

Big Data – Case Study

India Demonetization 2016: Tweet Sentiment Analysis

Objective

The objective of this case study is to analyze sentiments from twitter of a highly spoken-of event that took place on 8th November 2016 in India. To combat the black money problem in the country, the government decided to demonetize two bills of the highest denominations. We will analyze on how the users reacted to the event using a dictionary and further on, we will visualize the results obtained.

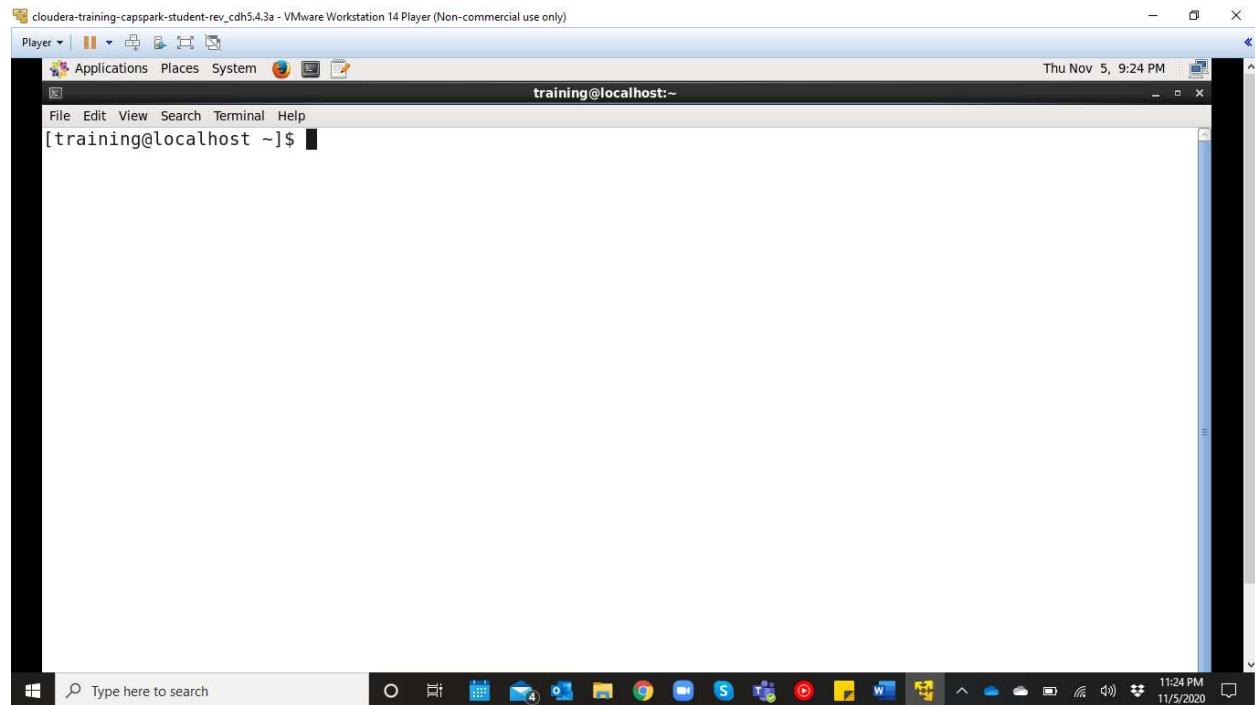
Table of Contents

Objective	1
Part 1: Setup	2
Part-2: Sentiment Analysis	6
Part-3: Visualization of Sentiments	15
Conclusion	22

Part 1: Setup

In this section we will set up the files and data that is required to progress throughout this case study.

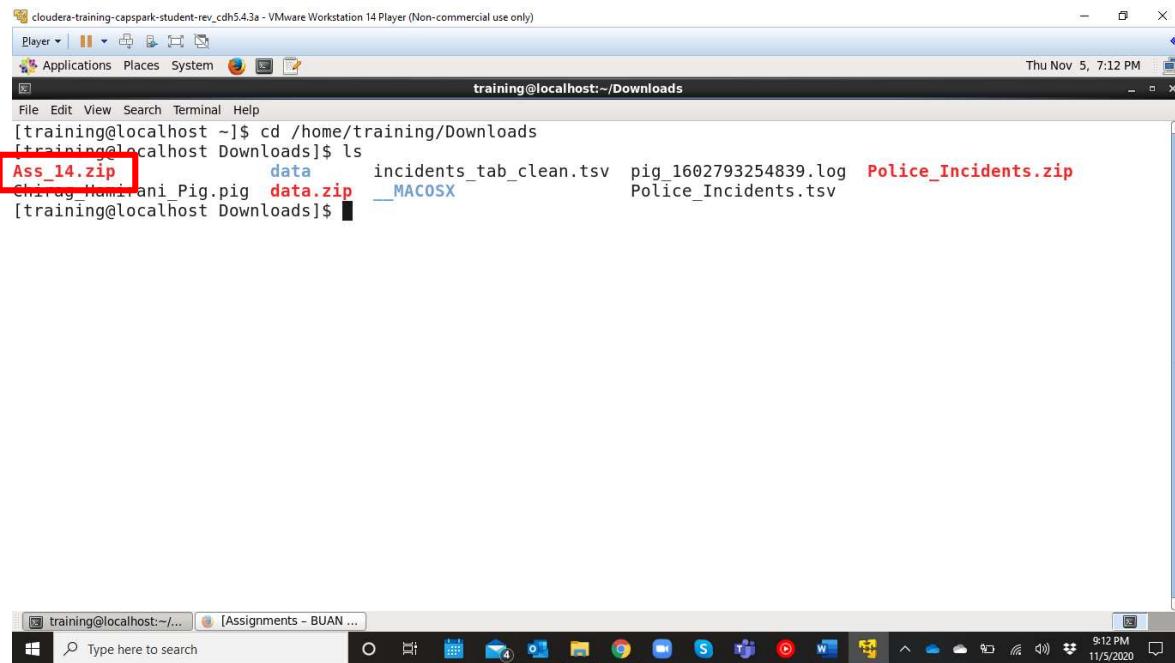
Step-1: Open the terminal application on the Cloudera VM desktop.



Step-2: Download the 'CS01_Dat_Dict_Files.zip' file into the downloads folder **IN** the Cloudera VM using Mozilla Firefox.

Step-3: Execute the following command:

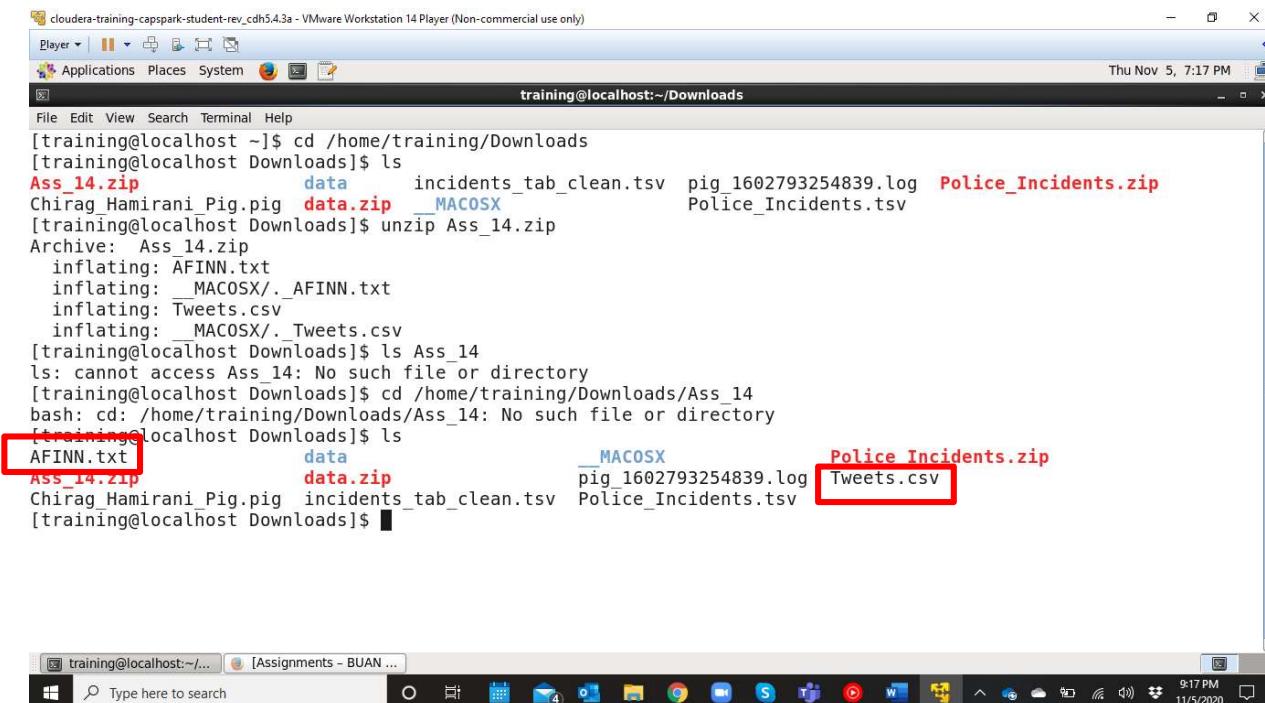
```
cd/home/training/Downloads  
ls
```



```
[training@localhost ~]$ cd /home/training/Downloads
[training@localhost Downloads]$ ls
Ass_14.zip          data      incidents_tab_clean.tsv  pig_1602793254839.log  Police_Incidents.zip
Chirag_Hamirani_Pig.pig  data.zip   __MACOSX             Police_Incidents.tsv
[training@localhost Downloads]$
```

Step-4: Execute the following commands at the shell prompt:

```
unzip CS01_Dat_Dict_Files.zip
ls CS01_Dat_Dict_Files
```

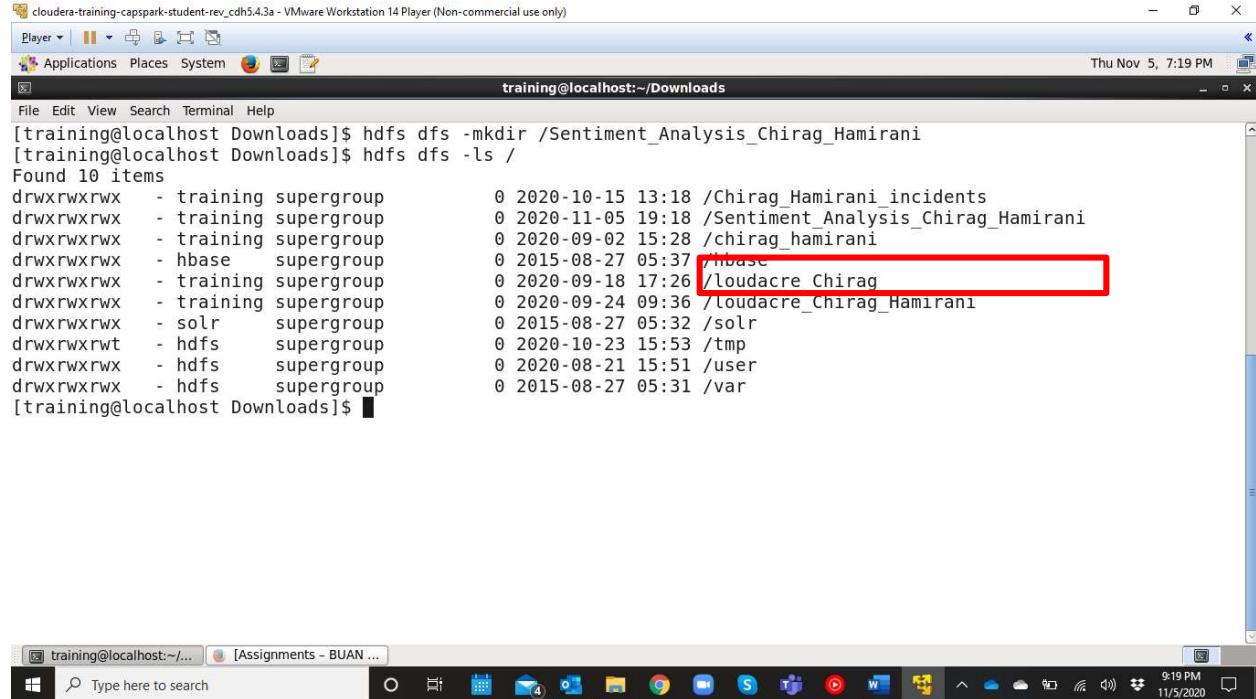


```
[training@localhost ~]$ cd /home/training/Downloads
[training@localhost Downloads]$ ls
Ass_14.zip          data      incidents_tab_clean.tsv  pig_1602793254839.log  Police_Incidents.zip
Chirag_Hamirani_Pig.pig  data.zip   __MACOSX             Police_Incidents.tsv
[training@localhost Downloads]$ unzip Ass_14.zip
Archive: Ass_14.zip
  inflating: AFINN.txt
  inflating: __MACOSX/.AFINN.txt
  inflating: Tweets.csv
  inflating: __MACOSX/.Tweets.csv
[training@localhost Downloads]$ ls Ass_14
ls: cannot access Ass_14: No such file or directory
[training@localhost Downloads]$ cd /home/training/Downloads/Ass_14
bash: cd: /home/training/Downloads/Ass_14: No such file or directory
[training@localhost Downloads]$ ls
AFINN.txt          data      MACOSX          Police_Incidents.zip
Ass_14.zip          data.zip   pig_1602793254839.log  Tweets.csv
Chirag_Hamirani_Pig.pig  incidents_tab_clean.tsv  Police_Incidents.tsv
[training@localhost Downloads]$
```

Explanation: The “unzip” command unzips the zip file.

