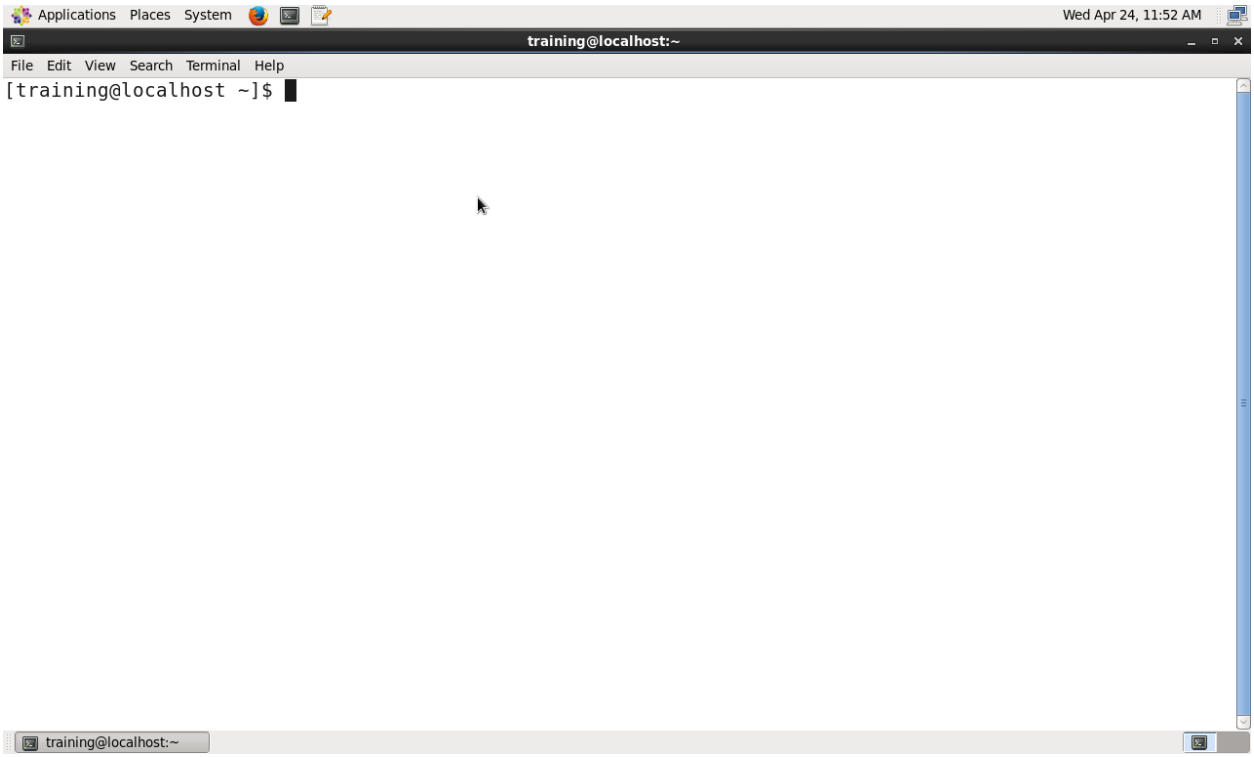
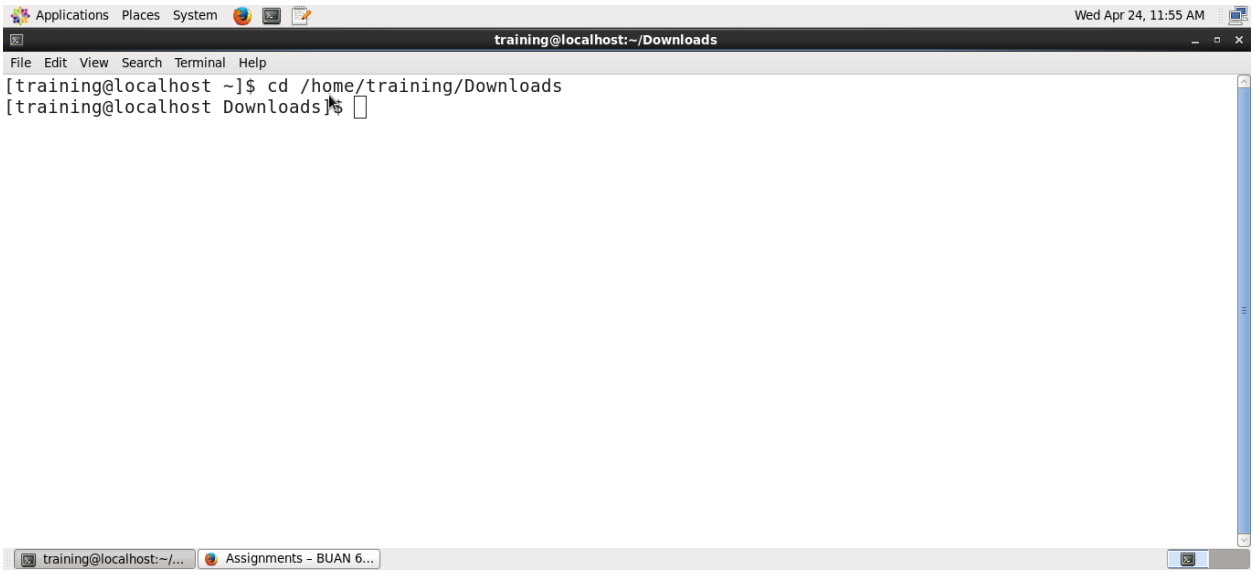
Sentiment Analysis Using Hadoop Tools

