**Principles of Big Data Management**

**CSEE5540**

**Phase 1 Report**

Team :

Chandrasekhar Pentakota (16267781)

Kartheek Katta (16273764)

Jaya Prakash Ravella (16271007)

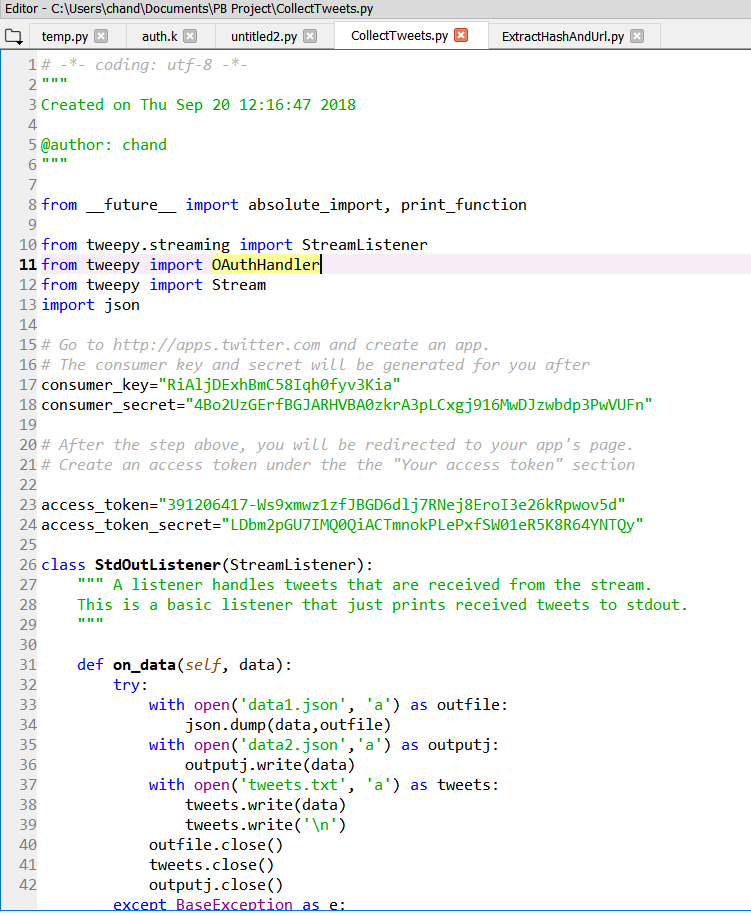
**Links :**

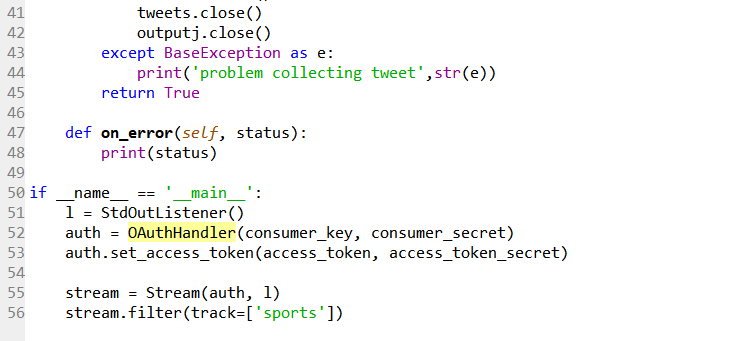
<https://github.com/chanduhub/BigDataProgramming/wiki>

**Task 1**

Collect tweets using twitters streaming API(tweepy)

