

**Principles Of Big Data**

# **Project: Twitter Data Analysis - Phase 1**

***Professor: Dr. Praveen Rao***

|  |  |  |
| --- | --- | --- |
| **Name** | **Email** | **Student ID** |
| Yamini Reddy Dontireddy | yddpw@mail.umkc.edu | 16293335 |
| Manideep Nadimpelli | mnpgc@mail.umkc.edu | 16292336 |
| Bhavya sri Sirasanagandla | Bs9r8@mail.umkc.edu | 16292335 |

Contents

[**Project: Twitter Data Analysis - Phase 1** 1](#_Toc19478628)

[**Version History** 2](#_Toc19478629)

[**Project Scope** 3](#_Toc19478630)

[**Potential Use cases** 3](#_Toc19478631)

[**Implementation Approach** 4](#_Toc19478632)

[**Implementation Steps** 4](#_Toc19478633)

[**Word Count Output (Spark & Hadoop)** 7](#_Toc19478634)

[**Project Summary** 8](#_Toc19478635)

[**Appendix - Documentation** 9](#_Toc19478636)

[**Appendix - References** 9](#_Toc19478637)

# **Version History**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Version number | Date Modified | Author | Reviewer | Comments |
| 01 | 9/15/2019 | Yamini Reddy/Manideep/Bhavya | PB -Grader | Initial copy – Phase 1 Submission |

# **Project Scope**

Scope of this project is to leverage the latest big data technologies such as Hadoop, Spark and APIs to analyze the Twitter data to gain any insights around a topic or trend or a media campaign.

Project is divided into three phases:

* A high-level analysis of specific hashtags and URLs in phase 1.
* Phase 2 requires a meaningful analysis of twitter data to analyze sentiments or real time trends or specific pattern/impact from a set of influencing twitter handles.
* A poster presentation with analytical queries performance metrics/details in Phase 3.

**Acronyms**

* WHO – World Health Organization
* CDC – Center for Disease Control
* NCD – Non Communicable Diseases
* AHA – American Heart Association
* CMS – Centre for Medicare & Medicaid Services
* API - Application Programming Interface
* JSON - JavaScript Object Notation
* URL – Uniform/Unique Resource Locator

# **Potential Use cases**

Apart from typical usage of twitter for customer engagement through key words or searches through about specific product or campaigns; twitter can be leveraged in many ways.

* Twitter Analytics As A Service for brand promotion, customer acquisition
* Twitter Analytics for Diagnostics & Policing
* Twitter based insights such as sentiments, trends, patterns
* Twitter based security products

Our project aims to leverage these technologies in healthcare in the area of continued medical education (CME) by providers and government entities like CMS and CDC.

NCD (Non Communicable Diseases) being the biggest of cause of human mortality; there is continued campaigns and education by organizations like WebMD, CDC, AHA and providers like Mayo and Cleveland about NCDs, lifestyles changes and preventive care.

Detailed analysis of campaigns would be performed in the next phase of the project to derive meaningful insights about the reach and discrepancy with actual NCD mortality rates.

# **Implementation Approach**

* Twitter API keys and access tokens needs to be created in order to collect the tweets. This is done from the twitter’s developer account.
* Post the keys generation, using the Python code, specific tweets are collected and stored in the JSON format .
* Python programs to extract all the relevant hashtags and URLs separately and concatenate into a single text file.
* Finally, run a Hadoop and Spark word count programs on the text file (extracted output of hastags & URLs ).

**Twitter API Keys**

##### Consumer API keys

dFjgu6BlFBOVec7LneZk64EMh (API key)

NT3nQ8WCxNcXVNyH6kFl8PwFjlN75aStlAstM4kxSGrpRbp2xY (API secret key)

##### Access token & access token secret

3933378373-0YIC5CkmCzpj3ddTneck5DQZtb00TWU37uWjn4g (Access token)

