CS5540: Principles to Big Data - SP2017

Project 1 - Team 2

**Project Overview:**

The objective of the project is to collect 100,000 tweets from Twitter and identify the top 10 hashtags using the Hadoop file system. This report is split into 5 sections, as follows:

1. Collection of tweets from twitter with Twitter API
2. Processing of tweets to categorize by hashtags and identify top 10
3. Screenshots of Hadoop File System
4. Usage of Map Reduce (Extra Requirement)
5. Appendix: References, Source Code, Etc

The sections will include text, diagrams and source code where applicable to illustrate the methods and results of this project.

From a high level perspective, the algorithm flows as follows:

Stage 1: Collection of Tweets

Stage 2: Processing and storage in Hadoop

Stage 2a: Processing and storage in Hadoop with MapReduce

TODO: Section 4 needs to be completed – will be finished by Saturday morning.

**Section 1.**

As dictated by the project requirements, 100K (100,000) tweets needed to be collected. To accomplish this Stage 1, this project uses the twitter API.By creating a twitter account and setting up an application, tweets can be collected via the Search API. For example, take this query:

https://api.twitter.com/1.1/search/tweets.json?q=the&count=100

This example would return 100 tweets containing the word “the”. The tweets will be returned in a JSON format and will need to be processed into t for stage 2.

From a high level perspective, the flow diagram looks like so:

Initialization

Pick a random word to search for

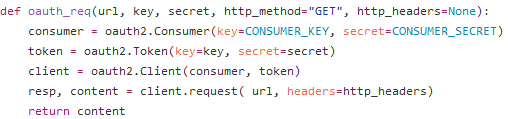
Using oauth2, connect to twitter and use search API to retrieve 100 tweets

Convert tweets into plaintext and write to output files.

Repeat to collect 100,000 Tweets

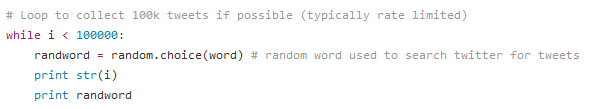
The full source code is available in the appendix.

During initialization several libraries are imported, one of which is oauth2. In order to use certain twitter API queries, authentication must first be established. For this project, oauth2 was used; a function, oauth\_req, is query twitter. The function will be explained later on.

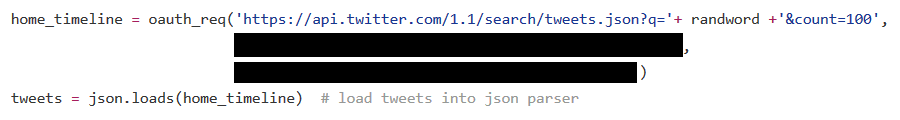


Output files are created and prepared to be written to and a dictionary is opened and stored in a list to facilitate the picking or a random word. In addition, the consumer key and consumer secret are set.

A while loop is started, based on increment variable i. i will increment each time a tweet is written to an output file. A random word is chosen from the list of words that were initialized earlier.



With the random word chosen, the program will now query Twitter.



The function will return the tweets to home\_timeline. Three arguments are passed into the oauth\_req function.

The first is a string concatenated from three parts:

'https://api.twitter.com/1.1/search/tweets.json?q='+ randword +'&count=100'

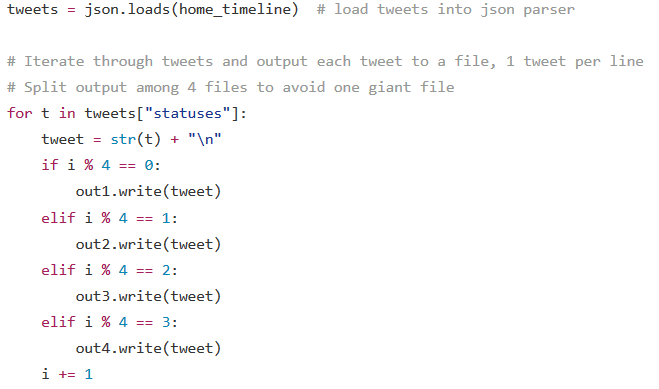
If for example randword was chosen to be “bye”, the concatenated string would be:

'https://api.twitter.com/1.1/search/tweets.json?q=bye&count=100'

