Special characters
  II. Images(jpg)
  III. HTTP/HTTPS URL links
  IV. Emoticons
  V. Lowering the alphabets in tweets (useful for analysis)
- After applying cleaning process, the required data from tweets such as tweet id, username, date(created_at), cleaned tweet, user id and location into csv file.
- I am using MongoDB which I installed on ubuntu on aws to store the data using csv file.

## 4. News Article Data Extraction & Transformation

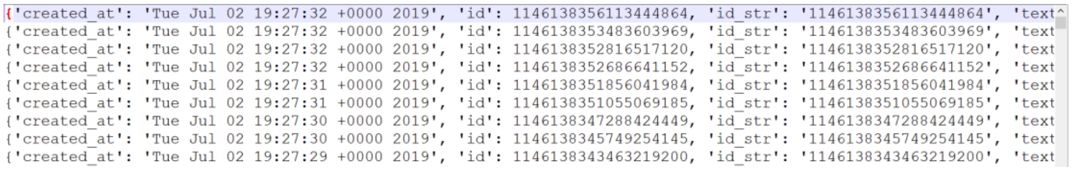
- In the second part of assignment two files are given for article extraction using tags.
- Article_extrction.py extracts the texts lying between two <TEXT> tags.
- It creates two separate folders "020" and "021" having separate news articles.

## 5. Data Processing (Map reduce)

- This works for counting word frequency from "searchedtweets.csv", "streamedtweets.csv" and extracted news articles.

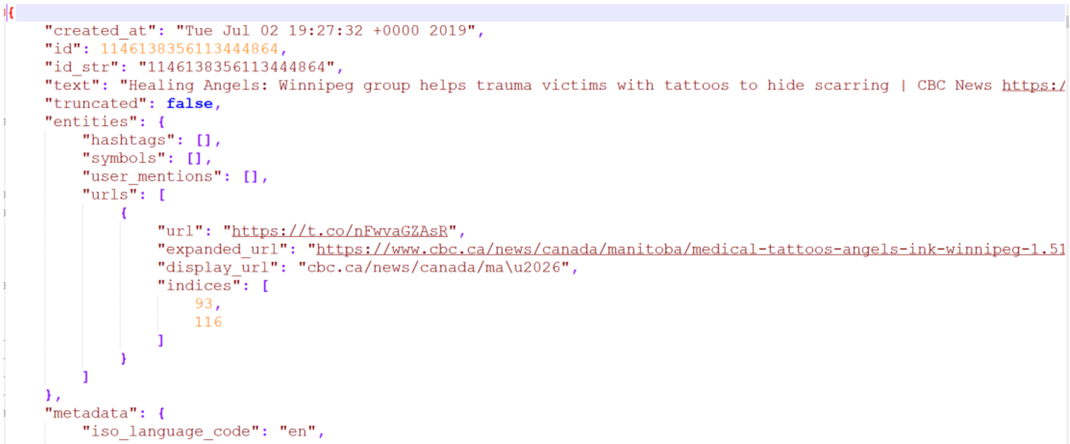
## 6. Sample of tweets in different file formats (.txt, JSON, CSV)

- I have stored raw tweets in text files and JSON files.
- Cleaned tweets are stored in CSV files for both searched and streamed tweets..