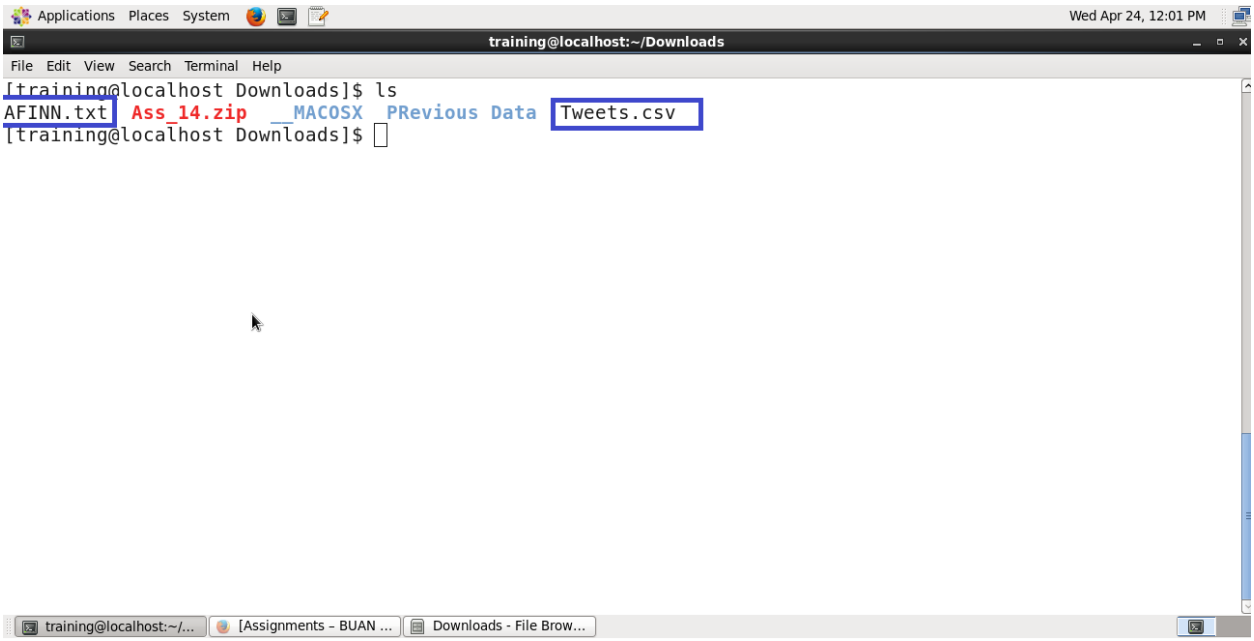
Step-1:



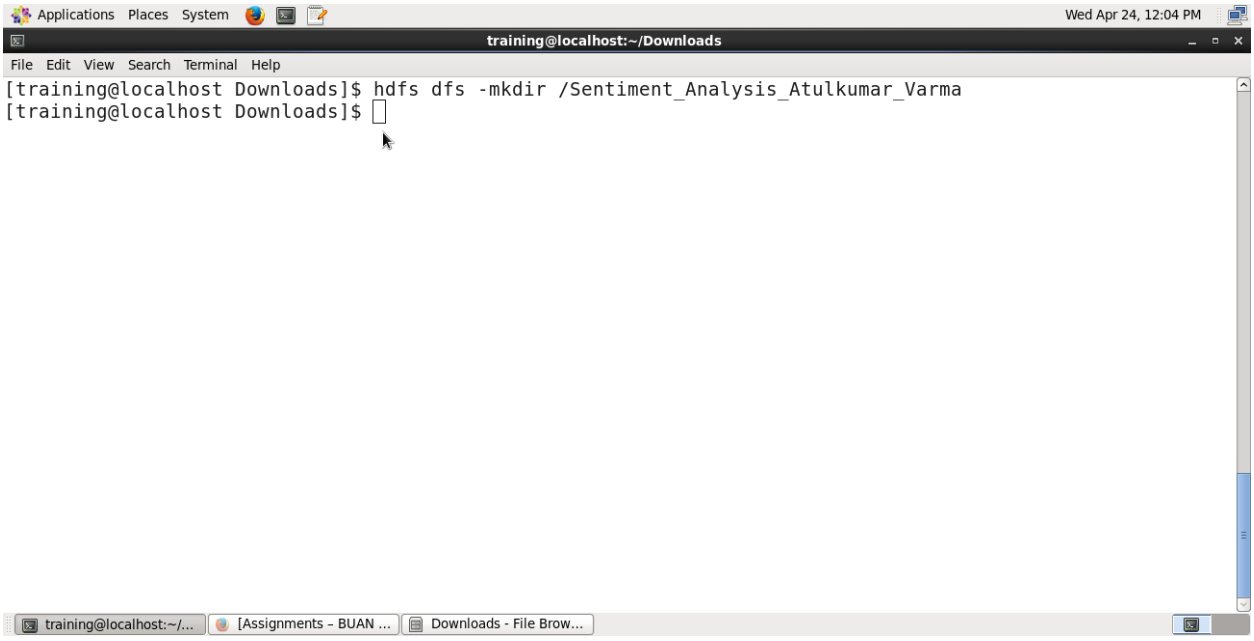
Step-2: Getting the Data from Download folder

