

**PRINCIPLES OF BIG DATA**

# **TWITTER DATA ANALYSIS - PHASE 1**

**INSTRUCTOR: Dr. PRAVEEN RAO**



Name ID

Navya G - 16282632

Anusha M - 16286311

Anusha K - 16285245

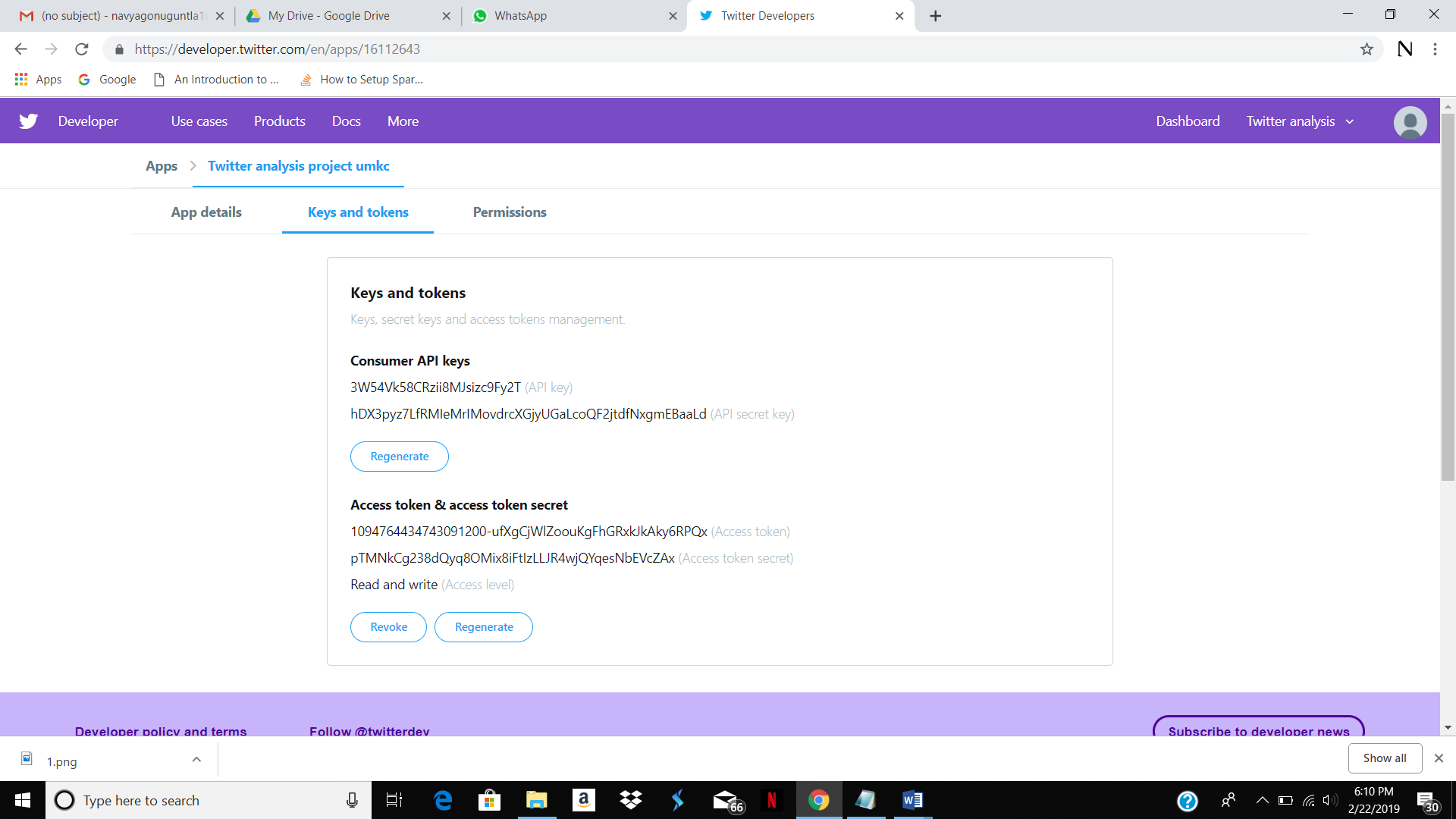
**GOAL**

The main aim of this phase is to extract tweets made by several twitter users all over the world. The extracted data is mainly concentrated on celebrities. In this phase we aim at extracting URL’S and Hashtags within these tweets and run wordcount program both in Apache Hadoop and Apache Spark. The final step is to collect the output of wordcount program as well as the log files from Hadoop.