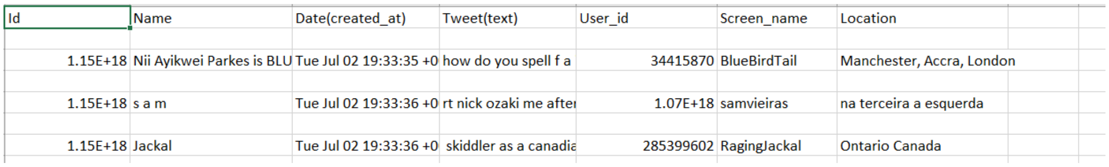
```
{'created_at': 'Tue Jul 02 19:27:32 +0000 2019', 'id': 1146138356113444864, 'id_str': '1146138356113444864', 'text
{'created_at': 'Tue Jul 02 19:27:32 +0000 2019', 'id': 1146138353483603969, 'id_str': '1146138353483603969', 'text
{'created_at': 'Tue Jul 02 19:27:32 +0000 2019', 'id': 1146138352816517120, 'id_str': '1146138352816517120', 'text
{'created_at': 'Tue Jul 02 19:27:32 +0000 2019', 'id': 1146138352686641152, 'id_str': '1146138352686641152', 'text
{'created_at': 'Tue Jul 02 19:27:31 +0000 2019', 'id': 1146138351856041984, 'id_str': '1146138351856041984', 'text
{'created_at': 'Tue Jul 02 19:27:31 +0000 2019', 'id': 1146138351055069185, 'id_str': '1146138351055069185', 'text
{'created_at': 'Tue Jul 02 19:27:30 +0000 2019', 'id': 1146138347288424449, 'id_str': '1146138347288424449', 'text
{'created_at': 'Tue Jul 02 19:27:30 +0000 2019', 'id': 1146138345749254145, 'id_str': '1146138345749254145', 'text
{'created_at': 'Tue Jul 02 19:27:29 +0000 2019', 'id': 1146138343463219200, 'id_str': '1146138343463219200', 'text
```

*Figure 1. Raw tweets in text file*



```
{
    "created_at": "Tue Jul 02 19:27:32 +0000 2019",
    "id": 1146138356113444864,
    "id_str": "1146138356113444864",
    "text": "Healing Angels: Winnipeg group helps trauma victims with tattoos to hide scarring | CBC News https:/
    "truncated": false,
    "entities": {
        "hashtags": [],
        "symbols": [],
        "user_mentions": [],
        "urls": [
            {
                "url": "https://t.co/nFwvaGZAsR",
                "expanded_url": "https://www.cbc.ca/news/canada/manitoba/medical-tattoos-angels-ink-winnipeg-1.51
                "display_url": "cbc.ca/news/canada/ma\u2026",
                "indices": [
                    93,
                    116
                ]
            }
        ]
    },
    "metadata": {
        "iso_language_code": "en",
```

*Figure 2 Raw tweets in JSON*



| Id | Name | Date(created_at) | Tweet(text) | User_id | Screen_name | Location |
|---|---|---|---|---|---|---|
| 1.15E+18 | Nii Ayikwei Parkes is BLU | Tue Jul 02 19:33:35 +0 | how do you spell f a | 34415870 | BlueBirdTail | Manchester, Accra, London |
| 1.15E+18 | s a m | Tue Jul 02 19:33:36 +0 | rt nick ozaki me after | 1.07E+18 | samvieiras | na terceira a esquerda |
| 1.15E+18 | Jackal | Tue Jul 02 19:33:36 +0 | skiddler as a canadia | 285399602 | RagingJackal | Ontario Canada |

*Figure 3. Cleaned tweets in CSV*

## References

[1]"Tweepy Documentation — tweepy 3.5.0 documentation", *Tweepy.readthedocs.io*, 2019. [Online]. Available: https://tweepy.readthedocs.io/en/v3.5.0/. [Accessed: 03- Jul- 2019].

[2]"Twitter Data Mining: A Guide to Big Data Analytics Using Python", *Toptal Engineering Blog*, 2019. [Online]. Available: https://www.toptal.com/python/twitter-data-mining-using-python. [Accessed: 03- Jul- 2019].

[3]"Twitter API with Python: Part 1 -- Streaming Live Tweets", *YouTube*, 2019. [Online]. Available: https://www.youtube.com/watch?v=wlnx-7cm4Gg&t=5s. [Accessed: 03- Jul- 2019].

[4]S. 3.6, C. Hanna and M. Tolonen, "Saving tweets to JSON file in Python 3.6", *Stack Overflow*, 2019. [Online]. Available: https://stackoverflow.com/questions/48157921/saving-tweets-to-json-file-in-python-3-6-. [Accessed: 03- Jul- 2019].

[5]p. Remove all special characters, A. White and G. Chauhan, "Remove all special characters, punctuation and spaces from string", *Stack Overflow*, 2019. [Online]. Available: https://stackoverflow.com/questions/5843518/remove-all-special-characters-punctuation-and-spaces-from-string. [Accessed: 03- Jul- 2019].

[6]G. file, R. Silberie and A. Elijah, "Get data from Twitter using Tweepy and store in csv file", *Stack Overflow*, 2019. [Online]. Available: https://stackoverflow.com/questions/21865413/get-data-from-twitter-using-tweepy-and-store-in-csv-file. [Accessed: 03- Jul- 2019].

