**Principles of Big Data Management Project (CS5540)**

**Project Phase 1**

**Extraction of Twitter Data**

**and Analysis**

**TEAM**

Pooja Reddy Gopu - 16297702

Sandeep Meesala - 16299481

Harini Reddy Anumandla - 16301823

**Instructor**

**Main Objective:**

The objective of this part of project is to store, analyze and visualize the twitter data. User Tweets are collected for different sports to compare popularity between various games globally.

**Objective of Phase 1:**

Collect Twitter tweets using the Twitter API in JSON format. The collected files in JSON format are stored into HDFS and then URLs and hashtags are extracted using Python script. Execute word count program on extracted URL’s and Hashtags in Hadoop and Spark.

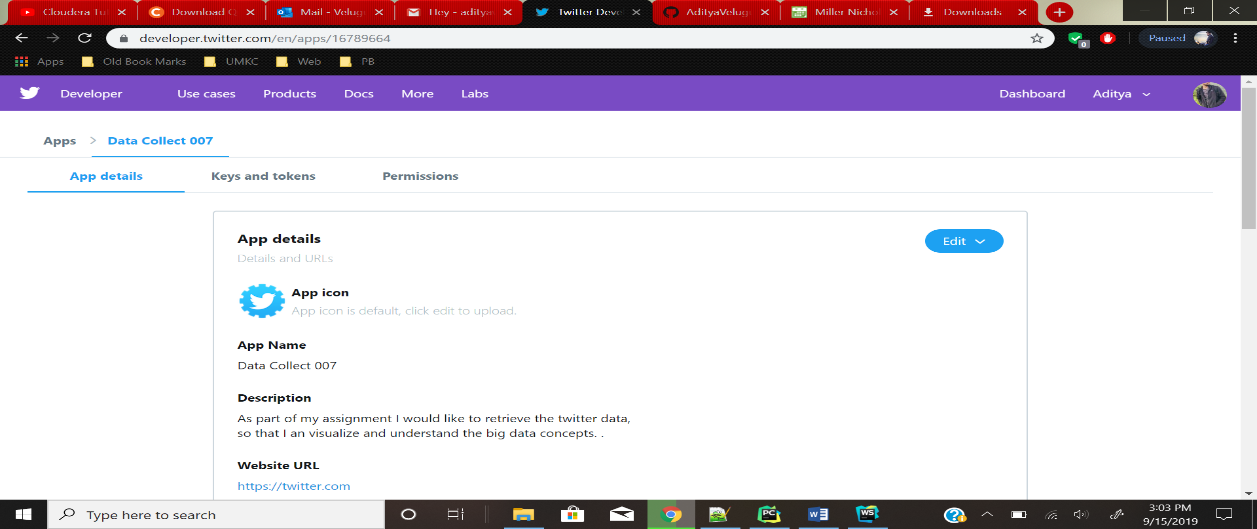
### **Technologies Used:**

Python scripting  
Apache Hadoop  
Apache Spark

**Step 1: Collecting tweets from twitter using Twitter streaming APIs.**

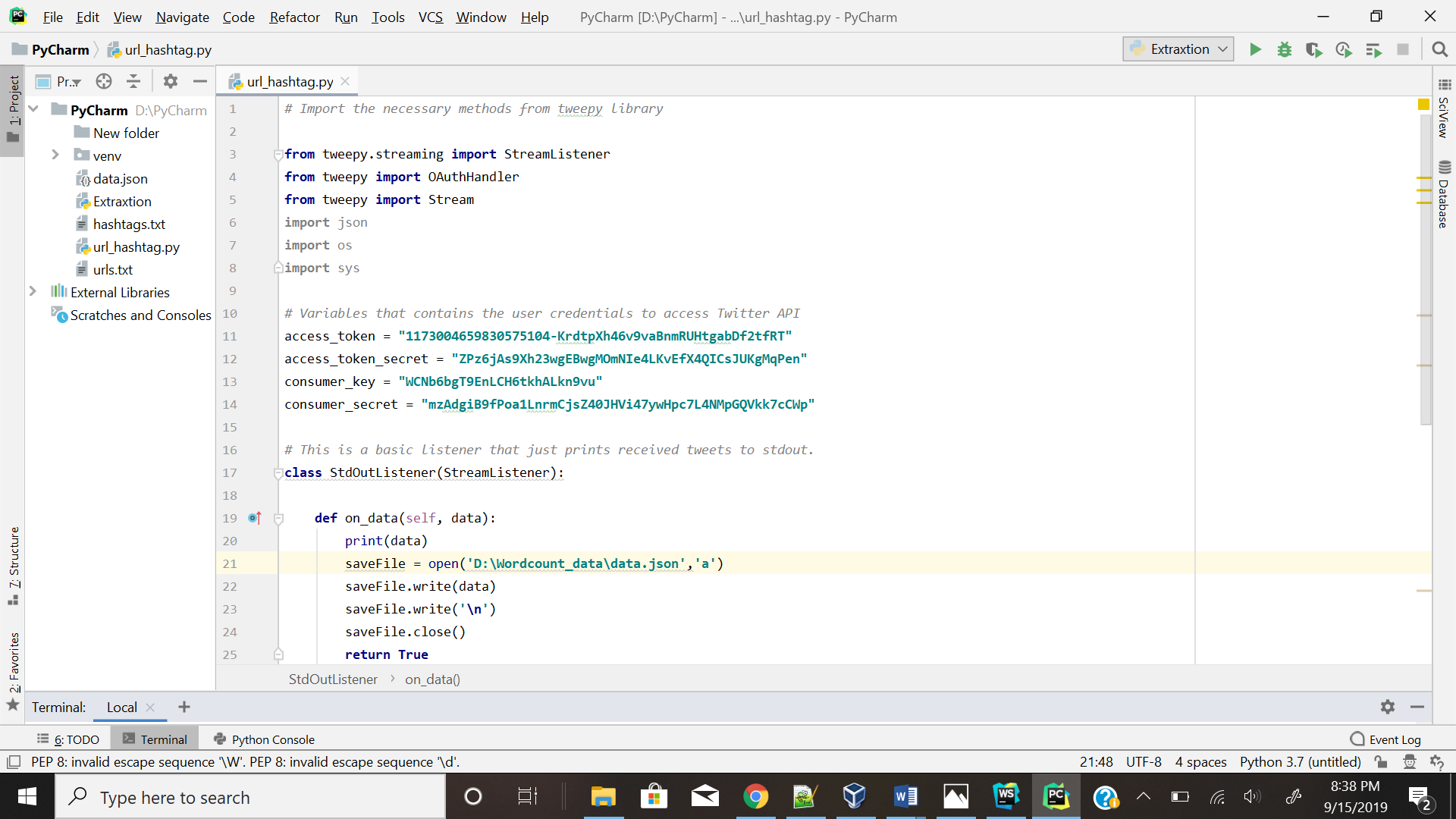
1. **Generating twitter streaming APIs.**