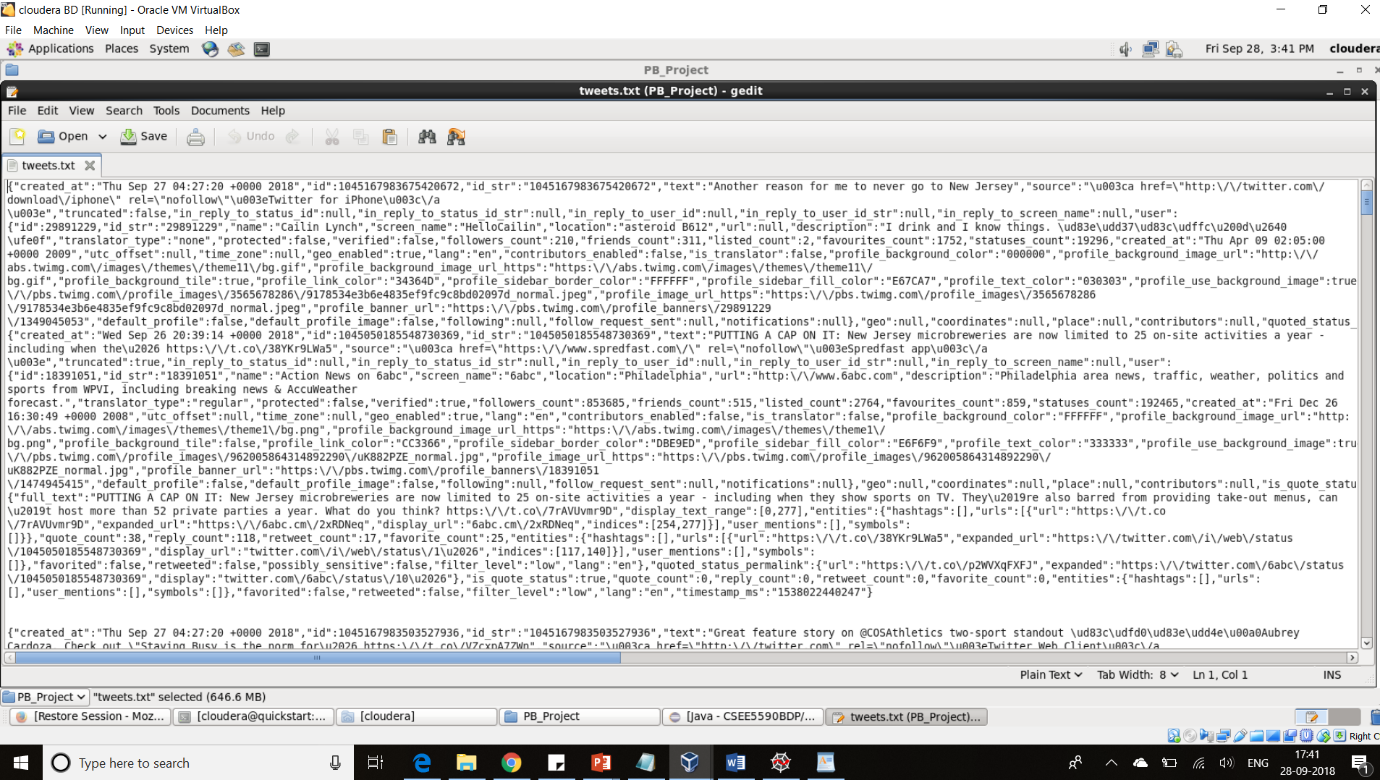
Python code for collecting tweets :





The above python code collects tweets into a text file by using Twitter’s streaming API Tweepy. In order to connect to the Twitter’s Streaming API, we need to authenticate using credentials from our twitter developer account. The program filters data by "Sports" Keyword. The data is appended to file ‘twitter\_data.txt’. We have collected around tweets.

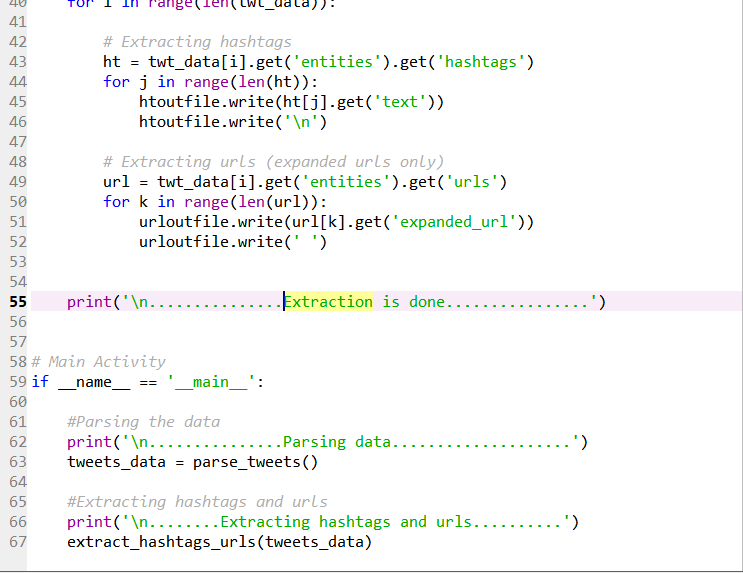
Output file generated is of size 646.6 MB



