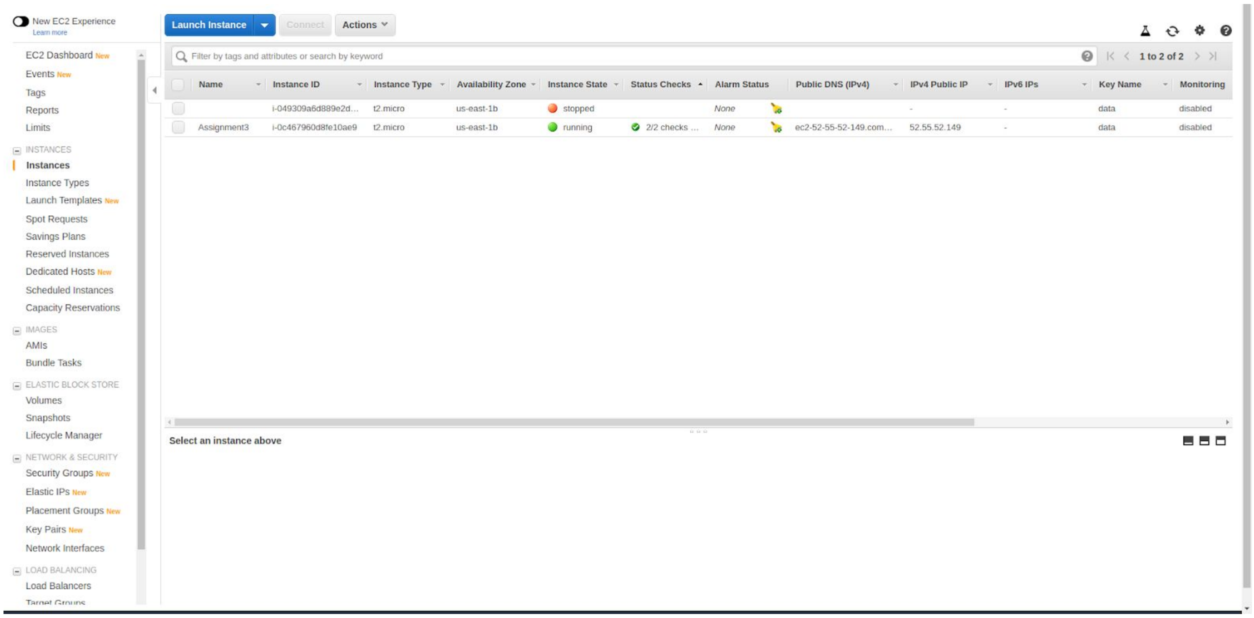
Akash Bharti
B00837970

CSCI 5408 - Data Management, Warehousing and Analytics
Assignment 3

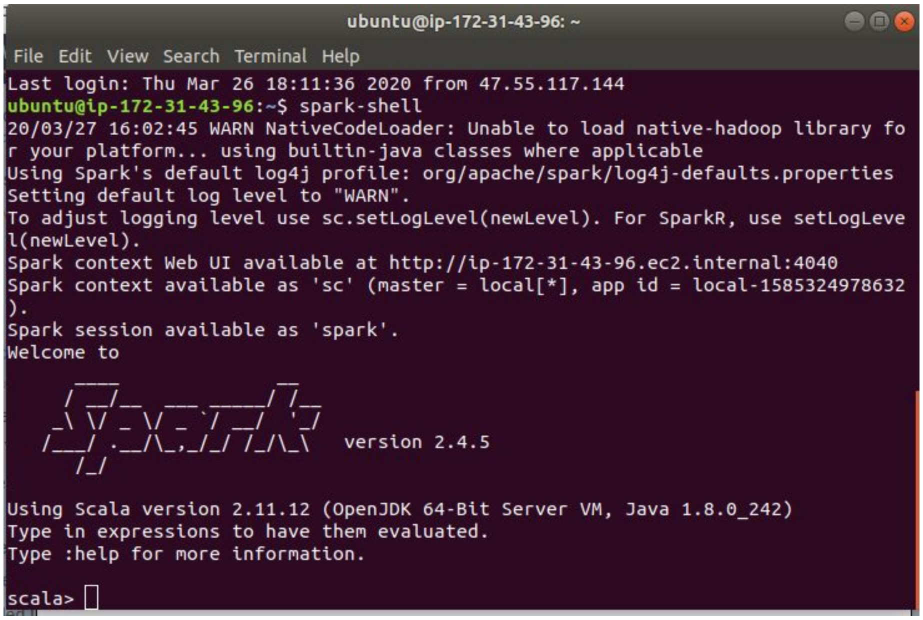
A. Cluster Setup:

I create an account on AWS [1] and then created an instance of EC2. Figure 1 shows the console of EC2 and running instances. I then installed Apache Spark [2] on the EC2 instance. Figure 2 shows the spark-shell running on the EC2. I then installed MongoDB [3] on the instance. Figure 3 shows the mongo shell running on the instance.



Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS (IPv4)	IPv4 Public IP	IPv6 IPs	Key Name	Monitoring
	i-049309a6d889e2d...	t2.micro	us-east-1b	● stopped		None		-	-	data	disabled
Assignment3	i-0c467960d8fe10ae9	t2.micro	us-east-1b	● running	✓ 2/2 checks ...	None	ec2-52-55-52-149.com...	52.55.52.149	-	data	disabled

