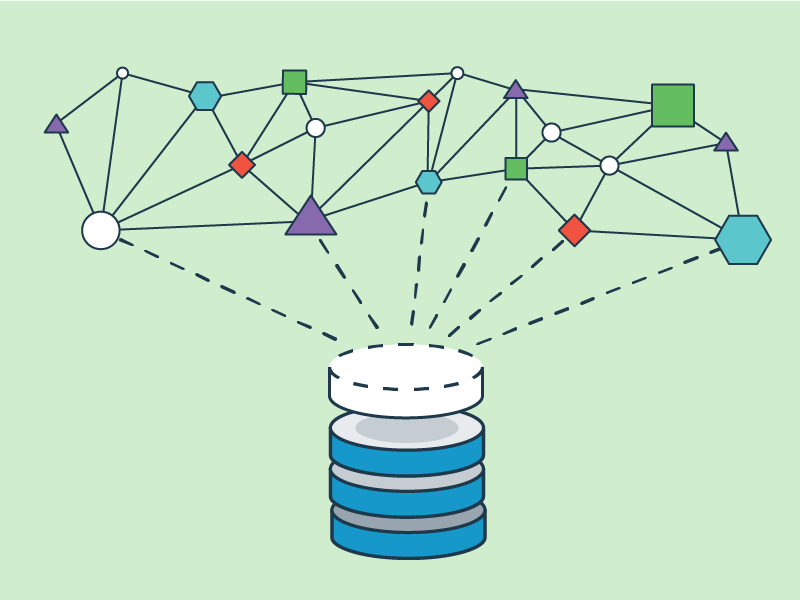
ASSIGNMENT 3 REPORT

*CSCI 5408 Data Management, Warehousing and Analytics*



ANAND BHADANIA

B00837967

# INTRODUCTION

Assignment 3 was given with several tasks which include the following steps:

1. Cluster Setup
2. Data Extraction and Transformation (Twitter and News Article)
3. Data Processing (Spark)

# CLUSTER SETUP

The cluster setup was done following Tutorial3 as shown below [1]:

1. Go to <https://aws.amazon.com> and select a Free Tier. Log into your AWS account using the appropriate credentials.
2. After signing in to your AWS account, click on “Launch a Virtual Machine with EC2”.
3. Select Free Tier only option and click on “Ubuntu Server 18.04 LTS (HVM), SSD Volume Type”.
4. Click on “Next: Configure Instance Details” and then click on “Next: Add Storage”.
5. Add Tags.
6. Provide Key-Value name as per your convenience.
7. Add SSH, MySQL and HTTP type of connections and configure the source IPs as 0.0.0.0/0 and ::/0 for all three types.
8. Create a new key pair and this will give you the private key (\*.pem) to connect to your VM.
9. Download the Key Pair.
10. Your instance is created successfully.

## For PuTTY Key Generator:

1. Download and Install Putty gen.
2. Using PuttyGen, create your .ppk file using your .pem file.
3. Open Putty config and copy your public DNS or IP from the VM instance.
4. Under Connection, go to SSH and click Auth.
5. Attach your (.ppk) file and Click on Open.

# DATA EXTRACTION AND TRANSFORMATION – TWITTER

The extraction of data from Twitter can be done with the help of a Twitter developer account. For that, first you need to have a Twitter account and then you can apply for a developer account.

To get the developer account’s access, you would need to get permission from Twitter. For that, go to Twitter developer account site [2]. Then it will ask you some questions regarding your usage of Twitter data and if it satisfies the terms and conditions, then Twitter will provide developer access right away.

I have used Python scripts to extract tweets from Twitter in which I used Tweepy to obtain existing and live tweets [3]. I have not attached the screenshot of streaming API script in here because the code is larger than one page so was not able to capture the screenshot.