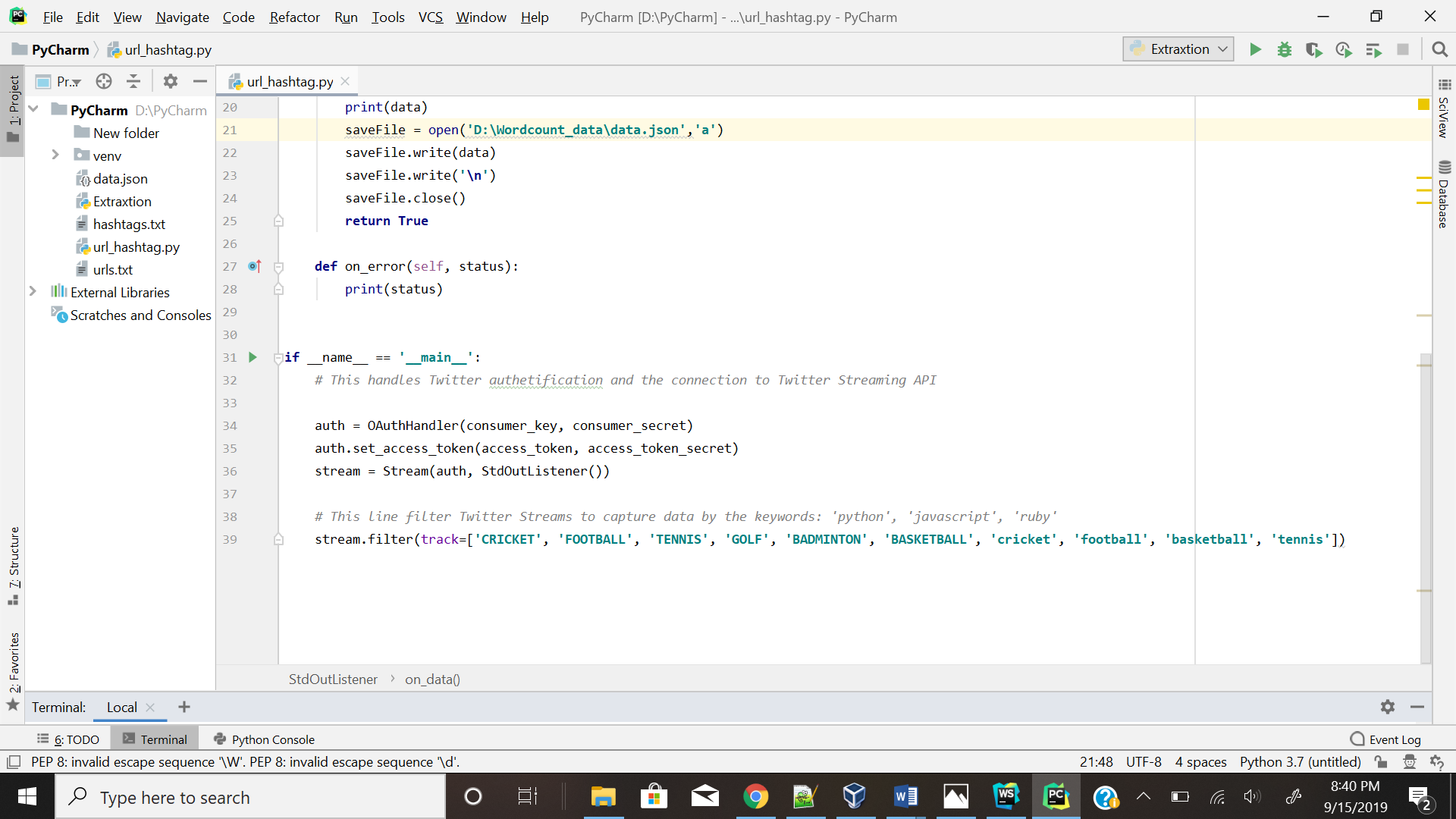
* Login to Twitter developer account( <https://developer.twitter.com/en/apps> ).
* Create an application and generate consumer and access API keys.