Figure 1

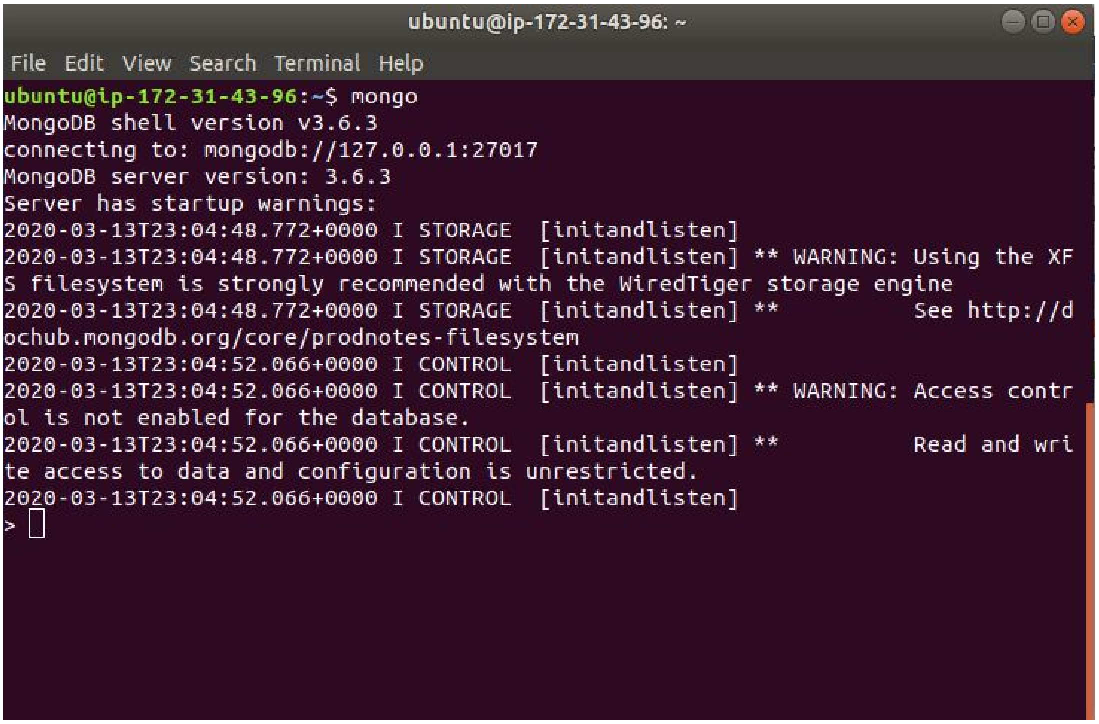

ubuntu@ip-172-31-43-96: ~
File Edit View Search Terminal Help

```
Last login: Thu Mar 26 18:11:36 2020 from 47.55.117.144
ubuntu@ip-172-31-43-96:~$ spark-shell
20/03/27 16:02:45 WARN NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLeve
l(newLevel).
Spark context Web UI available at http://ip-172-31-43-96.ec2.internal:4040
Spark context available as 'sc' (master = local[*], app id = local-1585324978632
).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.4.5
      /_/

Using Scala version 2.11.12 (OpenJDK 64-Bit Server VM, Java 1.8.0_242)
Type in expressions to have them evaluated.
Type :help for more information.

scala> 
```

Figure 2


ubuntu@ip-172-31-43-96: ~
File Edit View Search Terminal Help

```
ubuntu@ip-172-31-43-96:~$ mongo
MongoDB shell version v3.6.3
connecting to: mongodb://127.0.0.1:27017
MongoDB server version: 3.6.3
Server has startup warnings:
2020-03-13T23:04:48.772+0000 I STORAGE  [initandlisten]
2020-03-13T23:04:48.772+0000 I STORAGE  [initandlisten] ** WARNING: Using the XF
S filesystem is strongly recommended with the WiredTiger storage engine
2020-03-13T23:04:48.772+0000 I STORAGE  [initandlisten] **          See http://d
ochub.mongodb.org/core/prodnotes-filesystem
2020-03-13T23:04:52.066+0000 I CONTROL  [initandlisten]
2020-03-13T23:04:52.066+0000 I CONTROL  [initandlisten] ** WARNING: Access contr
ol is not enabled for the database.
2020-03-13T23:04:52.066+0000 I CONTROL  [initandlisten] **          Read and wri
te access to data and configuration is unrestricted.
2020-03-13T23:04:52.066+0000 I CONTROL  [initandlisten]
> 
```

Figure 3

B. (i)Twitter Data Extraction and Transformation:

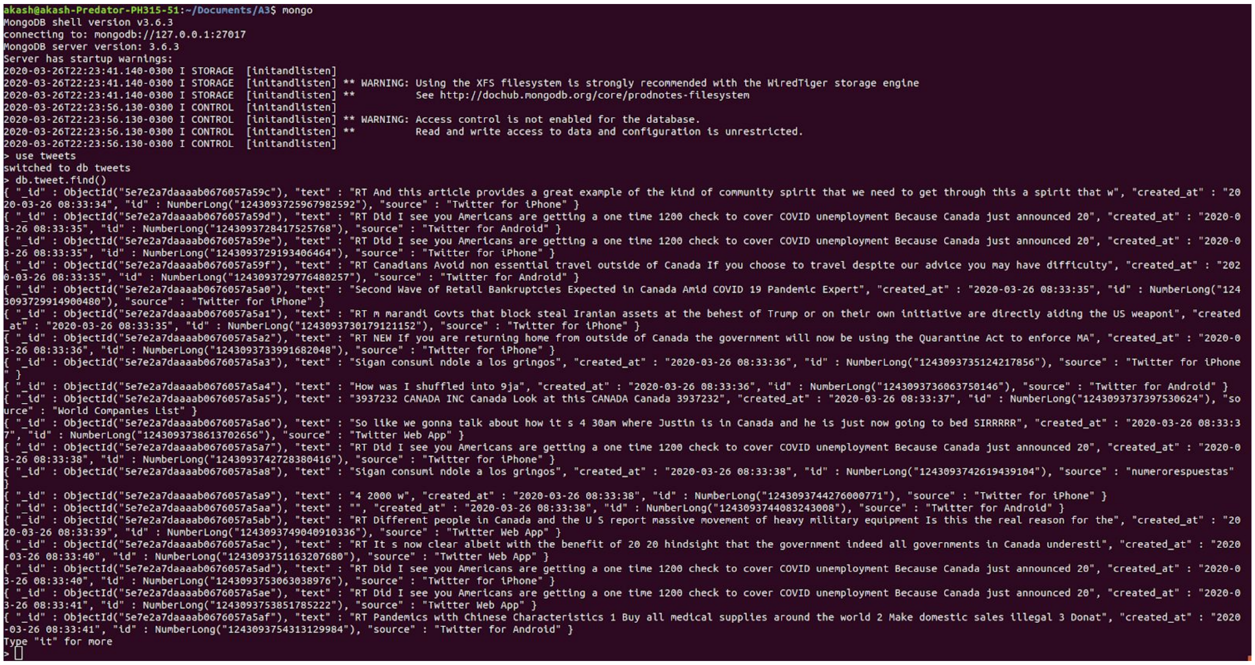
I created the twitter developer account and it was approved in 3 days. I used the tweepy [5] python package to extract the tweets from twitter APIs [4]. I used both the streaming and REST APIs to extract the tweets. I used the following scripts to extract the data:

1. *twitterStreamingApi.py* - This script is used to extract tweets for keywords “Canada” and “University”. The script uses twitter streaming API to extract the tweets. The scripts stores the extracted tweets in file tweets.json.