**APPROACH**

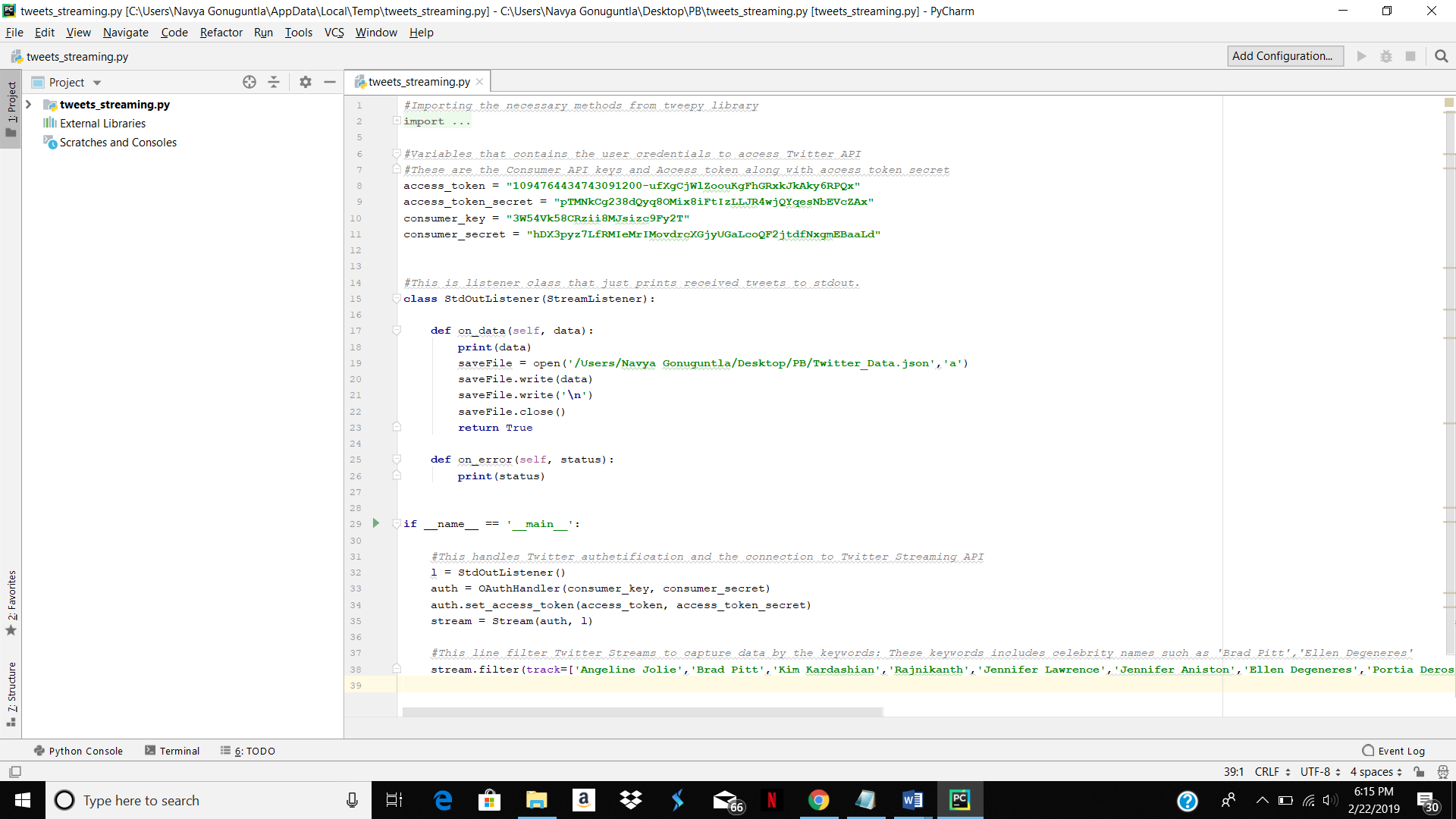
* In order to collect tweets from Twitter, It is required to create a developer account By creating a Twitter developer account, twitter API keys and access tokens are generated. These keys are required to collect tweets.
* Using these keys, required tweets are collected with the help of a python code. All these tweets are stored in a json file
* The json file is then given as an input to another Python code which is run to extract URL’S and Hashtags from tweets. The output of this program is saved as a text file.
* As a final step, WordCount program is run on both Apache Hadoop and Spark.