1. **Installed PyCharm and Anaconda for Twitter data collection.**

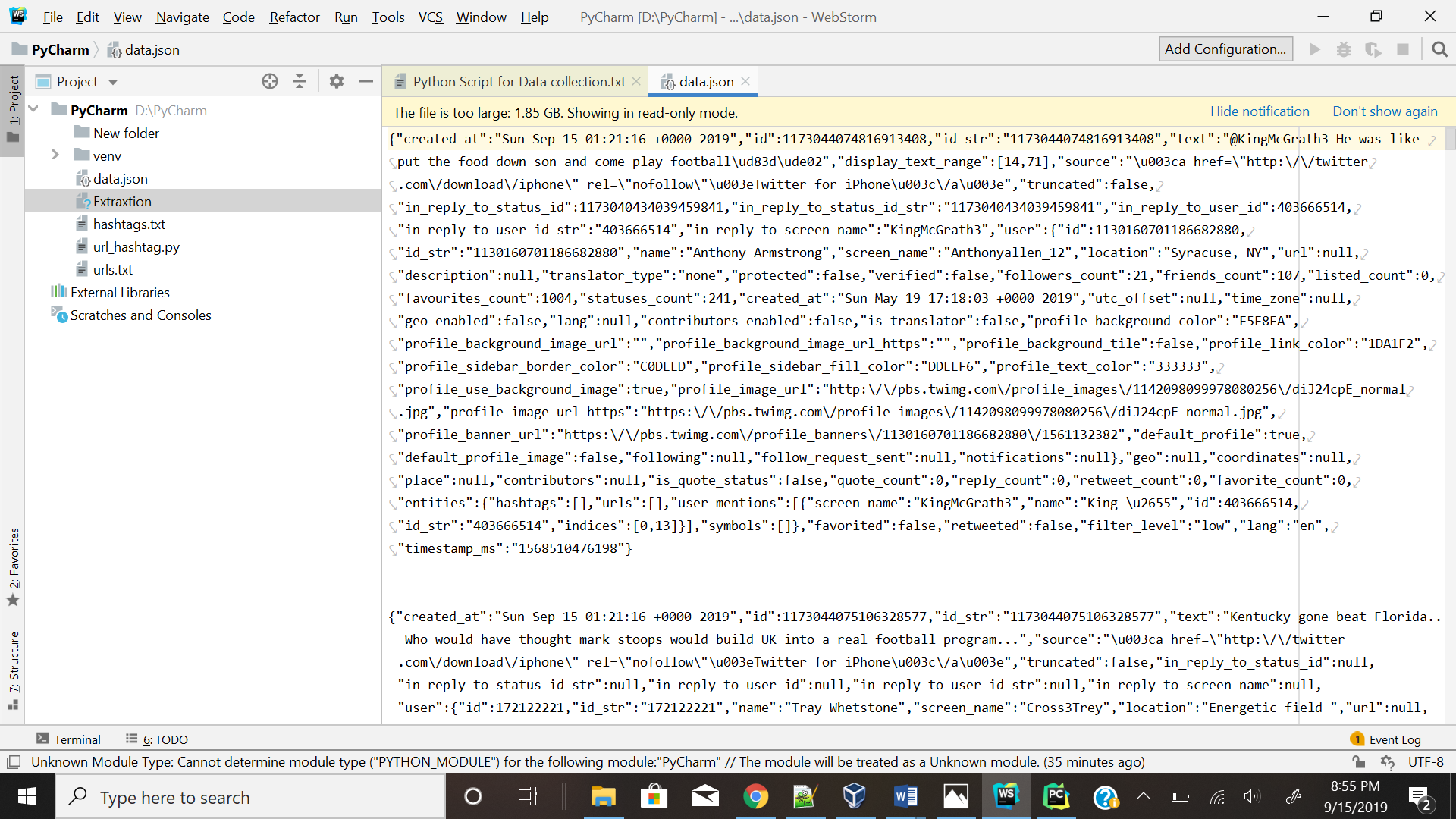
* Install libraries for JSON, OS and SYS.
* Executed Python script for extraction of twitter data using the access keys.





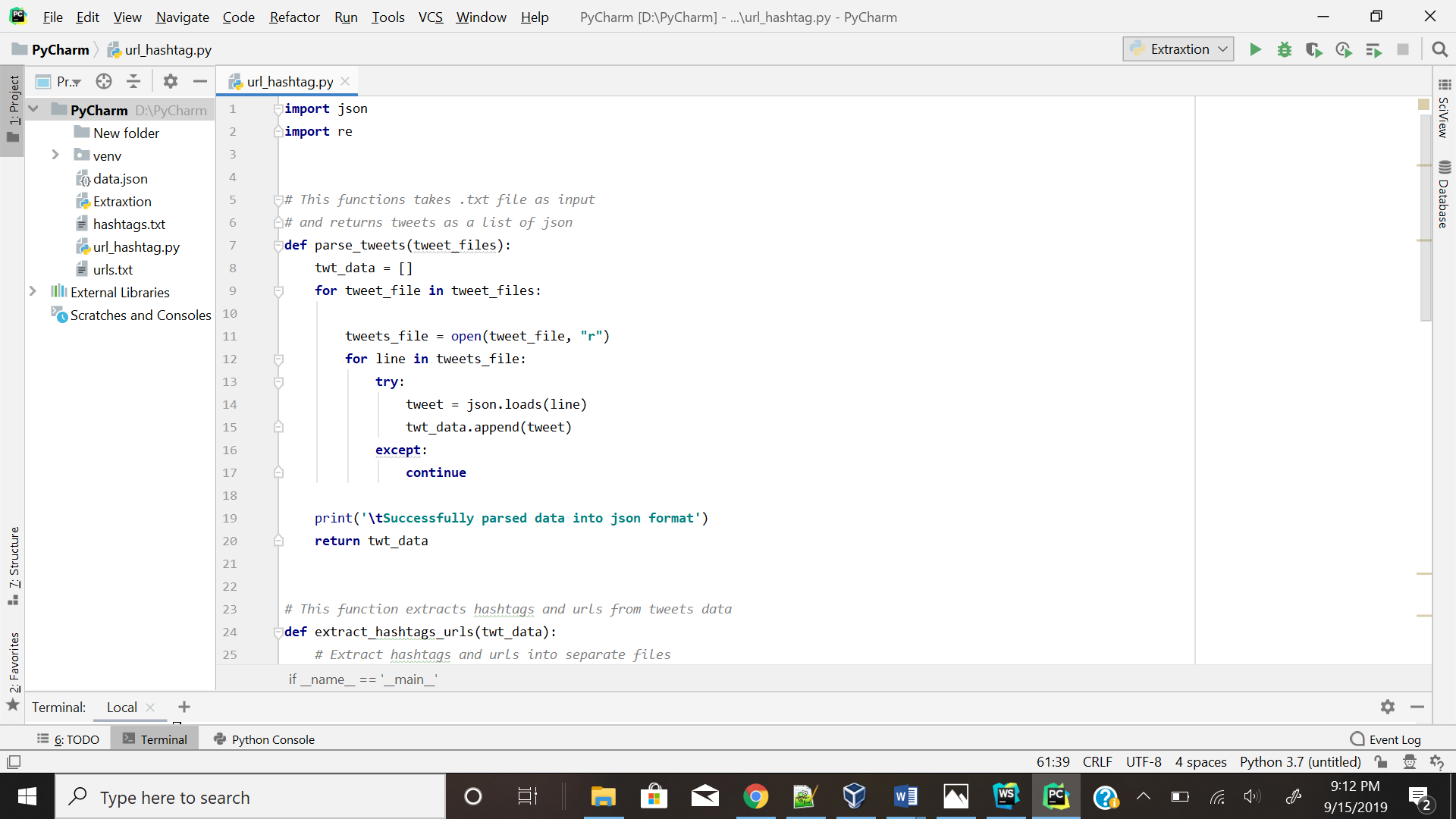
Note: The above two screenshots are the Python scripts used for the extraction of twitter data in JSON format.