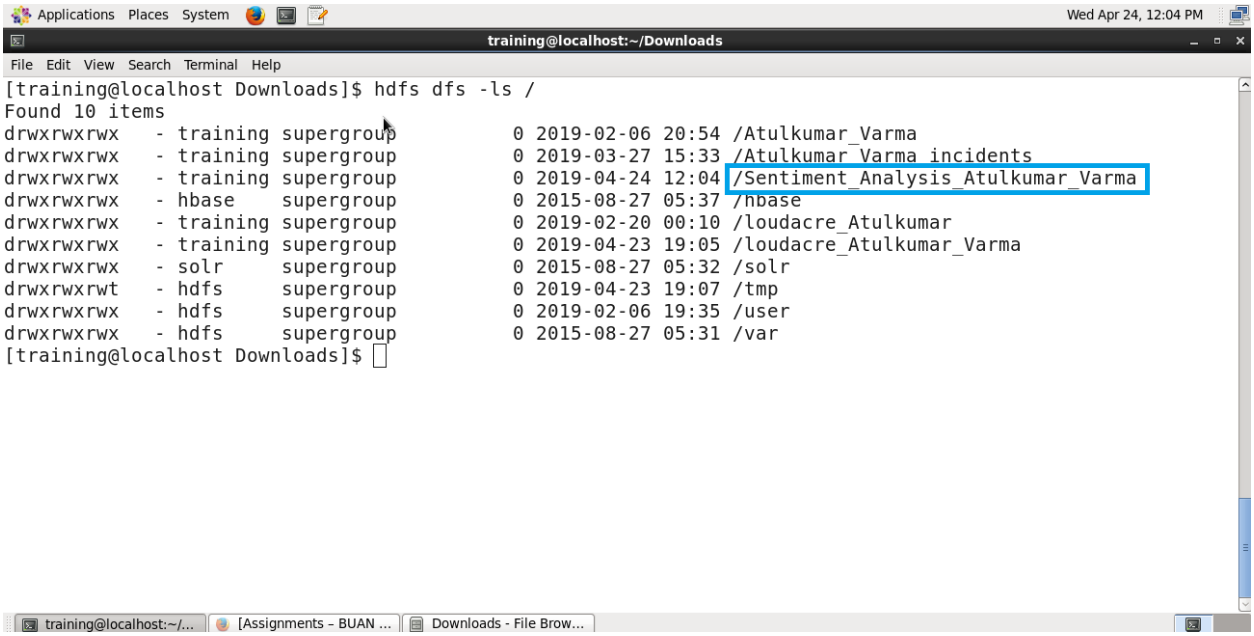


Step-3: Contents of the Downloads folder. Contains all the tweets file.