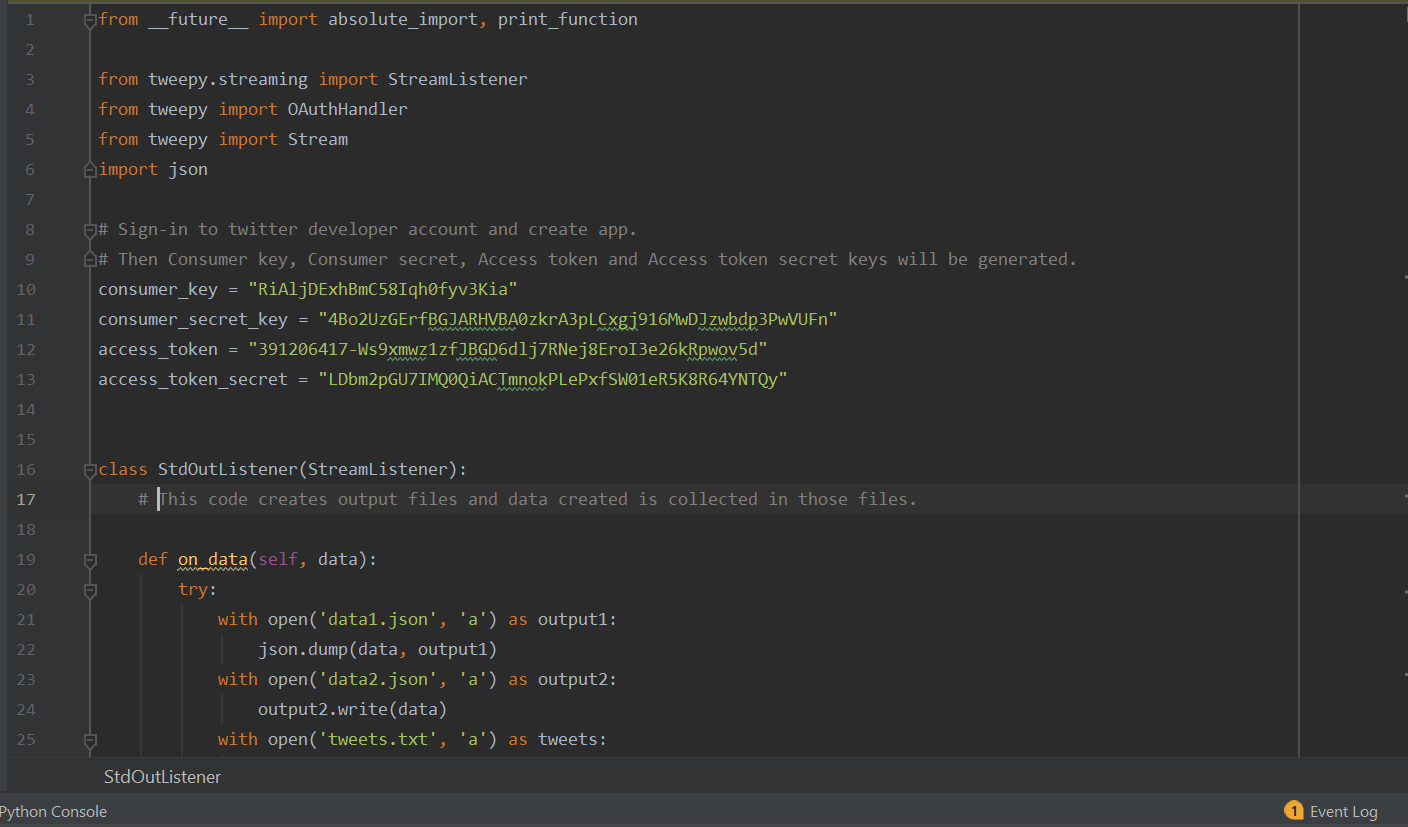
3u08vPrJtqG2RDElT7cT2U2GBlh6tyu6L0P017y1JzFD4 (Access token secret)

# **Implementation Steps**

* **Data Collection - Tweets (using the API keys)**

Below is the screen-print of the code that’s used to pull the tweets using the API:

Pulled the tweets from WebMd, WHO, Business wire -Health, Harvard Health, Cerner, CVS Health , BCBS , AHA, Maya Clinic, NCD etc.