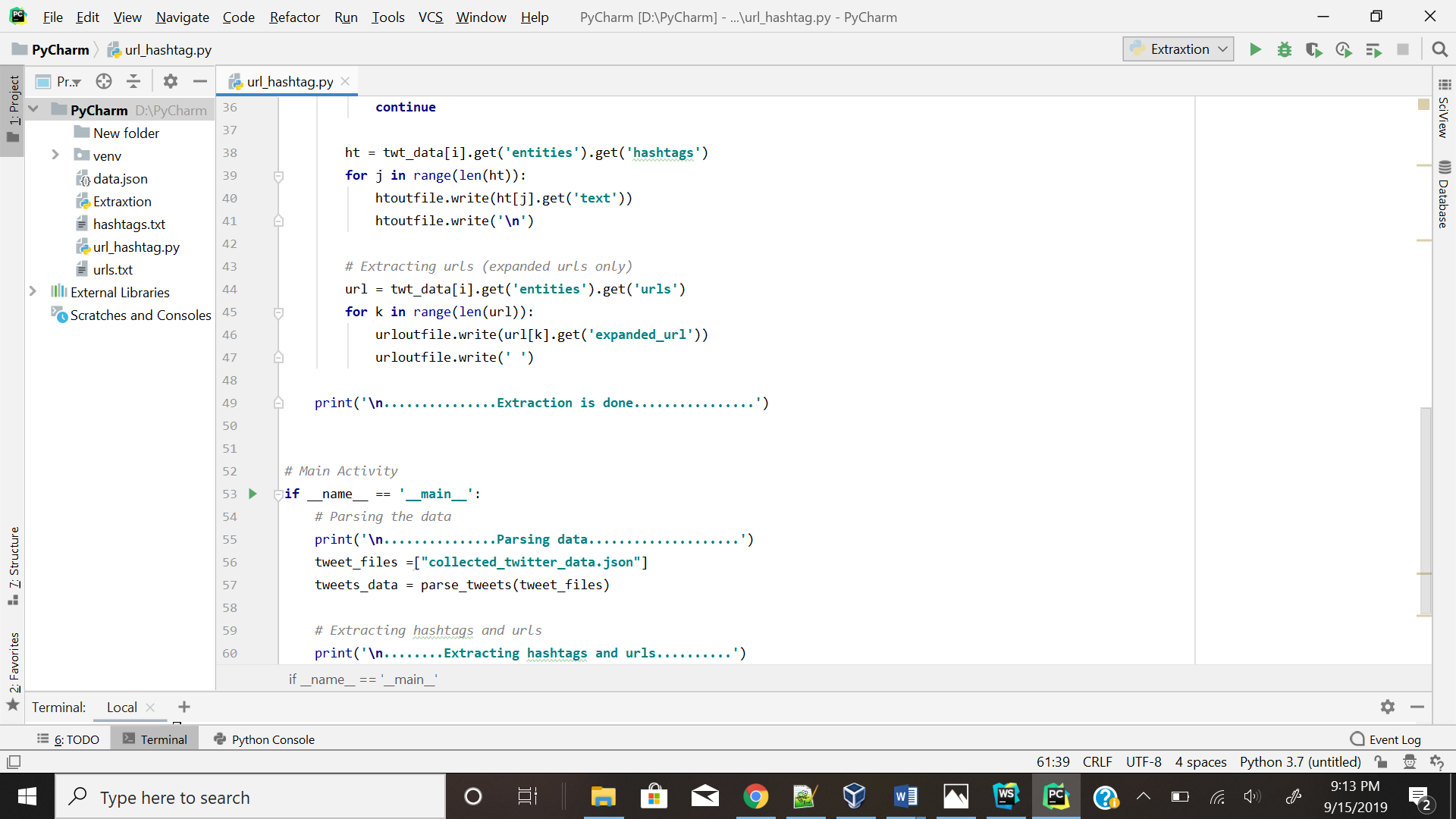
**Collected Twitter Data**



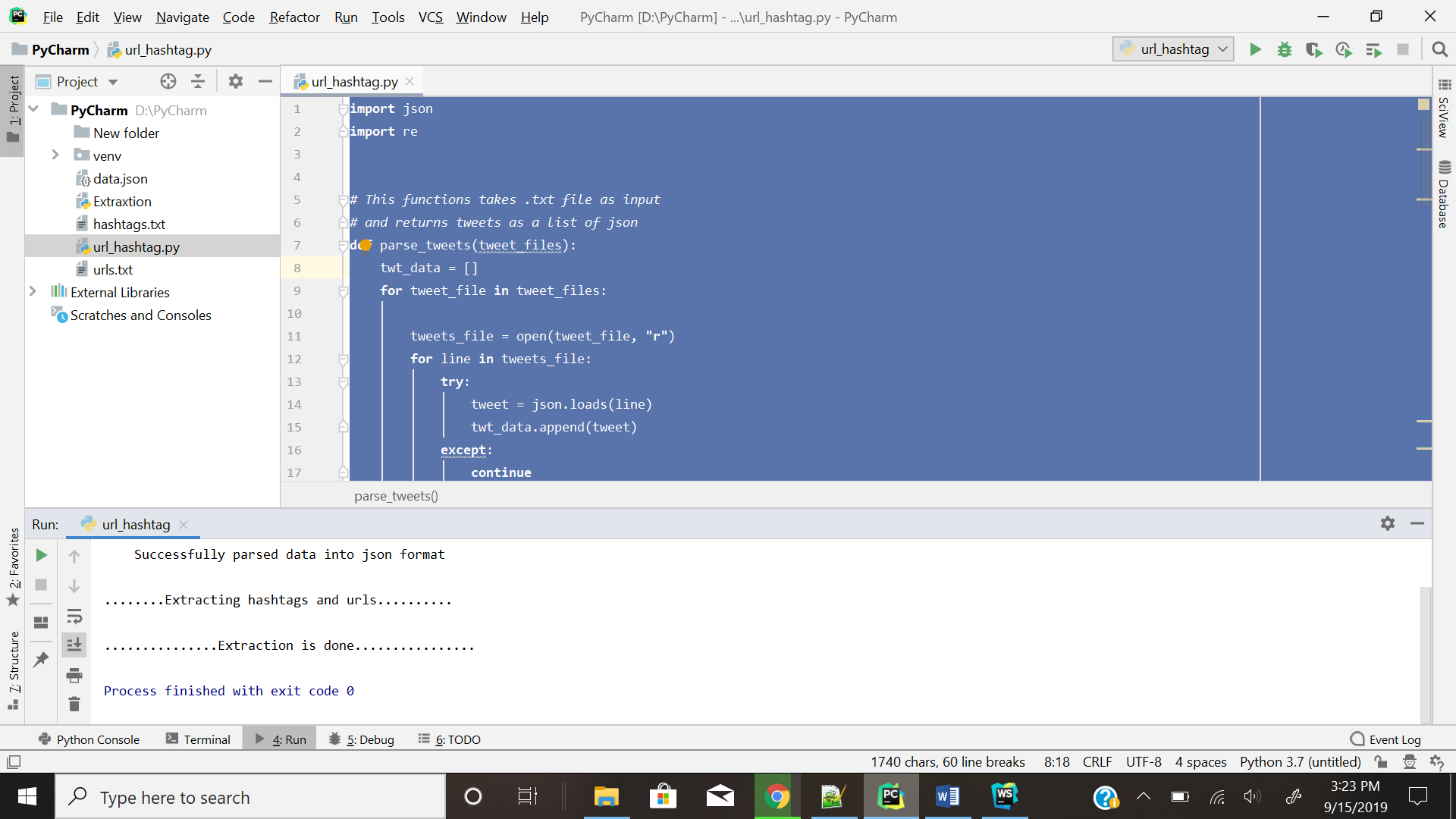
1. Generated Hashtag and URL files for wordcount.

* Python script is written to separate the Hashtags and URLs from the twitter data.
* The Hashtags and URL files are then used for the word count program.