Step-5: Execute the following command at the shell prompt:

```
hdfs dfs -mkdir /Sentiment_Analysis_Chirag_Hamirani  
hdfs dfs -ls /
```



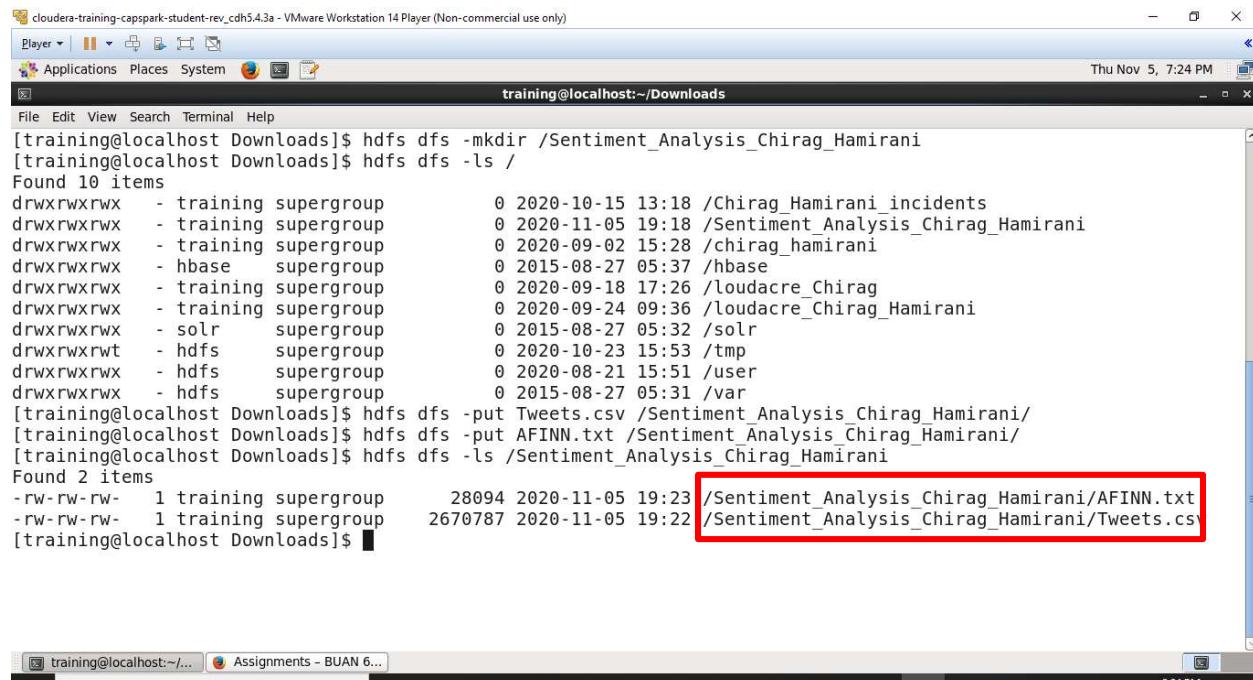
The screenshot shows a terminal window titled "training@localhost:~/Downloads". It displays the following command and its output:

```
[training@localhost Downloads]$ hdfs dfs -mkdir /Sentiment_Analysis_Chirag_Hamirani  
[training@localhost Downloads]$ hdfs dfs -ls /  
Found 10 items  
drwxrwxrwx  - training supergroup          0 2020-10-15 13:18 /Chirag_Hamirani_incidents  
drwxrwxrwx  - training supergroup          0 2020-11-05 19:18 /Sentiment_Analysis_Chirag_Hamirani  
drwxrwxrwx  - training supergroup          0 2020-09-02 15:28 /chirag_hamirani  
drwxrwxrwx  - hbase   supergroup          0 2015-08-27 05:37 /hbase  
drwxrwxrwx  - training supergroup          0 2020-09-18 17:26 /Loudacre_Chirag  
drwxrwxrwx  - training supergroup          0 2020-09-24 09:36 /Loudacre_Chirag_Hamirani  
drwxrwxrwx  - solr    supergroup          0 2015-08-27 05:32 /solr  
drwxrwxrwt  - hdfs   supergroup          0 2020-10-23 15:53 /tmp  
drwxrwxrwx  - hdfs   supergroup          0 2020-08-21 15:51 /user  
drwxrwxrwx  - hdfs   supergroup          0 2015-08-27 05:31 /var  
[training@localhost Downloads]$
```

Explanation: We are creating a directory at the root directory for this assignment. We will be using this directory throughout the assignment.

Step-6: Execute the following 2 commands at the shell prompt

```
hdfs dfs -put Tweet.csv /Sentiment_Analysis_Chirag_Hamirani/  
hdfs dfs -put AFINN.txt /Sentiment_Analysis_Chirag_Hamirani/  
hdfs dfs -ls /Sentiment_Analysis_Chirag_Hamirani/
```



The screenshot shows a terminal window titled "training@localhost:~/Downloads". The user has run several HDFS commands:

```
[training@localhost Downloads]$ hdfs dfs -mkdir /Sentiment_Analysis_Chirag_Hamirani  
[training@localhost Downloads]$ hdfs dfs -ls /  
Found 10 items  
drwxrwxrwx  - training supergroup          0 2020-10-15 13:18 /Chirag_Hamirani_incidents  
drwxrwxrwx  - training supergroup          0 2020-11-05 19:18 /Sentiment_Analysis_Chirag_Hamirani  
drwxrwxrwx  - training supergroup          0 2020-09-02 15:28 /chirag_hamirani  
drwxrwxrwx  - hbase   supergroup          0 2015-08-27 05:37 /hbase  
drwxrwxrwx  - training supergroup          0 2020-09-18 17:26 /loudacre_Chirag  
drwxrwxrwx  - training supergroup          0 2020-09-24 09:36 /loudacre_Chirag_Hamirani  
drwxrwxrwx  - solr    supergroup          0 2015-08-27 05:32 /solr  
drwxrwxrwt  - hdfs   supergroup          0 2020-10-23 15:53 /tmp  
drwxrwxrwx  - hdfs   supergroup          0 2020-08-21 15:51 /user  
drwxrwxrwx  - hdfs   supergroup          0 2015-08-27 05:31 /var  
[training@localhost Downloads]$ hdfs dfs -put Tweets.csv /Sentiment_Analysis_Chirag_Hamirani/  
[training@localhost Downloads]$ hdfs dfs -put AFINN.txt /Sentiment_Analysis_Chirag_Hamirani/  
[training@localhost Downloads]$ hdfs dfs -ls /Sentiment_Analysis_Chirag_Hamirani/  
Found 2 items  
-rw-rw-rw-  1 training supergroup      28094 2020-11-05 19:23 /Sentiment_Analysis_Chirag_Hamirani/AFINN.txt  
-rw-rw-rw-  1 training supergroup      2670787 2020-11-05 19:22 /Sentiment_Analysis_Chirag_Hamirani/Tweets.csv
```

A red box highlights the last two lines of output, which show the files "AFINN.txt" and "Tweets.csv" being listed in the directory.

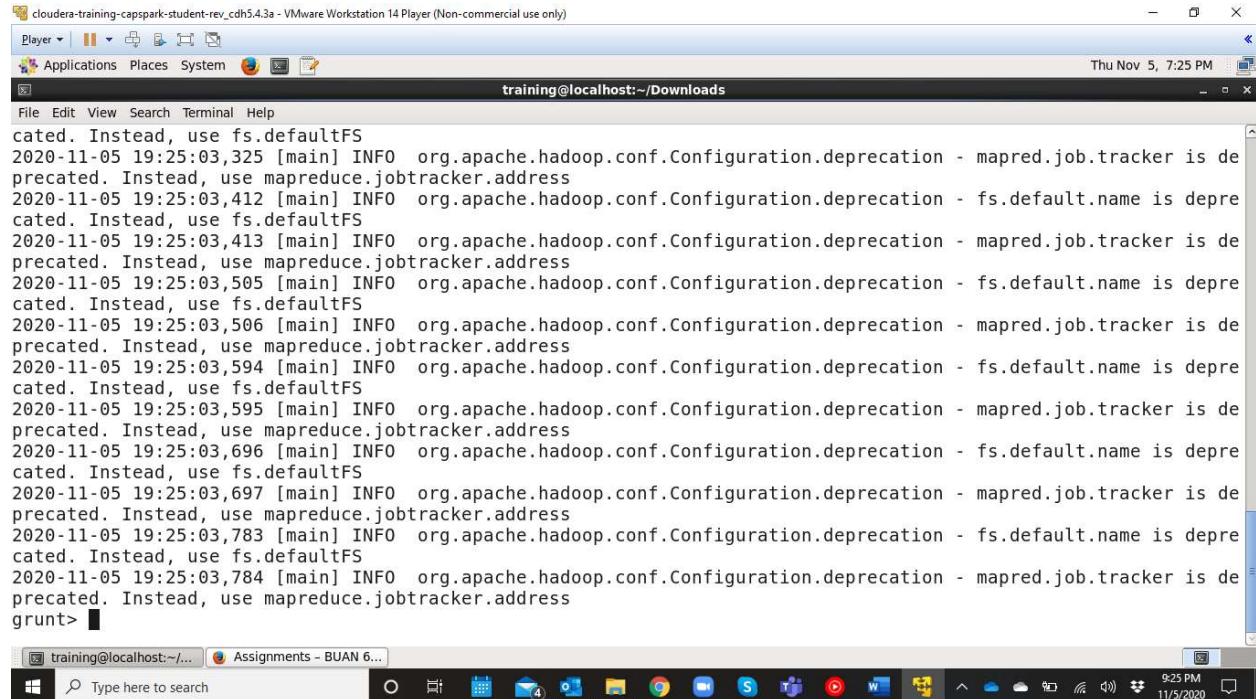
Explanation: We are copying the data files required for this exercise into the directory that we created in Step 5.

Part-2: Sentiment Analysis

In this section we will perform a sentiment analysis on tweets regarding Demonetization in India.