[7]r. Python, M. Jalal, M. Jalal, A. Adam and A. Tavakoli, "removing emojis from a string in Python", *Stack Overflow*, 2019. [Online]. Available: https://stackoverflow.com/questions/33404752/removing-emojis-from-a-string-in-python. [Accessed: 03- Jul- 2019].

[8]H. Python et al., "How to remove any URL within a string in Python", *Stack Overflow*, 2019. [Online]. Available: https://stackoverflow.com/questions/11331982/how-to-remove-any-url-within-a-string-in-python. [Accessed: 03- Jul- 2019].

[9]"Python Regex Remove URL from String", *YouTube*, 2019. [Online]. Available: https://www.youtube.com/watch?v=O2onA4r5UaY. [Accessed: 03- Jul- 2019].

[10]"Streaming With Tweepy — tweepy 3.3.0 documentation", *Docs.tweepy.org*, 2019. [Online]. Available: http://docs.tweepy.org/en/v3.4.0/streaming_how_to.html. [Accessed: 03- Jul- 2019].

[11]H. tweepy?, E. Yan, E. Yan and K. Rother, "How do I save streaming tweets in json via tweepy?", *Stack Overflow*, 2019. [Online]. Available: https://stackoverflow.com/questions/23531608/how-do-i-save-streaming-tweets-in-json-via-tweepy. [Accessed: 03- Jul- 2019].

[12]P. object, R. Knop, R. Knop, M. Pieters and P. Mandrekar, "Python accessing data in JSON object", *Stack Overflow*, 2019. [Online]. Available:

https://stackoverflow.com/questions/11241583/python-accessing-data-in-json-object. [Accessed: 03- Jul- 2019].

[13]H. keywords, D. Wei and R. Sharabani, "How to get tweets data that contain multiple keywords", *Stack Overflow*, 2019. [Online]. Available: https://stackoverflow.com/questions/49027297/how-to-get-tweets-data-that-contain-multiple-keywords. [Accessed: 03- Jul- 2019].

[14]R. tags, S. Hassan and P. Gupta, "Regex to find words between two tags", *Stack Overflow*, 2019. [Online]. Available: https://stackoverflow.com/questions/22247957/regex-to-find-words-between-two-tags. [Accessed: 03- Jul- 2019].

[15]"Map Reduce Word Count with Python", *YouTube*, 2019. [Online]. Available: https://www.youtube.com/watch?v=RxrMWh1lJ_k&t=649s. [Accessed: 03- Jul- 2019].

[16]"Word Count using PySpark", *YouTube*, 2019. [Online]. Available: https://www.youtube.com/watch?v=jg7Z8ctKpEs&t=330s. [Accessed: 03- Jul- 2019].

[17]2019. [Online]. Available: https://idevji.com/map-reduce-word-count-with-python/. [Accessed: 03- Jul- 2019].

[18]"zonca/python-wordcount-hadoop", *GitHub*, 2019. [Online]. Available: https://github.com/zonca/python-wordcount-hadoop. [Accessed: 03- Jul- 2019]

[19]"Hadoop Tutorial 2 -- Running WordCount in Python - dftwiki", *Science.smith.edu*, 2019. [Online]. Available: http://www.science.smith.edu/dftwiki/index.php/Hadoop_Tutorial_2_--_Running_WordCount_in_Python. [Accessed: 03- Jul- 2019].

[20]"Examples | Apache Spark", *Spark.apache.org*, 2019. [Online]. Available: https://spark.apache.org/examples.html. [Accessed: 03- Jul- 2019].

[21]"pyspark package — PySpark 2.1.0 documentation", *Spark.apache.org*, 2019. [Online]. Available: https://spark.apache.org/docs/2.1.0/api/python/pyspark.html. [Accessed: 03- Jul- 2019].

[22]H. PySpark?, C. Sobrino and C. Sobrino, "How to group by multiple columns and collect in list in PySpark?", *Stack Overflow*, 2019. [Online]. Available: https://stackoverflow.com/questions/46538991/how-to-group-by-multiple-columns-and-collect-in-list-in-pyspark. [Accessed: 03- Jul- 2019].