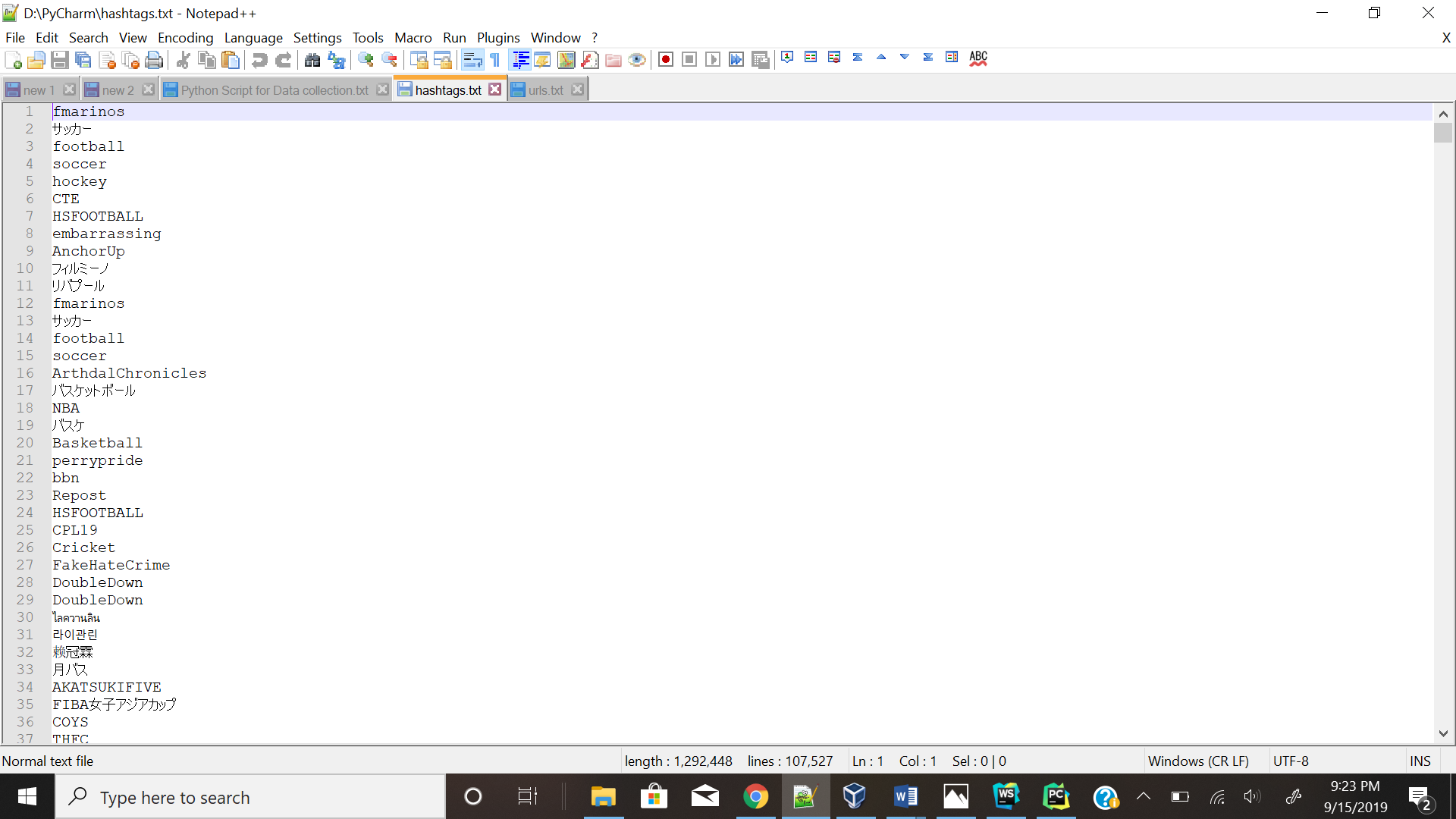


**Note: The above screenshot is the script for the extraction of Hashtag and URL files.**

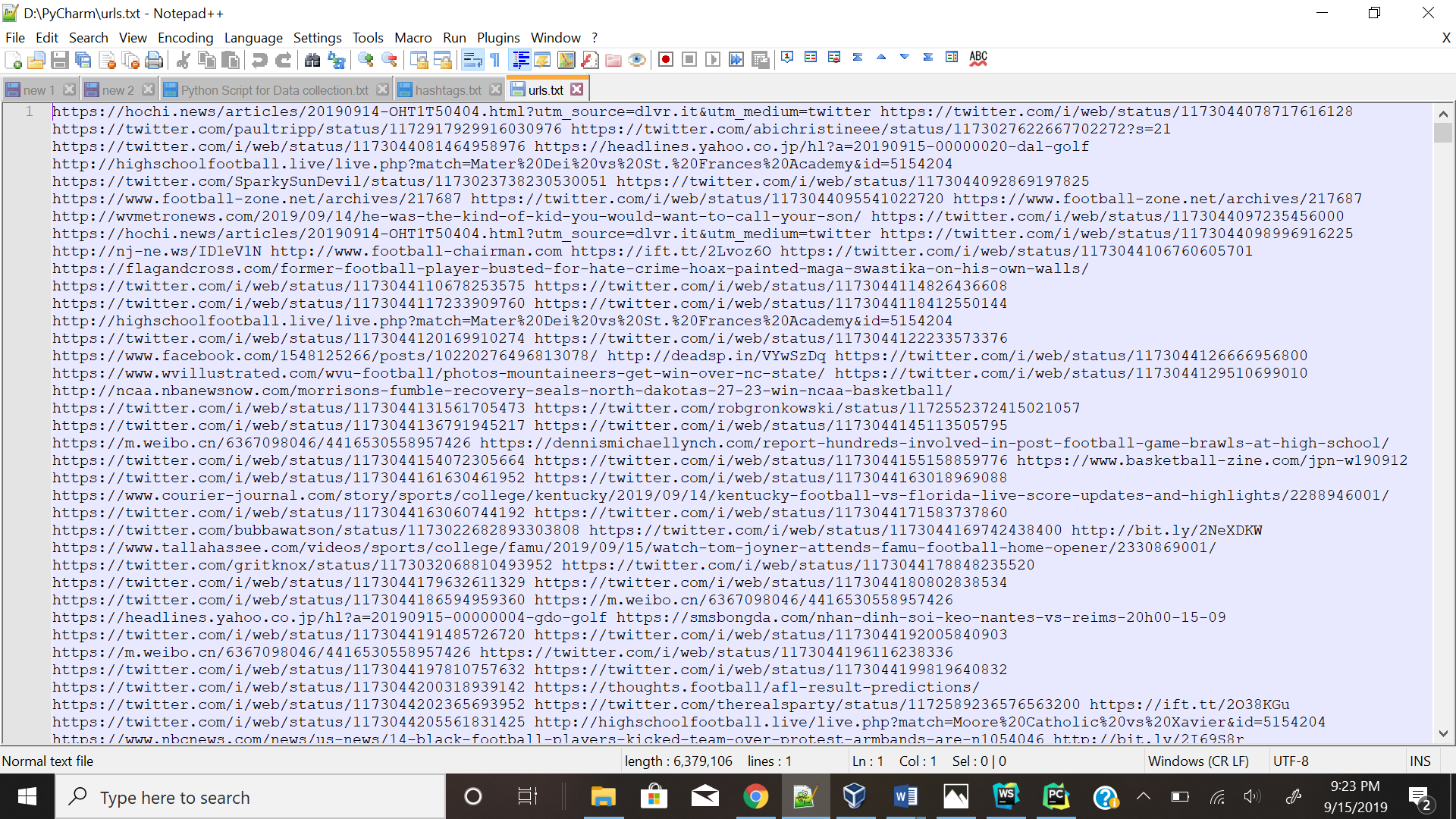


**Note: The above screenshot shows the URL and Hashtag files are generated.