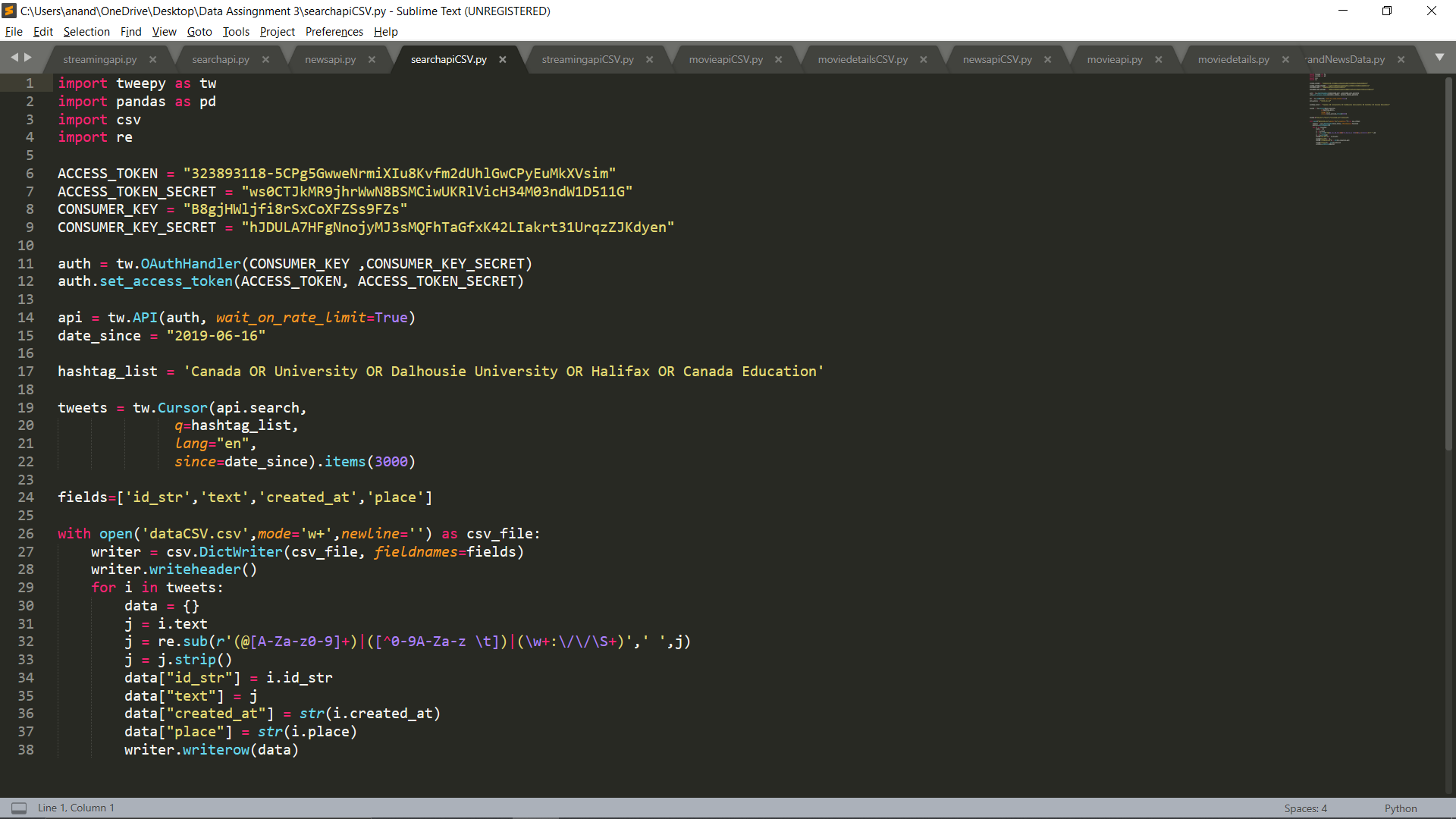


Figure 1 : search API python script | Source : Author

# DATA EXTRACTION AND TRANSFORMATION – NEWSAPI

To extract articles from NewsAPI you need to have a developer account [8]. Visit <https://newsapi.org/>, click on “Get API Key” and create an account. I have used python scripts to extract data from NewsAPI [4]. For performing the task of frequency count, I have saved both the data to same files.