2. *twitterRestApi.py* - This script is used to extract tweets for keywords “Dalhousie University”, “Halifax” and “Canada Education”. The script uses twitter REST API to extract the tweets. The scripts stores the extracted tweets in file tweets.json.

3. Clean_tweets.py - This script cleans the extracted tweets and then stores them in MongoDB cluster. The script uses regular expressions to clean the tweets. The script also saves the tweets in file *input_for_word_count.txt* which will be later used to process the word counts for different keywords.

Figure 4 shows the tweets stored in MongoDb cluster.

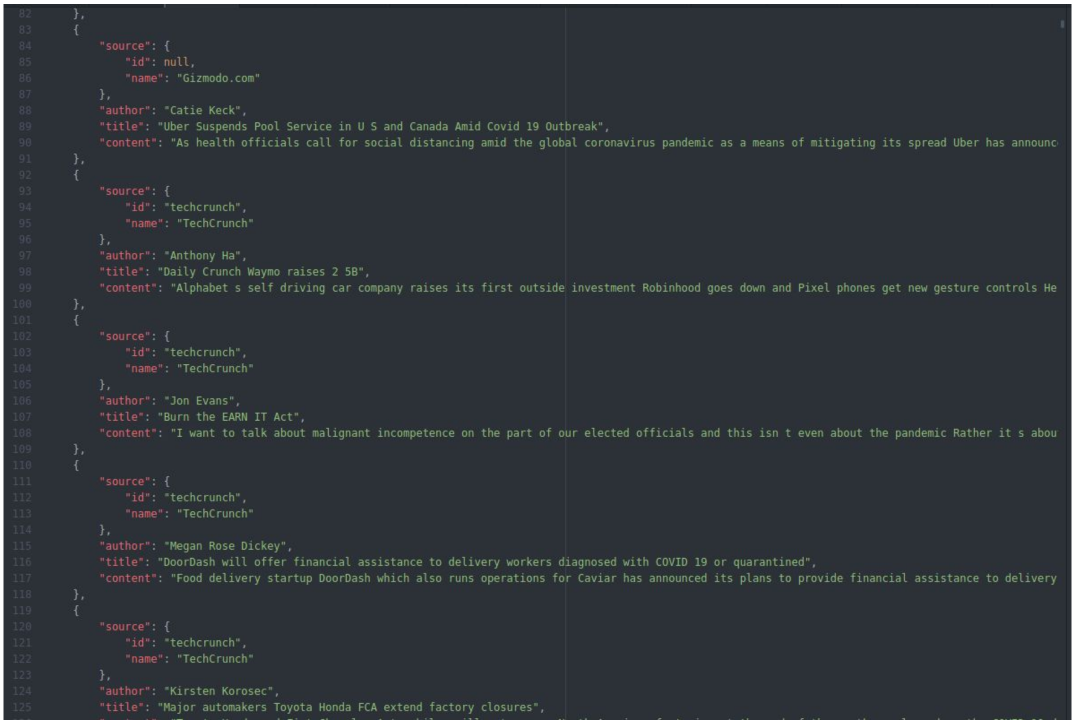


```
akash@akash-Predator-PH315-51:~/Documents/A3$ mongo
MongoDB shell version v3.6.3
connecting to: mongodb://127.0.0.1:27017
MongoDB server version: 3.6.3
Server has startup warnings:
2020-03-26T22:23:41.140-0300 I STORAGE  [initandlisten]
2020-03-26T22:23:41.140-0300 I STORAGE  [initandlisten] ** WARNING: Using the XFS filesystem is strongly recommended with the WiredTiger storage engine
2020-03-26T22:23:41.140-0300 I STORAGE  [initandlisten] **          See http://dochub.mongodb.org/core/prodnotes-filesystem
2020-03-26T22:23:56.130-0300 I CONTROL  [initandlisten]
2020-03-26T22:23:56.130-0300 I CONTROL  [initandlisten] ** WARNING: Access control is not enabled for the database.
2020-03-26T22:23:56.130-0300 I CONTROL  [initandlisten] **          Read and write access to data and configuration is unrestricted.
2020-03-26T22:23:56.130-0300 I CONTROL  [initandlisten]
> use tweets
switched to db tweets
> db.tweet.find()
{ "_id" : ObjectId("5e7e2a7daaaab0676057a59c"), "text" : "RT And this article provides a great example of the kind of community spirit that we need to get through this a spirit that w", "created_at" : "20
20-03-26 08:33:34", "id" : NumberLong("1243093725967982592"), "source" : "Twitter for iPhone" }
{ "_id" : ObjectId("5e7e2a7daaaab0676057a59d"), "text" : "RT Did I see you Americans are getting a one time 1200 check to cover COVID unemployment Because Canada just announced 20", "created_at" : "2020-0
3-26 08:33:35", "id" : NumberLong("1243093728417525768"), "source" : "Twitter for Android" }
{ "_id" : ObjectId("5e7e2a7daaaab0676057a59e"), "text" : "RT Did I see you Americans are getting a one time 1200 check to cover COVID unemployment Because Canada just announced 20", "created_at" : "2020-0
3-26 08:33:35", "id" : NumberLong("1243093729193406464"), "source" : "Twitter for iPhone" }
{ "_id" : ObjectId("5e7e2a7daaaab0676057a59f"), "text" : "RT Canadians Avoid non essential travel outside of Canada If you choose to travel despite our advice you may have difficulty", "created_at" : "202
0-03-26 08:33:35", "id" : NumberLong("1243093729776480257"), "source" : "Twitter for Android" }
{ "_id" : ObjectId("5e7e2a7daaaab0676057a5a0"), "text" : "Second Wave of Retail Bankruptcies Expected in Canada Amid COVID 19 Pandemic Expert", "created_at" : "2020-03-26 08:33:35", "id" : NumberLong("124
3093729914900480"), "source" : "Twitter for iPhone" }
{ "_id" : ObjectId("5e7e2a7daaaab0676057a5a1"), "text" : "RT m marandi Govts that block steal Iranian assets at the behest of Trump or on their own initiative are directly aiding the US weaponi", "created
_at" : "2020-03-26 08:33:35", "id" : NumberLong("1243093730179121152"), "source" : "Twitter for iPhone" }
{ "_id" : ObjectId("5e7e2a7daaaab0676057a5a2"), "text" : "RT NEW If you are returning home from outside of Canada the government will now be using the Quarantine Act to enforce MA", "created_at" : "2020-0