Step-7: Execute the following command at the shell prompt:

```
pig
```



```
cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 14 Player (Non-commercial use only)
Player | Applications Places System
File Edit View Search Terminal Help
cated. Instead, use fs.defaultFS
2020-11-05 19:25:03,325 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2020-11-05 19:25:03,412 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2020-11-05 19:25:03,413 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2020-11-05 19:25:03,505 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2020-11-05 19:25:03,506 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2020-11-05 19:25:03,594 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2020-11-05 19:25:03,595 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2020-11-05 19:25:03,696 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2020-11-05 19:25:03,697 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2020-11-05 19:25:03,783 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2020-11-05 19:25:03,784 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
grunt>
```

Explanation: The above command invokes the grunt shell. The grunt shell is an interactive shell to execute pig commands & scripts. The command above can be followed by various additional configurations but for our use case the stand grunt shell will suffice.

Step-8: Execute the following command at the grunt shell prompt:

```
Demonit_Tweets = LOAD '/Sentiment_Analysis_Chirag_Hamirani/Tweets.csv'
Using PigStorage(',');
dump Demonit_Tweets;
```

```
cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 14 Player (Non-commercial use only)
File Edit View Search Terminal Help
zation an all knows about it",FALSE,0,"quizderek","2016-11-22 10:58:12",FALSE,"801014696094539776","801016928496
013312","120965579","<a href=""http://twitter.com"" rel=""nofollow"">Twitter Web Client</a>","Devtech11",0,FALSE
,FALSE)
("7995","@baliramsingh2 So many restrictions. Not easy to avail the facility by anyone. Multiple U-turns by GOI
on the issue. #DeMonetization #RBI",FALSE,0,"baliramsingh2","2016-11-22 10:58:03",FALSE,"801011262805012480","80
1016889082163200","<a href=""http://twitter.com"" rel=""nofollow"">Twitter Web Client</a>","RahulCip
her",0,FALSE,FALSE)
("7997","FYI, I must tell and clear your doubts to enhance your GENERAL KNOWLEDGE that this wholesome experiment
of #DEMONETIZATION is only",FALSE,0,NA,"2016-11-22 10:57:59",FALSE,NA,"<a href=""http://www.twitter.com"" rel=""nofollow"">Twitter for Windows Phone</a>","Rishimathur18",0,FALSE,FALSE)
("7997","RT @sukanyaaiyer2: #DeMonetization AAP protests by marching Against Govts move over DeMonetization &
he is also detained as he Tried 2 March",FALSE,0,NA,"2016-11-22 10:57:58",FALSE,NA,"80101686747322776",NA,"<a href=""http://twitter.com/download/android"" rel=""nofollow"">Twitter for Android</a>","asitawasthi",2,TRUE,FA
LSE)
("7998","#demonetization will help combat terror because Pak won't be able to print new notes! And now, this. *s
low claps* https://t.co/p0p93sXlMn",FALSE,2,NA,"2016-11-22 10:57:44",FALSE,NA,"801016810006945792",NA,"<a href=""http://twitter.com/download/iphone"" rel=""nofollow"">Twitter for iPhone</a>","Modern_Gypsy",1,FALSE,FALSE)
("7999","#DeMonetization Positive Effect: Cigarette Sales Slashed To Almost Half After Demonetization:Report @de
epparihar @Tabassumzia @AshSharmail02",FALSE,2,NA,"2016-11-22 10:57:43",FALSE,NA,"801016807825940481",NA,"<a href=""http://twitter.com"" rel=""nofollow"">Twitter Web Client</a>","HP_Journo",1,FALSE,FALSE)
("8000","RT @UnSubtleDesi: Kejriwal posts pic of dead robber and claims it's #demonetization related death? How
shameless has this man become? https",FALSE,0,NA,"2016-11-22 10:57:33",FALSE,NA,"801016763722797057",NA,"<a href=""http://twitter.com"" rel=""nofollow"">Twitter Web Client</a>","prashik2107",897,TRUE,FALSE)
grunt> ■
```

Explanation: The dump command displays the data for the variable specified in the grunt shell.

Step-9: Execute the following command at the grunt shell prompt:

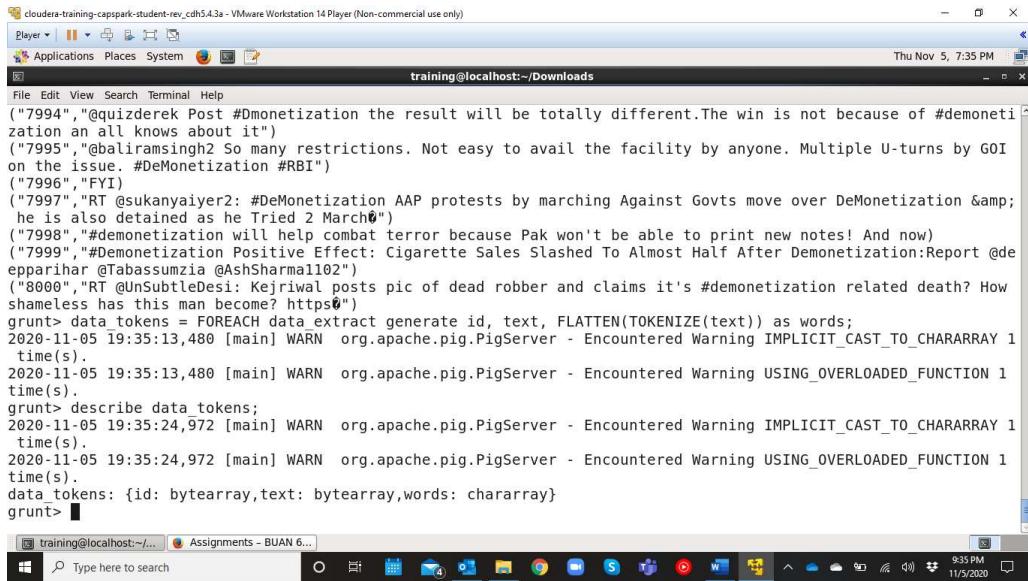
```
data_extract = FOREACH Demonit_Tweets GENERATE $0 as id, $1 as text;
dump data_extract;
```

```
cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 14 Player (Non-commercial use only)
File Edit View Search Terminal Help
fishy and requires full disclosure &#039;
("7990","All weddings now need to be approved by RBI... Amazing times #demonetization isn't that what we are und
erstanding")
("7991","RT @DrKumarVishwas: And the Oscar goes to ""Mr.<U+092D><U+093E><U+0935><U+0941><U+0915>"" <ed><U+00A0><
U+00BD><ed><U+0088><U+00A9><ed><U+00A0><U+00BD><ed><U+00B8><U+00AD><ed><U+00A0><U+00BD><ed><U+00B8><U+00A2><ed><
U+00A0><U+00BD><ed><U+00B8><U+00AD>#demonetization https://t.co/0bqhrhLNSL6")
("7992","RT @jan14anurag: Terrorists raided a bank to get new <U+20B9>2000 notes =&gt; their old finances are i
ndeed squeezed tight #Demonetization")
(<https://t&gt;,FALSE)
("7993","RT @jackerhack: Indore's collector would like you to shut up about #demonetization. At @internetfreedom
we think that is a problem. https://t&gt;")
("7994","@quizderek Post #Monetization the result will be totally different.The win is not because of #demone
tization an all knows about it")
("7995","@baliramsingh2 So many restrictions. Not easy to avail the facility by anyone. Multiple U-turns by GOI
on the issue. #DeMonetization #RBI")
("7996","FYI")
("7997","RT @sukanyaaiyer2: #DeMonetization AAP protests by marching Against Govts move over DeMonetization &
he is also detained as he Tried 2 March")
("7998","#demonetization will help combat terror because Pak won't be able to print new notes! And now)
("7999","#DeMonetization Positive Effect: Cigarette Sales Slashed To Almost Half After Demonetization:Report @de
epparihar @Tabassumzia @AshSharmail02")
("8000","RT @UnSubtleDesi: Kejriwal posts pic of dead robber and claims it's #demonetization related death? How
shameless has this man become? https")
grunt> ■
```

Explanation: We are defining and collecting the first two data elements in the csv file and dumping the data at the grunt shell.

Step-10: Execute the following command at the grunt shell prompt:

```
data_tokens = FOREACH data_extract generate id, text, FLATTEN(TOKENIZE(TEXT)) as words;
describe data_token;
```

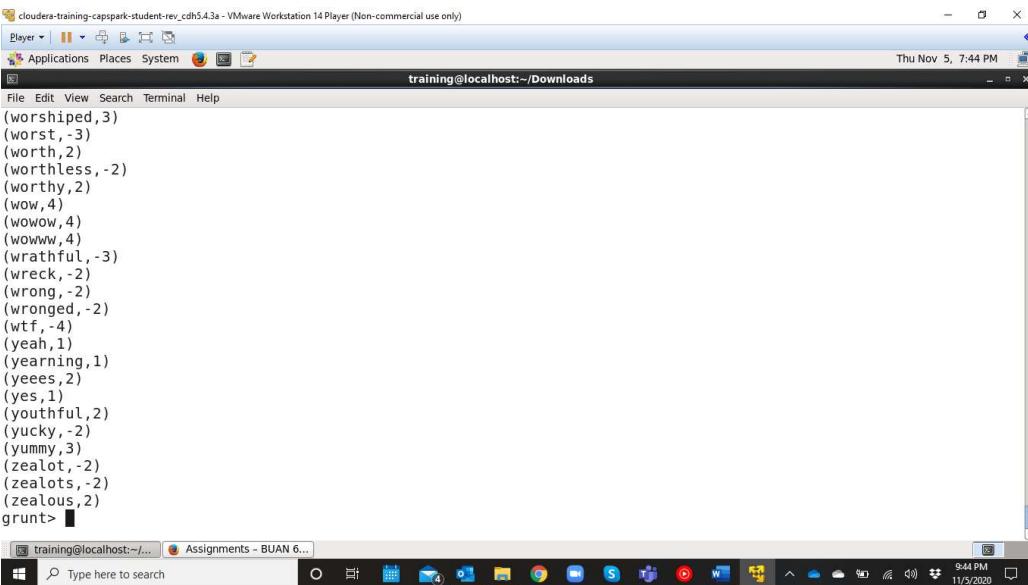


```
(7994,"@quizderek Post #Dmonetization the result will be totally different.The win is not because of #demonetization an all knows about it")
(7995,"@baliramsingh2 So many restrictions. Not easy to avail the facility by anyone. Multiple U-turns by GOI on the issue. #DeMonetization #RBI")
(7996,"FYI")
(7997,"RT @sukanyaiyer2: #DeMonetization AAP protests by marching Against Govts move over DeMonetization & he is also detained as he Tried 2 March")
(7998,"#demonetization will help combat terror because Pak won't be able to print new notes! And now)
(7999,"#Demonetization Positive Effect: Cigarette Sales Slashed To Almost Half After Demonetization:Report @de epparihar @Tabassumzia @AshSharmail02")
(8000,"RT @UnSubtleDesi: Kejriwal posts pic of dead robber and claims it's #demonetization related death? How shameless has this man become? https")
grunt> data_tokens = FOREACH data_extract generate id, text, FLATTEN(TOKENIZE(text)) as words;
2020-11-05 19:35:13,480 [main] WARN org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 1 time(s).
2020-11-05 19:35:13,480 [main] WARN org.apache.pig.PigServer - Encountered Warning USING_OVERLOADED_FUNCTION 1 time(s).
grunt> describe data_tokens;
2020-11-05 19:35:24,972 [main] WARN org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 1 time(s).
2020-11-05 19:35:24,972 [main] WARN org.apache.pig.PigServer - Encountered Warning USING_OVERLOADED_FUNCTION 1 time(s).
data_tokens: {id: bytearray, text: bytearray, words: chararray}
grunt> ■
```

Explanation: We are defining a third element by breaking down the text information in the variable ‘text’ into words. The describe operation shows us the variable names and types of the data in the variable specified.