**TWITTER API KEYS**



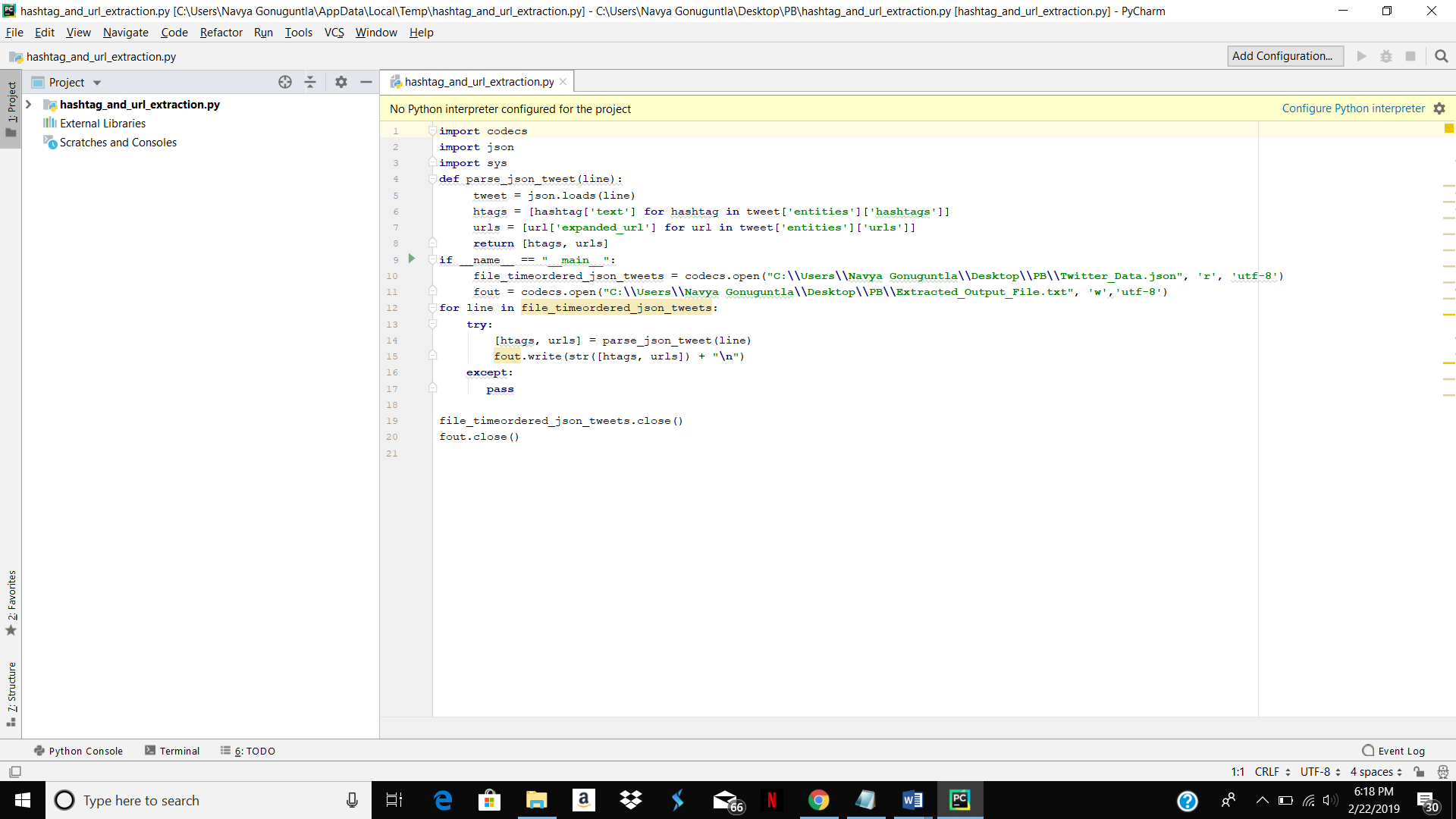
The following are the screenshots of several codes used and the output files.