**

**Extracted URLs and Hashtags Files:**



**Hashtags**



**URLs**

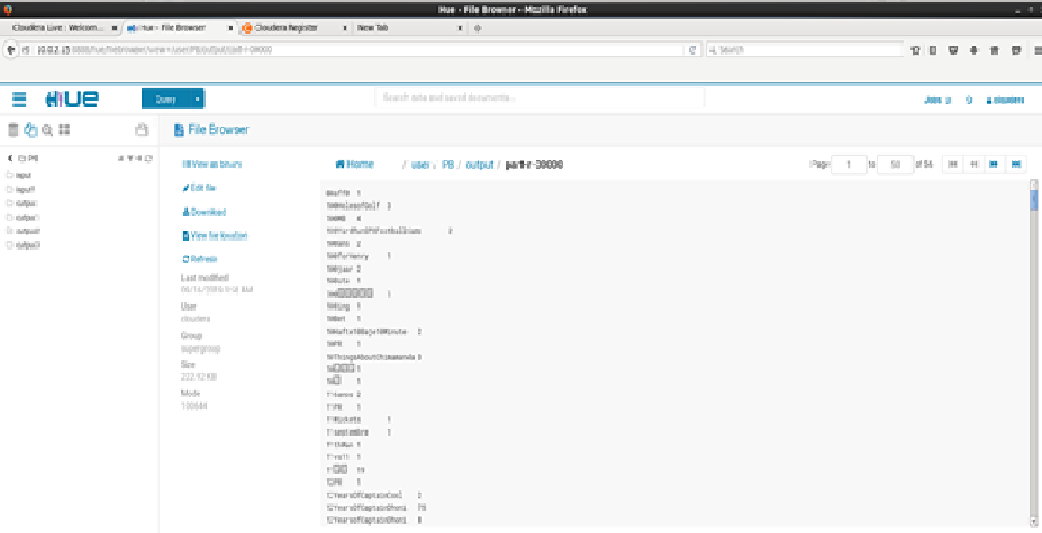
### **Running Word Count program** **using** **MapReduce.**

* The URL and Hashtag files are placed on HDFS in a directory.
* The wordcount program is executed on both url and hashtag files.