Step-11: Execute the following command at the grunt shell prompt:

```
word_dict = LOAD '/Sentiment_Analysis_Chirag_Hamirani/AFINN.txt'
USING PigStorage('\t') As(word:chararray, rating:int);
dump word_dict
```



```
(worshipped,3)
(worst,-3)
(worth,2)
(worthless,-2)
(worthy,2)
(wow,4)
(wowow,4)
(wowwww,4)
(wrathful,-3)
(wreck,-2)
(wrong,-2)
(wronged,-2)
(wtf,-4)
(yeah,1)
(yearning,1)
(yees,2)
(yes,1)
(youthful,2)
(yucky,-2)
(yummy,3)
(zealot,-2)
(zealots,-2)
(zealous,2)
grunt> ■
```

Explanation: AFINN dictionary is a dictionary consisting of 2500 words that are rated on a scale from -5 to +5 depending on the sentiment the word conveys. We will be using this dictionary to perform the sentiment analysis on our tweet data.

Step-12: Execute the following command at the grunt shell prompt:

```
sent_join = join data_tokens by words left outer, word_dict by word using 'replicated';
describe sent_join;
```

```
(wtf,-4)
(yeah,1)
(yearning,1)
(yees,2)
(yes,1)
(youthful,2)
(yucky,-2)
(yummy,3)
(zealot,-2)
(zealots,-2)
(zealous,2)
grunt> sent_join = join data_tokens by words left outer, word_dict by word using 'replicated';
2020-11-05 19:46:11,266 [main] WARN org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 1 time(s).
2020-11-05 19:46:11,266 [main] WARN org.apache.pig.PigServer - Encountered Warning USING_OVERLOADED_FUNCTION 1 time(s).
grunt> describe sent_join;
2020-11-05 19:46:22,403 [main] WARN org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 1 time(s).
2020-11-05 19:46:22,403 [main] WARN org.apache.pig.PigServer - Encountered Warning USING_OVERLOADED_FUNCTION 1 time(s).
sent_join: {data_tokens::id: bytearray,data_tokens::text: bytearray,data_tokens::words: chararray,word_dict::word: chararray,word_dict::rating: int}
grunt> ■
```

Explanation: We are joining the sentiments values from the AFINN dictionary with the words in the text from each tweet.

Step-13: Execute the following command at the grunt shell prompt:

```
sent_extract = FOREACH sent_join generate data_tokens::id as id, data_tokens::text as text, word_dict::rating as rate;
describe sent_extract;
```

```
grunt> sent_extract = FOREACH sent_join generate data_tokens::id as id, data_tokens::text as text, dict::rating as rate;
2020-11-05 19:49:04,393 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1025:
<line 8, column 92> Invalid field projection. Projected field [dict::rating] does not exist in schema: data_tokens::id:bytearray,data_tokens::text:bytearray,data_tokens::words:chararray,word_dict::word:chararray,word_dict::rating:int.
Details at logfile: /home/training/Downloads/pig_1604633101019.log
grunt> describe sent_extract;
2020-11-05 19:49:14,219 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1003: Unable to find an operator f or alias sent_extract
Details at logfile: /home/training/Downloads/pig_1604633101019.log
grunt> sent_extract = FOREACH sent_join generate data_tokens::id as id, data_tokens::text as text, word_dict::ra ting as rate;
2020-11-05 19:50:09,247 [main] WARN org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 1 time(s).
2020-11-05 19:50:09,247 [main] WARN org.apache.pig.PigServer - Encountered Warning USING_OVERLOADED_FUNCTION 1 time(s).
grunt> describe sent_extract;
2020-11-05 19:50:13,939 [main] WARN org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 1 time(s).
2020-11-05 19:50:13,939 [main] WARN org.apache.pig.PigServer - Encountered Warning USING_OVERLOADED_FUNCTION 1 time(s).
sent_extract: {id: bytearray,text: bytearray,rate: int}
grunt> ■
```

Explanation: We are now extracting the data from the variable in step 12 that we need to complete our sentiment analysis.

Step-14: Execute the following command at the grunt shell prompt:

```
word_grouping = group sent_extract by (id,text);
dump word_grouping;
```

Explanation: We are grouping each record by its id and text.

Step-15: Execute the following command at the grunt shell prompt:

```
average_rating = FOREACH word_grouping generate group, AVG(sent_extract.rate) as tweet_rate;  
dump average_rating;
```

```
cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 14 Player (Non-commercial use only)
Player || Applications Places System Firefox
training@localhost:~/Downloads
File Edit View Search Terminal Help
((If modi.withdrawals #demonetization scheme Today.<ed><U+00A0><U+00BD><ed><U+00B8><U+0082><ed><U+00A0><U+00BD><e
d><U+00B8><U+0082> https://t.co/000MptVdQT", FALSE), )
((<U+0915><U+0939><U+0940><U+0902> <U+092F><U+0947> <U+092D><U+093E><U+0921><U+093C><U+0947> <U+0915><U+0947> <U+0924><U+094B> <U+0928><U+0939><U+0940><U+0902>! , ), )
((<ed><U+00A0><U+00BD><ed><U+00B1><U+008C><ed><U+00A0><U+00BC><ed><U+00BF><U+00BB><ed><U+00A0><U+00BD><ed><U+00B
9><U+008F><ed><U+00A0><U+00BD><ed><U+00B1><U+0087> , ), )
(( <ed><U+00A0><U+00BD><ed><U+00B8><U+00B1><ed><U+00A0><U+00BD><ed><U+00B8><U+00B1><ed><U+00A0><U+00BD><ed><U+00
B8><U+00B1><ed><U+00A0><U+00BD><ed><U+00B8><U+00B1>, ), )
((<ed><U+00A0><U+00BC><ed><U+00B6><U+0098> <ed><U+00A0><U+00BC><ed><U+00B6><U+0098> <ed><U+00A0><U+00BC><ed><U+0
0B6><U+0098> <ed><U+00A0><U+00BC><ed><U+00B6><U+0098>" , FALSE), )
((cu of <ed><U+00A0><U+00BD><ed><U+00B9><U+008F> @narendramodi <ed><U+00A0><U+00BD><ed><U+00B9><U+008F> #demonet
ization <ed><U+00A0><U+00BD><ed><U+00B1><U+0089>Barbadi done of , ), )
(@DonMufflerMan <ed><U+00A0><U+00BD><ed><U+00B8><U+0082><ed><U+00A0><U+00BD><ed><U+00B8><U+0080><ed><U+00A0><U+
00BD><ed><U+00B8><U+00A0><ed><U+00A0><U+00BD><ed><U+00B8><U+00AE", FALSE), )
(( <NoteNaHimiPBMadlo <U+0938><U+0902><U+0926> <U+092A><U+0941><U+0915><U+093E><U+0930><U+0947>_PM <U+092
A><U+094D><U+092F><U+093E><U+0930><U+0947> #MayawatiNextUPCM #BSPB @invito", FALSE), )
((But best thing was comments that followed it.. <ed><U+00A0><U+00BD><ed><U+00B8><U+0082> <ed><U+00A0><U+00BD><e
d><U+00B8><U+0082> <ed><U+00A0><U+00BD><ed><U+00B8><U+0082> https://t.co/o9KrTKZffq", FALSE), )
((Kejri & Mamta has lost thousands of crores and u want then to keep calm? <ed><U+00A0><U+00BD><ed><U+00B8><U+008B> , ), )
(<ed><U+00A0><U+00BD><ed><U+00B8><U+0082><ed><U+00A0><U+00BD><ed><U+00B8><U+0082><ed><U+00A0><U+00BD><ed><U+00B
8><U+0082><ed><U+00A0><U+00BD><ed><U+00B8><U+0082>, FALSE), )
((), )
grunt
```

Explanation: We are averaging the rating for all the words in each tweet and storing the result in average rating.

Step-16: Execute the following command at the grunt shell prompt:

```
final_tweets = filter average_rating by tweet_rate >= -5;
dump final_tweets;
```

```
(( "7995" , "@baliramsingh2 So many restrictions. Not easy to avail the facility by anyone. Multiple U-turns by GOI on the issue. #DeMonetization #RBI" ),1.0)
(( "7997" , "RT @sukanyaiyer2: #DeMonetization AAP protests by marching Against Govts move over DeMonetization & ; he is also detained as he Tried 2 March#"),-2.0)
(( "7998" , "#demonetization will help combat terror because Pak won't be able to print new notes! And now),-0.666666666666)
(( "8000" , "RT @UnSubtleDesi: Kejriwal posts pic of dead robber and claims it's #demonetization related death? How shameless has this man become? https#"),-2.5)
((How long, successful and sustainable will be this strategic game of #DeMonetization against Demons?"),3.0)
((Only noise, chaos & disruptions by obstructionist #),-2.0)
((Only noise, chaos & disruptions by obstructio# https://t.co/zVE7MYt04G"),-2.0)
((% bad idea, poor implementation"),-2.0)
((25% good idea, poor implementation),-2.0)
((No there r many, we cal them by many names like C#%),2.0)
((Akhilesh-not good,black money is good),3.0)
((If not for Aam Aadmi, listen to them no PM Modi?"),-1.0)
((And respect their decision,but support oppositio#),2.0)
((And respect their decision,but support opposition just b'coz of party"),2.0)
((Aim of #demonetization laudable, but Govt has no road map2create... https://t.co/A4Geu9ch0v"),-1.0)
((Enough jokes on #Demonetization, also no more posts on politics or social affairs...),-1.0)
((RT @kanimozh: Everyone seems to hate the rich, even the rich hates richer and the richer hates the richest. # Demonetization"),-1.33333333333333)
(( the avg indian wants corrupt free india.. So in d name of black money, everybody agrees),1.0)
grunt> ■
```

Explanation: We are filtering all the tweets for those tweets that have an average rating ≥ -5 .

Step-17: Execute the following command at the grunt shell prompt:

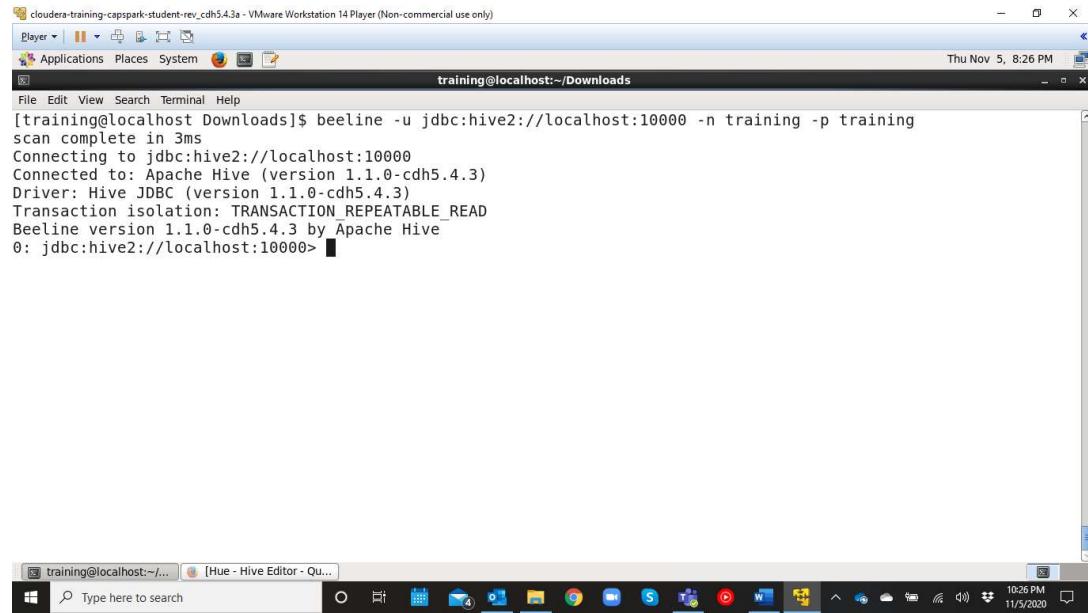
```
STORE final_tweets INTO
'/Sentiment_Analysis_Chirag_Hamirani/Analyzed_Tweets';
```

ACTIONS	INFO
View as text Download View file location Refresh	Last modified Nov. 5, 2020 8:04 p.m. User training Group supergroup Size
/ Sentiment_Analysis_Chirag_Hamirani / Analyzed_Tweets / part-r-00000	
000000: 28 53 6f 2c 20 69 66 20 79 6f 75 20 72 65 61 6c ... 000001: 6c 79 20 74 68 69 6e 6b 20 74 68 61 74 20 74 68 ... 000020: 69 73 20 23 44 65 6d 6f 6e 65 74 69 7a 61 74 69 ... 000030: 6f 6e 20 6d 6f 76 65 20 68 61 73 20 73 74 72 75 ... 000040: 63 6b 20 61 74 20 66 61 6b 65 85 22 29 09 2d 31 ... 000050: 2e 39 0a 28 22 31 22 22 52 54 20 40 72 73 73 ... 000060: 75 72 6a 65 77 61 6c 61 3a 20 43 72 69 74 6a 63 ... 000070: 61 6c 20 71 75 65 73 74 69 6f 6e 3a 20 57 61 73 ... 000080: 29 59 61 79 54 4d 20 69 6e 66 6f 72 6d 65 64 29 ... 000090: 61 62 6f 75 74 20 23 44 65 6d 6f 6e 65 74 69 7a ... 0000a0: 61 74 69 6f 6e 20 65 64 69 63 74 20 62 79 20 50 ... 0000b0: 4d 3f 20 49 74 27 73 29 63 6c 65 61 72 6c 79 20 ...	