**CODE TO COLLECT TWEETS**



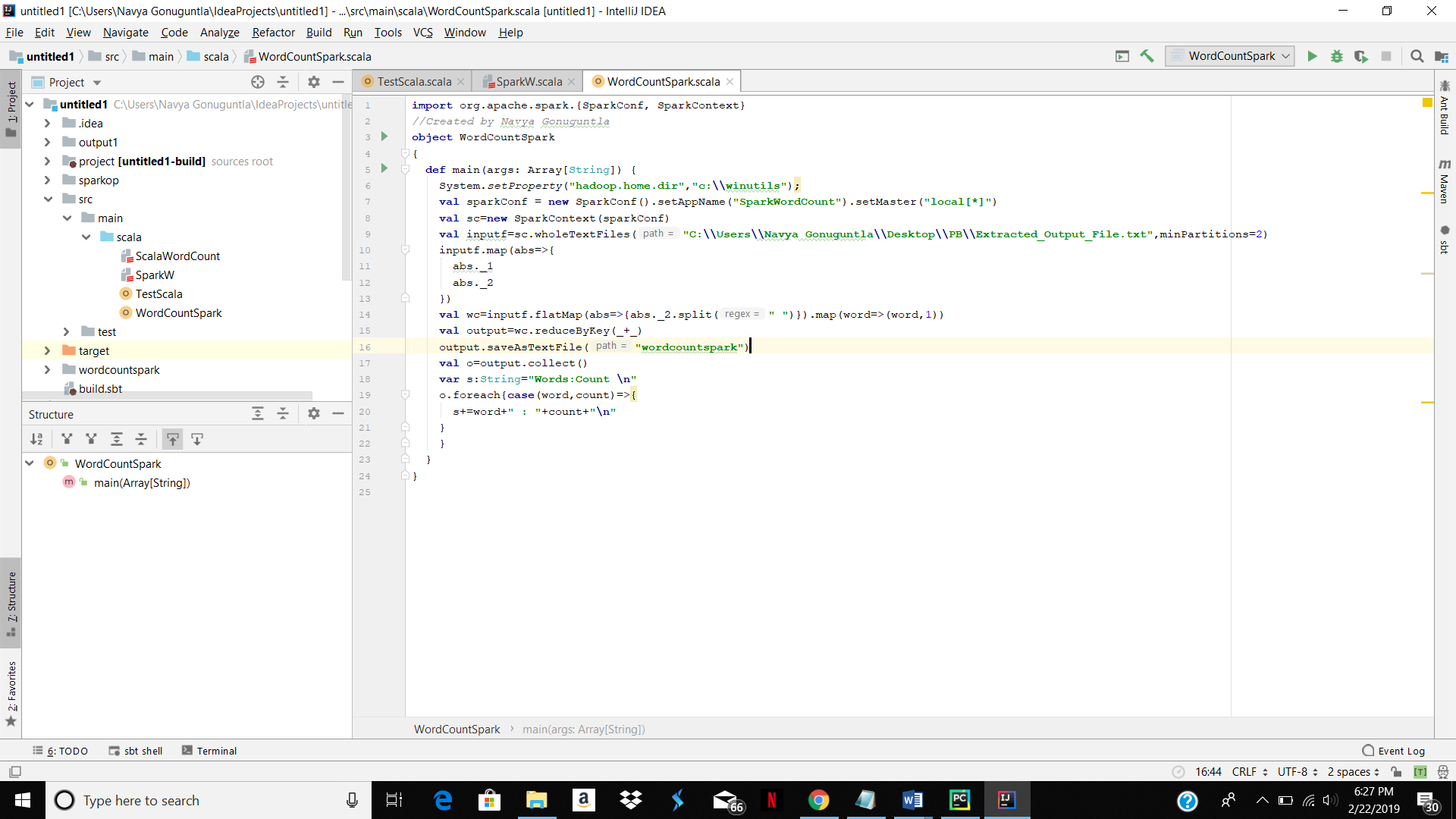
The above code is a python program used to collect tweets using Twitter API’s. the output of this program is a json file containing tweets which includes given keywords.

**CODE TO EXTRACT HASHTAGS/URL’s**



The above python code once executed will result a text file which extracts all the hashtags and URL’s from tweets present in json file. A wordcount program will now be run on this output file.