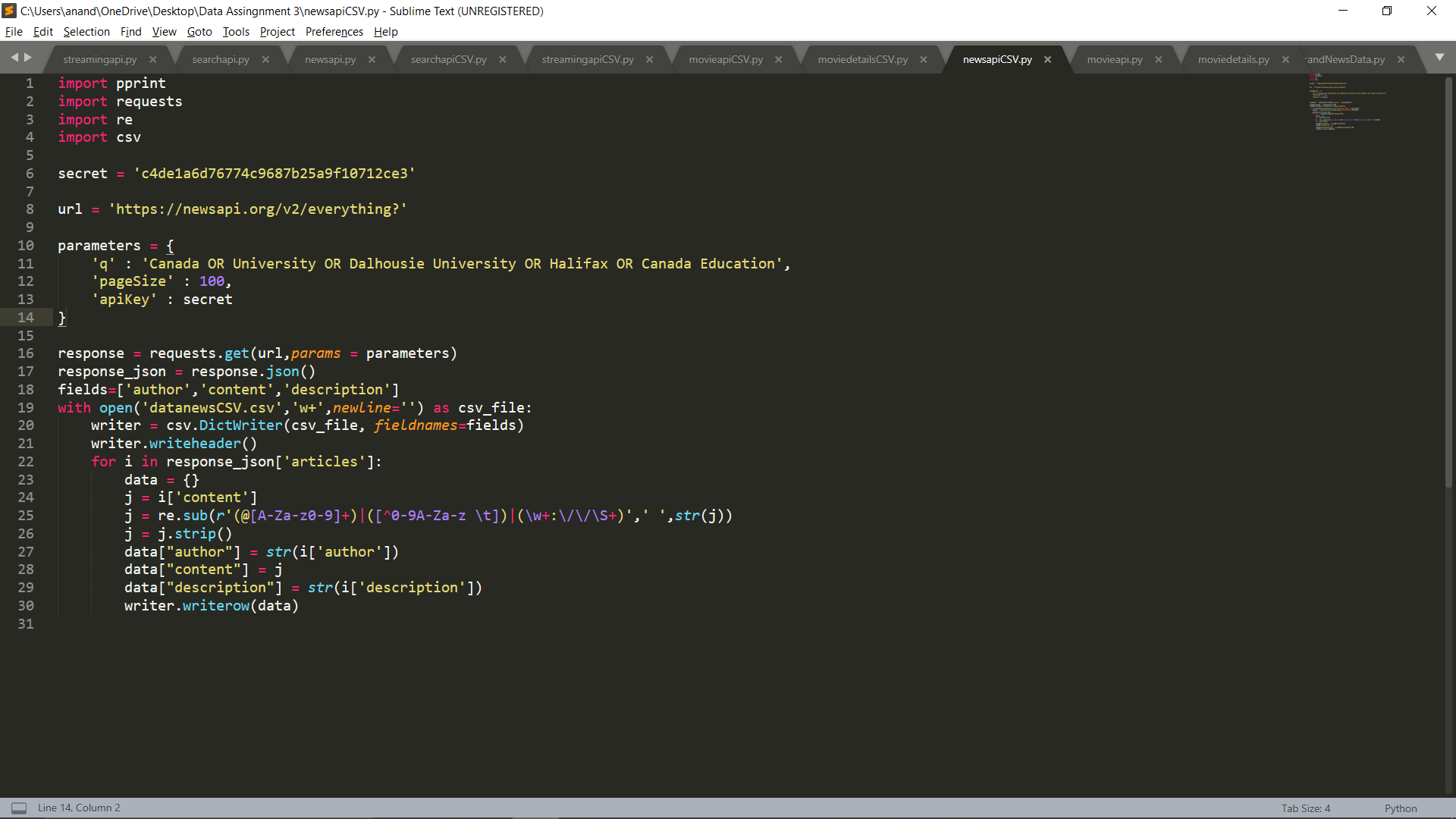


Figure 2 : Python Script to extract data from NewsAPI | Source: Author

# DATA EXTRACTION AND TRANSFORMATION – OMDBAPI

To extract movie data from omdb API, visit <http://www.omdbapi.com/>, click on “API Key” [7]. Select account type as “free” and enter your email address. You will receive your Key on your entered email address. I have used python scripts to extract data from omdb API [9].