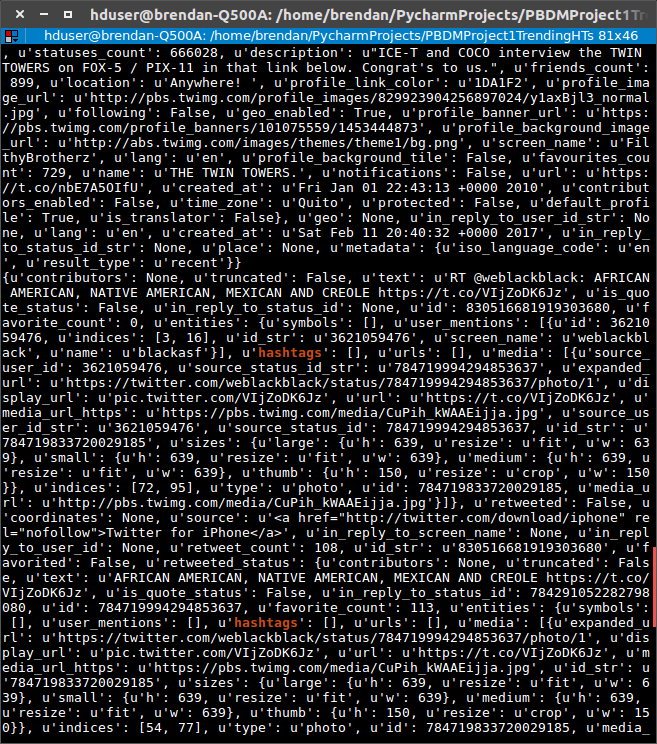
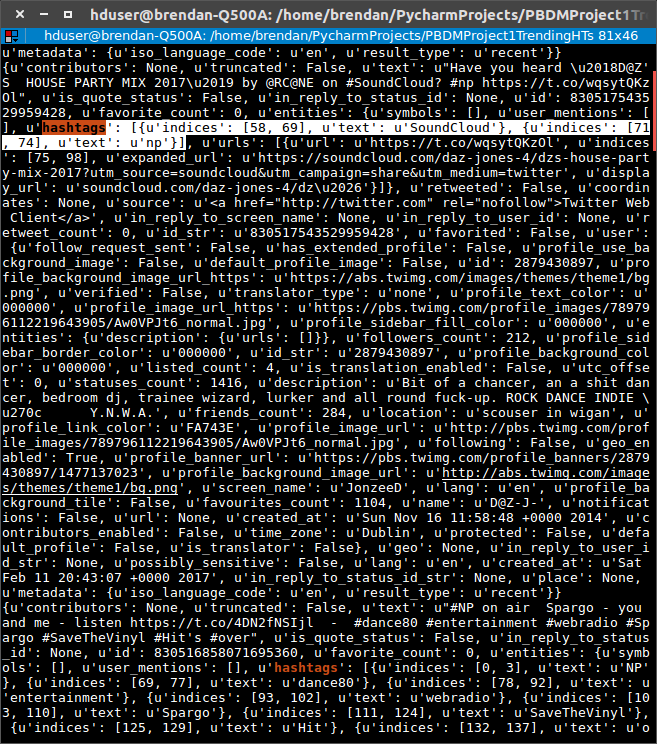
The twitter API will understand this to be a request for 100 tweets containing the text “bye”. The other two arguments, blacked out, are the token key and token secret. This program is meant to work with a Twitter App, and requires the key and secret for authentication. The earlier defined consumer secret and consumer token are also used for authentication.

home\_timeline is then loaded into a JSON parser and stored in tweets. The tweets are converted to string type and then written to output files, one tweet per line. This process, random word, query and write, are repeated until 100,000 tweets are collected.

Write code snippet:



Examples of collected tweets. Right hand side one has no hashtags.



**Section 2.**

With the tweets collected, they must now be processed and sorted. Based on the top 10 most common hashtags, 10 categories will be created in the HDFS: one for each hashtag. Two additional categories will also be created to store tweets that either have no hashtags or have a less common hashtag.