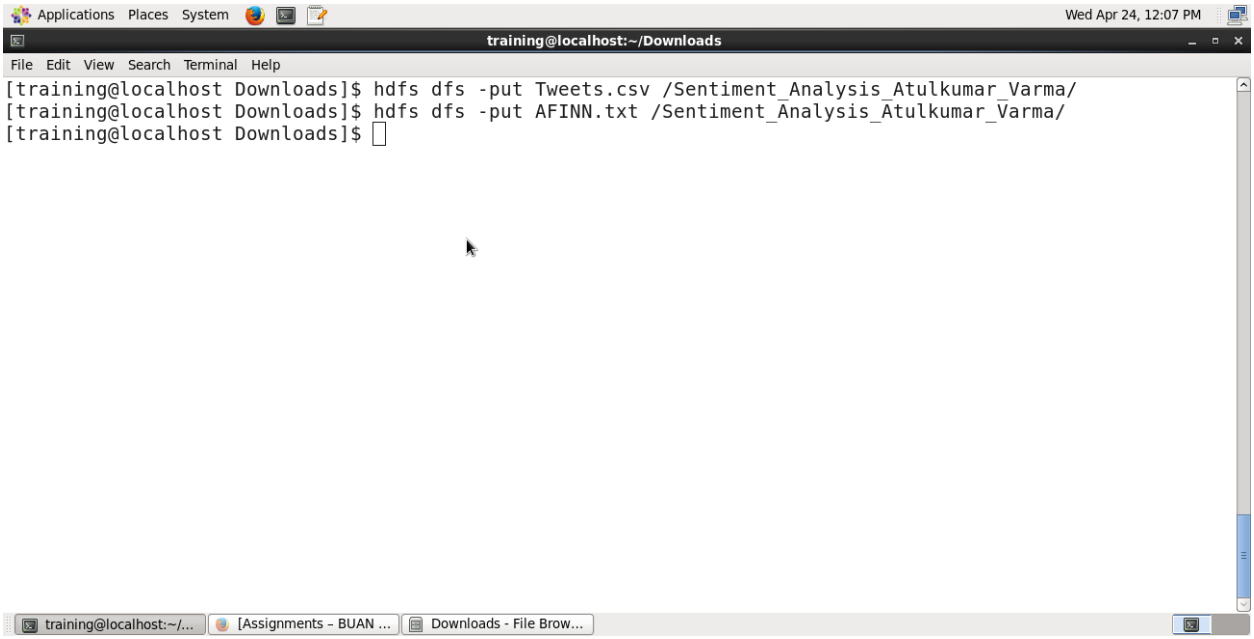


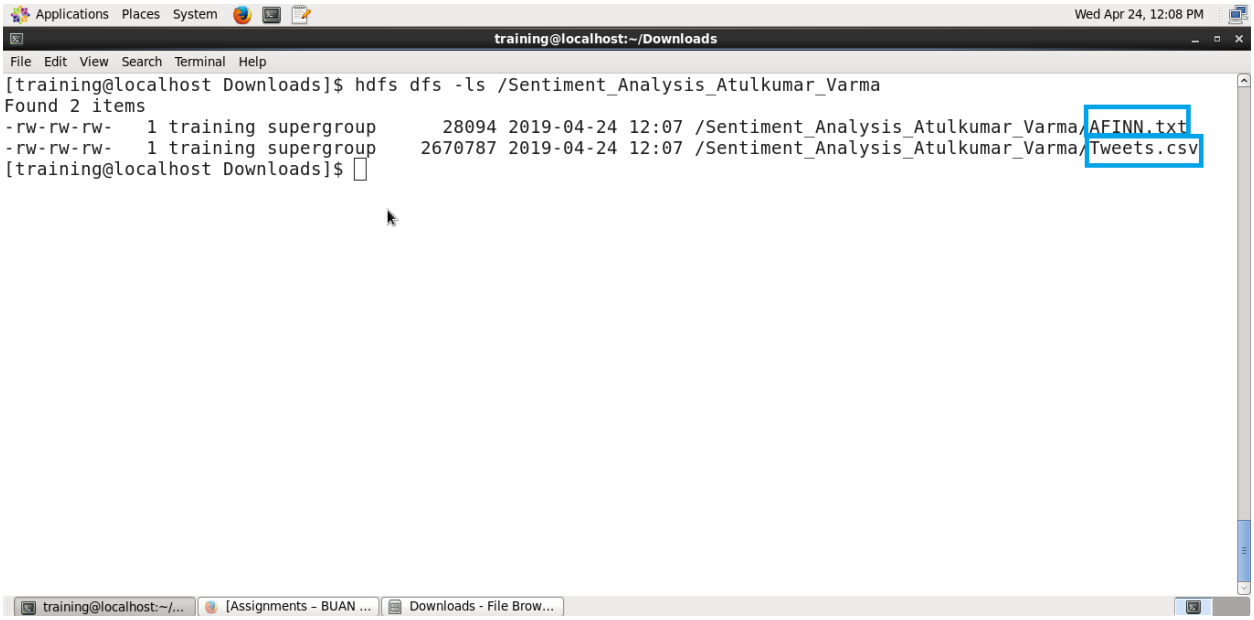
Step-4: We are creating a directory at the root directory.