**Please find below screenshots for Hashtag using MapReduce.**



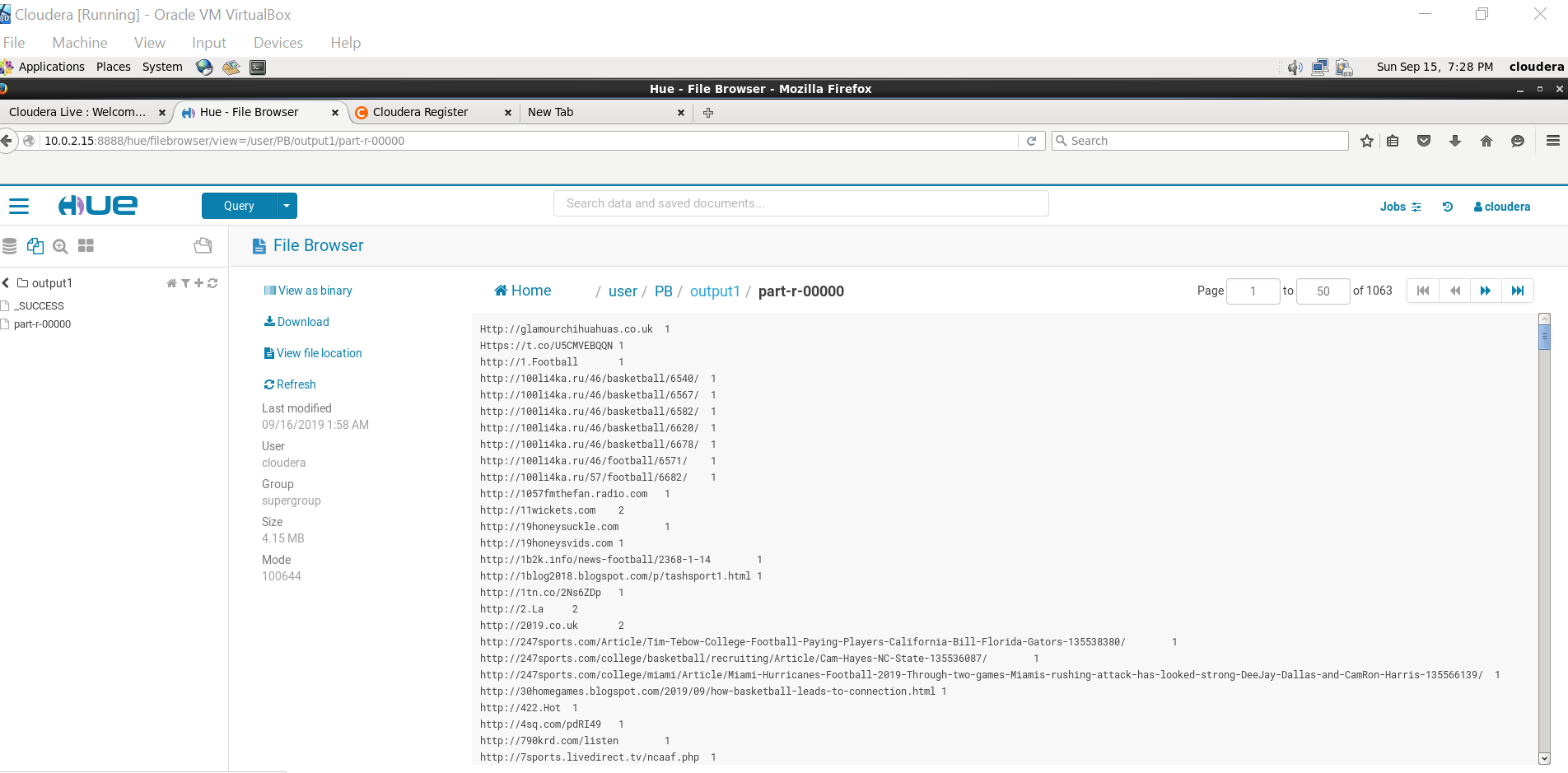
**Output of Hashtag Count**



**Please find below screenshots for URL.**



**Output of URL Count**

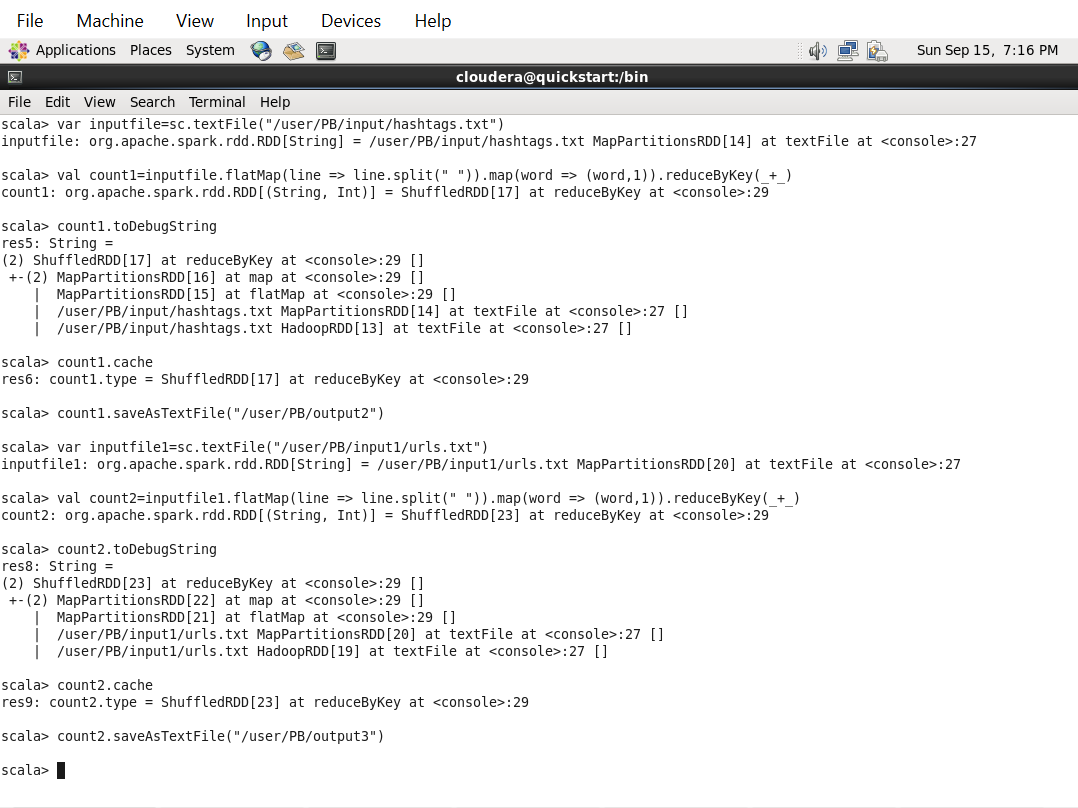


1. **Running Word Count program** **using Spark.**

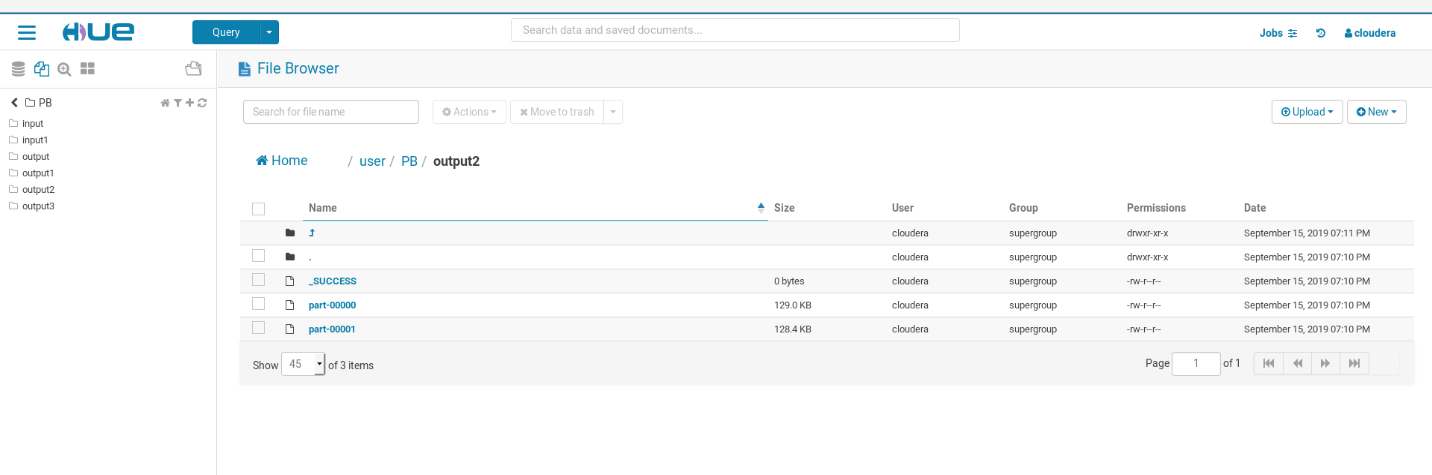
* Variable’s are assigned for Hashtag and URL input files and then Wordcount program is executed. The output files are stored.

**Please find the below screenshots for input and output**.

**Spark Execution**



**Spark Output**



**The Data extracted is uploaded on the below google drive link:**

[**https://drive.google.com/drive/folders/1LrvTGX4qBXBoTMu855l4jxrUKkA\_z3GU**](https://drive.google.com/drive/folders/1LrvTGX4qBXBoTMu855l4jxrUKkA_z3GU)

**The output files can be found in the following github location:**

[**https://github.com/AdityaVelugula/Prin-of-Big-Data-Mgmt**](https://github.com/AdityaVelugula/Prin-of-Big-Data-Mgmt)