3-26 08:33:36", "id" : NumberLong("1243093733991682048"), "source" : "Twitter for iPhone" }
{ "_id" : ObjectId("5e7e2a7daaaab0676057a5a3"), "text" : "Sigan consumi ndole a los gringos", "created_at" : "2020-03-26 08:33:36", "id" : NumberLong("1243093735124217856"), "source" : "Twitter for iPhone
" }
{ "_id" : ObjectId("5e7e2a7daaaab0676057a5a4"), "text" : "How was I shuffled into 9ja", "created_at" : "2020-03-26 08:33:36", "id" : NumberLong("1243093736063750146"), "source" : "Twitter for Android" }
{ "_id" : ObjectId("5e7e2a7daaaab0676057a5a5"), "text" : "3937232 CANADA INC Canada Look at this CANADA Canada 3937232", "created_at" : "2020-03-26 08:33:37", "id" : NumberLong("1243093737397530624"), "so
urce" : "World Companies List" }
{ "_id" : ObjectId("5e7e2a7daaaab0676057a5a6"), "text" : "So like we gonna talk about how it s 4 30am where Justin is in Canada and he is just now going to bed SIRRRRR", "created_at" : "2020-03-26 08:33:3
7", "id" : NumberLong("1243093738613702656"), "source" : "Twitter Web App" }
{ "_id" : ObjectId("5e7e2a7daaaab0676057a5a7"), "text" : "RT Did I see you Americans are getting a one time 1200 check to cover COVID unemployment Because Canada just announced 20", "created_at" : "2020-0
3-26 08:33:38", "id" : NumberLong("1243093742728380416"), "source" : "Twitter for iPhone" }
{ "_id" : ObjectId("5e7e2a7daaaab0676057a5a8"), "text" : "Sigan consumi ndole a los gringos", "created_at" : "2020-03-26 08:33:38", "id" : NumberLong("1243093742619439104"), "source" : "numerorespuestas"
}
{ "_id" : ObjectId("5e7e2a7daaaab0676057a5a9"), "text" : "4 2000 w", "created_at" : "2020-03-26 08:33:38", "id" : NumberLong("1243093744276000771"), "source" : "Twitter for iPhone" }
{ "_id" : ObjectId("5e7e2a7daaaab0676057a5aa"), "text" : "", "created_at" : "2020-03-26 08:33:38", "id" : NumberLong("1243093744083243008"), "source" : "Twitter for Android" }
{ "_id" : ObjectId("5e7e2a7daaaab0676057a5ab"), "text" : "RT Different people in Canada and the U S report massive movement of heavy military equipment Is this the real reason for the", "created_at" : "20
20-03-26 08:33:39", "id" : NumberLong("1243093749040910336"), "source" : "Twitter Web App" }
{ "_id" : ObjectId("5e7e2a7daaaab0676057a5ac"), "text" : "RT It s now clear albeit with the benefit of 20 20 hindsight that the government indeed all governments in Canada underesti", "created_at" : "2020
-03-26 08:33:40", "id" : NumberLong("1243093751163207680"), "source" : "Twitter Web App" }
{ "_id" : ObjectId("5e7e2a7daaaab0676057a5ad"), "text" : "RT Did I see you Americans are getting a one time 1200 check to cover COVID unemployment Because Canada just announced 20", "created_at" : "2020-0
3-26 08:33:40", "id" : NumberLong("1243093753063038976"), "source" : "Twitter for iPhone" }
{ "_id" : ObjectId("5e7e2a7daaaab0676057a5ae"), "text" : "RT Did I see you Americans are getting a one time 1200 check to cover COVID unemployment Because Canada just announced 20", "created_at" : "2020-0
3-26 08:33:41", "id" : NumberLong("1243093753851785222"), "source" : "Twitter Web App" }
{ "_id" : ObjectId("5e7e2a7daaaab0676057a5af"), "text" : "RT Pandemics with Chinese Characteristics 1 Buy all medical supplies around the world 2 Make domestic sales illegal 3 Donat", "created_at" : "2020
-03-26 08:33:41", "id" : NumberLong("1243093754313129984"), "source" : "Twitter for Android" }
Type "it" for more
> 
```