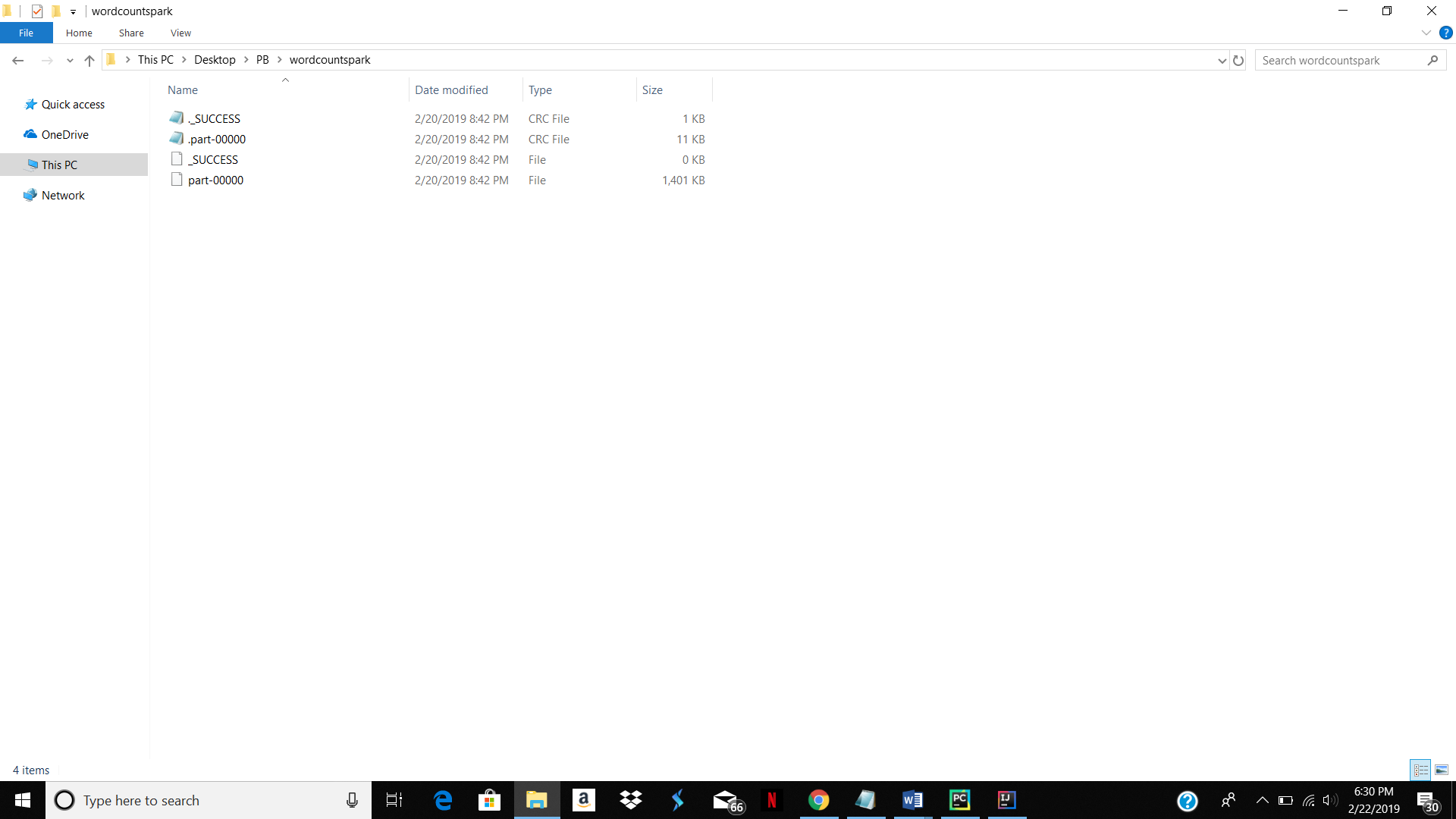
**WORD COUNT IN SPARK**