From a high level perspective, the flow diagram looks like so:

Initialization

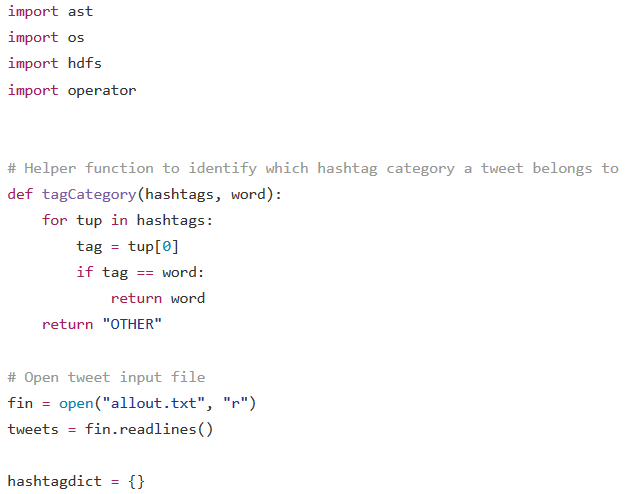
Send to Hadoop and close

Write each tweet to respective category

Connect to Hadoop and create output files for the 12 categories

Identify top 10 hashtags

Initialization will import relevant libraries, such as hdfs which is needed to interface with Hadoop.



The function tagCategory will be explained later. The file containing the tweets is opened and stored in a list. Finally a dictionary is initialized. This dictionary, hashtagdict, will store a count of hashtag occurrences to identify the top 10.

Using a for loop, each tweet’s ‘hashtags’ property is stored in ht. Since a tweet can have multiple hashtags, a for-loop is used to check each one. If the hashtag is already in the dictionary, simply increment the value. If there is no entry for that hashtag, then set it to one.

After going through every hashtag, in every tweet, the dictionary is then sorted so that the entries with the most occurrences are placed at the front. Store those first ten into a list called topten.

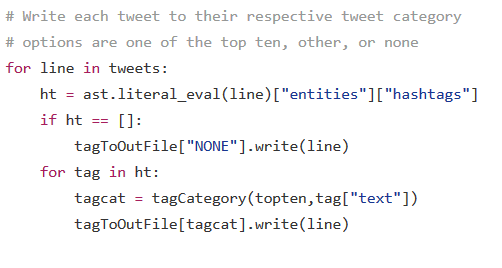


With both a sorted list of hashtags by occurrences and the top ten hashtags, the tweets must now be placed into the Hadoop File system. This is accomplished using the hdfs library.

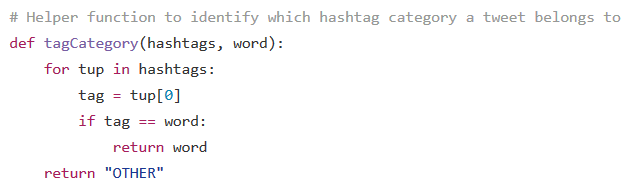
First a connection is established to the HDFS server, which is run locally. Using the topten list, ten directories are created in the HDFS system; a map from the tags to the directories are also created. Two additional directories are set to complete the 12 categories. Output files are then set to store the tweets. For categories of “OTHER” and “NONE”, the tags are set directly to the text file.



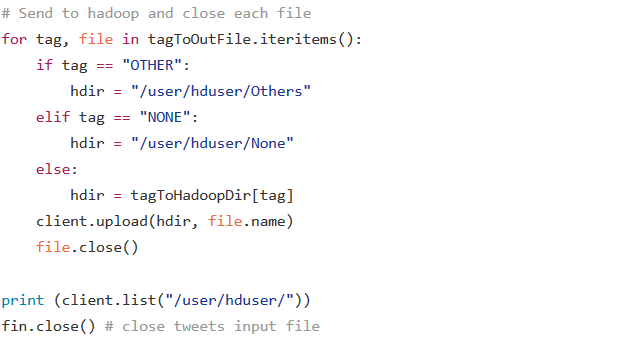
The program now begins writing tweets to the appropriate categories. Similar to before, the tweet’s hashtags are stored in ht. If ht is null (no hashtag) then write to the file tagged as NONE. This is accomplished using the tagToOutFile function. Otherwise, check each hashtag and write to the appropriate file.



A key function is the tagCategory function, which was defined earlier. If the tweet’s hashtag is among the top ten, it will be returned to be used as a tag. If the tweet’s hashtag is not among the top ten, “OTHER” will be returned to be used as a tag.