Figure 4

(ii) News Articles Data Extraction and Transformation:

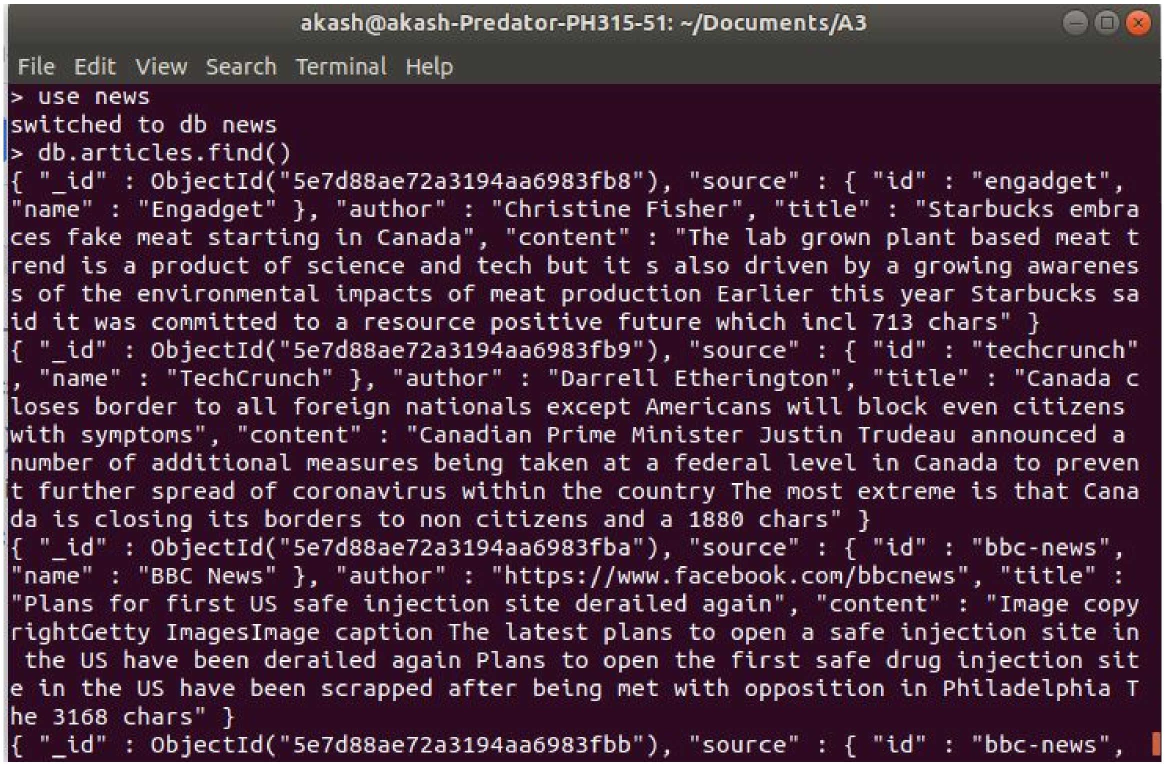
The *news_api.py* script extracts the news articles from News API [6] for all the keywords required, cleans them, stores them on the MongoDB cluster and also saves them in news_articles.json file. Figure 5 shows the contents of the news_articles.json file.



```
    },
    {
        "source": {
            "id": null,
            "name": "Gizmodo.com"
        },
        "author": "Catie Keck",
        "title": "Uber Suspends Pool Service in U S and Canada Amid Covid 19 Outbreak",
        "content": "As health officials call for social distancing amid the global coronavirus pandemic as a means of mitigating its spread Uber has announc
    },
    {
        "source": {
            "id": "techcrunch",
            "name": "TechCrunch"
        },
        "author": "Anthony Ha",
        "title": "Daily Crunch Waymo raises 2 5B",
        "content": "Alphabet s self driving car company raises its first outside investment Robinhood goes down and Pixel phones get new gesture controls He
    },
    {
        "source": {
            "id": "techcrunch",
            "name": "TechCrunch"
        },
        "author": "Jon Evans",
        "title": "Burn the EARN IT Act",
        "content": "I want to talk about malignant incompetence on the part of our elected officials and this isn t even about the pandemic Rather it s abou
    },
    {
        "source": {
            "id": "techcrunch",
            "name": "TechCrunch"
        },
        "author": "Megan Rose Dickey",
        "title": "DoorDash will offer financial assistance to delivery workers diagnosed with COVID 19 or quarantined",
        "content": "Food delivery startup DoorDash which also runs operations for Caviar has announced its plans to provide financial assistance to delivery
    },
    {
        "source": {
            "id": "techcrunch",
            "name": "TechCrunch"
        },
        "author": "Kirsten Korosec",
        "title": "Major automakers Toyota Honda FCA extend factory closures",
```

Figure 5

Figure 6 shows the news articles stored in MongoDb cluster.