**Task 2**

Hashtags and URLs extraction from the twitter data

****

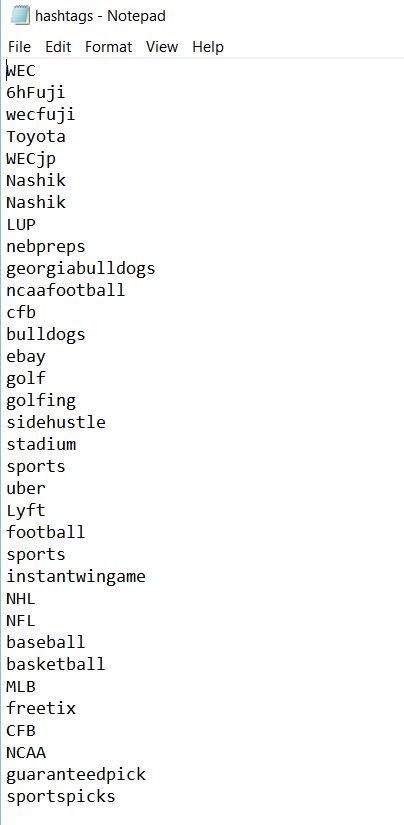
****

The above program parses tweets file into json format and extracts required Hashtags and urls and writes to two different files.

Outputfiles :