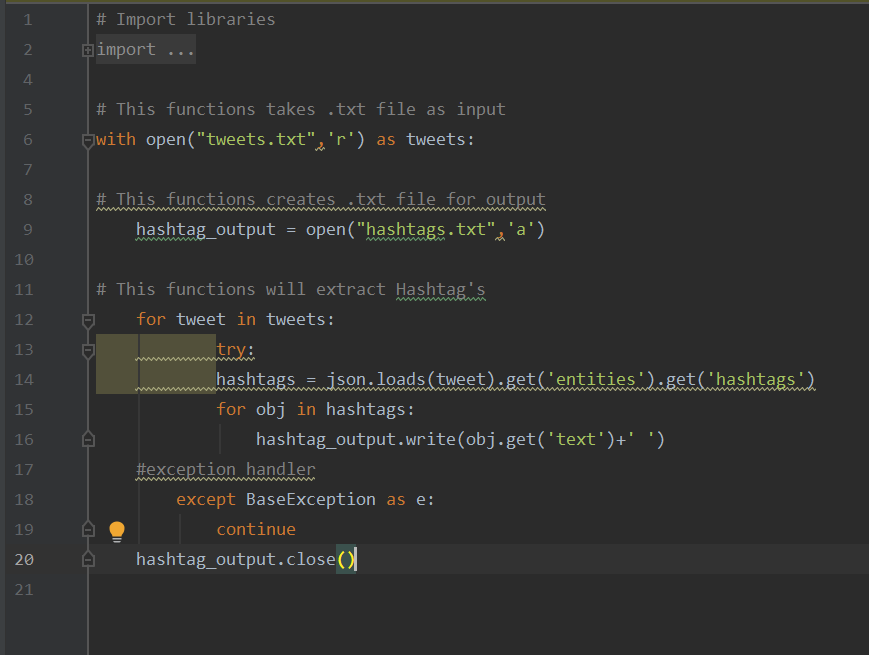
A screenshot of a computer

Description automatically generated

* **Data Refinement (Extraction of Hashtags/URLs)**

Below is the screen-print of the code that is used to extract the Hashtags and URLs from the collected twitter data.

1. **Hashtag Extraction:**



1. **URLs Extraction:**

A screenshot of a cell phone

Description automatically generated

The output of the above 2 codes is concatenated into a single text file.

* **Data Processing and Analysis (Word Count in Spark and Hadoop)**

Below is the screen-print of the Spark code that is used to get the word count on the extracted twitter URL’s and hashtags.