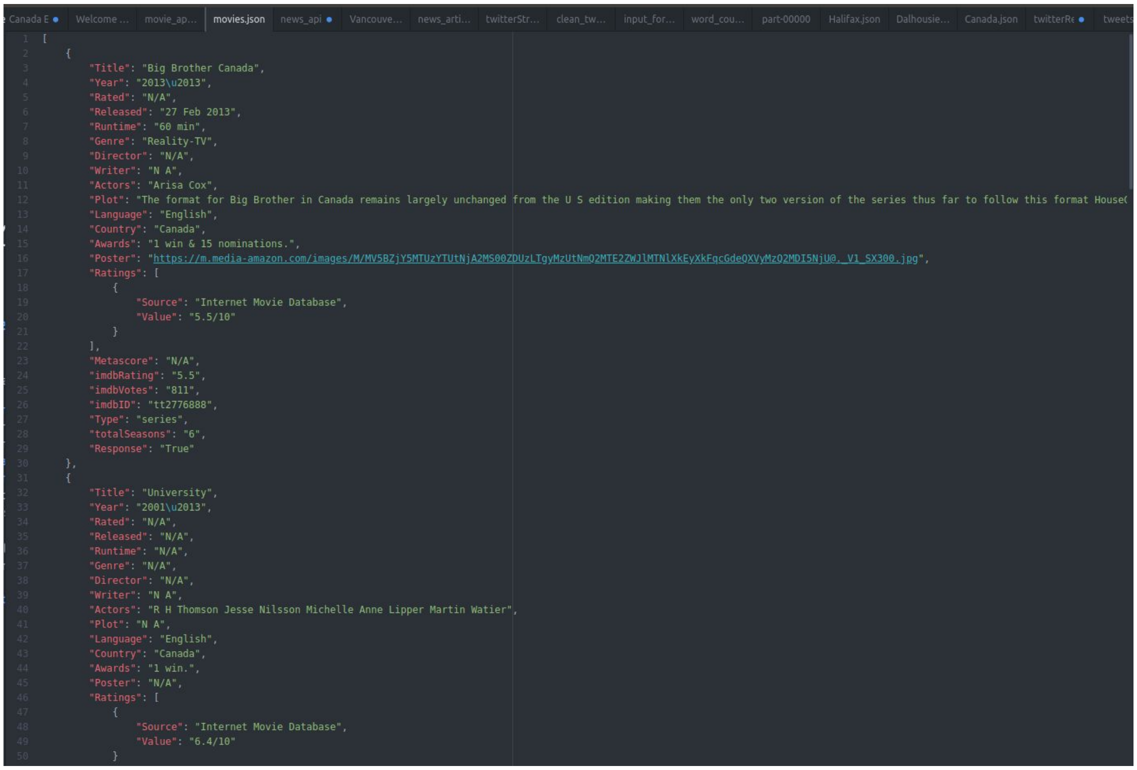

akash@akash-Predator-PH315-51: ~/Documents/A3
File Edit View Search Terminal Help

```
> use news
switched to db news
> db.articles.find()
{ "_id" : ObjectId("5e7d88ae72a3194aa6983fb8"), "source" : { "id" : "engadget",
"name" : "Engadget" }, "author" : "Christine Fisher", "title" : "Starbucks embra
ces fake meat starting in Canada", "content" : "The lab grown plant based meat t
rend is a product of science and tech but it s also driven by a growing awarenes
s of the environmental impacts of meat production Earlier this year Starbucks sa
id it was committed to a resource positive future which incl 713 chars" }
{ "_id" : ObjectId("5e7d88ae72a3194aa6983fb9"), "source" : { "id" : "techcrunch"
, "name" : "TechCrunch" }, "author" : "Darrell Etherington", "title" : "Canada c
loses border to all foreign nationals except Americans will block even citizens
with symptoms", "content" : "Canadian Prime Minister Justin Trudeau announced a
number of additional measures being taken at a federal level in Canada to preven
t further spread of coronavirus within the country The most extreme is that Cana
da is closing its borders to non citizens and a 1880 chars" }
{ "_id" : ObjectId("5e7d88ae72a3194aa6983fba"), "source" : { "id" : "bbc-news",
"name" : "BBC News" }, "author" : "https://www.facebook.com/bbcnews", "title" :
"Plans for first US safe injection site derailed again", "content" : "Image copy
rightGetty ImagesImage caption The latest plans to open a safe injection site in
 the US have been derailed again Plans to open the first safe drug injection sit
e in the US have been scrapped after being met with opposition in Philadelphia T
he 3168 chars" }
{ "_id" : ObjectId("5e7d88ae72a3194aa6983fbb"), "source" : { "id" : "bbc-news",
```

Figure 6

(iii) Movie Data Extraction and Transformation:

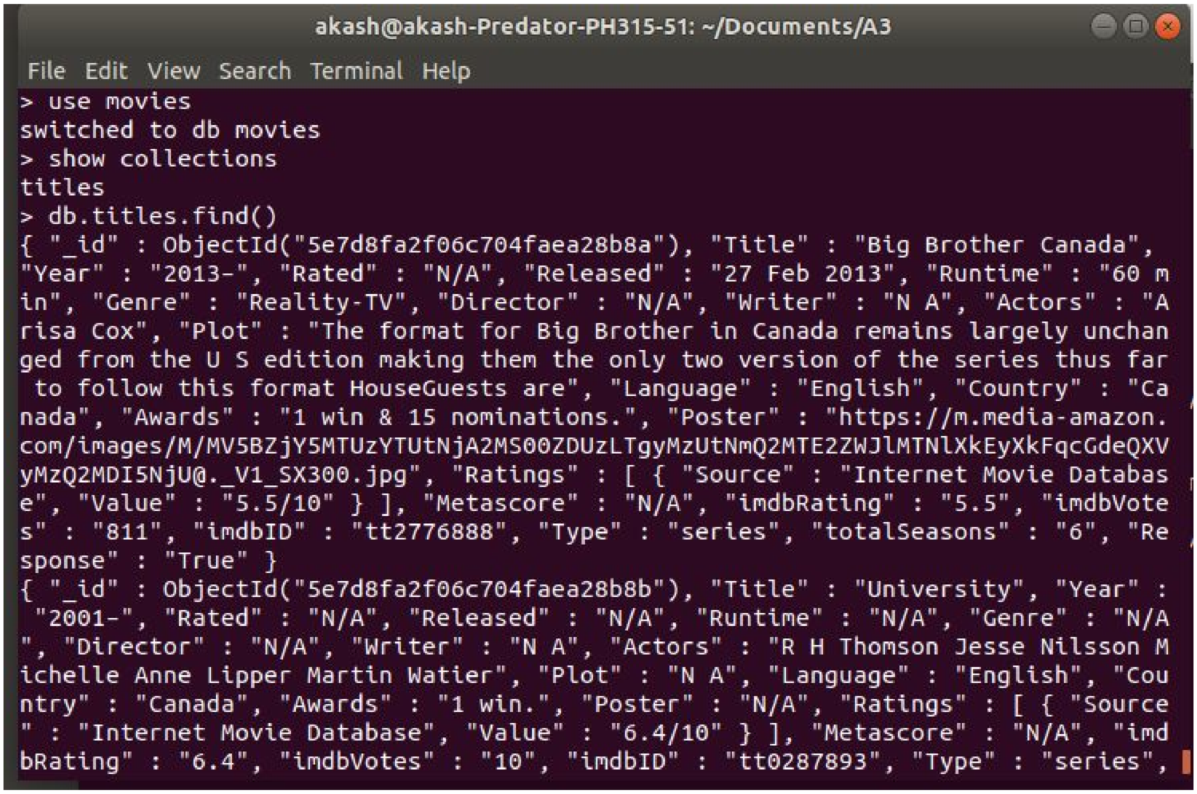
The *movies_api.py* script extracts the movies data from OMDb API [7] for all the keywords required, cleans them, stores them on the MongoDB cluster and also saves them in movies.json file. Figure 7 shows the contents of the movies.json file.



```
[
    {
        "Title": "Big Brother Canada",
        "Year": "2013\u2013",
        "Rated": "N/A",
        "Released": "27 Feb 2013",
        "Runtime": "60 min",
        "Genre": "Reality-TV",
        "Director": "N/A",
        "Writer": "N A",
        "Actors": "Arisa Cox",
        "Plot": "The format for Big Brother in Canada remains largely unchanged from the U S edition making them the only two version of the series thus far to follow this format House(
        "Language": "English",
        "Country": "Canada",
        "Awards": "1 win & 15 nominations.",
        "Poster": "https://m.media-amazon.com/images/M/MV5BZjY5MTUzYTUtNjA2MS00ZDUzLTgyMzUtNmQ2MTE2ZWJlMTNlXkEyXkFqcGdeQXVyMzQ2MDI5NjU@._V1_SX300.jpg",
        "Ratings": [
            {
                "Source": "Internet Movie Database",
                "Value": "5.5/10"
            }
        ],
        "Metascore": "N/A",
        "imdbRating": "5.5",
        "imdbVotes": "811",
        "imdbID": "tt2776888",
        "Type": "series",
        "totalSeasons": "6",
        "Response": "True"
    },
    {
        "Title": "University",
        "Year": "2001\u2013",
        "Rated": "N/A",
        "Released": "N/A",
        "Runtime": "N/A",
        "Genre": "N/A",
        "Director": "N/A",
        "Writer": "N A",
        "Actors": "R H Thomson Jesse Nilsson Michelle Anne Lipper Martin Watier",
        "Plot": "N A",
        "Language": "English",
        "Country": "Canada",
        "Awards": "1 win.",
        "Poster": "N/A",
        "Ratings": [
            {
                "Source": "Internet Movie Database",
                "Value": "6.4/10"
            }
```

Figure 7

Figure 8 shows the movies data stores in MongoDb cluster.