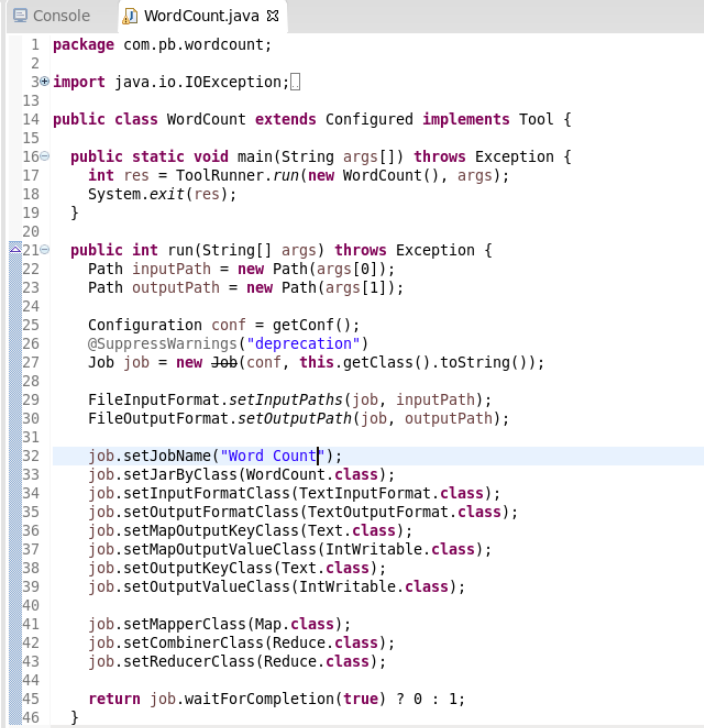
urlsoutput 

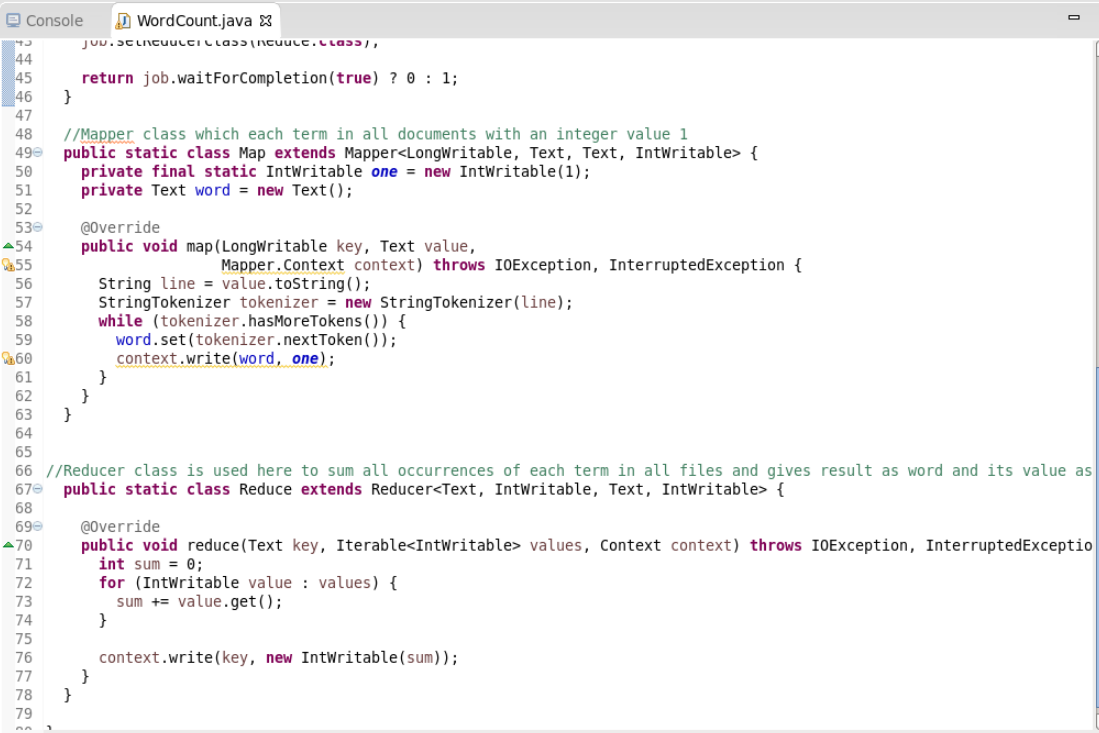
Hashtagsoutput