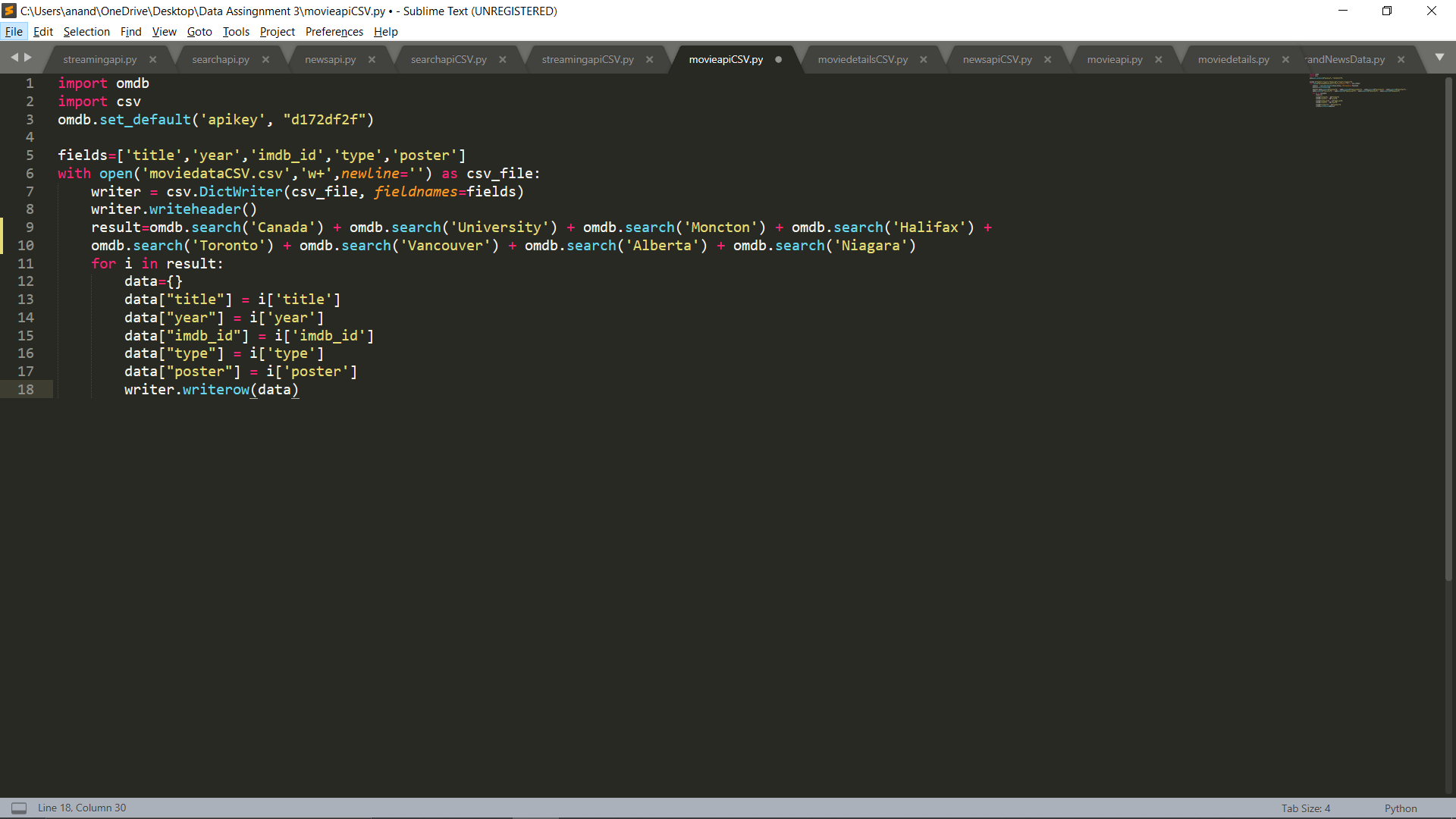


Figure 3 : Python script to extract data from omdb API | Source : Author

# DATA EXTRACTION AND TRANSFORMATION – Additional Movie Data

To extract movie rating, genre and plot from movie data collected from omdb API, I have python script.