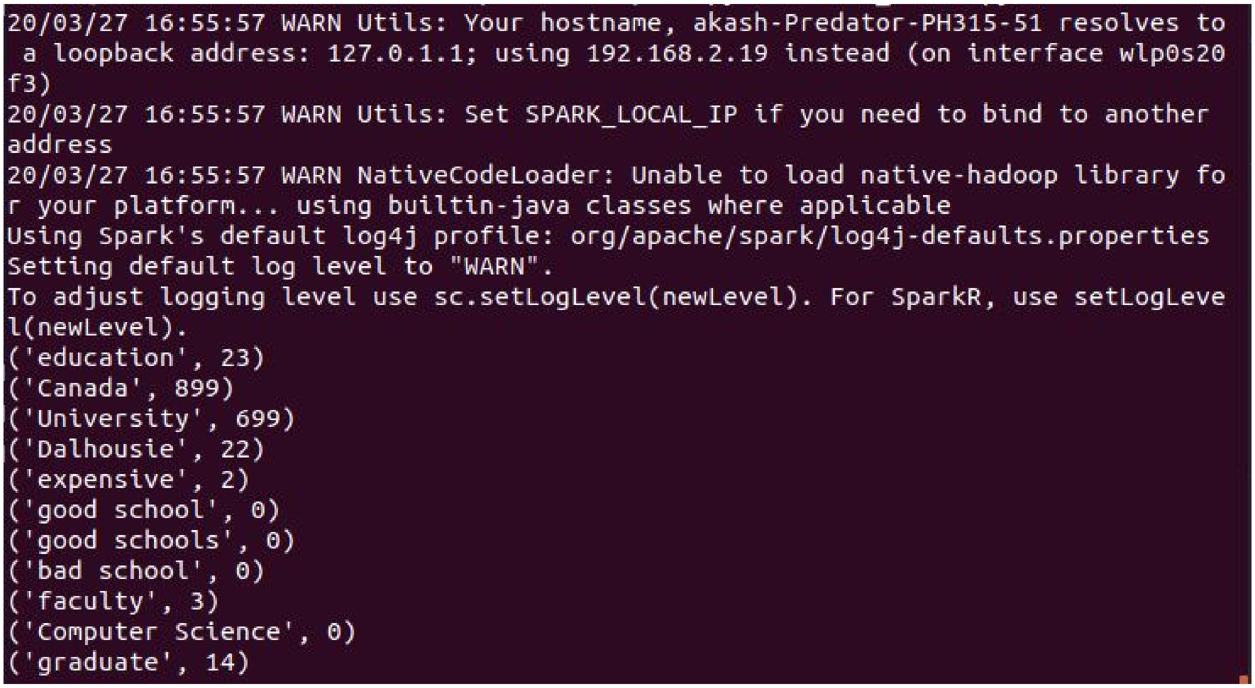

akash@akash-Predator-PH315-51: ~/Documents/A3
File Edit View Search Terminal Help

```
> use movies
switched to db movies
> show collections
titles
> db.titles.find()
{ "_id" : ObjectId("5e7d8fa2f06c704faea28b8a"), "Title" : "Big Brother Canada",
"Year" : "2013-", "Rated" : "N/A", "Released" : "27 Feb 2013", "Runtime" : "60 m
in", "Genre" : "Reality-TV", "Director" : "N/A", "Writer" : "N A", "Actors" : "A
risa Cox", "Plot" : "The format for Big Brother in Canada remains largely unchan
ged from the U S edition making them the only two version of the series thus far
 to follow this format HouseGuests are", "Language" : "English", "Country" : "Ca
nada", "Awards" : "1 win & 15 nominations.", "Poster" : "https://m.media-amazon.
com/images/M/MV5BZjY5MTUzYTUtNjA2MS00ZDUzLTgyMzUtNmQ2MTE2ZWJlMTNlXkEyXkFqcGdeQXV
yMzQ2MDI5NjU@._V1_SX300.jpg", "Ratings" : [ { "Source" : "Internet Movie Databas
e", "Value" : "5.5/10" } ], "Metascore" : "N/A", "imdbRating" : "5.5", "imdbVote
s" : "811", "imdbID" : "tt2776888", "Type" : "series", "totalSeasons" : "6", "Re
sponse" : "True" }
{ "_id" : ObjectId("5e7d8fa2f06c704faea28b8b"), "Title" : "University", "Year" :
 "2001-", "Rated" : "N/A", "Released" : "N/A", "Runtime" : "N/A", "Genre" : "N/A
", "Director" : "N/A", "Writer" : "N A", "Actors" : "R H Thomson Jesse Nilsson M
ichelle Anne Lipper Martin Watier", "Plot" : "N A", "Language" : "English", "Cou
ntry" : "Canada", "Awards" : "1 win.", "Poster" : "N/A", "Ratings" : [ { "Source
" : "Internet Movie Database", "Value" : "6.4/10" } ], "Metascore" : "N/A", "imd
bRating" : "6.4", "imdbVotes" : "10", "imdbID" : "tt0287893", "Type" : "series",
```

Figure 8

C. Data Processing (Spark):

The script *word_count.py* reads cleaned tweets and news articles from file *input_for_word_count.txt* and uses pyspark python package [8] to calculate the word count of the provided keywords. The script initializes a spark context and then uses map-reduce to print the word count of the provided keyword. Figure 9 shows the output of the script.