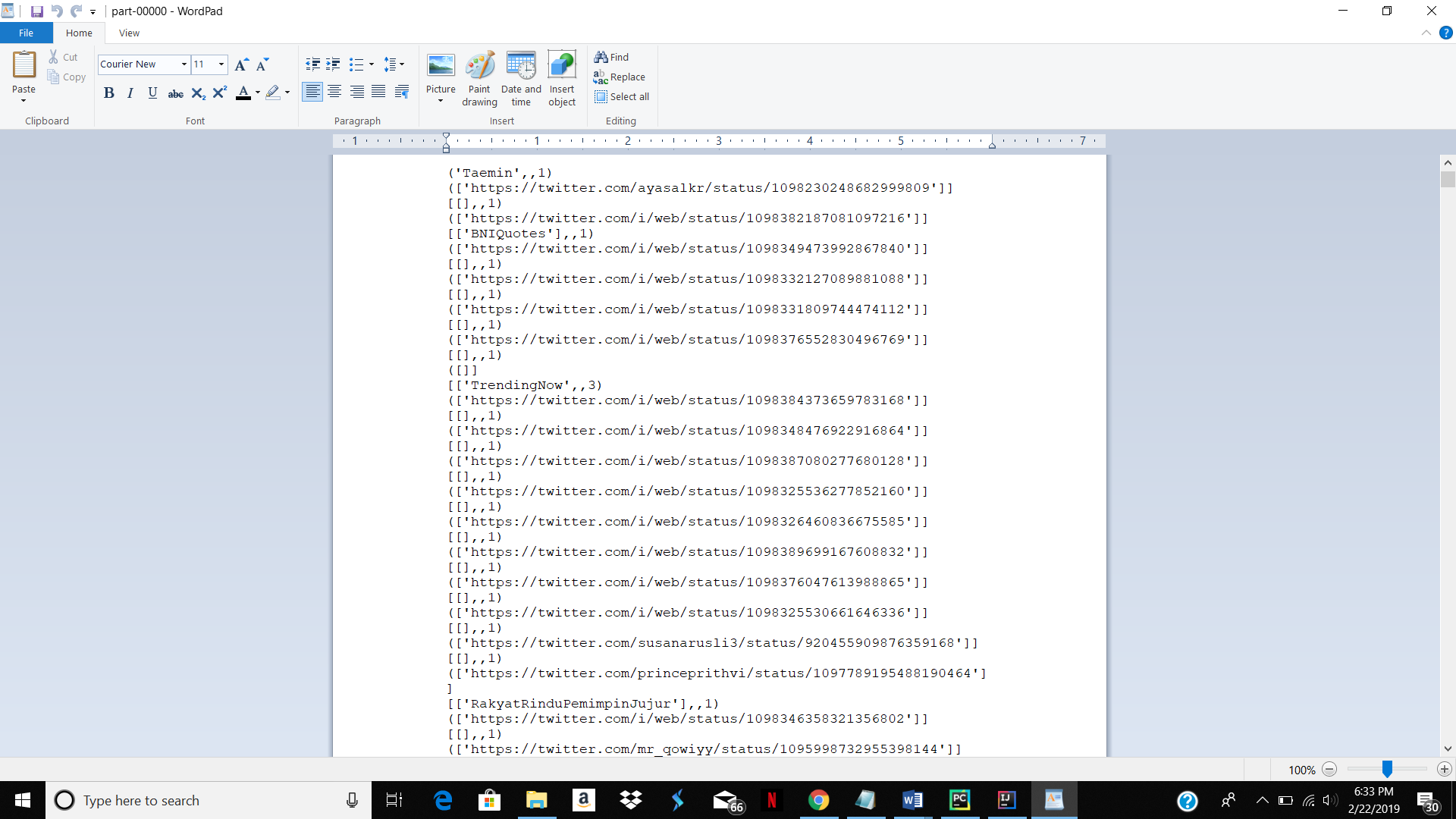


Above code when run will generate a wordcount file.