Explanation: We are viewing the output of the Analyzed_Tweets file in HUE.

Step-18: Execute the following command at the shell prompt:

```
sudo service zookeeper-server start  
sudo service hive-server2 start  
sudo service hue restart  
beeline -u jdbc:hive2://localhost:10000 -n training -p training
```



The screenshot shows a terminal window titled "training@localhost:~/Downloads". The window displays the following Beeline startup logs:

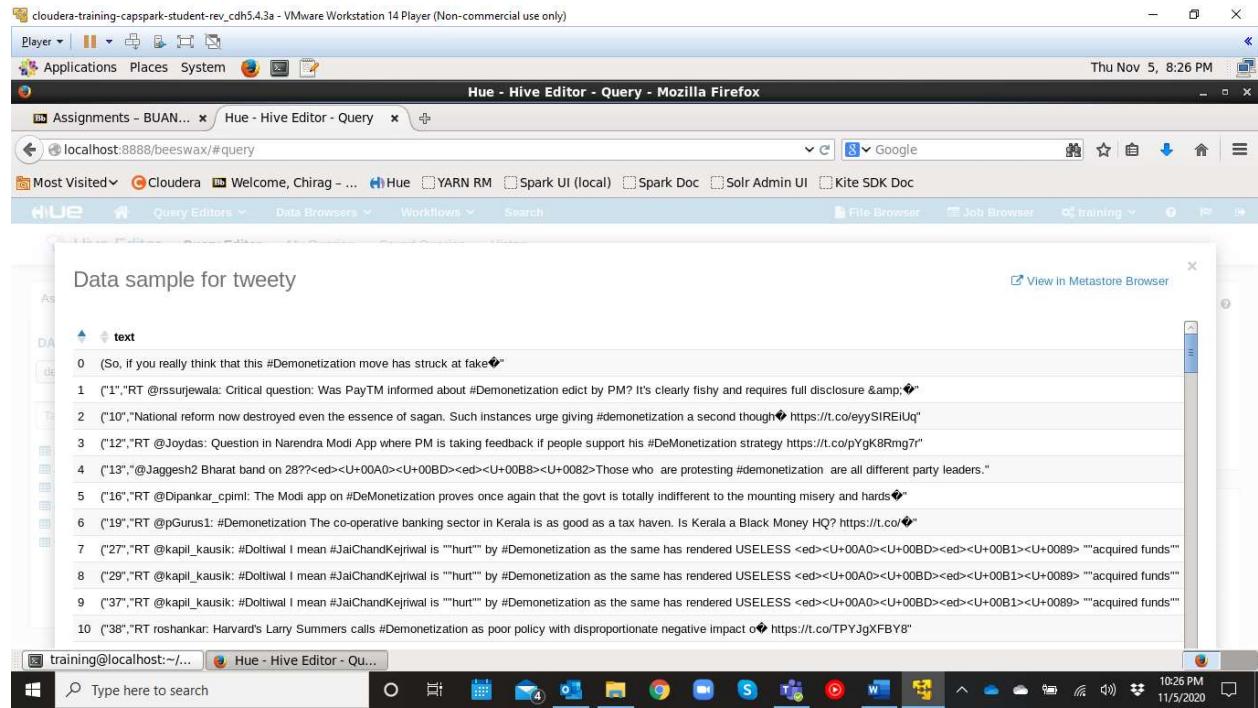
```
[training@localhost Downloads]$ beeline -u jdbc:hive2://localhost:10000 -n training -p training  
scan complete in 3ms  
Connecting to jdbc:hive2://localhost:10000  
Connected to: Apache Hive (version 1.1.0-cdh5.4.3)  
Driver: Hive JDBC (version 1.1.0-cdh5.4.3)  
Transaction isolation: TRANSACTION_REPEATABLE_READ  
Beeline version 1.1.0-cdh5.4.3 by Apache Hive  
0: jdbc:hive2://localhost:10000>
```

Explanation: We are starting up beeline, to execute the hive commands. Beeline shell is an interactive shell based on the SQLLine utility.

Step-19: Execute the following command at the beeline shell:

```
CREATE EXTERNAL TABLE tweety (text STRING, tweet_rating DOUBLE) ROW FORMAT DELIMITED FIELDS TERMINATED BY ')' LOCATION
'Sentiment_Analysis_Chirag_Hamirani/Analyzed_Tweets';
```

Open the Hive Query Editor and click on the refresh button in the database panel, then click on the 'Preview Sample Data' button next to the tweety table. Take a screenshot of the sample data and paste it below.



The screenshot shows the Hue - Hive Editor interface. The title bar says "Hue - Hive Editor - Query - Mozilla Firefox". The URL in the address bar is "localhost:8888/beeswax/#query". Below the address bar, there's a navigation bar with links like "Most Visited", "Cloudera", "Welcome, Chirag", "Hue", "YARN RM", "Spark UI (local)", "Spark Doc", "Solr Admin UI", and "Kite SDK Doc". The main content area is titled "Data sample for tweety". It shows a list of 10 tweets from the tweety table. Each tweet is represented by a JSON object with fields "text" and "tweet_rating". The "text" field contains the tweet content, and the "tweet_rating" field contains a double value. The tweets are numbered from 0 to 9. The interface includes a sidebar on the left with icons for different data types and a toolbar at the bottom.

Index	Text	Rating
0	(So, if you really think that this #Demonetization move has struck at fake...	
1	("1","RT @rssurjewala: Critical question: Was PayTM informed about #Demonetization edict by PM? It's clearly fishy and requires full disclosure &..."	
2	("10","National reform now destroyed even the essence of sagan. Such instances urge giving #demonetization a second thought https://t.co/eyySIREIUq"	
3	("12","RT @Joydas: Question in Narendra Modi App where PM is taking feedback if people support his #DeMonetization strategy https://t.co/pYgk8Rmg7?"	
4	("13","@Jaggesh2 Bharat band on 28??<ed><U+00A0><U+00BD><ed><U+00B8><U+0082>Those who are protesting #demonetization are all different party leaders."	
5	("16","RT @Dipankar_cpmi: The Modi app on #DeMonetization proves once again that the govt is totally indifferent to the mounting misery and hards..."	
6	("19","RT @pGurus1: #Demonetization The co-operative banking sector in Kerala is as good as a tax haven. Is Kerala a Black Money HQ? https://t.co/..."	
7	("27","RT @kapil_kausik: #Doltiwal I mean #JaiChandKejriwal is ""hurt"" by #Demonetization as the same has rendered USELESS <ed><U+00A0><U+00BD><ed><U+00B1><U+0089> ""acquired funds...""	
8	("29","RT @kapil_kausik: #Doltiwal I mean #JaiChandKejriwal is ""hurt"" by #Demonetization as the same has rendered USELESS <ed><U+00A0><U+00BD><ed><U+00B1><U+0089> ""acquired funds...""	
9	("37","RT @kapil_kausik: #Doltiwal I mean #JaiChandKejriwal is ""hurt"" by #Demonetization as the same has rendered USELESS <ed><U+00A0><U+00BD><ed><U+00B1><U+0089> ""acquired funds...""	
10	("38","RT roshankar_Harvard's Larry Summers calls #Demonetization as poor policy with disproportionate negative impact ohttps://t.co/TPYJgXFBy8"	

Explanation: We are creating a hive table to structure our unstructured data.

Step-20: Execute the following command at the beeline shell:

```
INSERT OVERWRITE LOCAL DIRECTORY 'Sentiment_Analysis_Chirag_Hamirani/Analyzed_Tweets' ROW FORMAT  
DELIMITED FIELDS TERMINATED BY '\t' select * from tweety;
```

Open Hue and view the file we just created, click on “View as text” on the left under ACTIONS. Take a screenshot of the contents and paste it below.

The screenshot shows a Mozilla Firefox window titled "Hue - File Browser - part-r-00000 - File Viewer - Mozilla Firefox". The URL in the address bar is "localhost:8888/filebrowser/view/Sentiment_Analysis_Chirag_Hamirani/Analyzed_Tweets/part-r-00000?mode=text&compress". The browser toolbar includes icons for Home, Stop, Back, Forward, Refresh, Stop, and a search bar. The main content area displays a text file with the following content:

```
(So, if you really think that this #Demonetization move has struck at fake#) -1.0
("1","RT @rssurjewala: Critical question: Was PayTM informed about #Demonetization edict by PM? It's clearly fishy and requires full disclosure &#038;#039;) 1.0
("10","National reform now destroyed even the essence of sagan. Such instances urge giving #demonetization a second thought# https://t.co/eyySIREiUq") -3.0
("12","RT @Joydas: Question in Narendra Modi App where PM is taking feedback if people support his #DeMonetization strategy https://t.co/pYgK8Rmg7r") 2.0
("13","@Jaggesh2 Bharat band on 28??<ed><U+00A0><ed><U+00BD><ed><U+00B8><U+0082>Those who are protesting #demonetization are all different party leaders.") -2.0
("16","RT @Dipankar_cpiml: The Modi app on #DeMonetization proves once again that the govt is totally indifferent to the mounting misery and har ds#") -2.0
```

The left sidebar contains an "ACTIONS" menu with options: View as binary, Download, View file location, and Refresh. Below that is an "INFO" section showing Last modified (Nov. 5, 2020 8:04 p.m.), User (training), Group (supergroup), and Size. The bottom of the screen shows a Windows taskbar with various icons and the system tray indicating the date and time as 11/5/2020 10:33 PM.

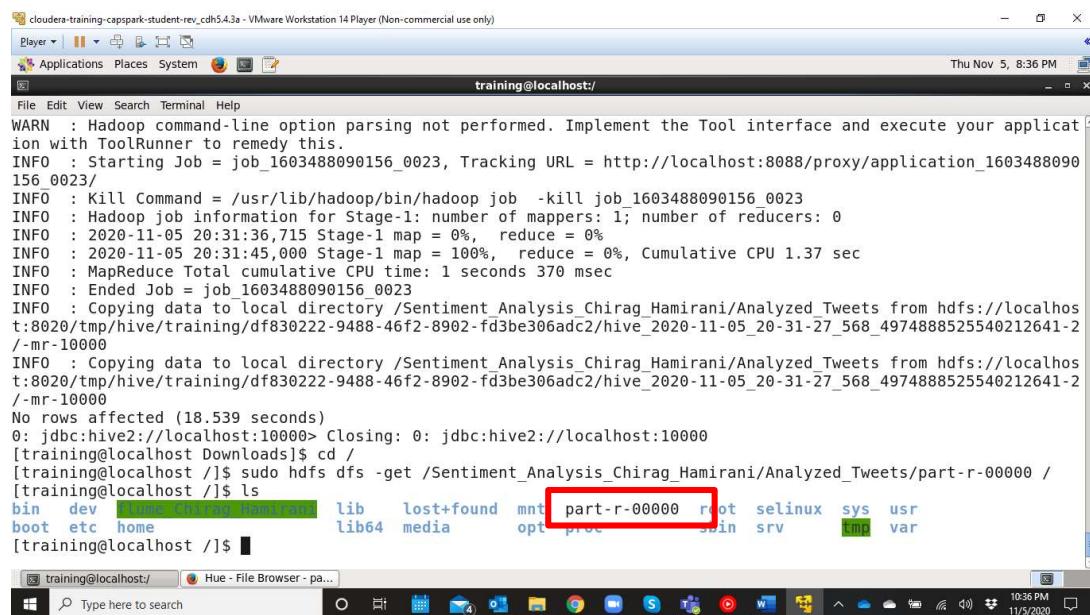
Explanation: We are now creating a tsv file from the table that we created in Step-20.