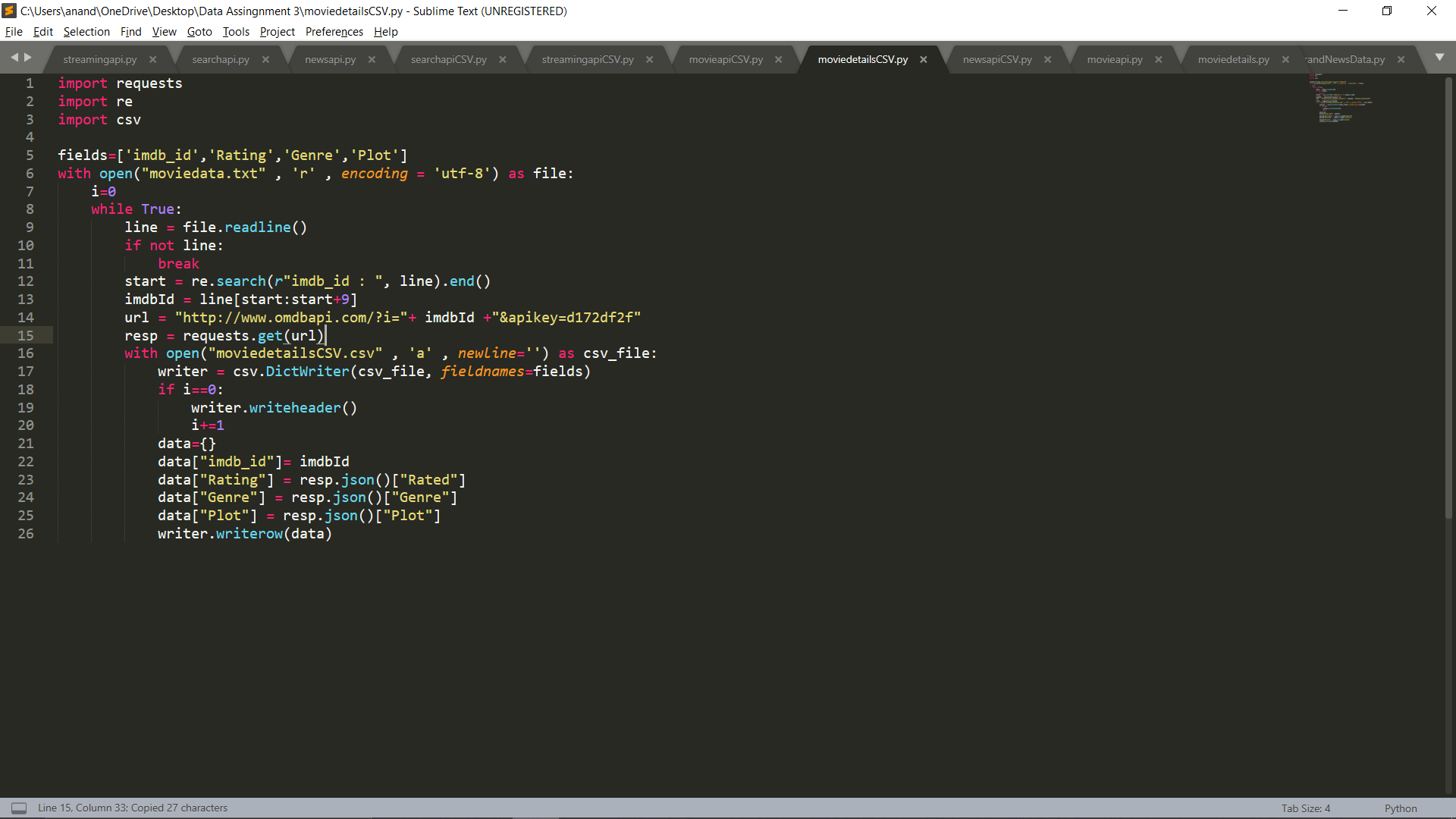


Figure 4 : python script to extract additional movie data | Source : Author

# CLEANING PROCESS

For the cleaning process, I have used regular expressions to remove emoticons, special characters, URLs etc. for both Twitter and NewsAPI data extraction. For omdb API data extraction, I have decided to keep the url link of poster and apart from that, there was no cleaning required.