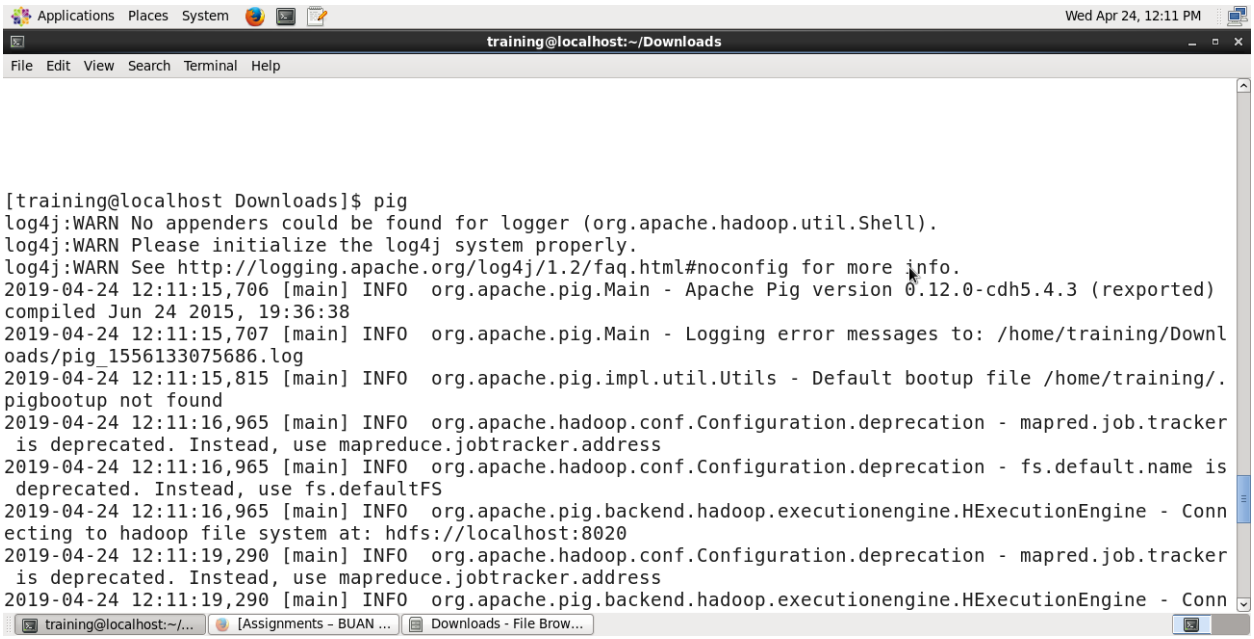


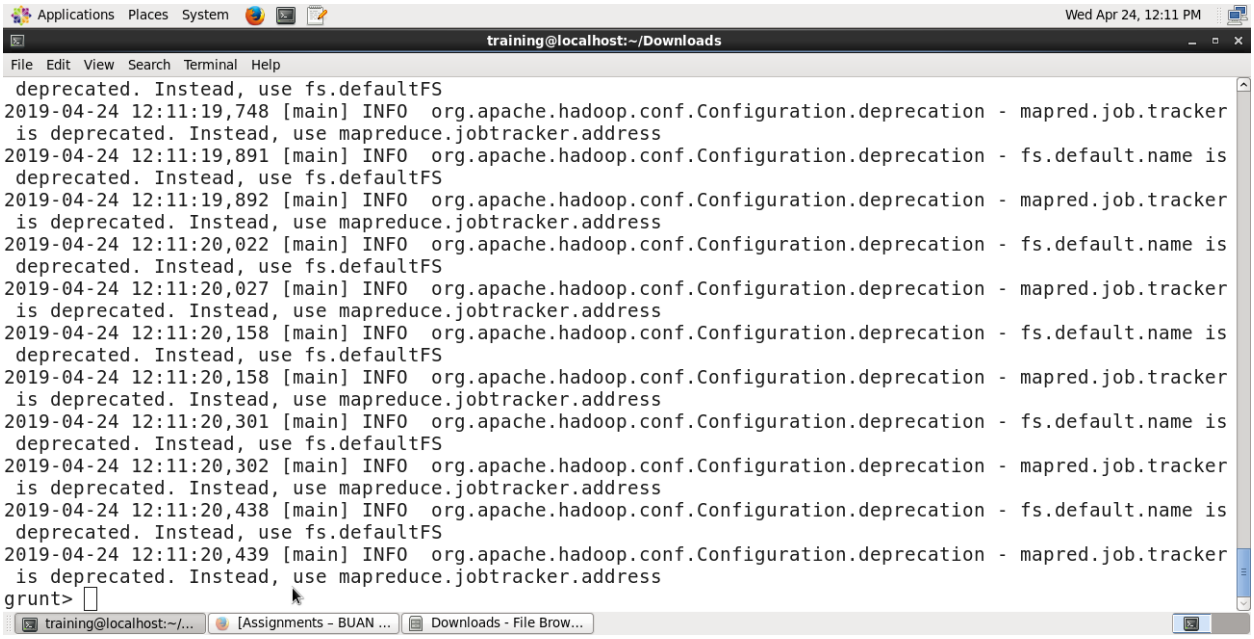
Step-5: Putting the data. We are copying the data files required into the directory that we created. AFINN dictionary is a dictionary consisting of 2500 words that are rated on a scale from -5 to +5 depending on the sentiment the word conveys.