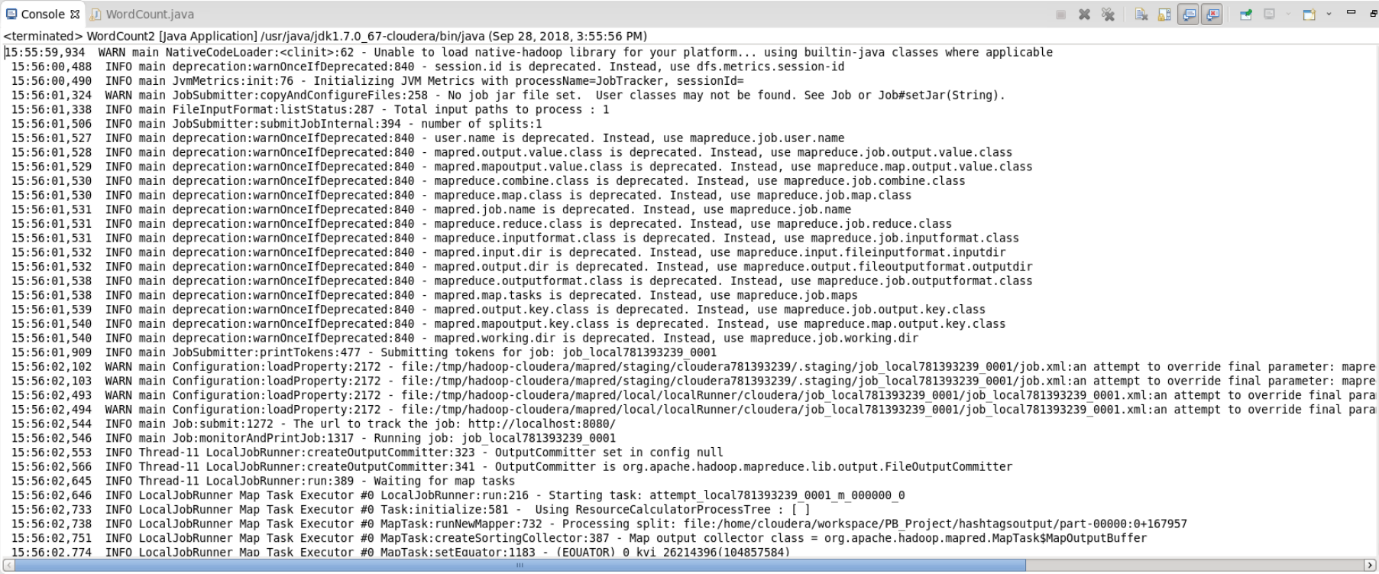
**Task 3**

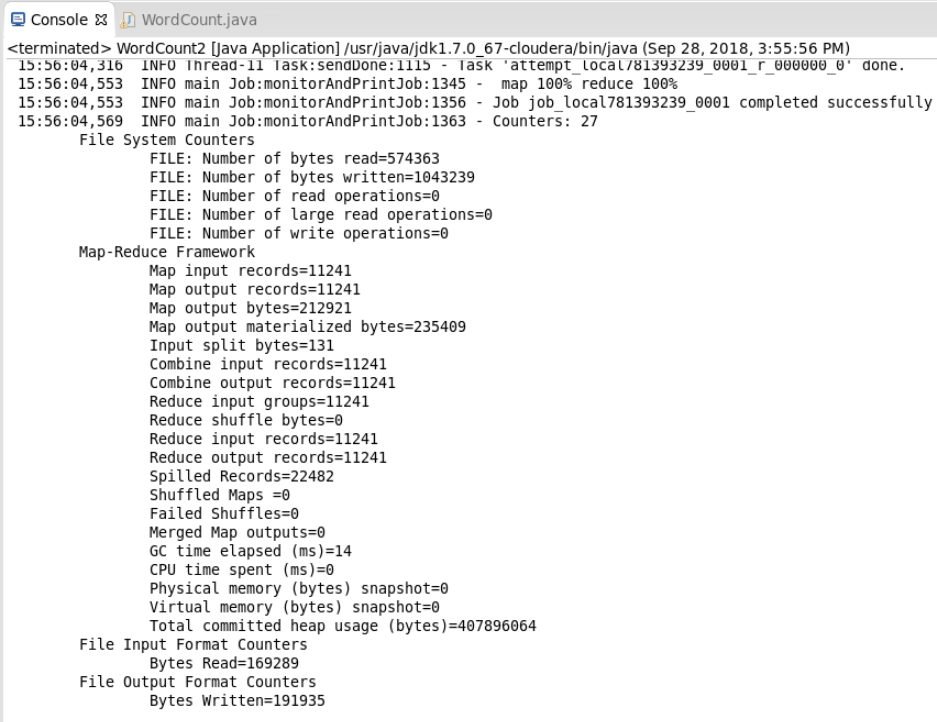
WordCount Program Using Apache Hadoop.