Part-3: Visualization of Sentiments

In this section we will visualize the results of our sentiment analysis using Jupyter notebooks.

Step-21: Exit Beeline then execute the following command at the shell prompt:

```
cd/
sudo hdfs dfs -get /Sentiment_Analysis_Chirag_Hamirani/Analyzed_Tweets/part-r-00000 /
ls
```

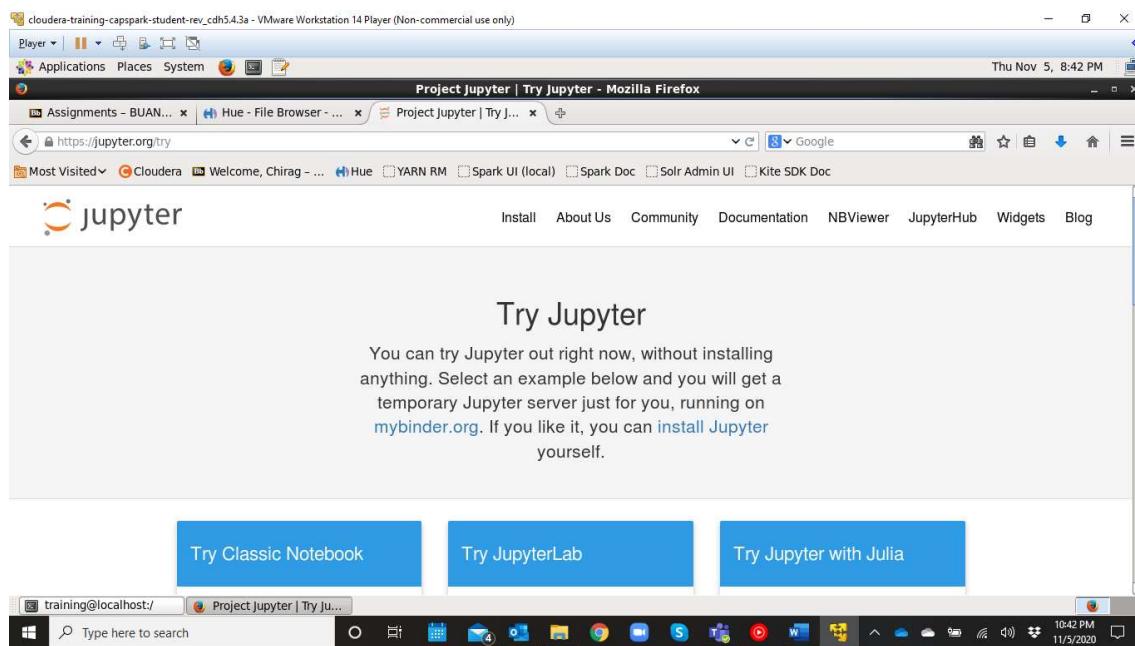


```

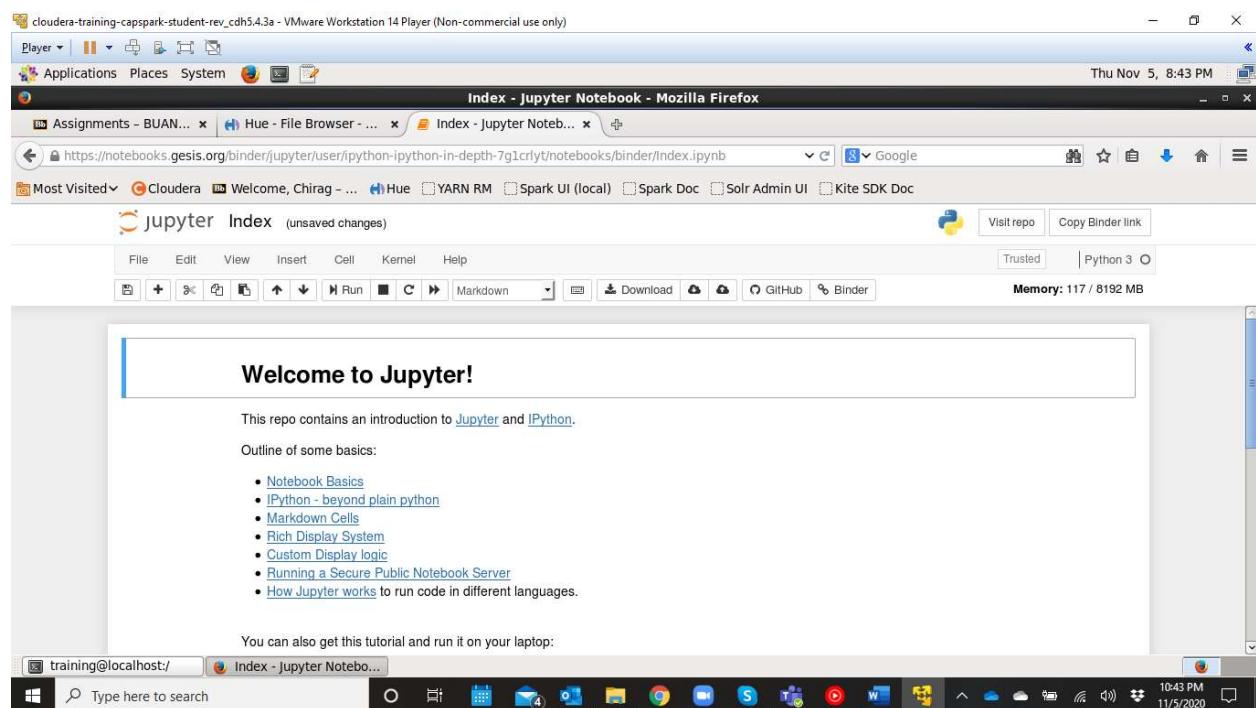
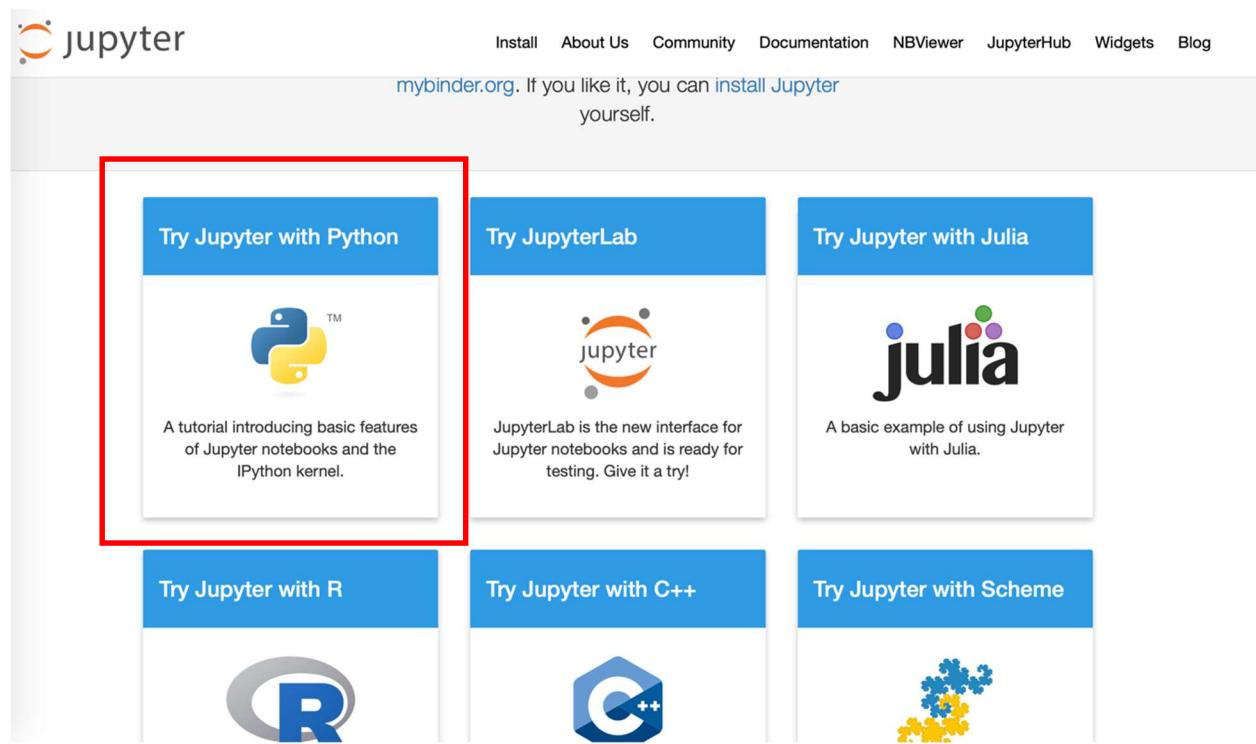
cloudera-training-capspark-student-rev_cdh5.4.3a - VMware Workstation 14 Player (Non-commercial use only)
Player | II | X
Applications Places System
File Edit View Search Terminal Help
WARN : Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
INFO : Starting Job = job_1603488090156_0023, Tracking URL = http://localhost:8088/proxy/application_1603488090156_0023/
INFO : Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1603488090156_0023
INFO : Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
INFO : 2020-11-05 20:31:36,715 Stage-1 map = 0%, reduce = 0%
INFO : 2020-11-05 20:31:45,000 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 1.37 sec
INFO : MapReduce Total cumulative CPU time: 1 seconds 370 msec
INFO : Ended Job = job_1603488090156_0023
INFO : Copying data to local directory /Sentiment_Analysis_Chirag_Hamirani/Analyzed_Tweets from hdfs://localhost:8020/tmp/hive/training/df830222-9488-46f2-8902-fd3be306adc2/hive_2020-11-05_20-31-27_568_4974888525540212641-2/-mr-10000
INFO : Copying data to local directory /Sentiment_Analysis_Chirag_Hamirani/Analyzed_Tweets from hdfs://localhost:8020/tmp/hive/training/df830222-9488-46f2-8902-fd3be306adc2/hive_2020-11-05_20-31-27_568_4974888525540212641-2/-mr-10000
No rows affected (18.539 seconds)
0: jdbc:hive2://localhost:10000> Closing: 0: jdbc:hive2://localhost:10000
[training@localhost Downloads]$ cd /
[training@localhost /]$ sudo hdfs dfs -get /Sentiment_Analysis_Chirag_Hamirani/Analyzed_Tweets/part-r-00000 /
[training@localhost /]$ ls
bin dev lib lost+found mnt part-r-00000 root selinux sys usr
boot etc lib64 media opt proc sbin srv var
[training@localhost /]$
```

Explanation: We are copying the results of our sentiment analysis back on the local file system to visualize.