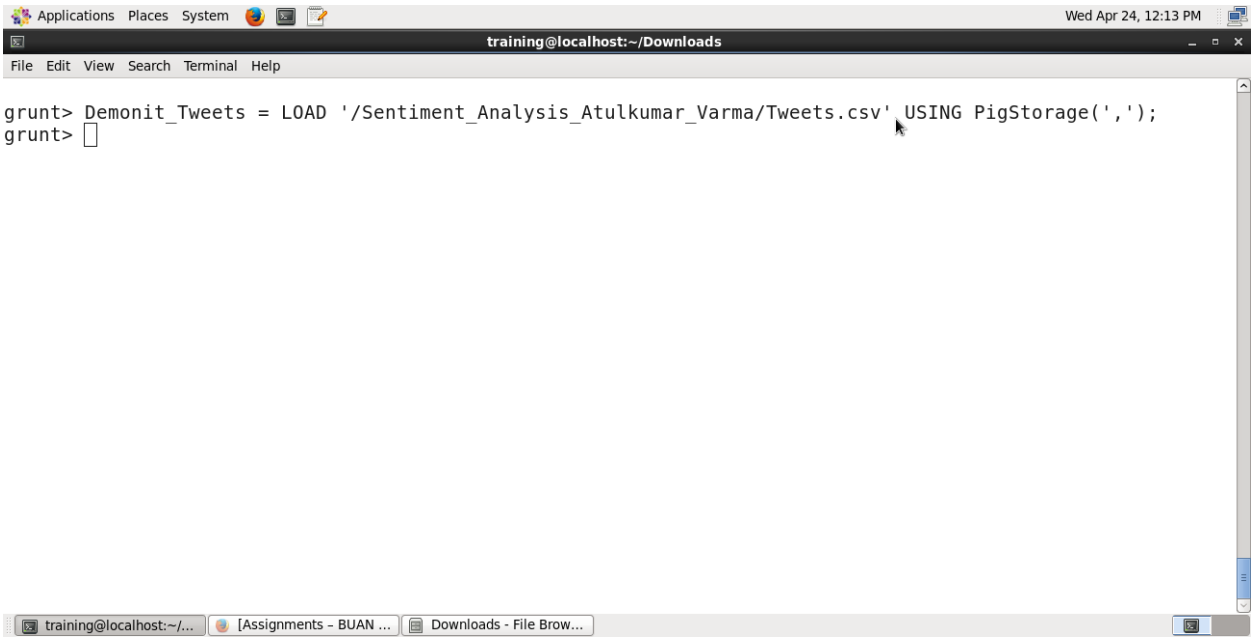


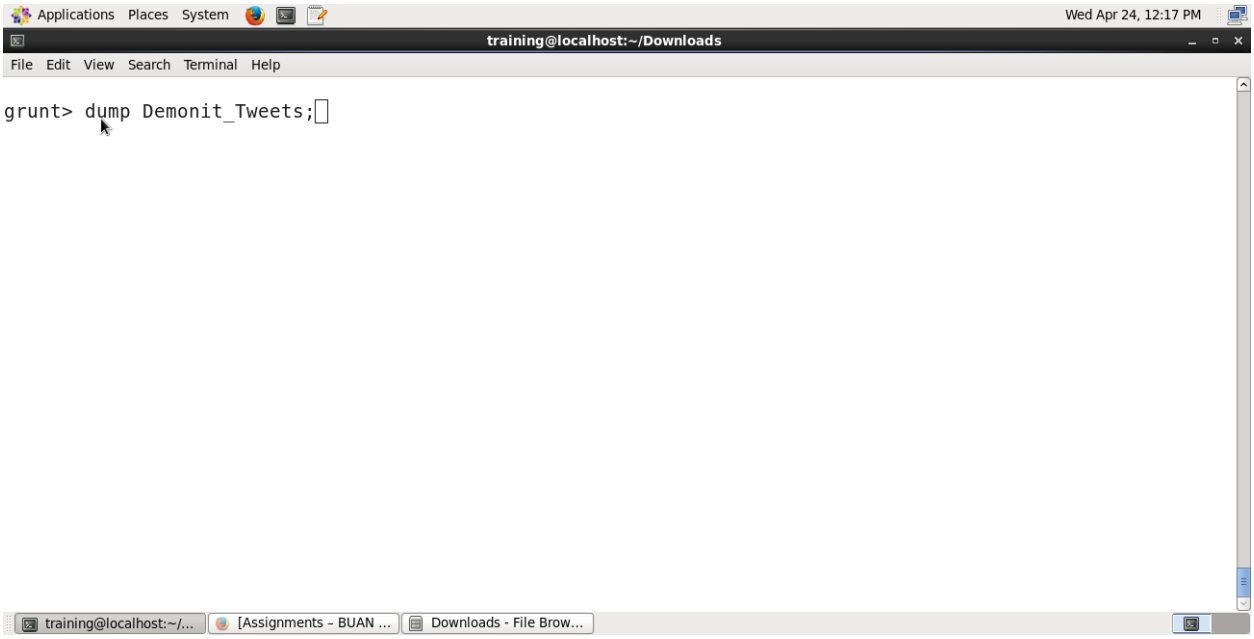
Step-6: Pig command invokes the grunt shell.