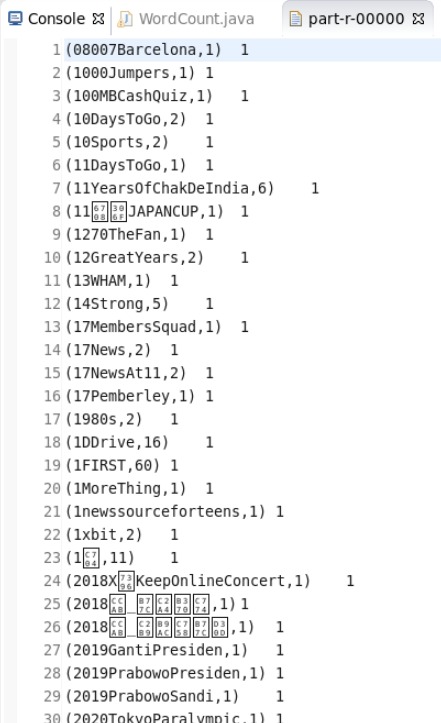


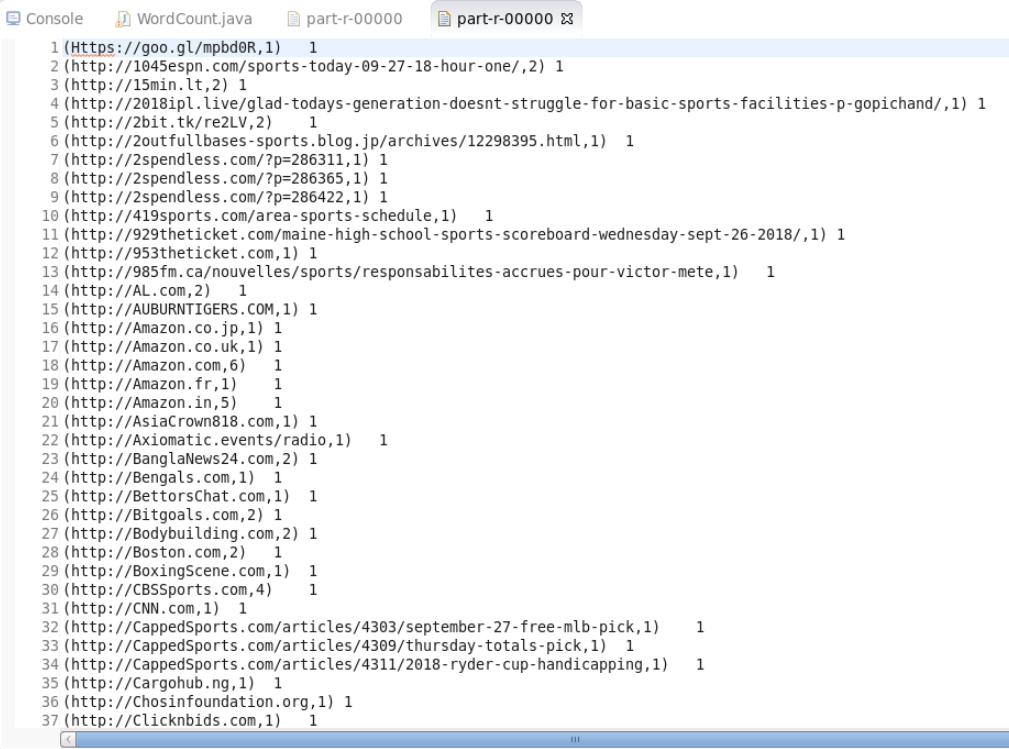
Log generated by running above program on both hashtags file and url file