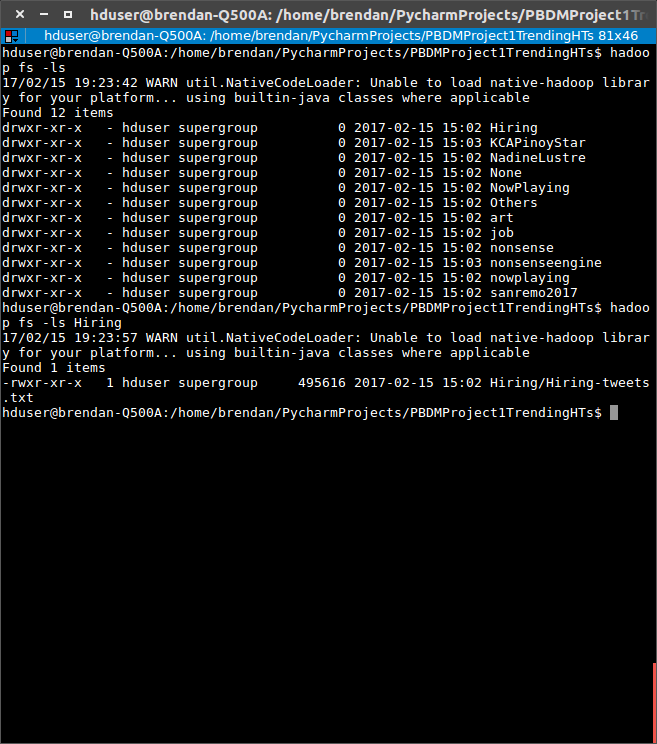
After all tweets have been written to the appropriate files, the program will tell Hadoop to close the files and finish up.



With the program complete, the project’s main requirement has been fulfilled.

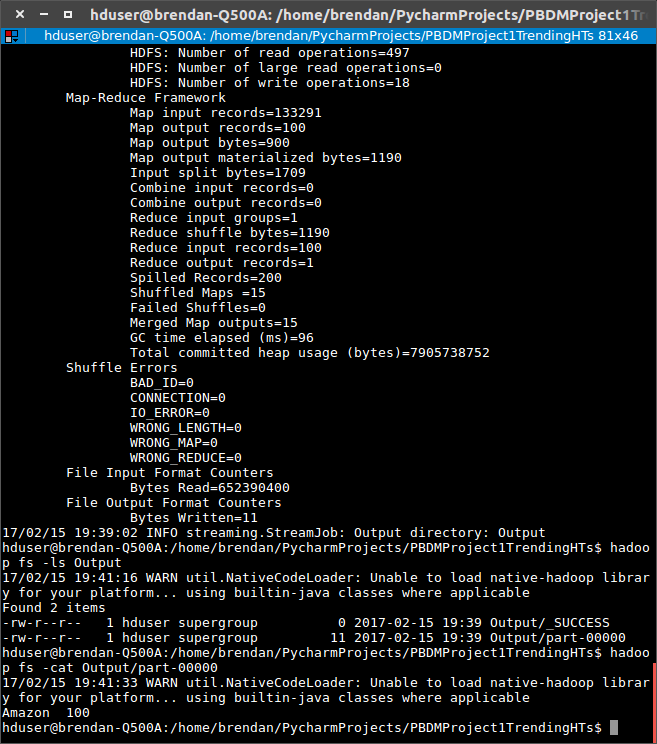
**Section 3.**

Screenshots of the HDFS after main program completion.



**Section 4.**





**Section 5.**

References:

* Twitter API:
  + <https://dev.twitter.com/docs>
* Twitter Search API:
  + <https://dev.twitter.com/rest/public/search>
* Twitter Authentication API:
  + <https://dev.twitter.com/oauth/application-only>
* OAuth2 Python Library:
  + <https://github.com/joestump/python-oauth2>
* OAuth example:
* <https://dev.twitter.com/oauth/overview/single-user>
* HDFS Install:
  + <http://www.bogotobogo.com/Hadoop/BigData_hadoop_Install_on_ubuntu_single_node_cluster.php>
* HDFS Quickstart:
  + <https://hdfscli.readthedocs.io/en/latest/quickstart.html#configuration>
* Check if Directory Exists:
  + http://stackoverflow.com/questions/273192/how-to-check-if-a-directory-exists-and-create-it-if-necessary
* Sorting Dictionary:
  + [http://stackoverflow.com/questions/613183/sort-a-python-dictionary-by-v`alue](http://stackoverflow.com/questions/613183/sort-a-python-dictionary-by-v%60alue)
* Hadoop MapReduce
  + <https://hadoop.apache.org/docs/r2.6.0/hadoop-mapreduce-client/hadoop-mapreduce-client-core/HadoopStreaming.html>
* Python/Hadoop MapReduce
  + <http://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/>