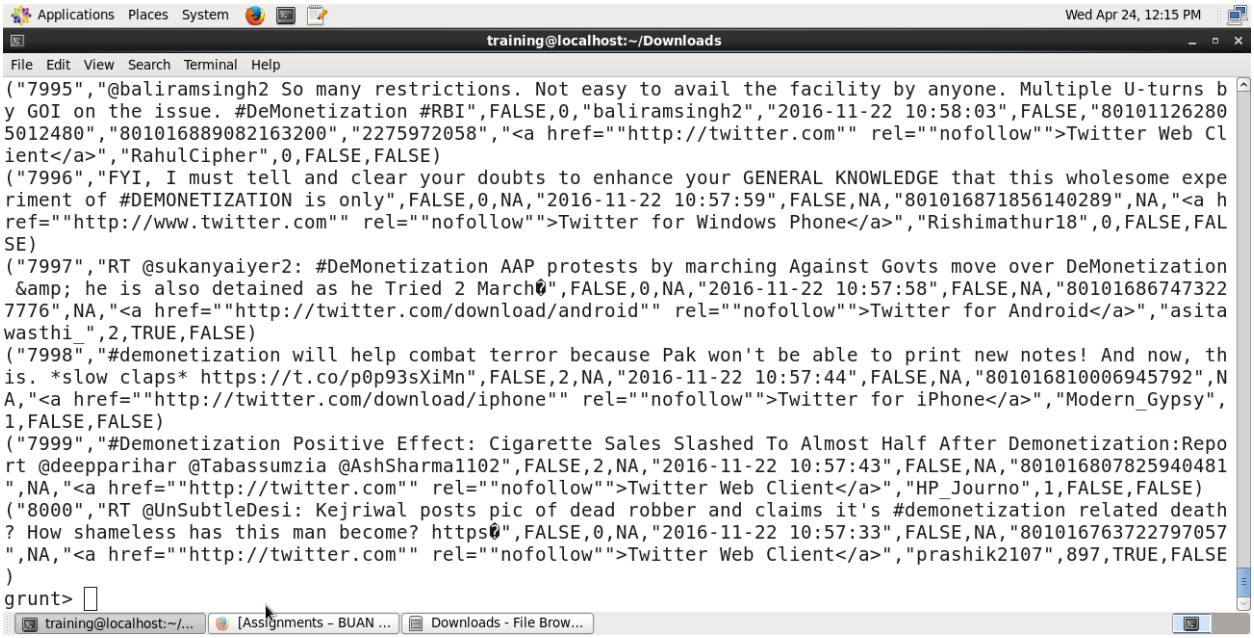




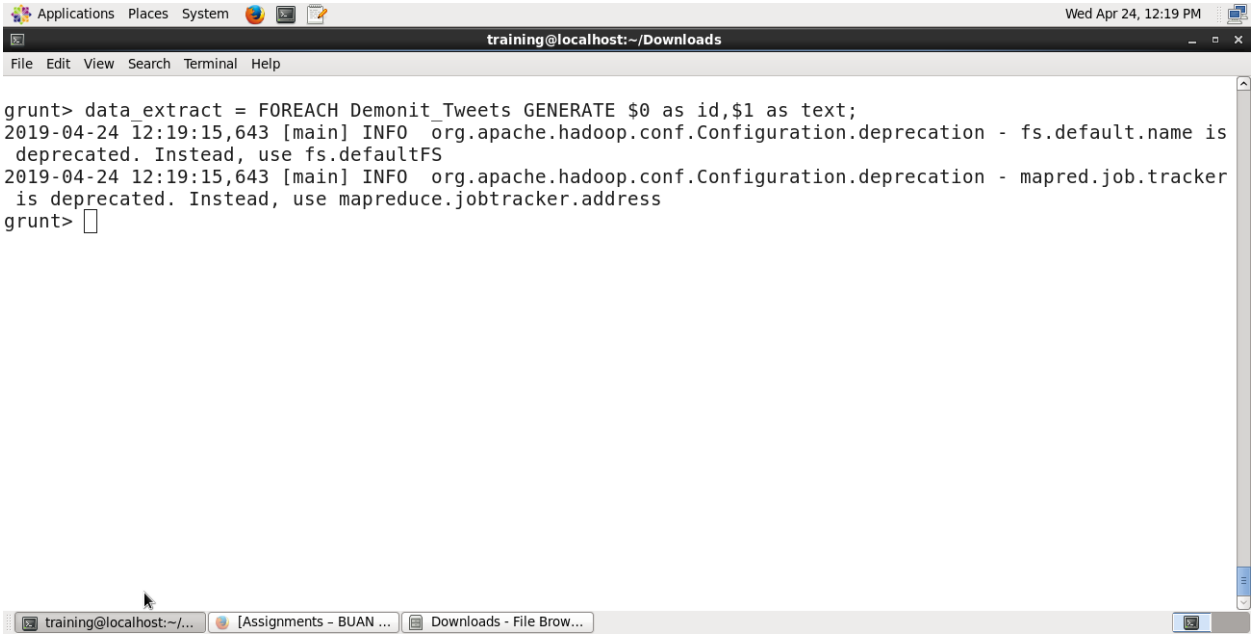
Step-7: The dump command displays the data for the variable Demonit_Tweets specified in the grunt shell.