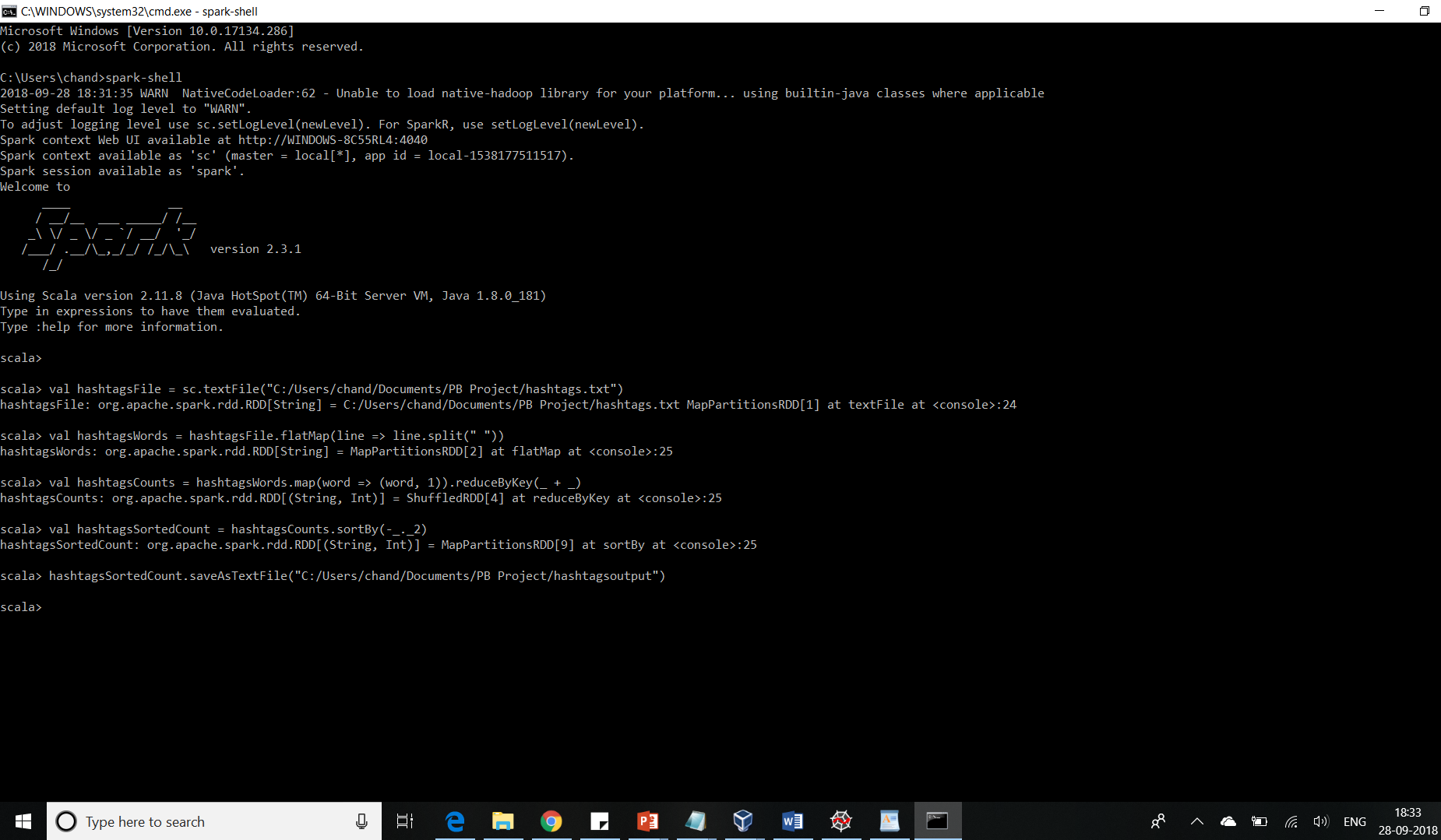
Output file showing hastags, urls and their counts