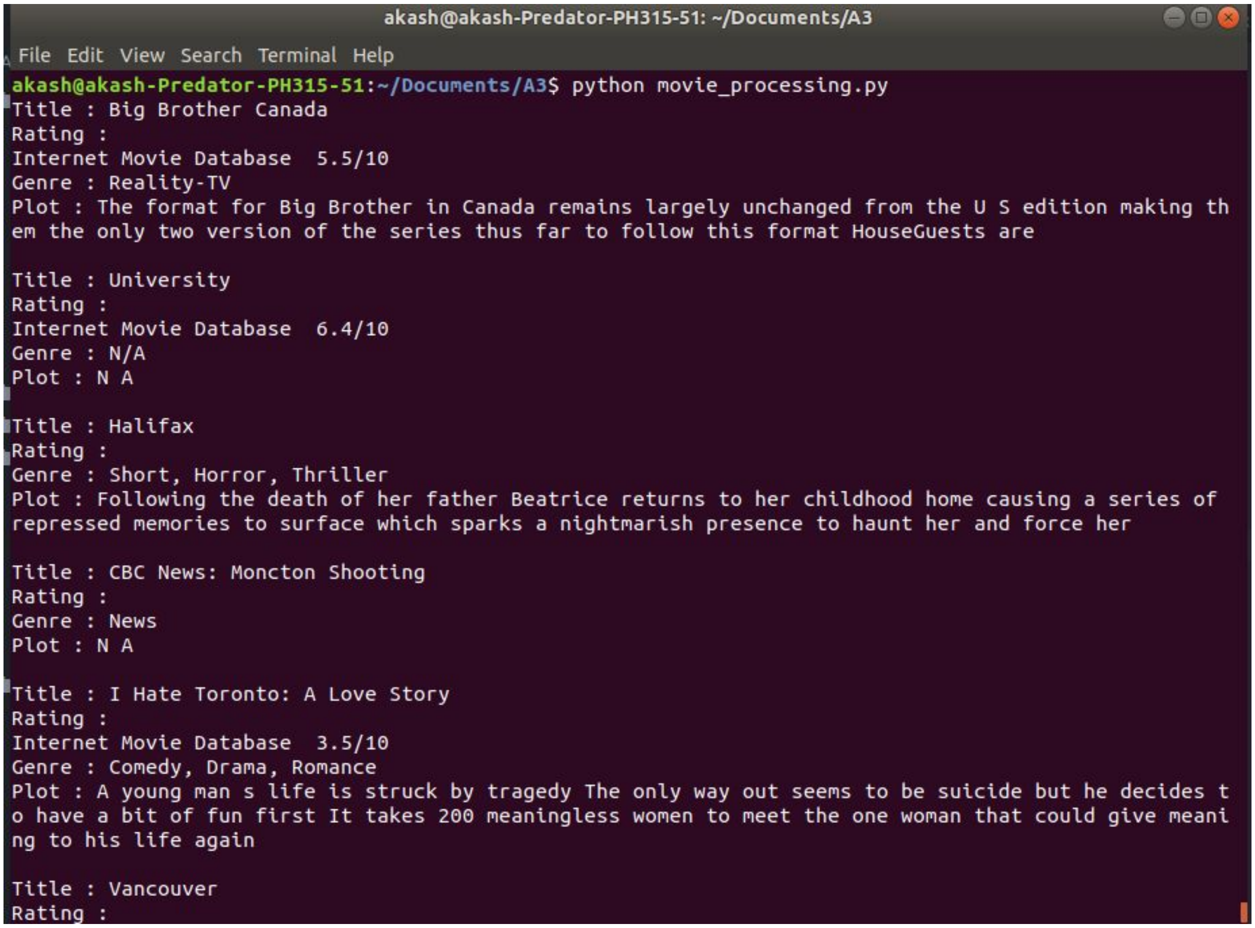


```
20/03/27 16:55:57 WARN Utils: Your hostname, akash-Predator-PH315-51 resolves to
 a loopback address: 127.0.1.1; using 192.168.2.19 instead (on interface wlp0s20
f3)
20/03/27 16:55:57 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another
address
20/03/27 16:55:57 WARN NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLeve
l(newLevel).
('education', 23)
('Canada', 899)
('University', 699)
('Dalhousie', 22)
('expensive', 2)
('good school', 0)
('good schools', 0)
('bad school', 0)
('faculty', 3)
('Computer Science', 0)
('graduate', 14)
```

Figure 9

Extracting movie Information:

The script *movies_processing.py* fetch the movies data from mongoDb and then displays the movie title, rating, genre and plot on the console. Figure 10 shows the output of the script.


akash@akash-Predator-PH315-51: ~/Documents/A3
File Edit View Search Terminal Help

```
akash@akash-Predator-PH315-51:~/Documents/A3$ python movie_processing.py
Title : Big Brother Canada
Rating :
Internet Movie Database  5.5/10
Genre : Reality-TV
Plot : The format for Big Brother in Canada remains largely unchanged from the U S edition making th
em the only two version of the series thus far to follow this format HouseGuests are

Title : University
Rating :
Internet Movie Database  6.4/10
Genre : N/A
Plot : N A

Title : Halifax
Rating :
Genre : Short, Horror, Thriller
Plot : Following the death of her father Beatrice returns to her childhood home causing a series of
repressed memories to surface which sparks a nightmarish presence to haunt her and force her

Title : CBC News: Moncton Shooting
Rating :
Genre : News
Plot : N A

Title : I Hate Toronto: A Love Story
Rating :
Internet Movie Database  3.5/10
Genre : Comedy, Drama, Romance
Plot : A young man s life is struck by tragedy The only way out seems to be suicide but he decides t
o have a bit of fun first It takes 200 meaningless women to meet the one woman that could give meani
ng to his life again

Title : Vancouver
Rating :
```

Figure 10

References:

[1] “What Is Amazon EC2? - Amazon Elastic Compute Cloud,” *Amazon.com*, 2020. [Online]. Available: https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/concepts.html. [Accessed: 27-Mar-2020].

[2] “Overview - Spark 2.4.5 Documentation,” *Apache.org*, 2020. [Online]. Available: https://spark.apache.org/docs/latest/. [Accessed: 27-Mar-2020].

[3] “MongoDB Documentation,” *Mongodb.com*, 2020. [Online]. Available: https://docs.mongodb.com/. [Accessed: 27-Mar-2020].

[4] “Docs,” *Twitter.com*, 2020. [Online]. Available: https://developer.twitter.com/en/docs. [Accessed: 27-Mar-2020].

[5] “Tweepy Documentation — tweepy 3.8.0 documentation,” *Tweepy.org*, 2020. [Online]. Available: http://docs.tweepy.org/en/latest/. [Accessed: 27-Mar-2020].

[6] Documentation - News API, “Documentation - News API,” *Newsapi.org*, 2020. [Online]. Available: https://newsapi.org/docs. [Accessed: 27-Mar-2020].

[7] “OMDb API - The Open Movie Database,” *Omdbapi.com*, 2020. [Online]. Available: http://www.omdbapi.com/. [Accessed: 27-Mar-2020].

[8] “Welcome to Spark Python API Docs! — PySpark 2.4.5 documentation,” *Apache.org*, 2020. [Online]. Available: https://spark.apache.org/docs/latest/api/python/index.html. [Accessed: 27-Mar-2020].