**OUTPUT**

The output file contains the following

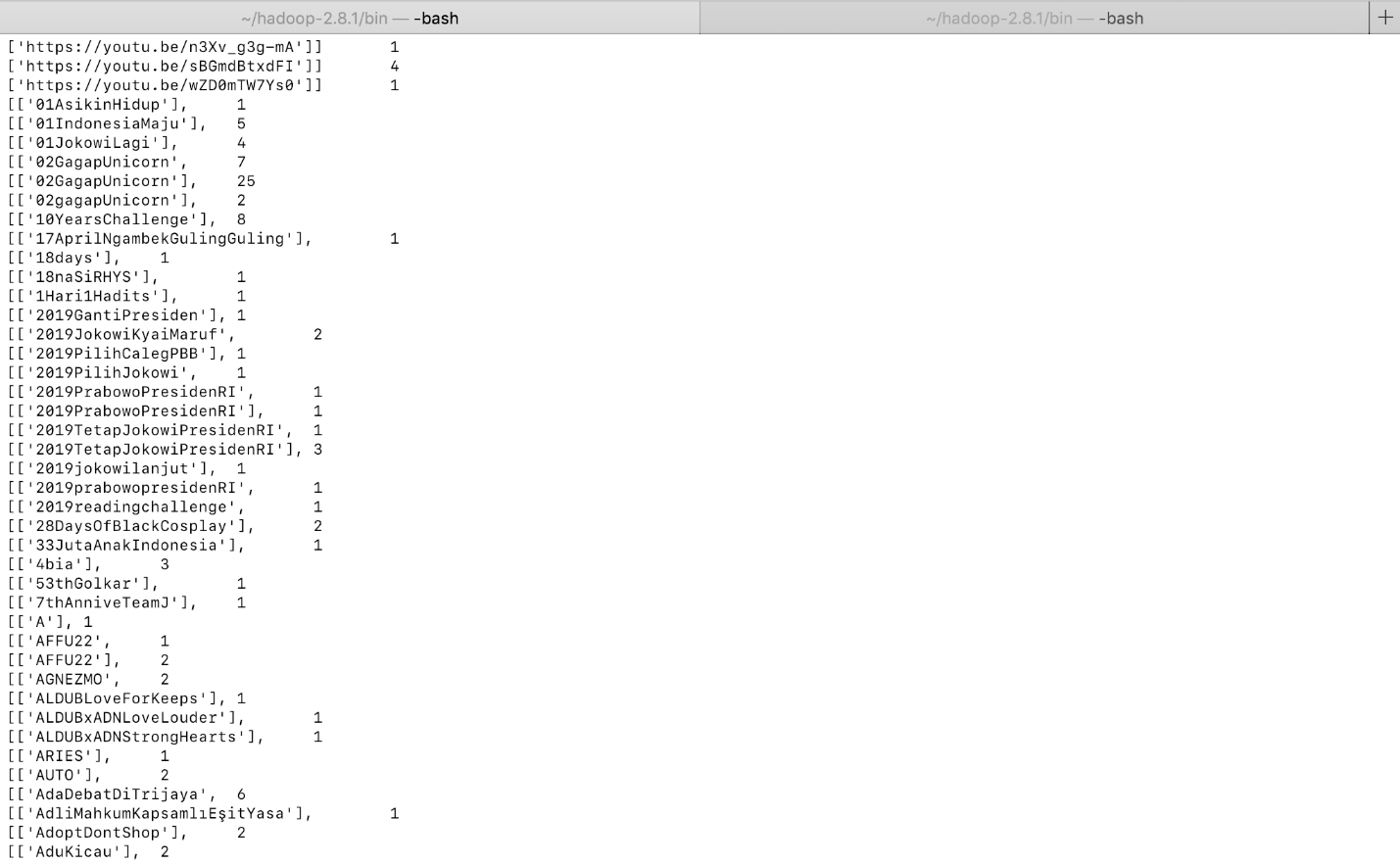




The above file is the output of Spark wordcount code. This gives the count of URL’s as well as Hashtags.

**WORD COUNT IN HADOOP**

The following is the wordcount output in Hadoop



**GOOGLE DRIVE LINK FOR ALL DOCUMENTS**

<https://drive.google.com/open?id=124fnYQuMWmifYkYJ77EJVoKtv2k7-1Lg>

**GITHUB LINK**

[**https://github.com/navyagonug/Twitter-Analysis**](https://github.com/navyagonug/Twitter-Analysis)

**REFERENCES**

* [www.google.com](http://www.google.com)
* <https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html>
* <http://adilmoujahid.com/posts/2014/07/twitter-analytics/>
* <https://www.youtube.com/watch?v=KDoZ6TsQHEg>
* <https://www.supergloo.com/fieldnotes/intellij-scala-spark/>
* https://www.cloudera.com/documentation/enterprise/5-5-x/topics/spark\_develop\_run.html
* https://developer.twitter.com/en/apps/16112643