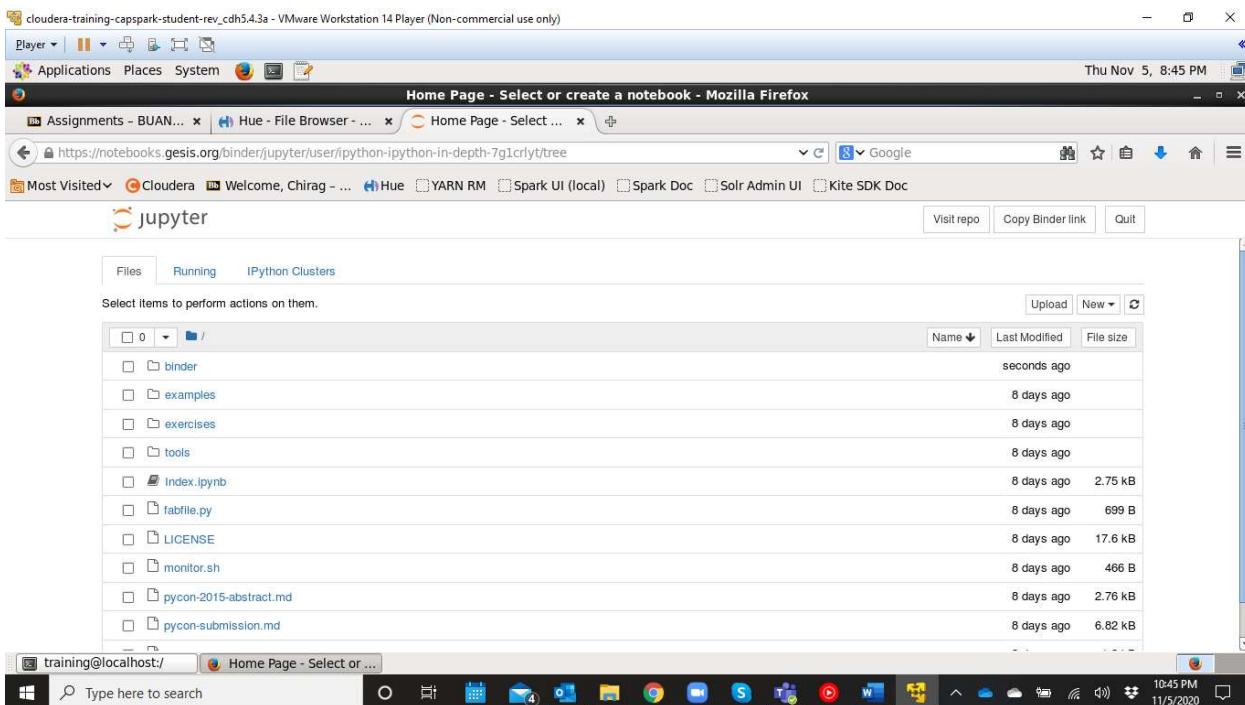
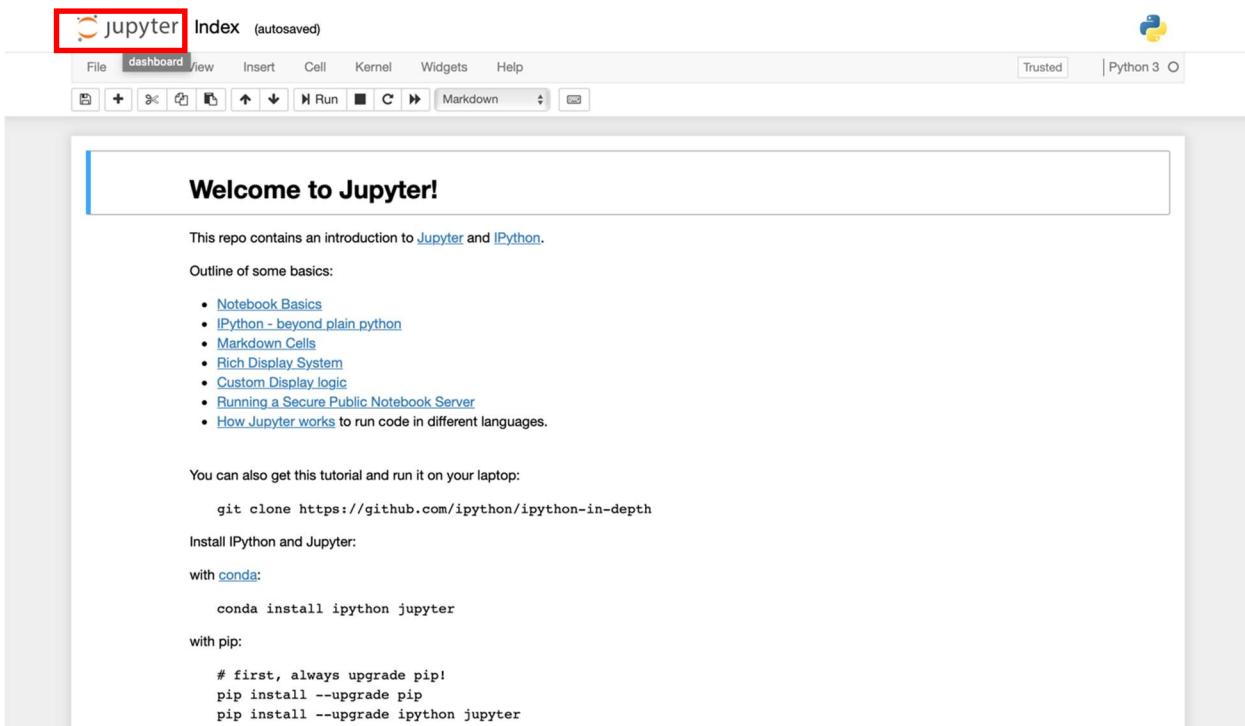
Step-22: Using Firefox, browse to the following address: <https://jupyter.org/try>



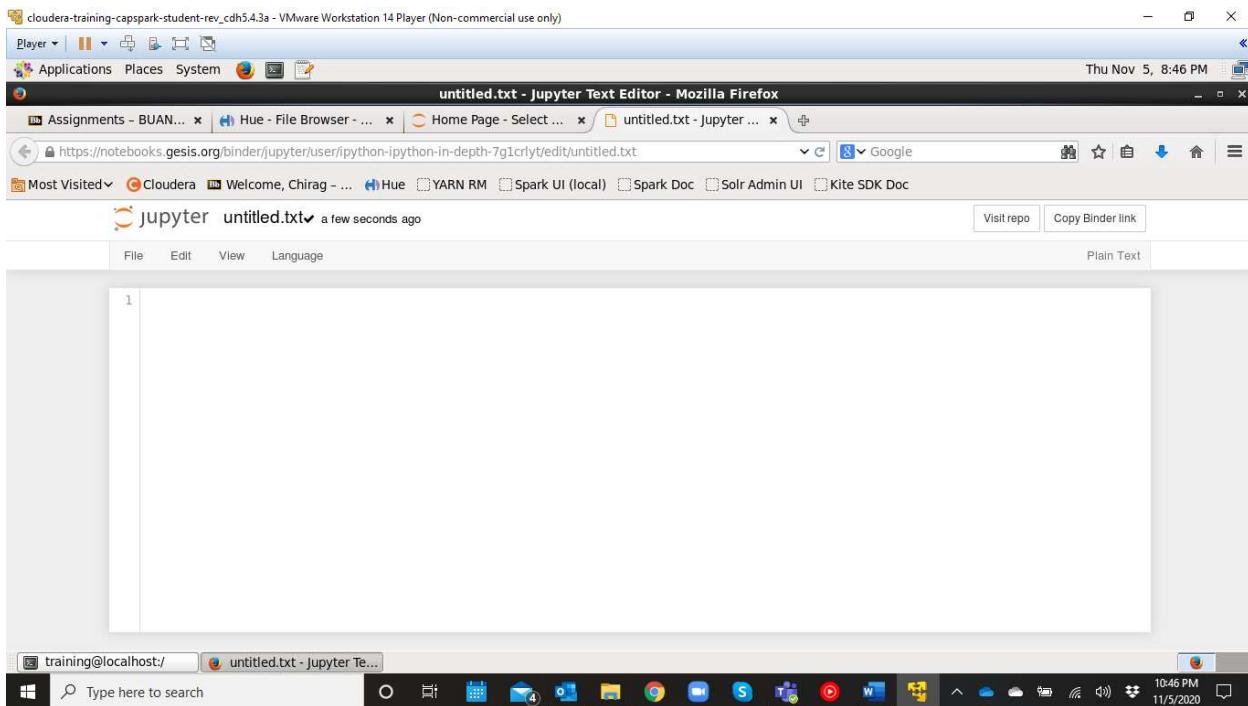
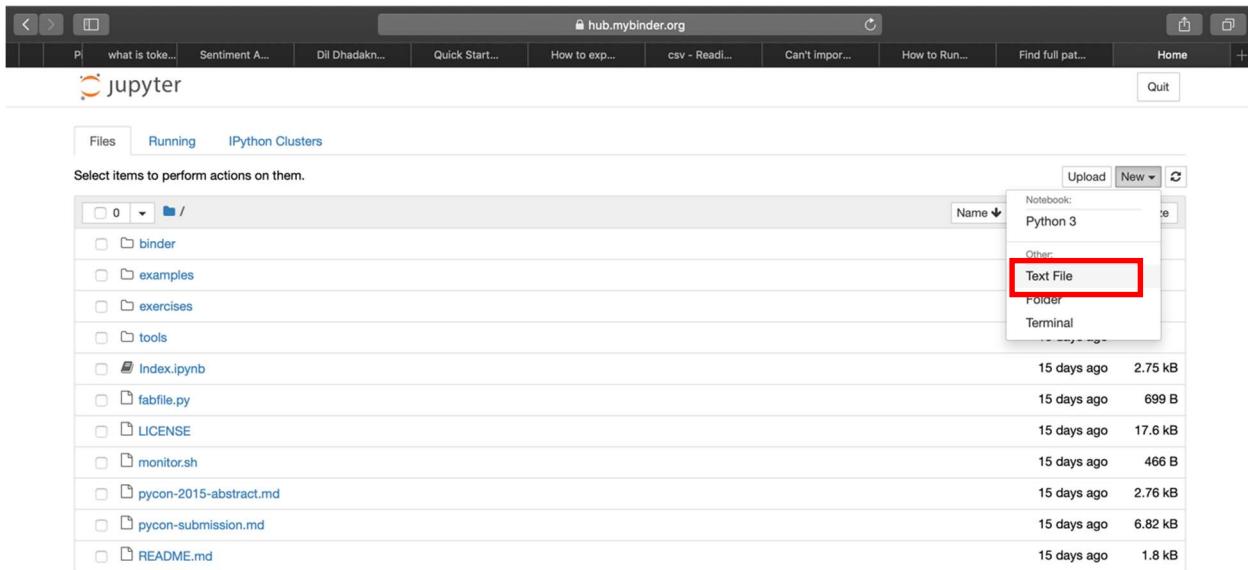
Step-23: On the Jupyter homepage click on the “Try Jupyter with Python” icon that is highlighted in the screenshot below:



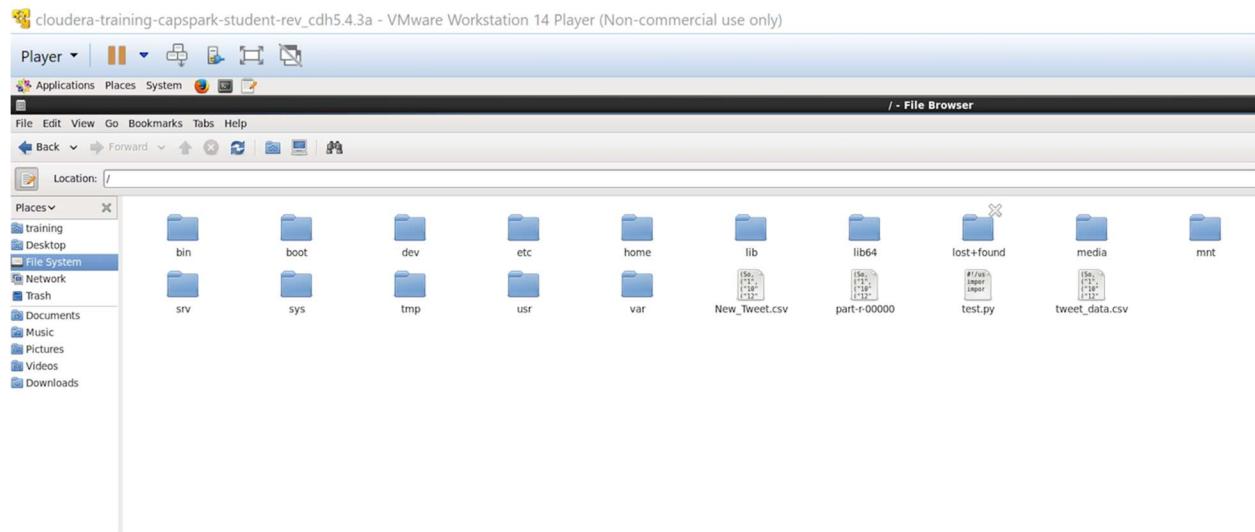
Step-24: Now click on the homepage button highlighted in the screenshot below.