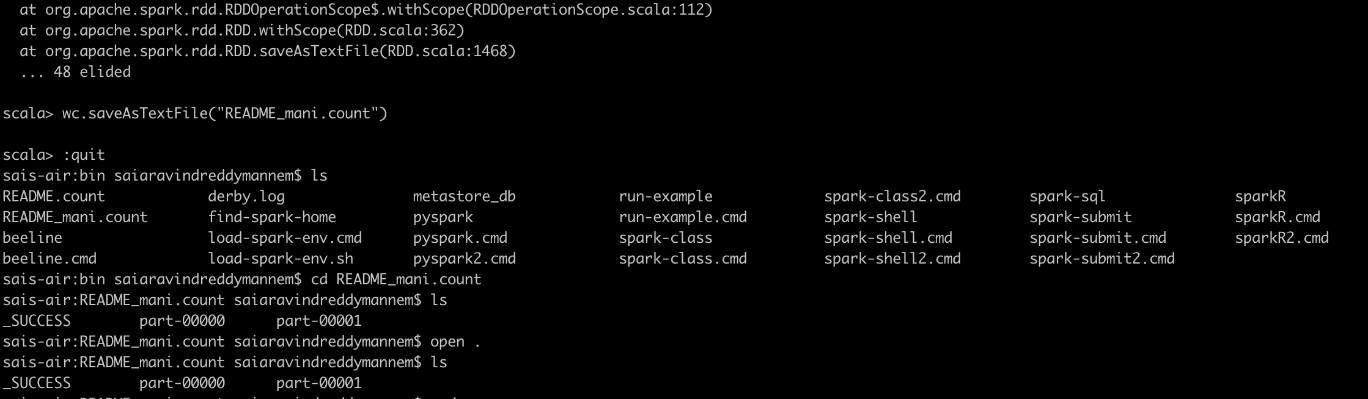
**Scala wordcount code:**

A screenshot of a cell phone

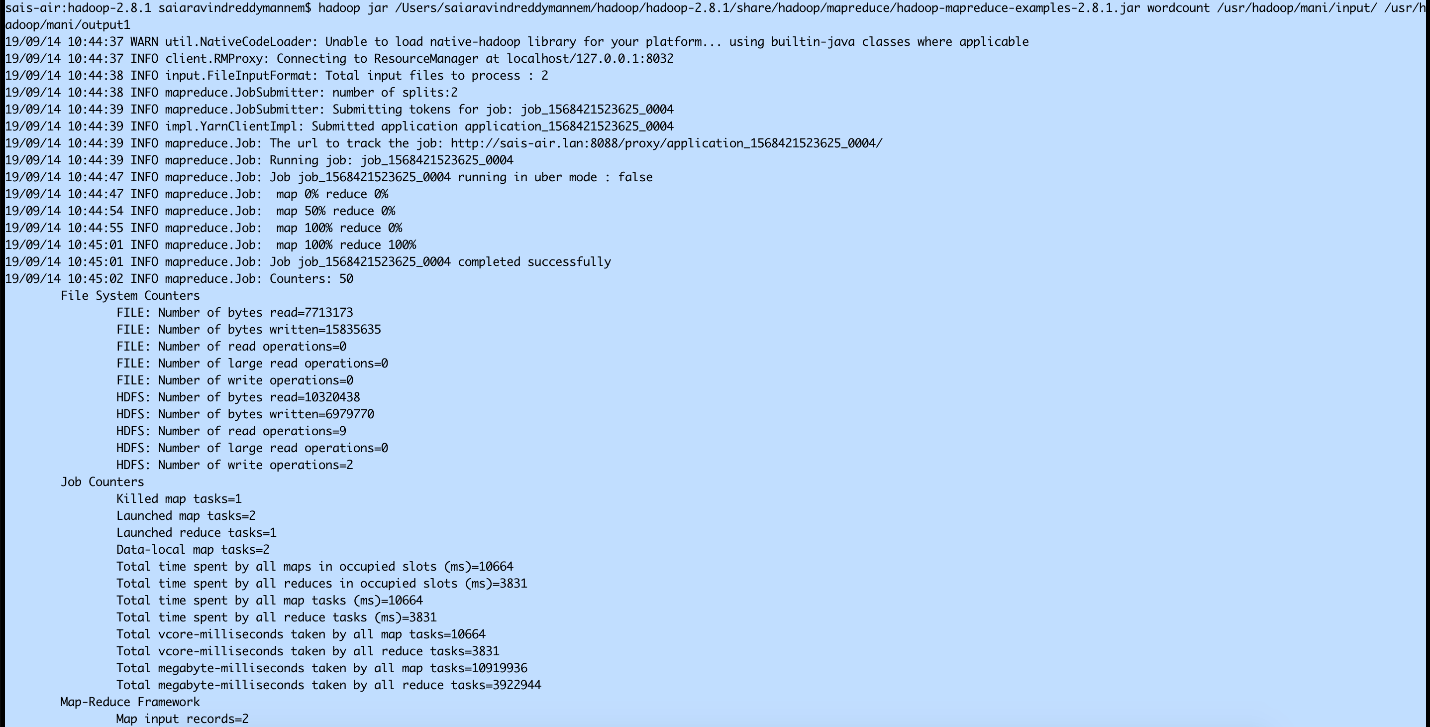
Description automatically generated

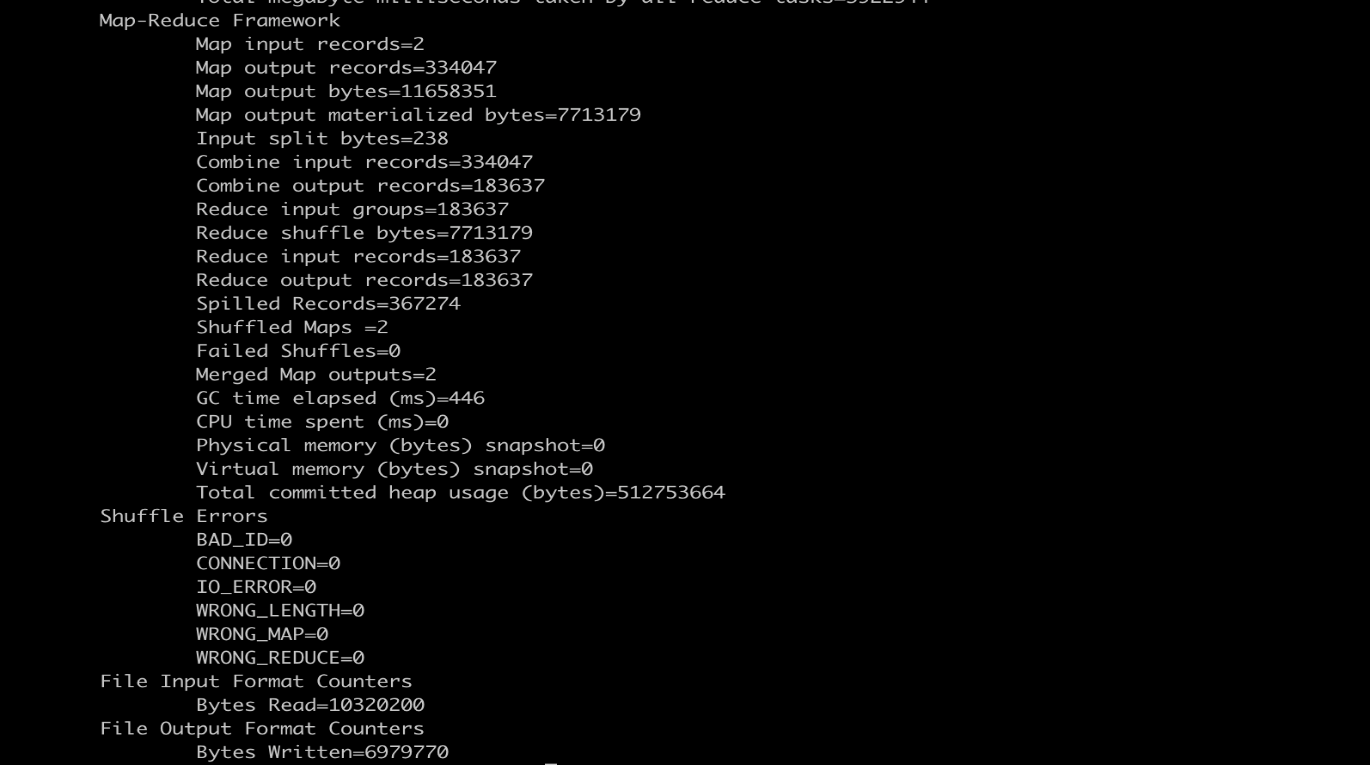
**Scala code running sippet:**





Hadoop file running snippet :





# **Word Count Output (Spark & Hadoop)**

1. Output files – Spark



1. Output files - Hadoop



# **Project Summary**

In this phase, the word count examples was successfully computed in Hadoop and Spark on the twitter data. More insights will be covered in the next phase.

# **Appendix - Documentation**

[**https://mailmissouri-my.sharepoint.com/:u:/r/personal/yddpw\_mail\_umkc\_edu/Documents/PB%20Project-%20Phase%201.zip?csf=1&e=JcwYo8**](https://drive.google.com/open?id=1WQvPmoE_H5Pzdj1m3CJs5lMuPCFGeRNd)

# **Appendix - References**

* [https://developer.twitter.com](https://developer.twitter.com/)
* <http://adilmoujahid.com/posts/2014/07/twitter-analytics/>
* https://github.com/jakekemple/Hadoop-Tweet-Wordcounter
* https://www.tutorialkart.com/apache-spark/scala-spark-shell-example/