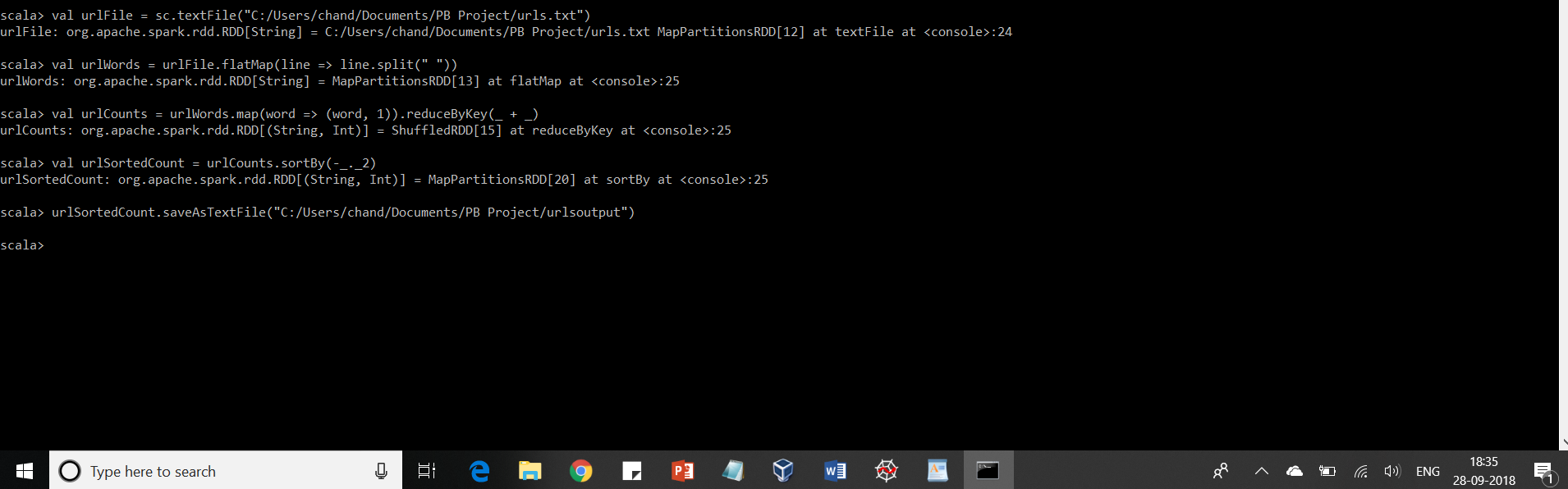




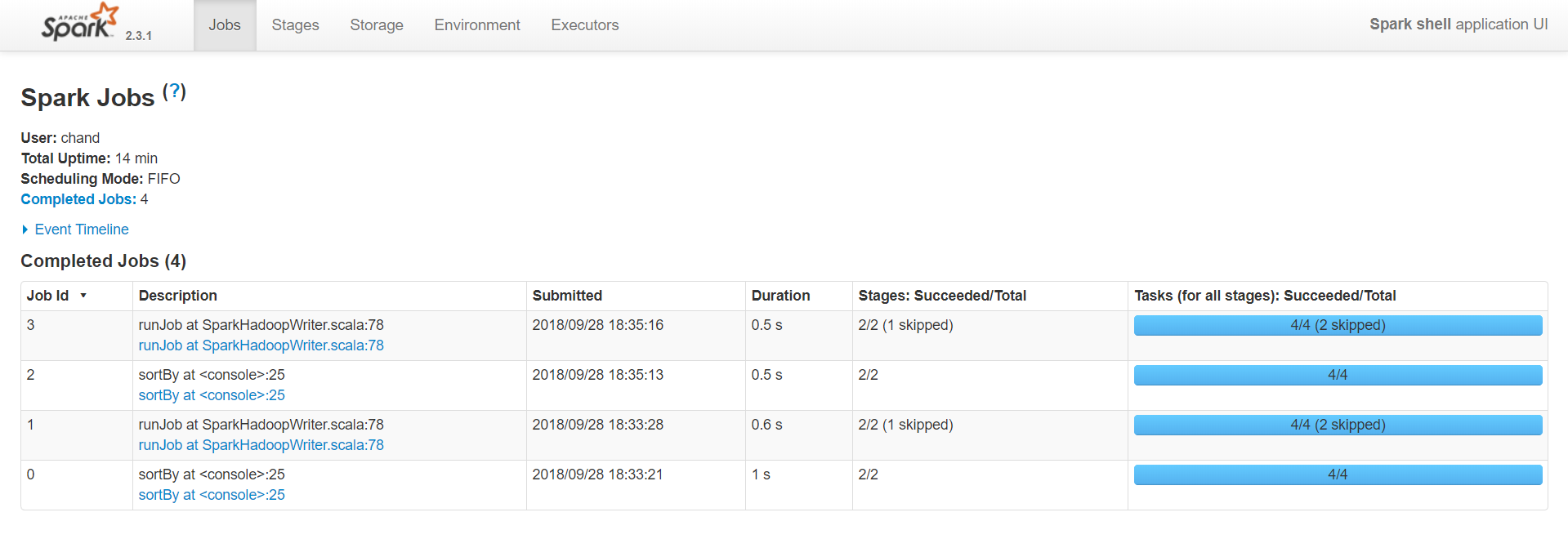
Here Map function gets input as byte offset as key and each line as value. StringTokenizer breaks each line into words splitting with space character. Each word in all documents is combined with an integer value one and sent to combiner and reducer as value with the key as word. Each Reducer takes Similar key values from the mapper, and counts all the values of unique key and send it as output value. The result from the reducer is word and the number of its occurrences from all the documents.

WordCount Program Using Apache Spark.





In above spark code, first rdd is created by using textFile function of the sparkcontext from the hashtags and urls textfile. By using flatMap function, we generated words from each line by taking space as delimiter and appended one as value with the word as key. Using reduceByKey function we are adding values by the key and the output generated is the words and their sum is saved to a textfile.



Outputfiles generated from the above spark jobs for both hashtags and urls