Step-25: Now click on the “Text File” button highlighted in the screenshot below.

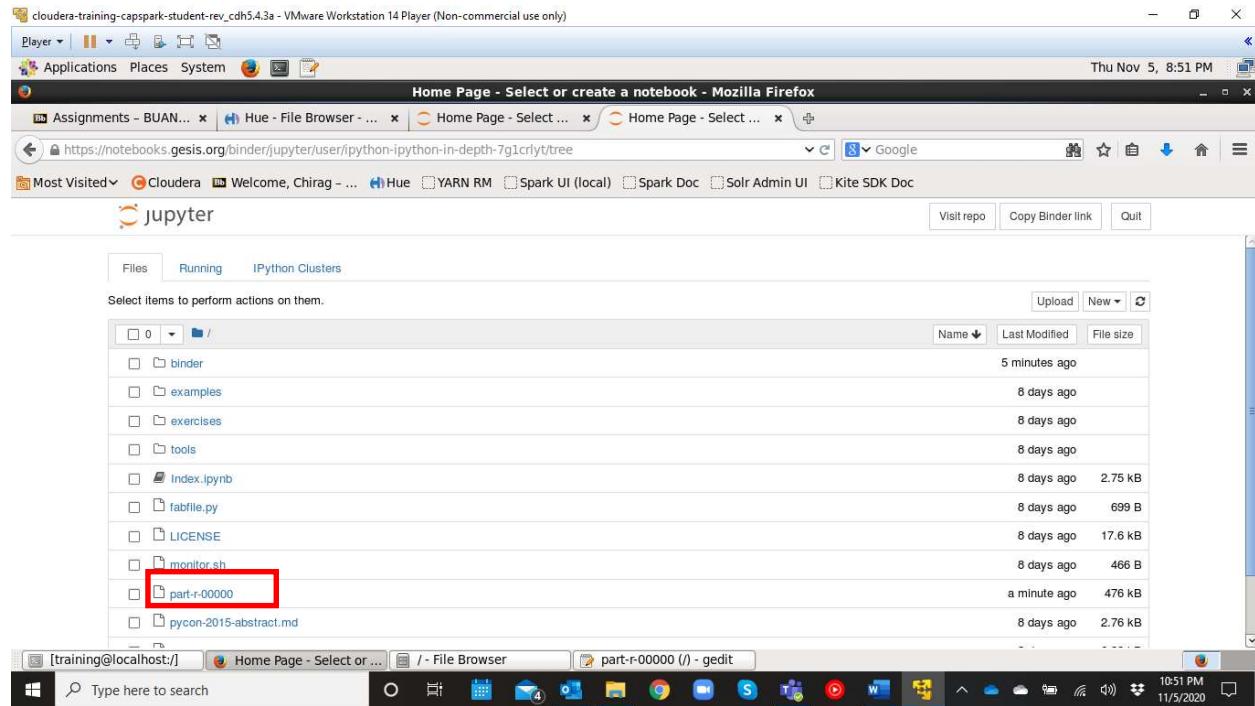


Step-26: Now open the file containing the calculated sentiments in the computer application. After opening the computer application click on the filesystem folder highlighted in the screenshot below.

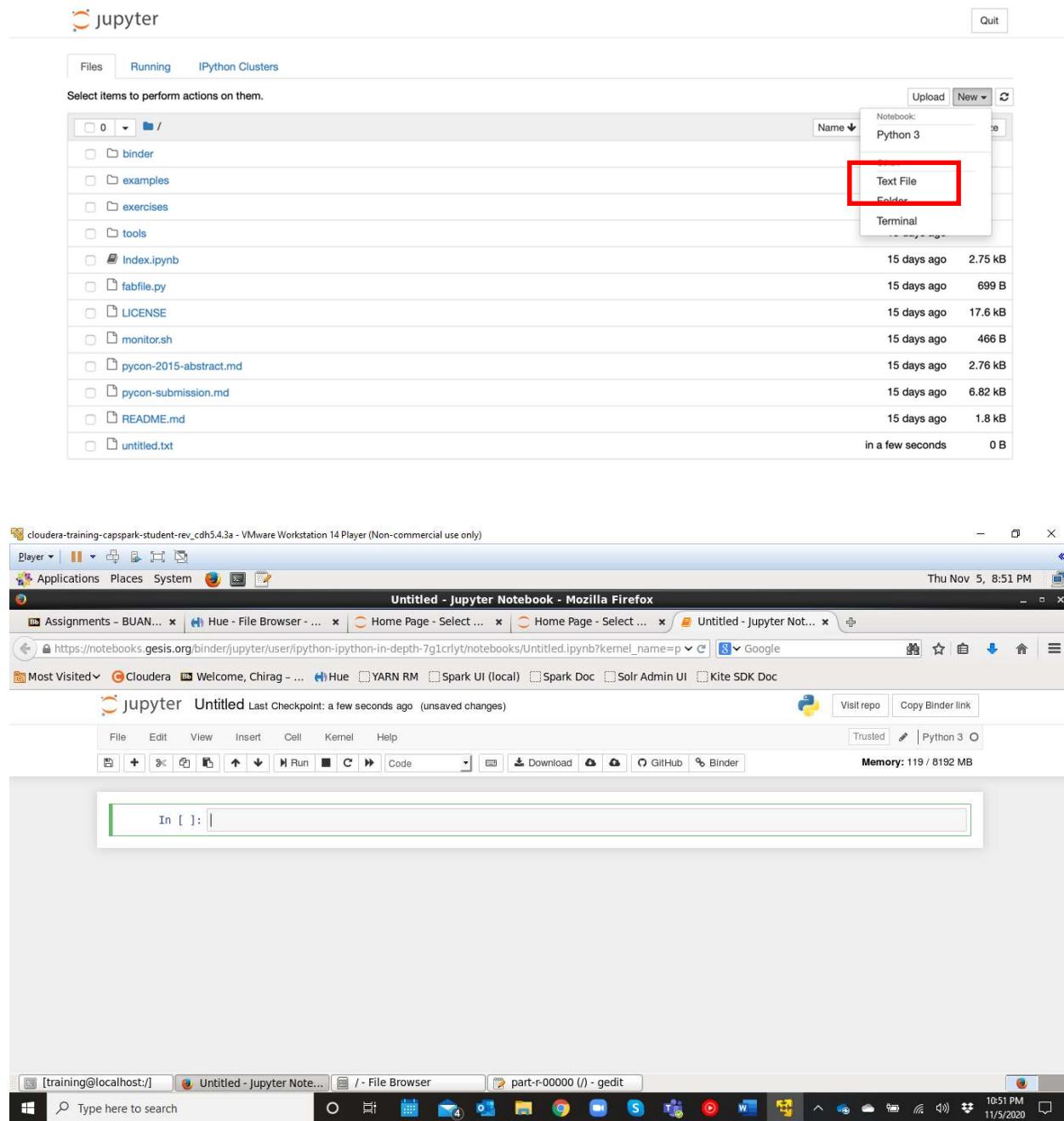


Open the part-r-00000 and copy the contents and paste it in the browser running Jupyter we opened in Step-25. Save the file as “part-r-00000”. Force save the file.

Click on the Jupyter icon like in Step 24 and highlight the new file you just created.



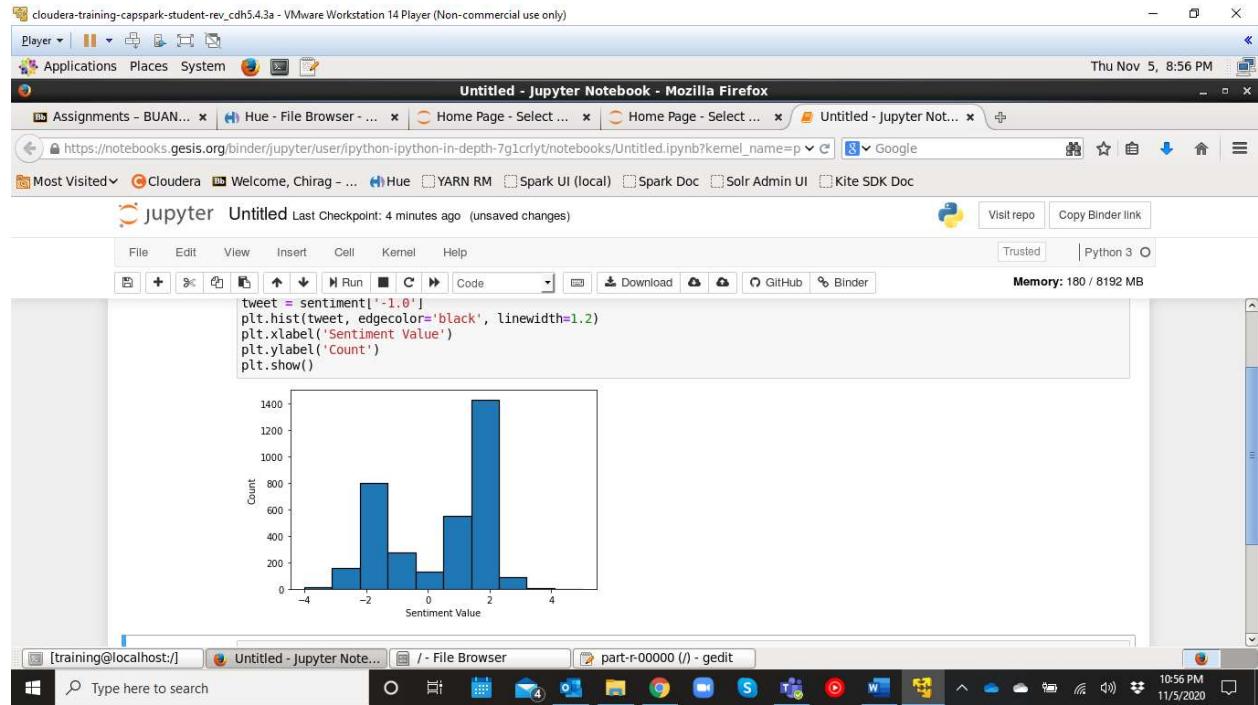
Step-27: Now create a Python 3 Jupyter Notebook file by clicking on the “Python 3” button highlighted in the screenshot below.



Step-28: Now type the code in the screenshot below at the prompt (after In []:) then hit Run.

```
import matplotlib.pyplot as plt
import pandas as pd
sentiment = pd.read_csv('part-r-00000', delimiter = '\t')
tweet = sentiment['-1.0']
plt.hist(tweet, edgecolor='black', linewidth=1.2)
plt.xlabel('Sentiment Value')
plt.ylabel('Count')
plt.show()
```

Conclusion



The tweets seem to be divided in two normal distribution groups on both sides, positive and negative. Tweets with highest frequency are moderate negative (-2) and moderate positive (+2). However, the distribution clearly shows that more tweets are assessed as positive. Note that number of tweets with positive 1 and 2 are almost double compared to tweets with negative 1 and 2 respectively. This clearly shows that most people (out of given sample) are in favor of demonetization despite certain unrest.