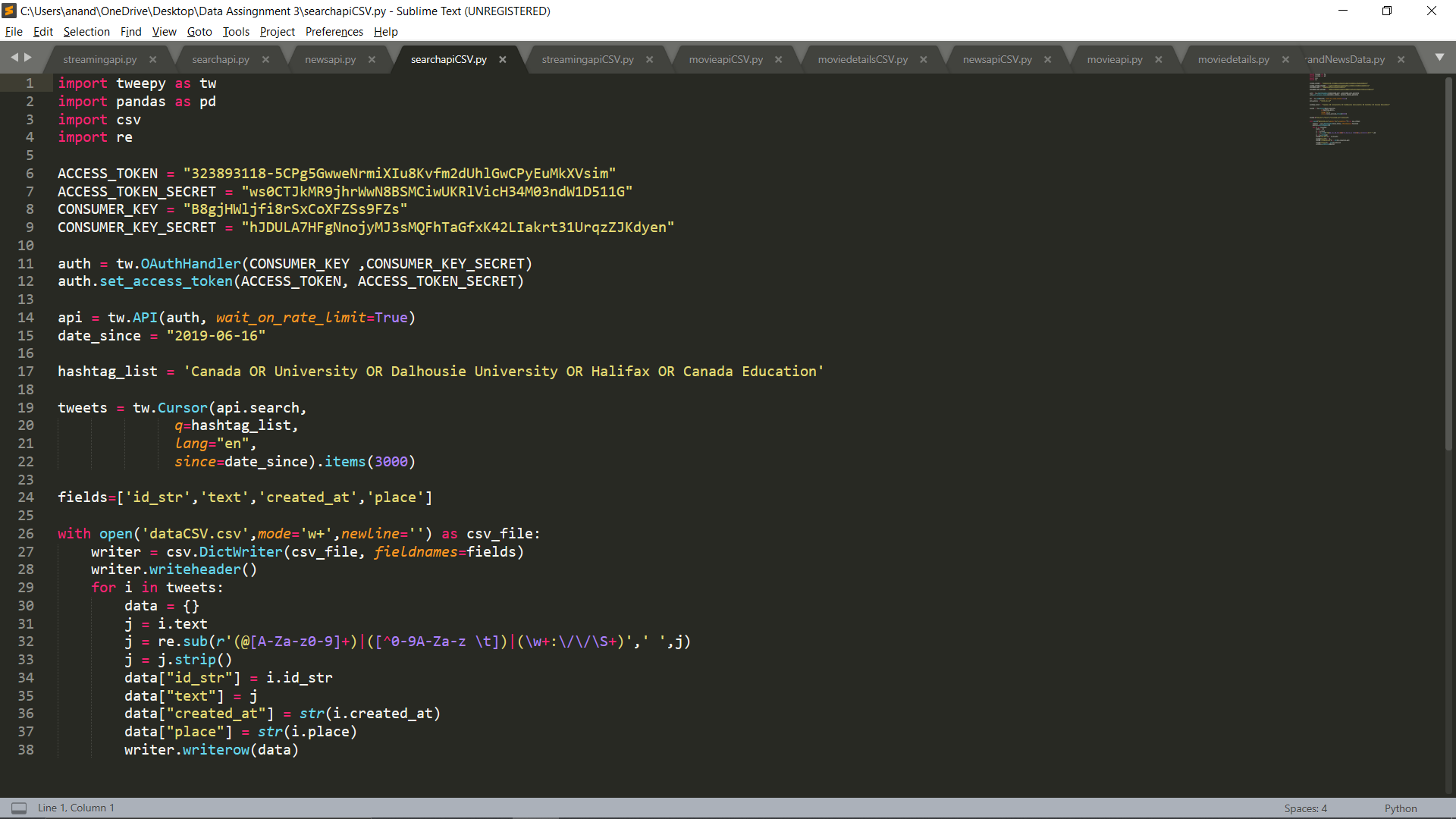


Figure 5 : regular expression to clean tweets and articles from news API | Source : Author

# UPLOADING DATA TO MONGODB

After extracting and cleaning the data, I uploaded the data files to MongoDB.

A screenshot of a computer

Description automatically generated

Figure 6 : Files uploaded on Mongo DB | Source : Author

A screenshot of a cell phone

Description automatically generated

Figure 7 : content of one fo the file on mongo DB | Source : Author

A screenshot of a computer

Description automatically generated

Figure 8 : File successfully uploaded to mongo DB | Source : Author

# DATA PROCESSING (Spark)

I followed Tutorial5 to start Master and Slave node and also to initialize the spark shell [5]. In Spark, I have written commands that will provide frequency count for specified set of words [6].

A screenshot of a computer screen

Description automatically generated

Figure 9 : Cloud Dashboard |Source: AWS

A screenshot of a computer

Description automatically generated

Figure 10: Count Output | Source: Author

CONCLUSION

The conclusion we can derive from this assignment is how to set up Cloud service and create Master-Slave instances and initiate the Spark shell. How to extract tweets from the twitter, articles from NewsAPI and movie data from omdb API. How to transfer file to MondoDB and perform task of frequency count using Map Reduce.

\* *For this assignment, all the scripts and images are available in the folder.*

REFERENCES

1. “Tutorial3 - CSCI5408 - Data Mgmt  Warhsng Analytics (Sec 1) - 2020 Winter,” *Brightspace.com*, 2020. [Online]. Available: https://dal.brightspace.com/d2l/le/content/110354/viewContent/1596093/View. [Accessed: 05-Mar-2020].
2. Developer.twitter.com. (2019). *Apply for access*. [online] Available: <https://developer.twitter.com/en/apply-for-access> [Accessed 05-Mar-2020].
3. J[ayesh Srivastava](https://medium.com/@jayeshsrivastava470?source=post_page-----47dd07f4e8e7----------------------) (2019). **How to Extract Tweets from Twitter in Python. [online] Available:**

<https://medium.com/@jayeshsrivastava470/how-to-extract-tweets-from-twitter-in-python-47dd07f4e8e7> [Accessed 09-Mar-2020].

1. Accessing the NewsAPI in Python. [online] Available: <https://www.datacareer.de/blog/accessing-the-news-api-in-python/> [Accessed 09-Mar-2020].
2. “Tutorial-5 - CSCI5408 - Data Mgmt  Warhsng Analytics (Sec 1) - 2020 Winter,” *Brightspace.com*, 2020. [Online]. Available: https://dal.brightspace.com/d2l/le/content/110354/viewContent/1608781/View. [Accessed: 05-Mar-2020].
3. Michael Galarnyk (2017). Word Count of Pride and Prejudice using PySpark. [online] Available: <https://github.com/mGalarnyk/Python_Tutorials/blob/master/PySpark_Basics/PySpark_Part1_Word_Count_Removing_Punctuation_Pride_Prejudice.ipynb> [Accessed 20-Mar-2020].
4. “OMDb API - The Open Movie Database,” *Omdbapi.com*, 2020. [Online]. Available: http://www.omdbapi.com/apikey.aspx. [Accessed: 12-Mar-2020].
5. News API - A JSON API for live news and blog articles, “News API - A JSON API for live news and blog articles,” *Newsapi.org*, 2020. [Online]. Available: https://newsapi.org/. [Accessed: 09-Mar-2020].
6. “omdb,” *PyPI*, 11-Dec-2018. [Online]. Available: https://pypi.org/project/omdb/. [Accessed: 12-Mar-2020].

‌