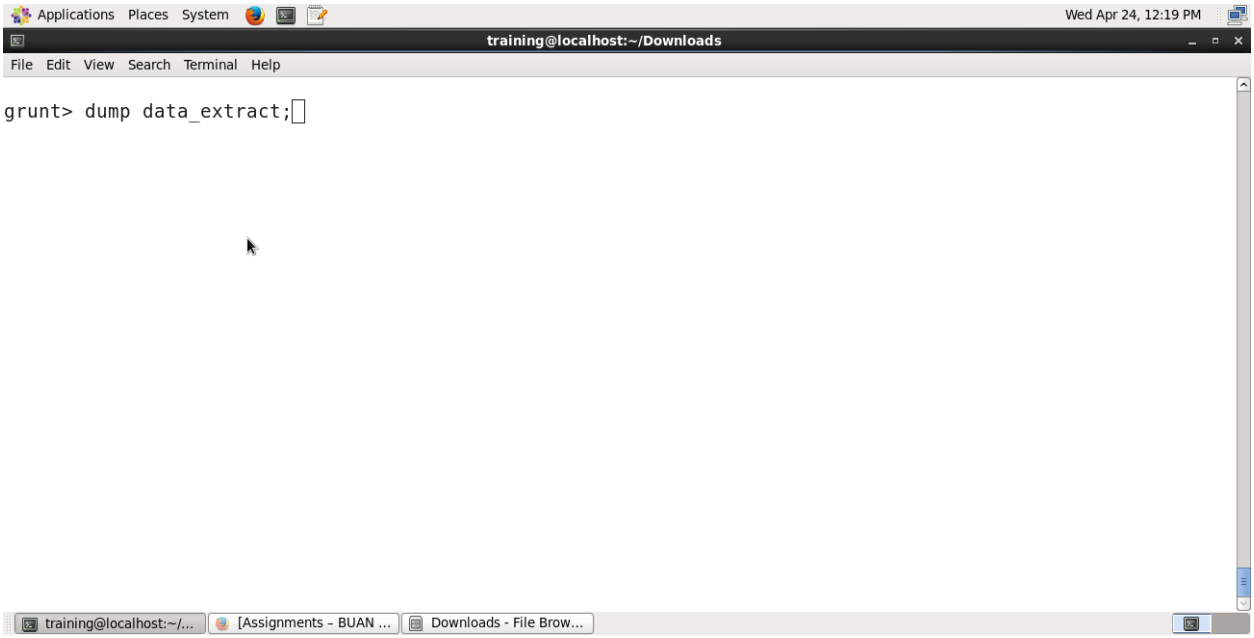


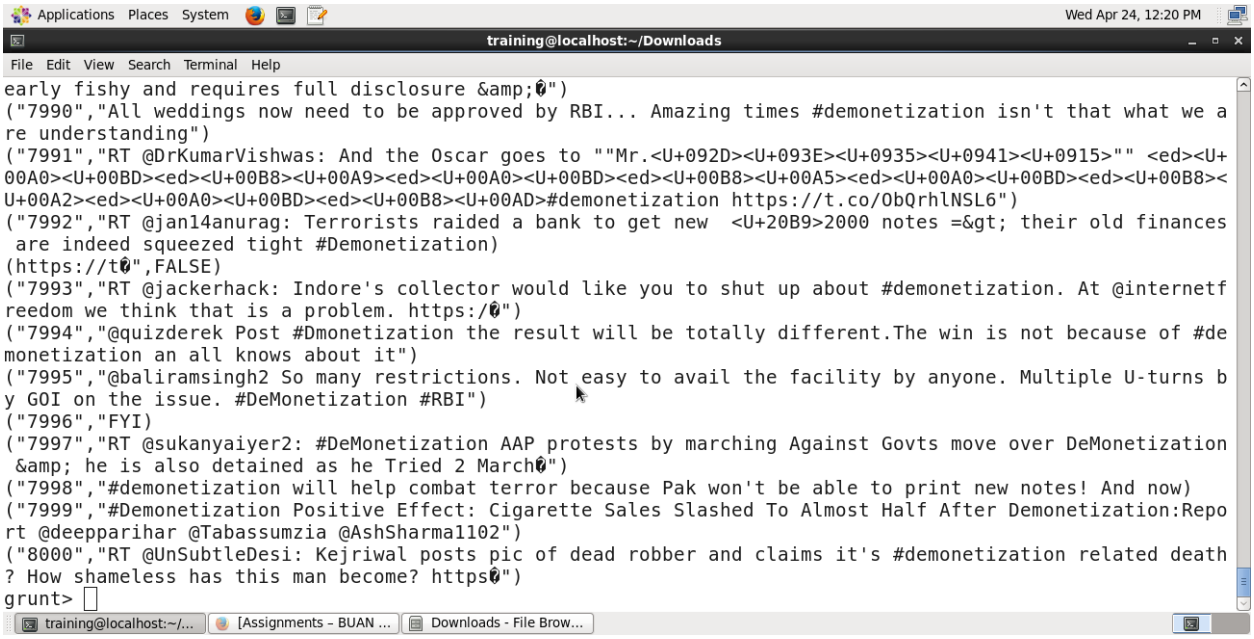

Applications Places System
Wed Apr 24, 12:17 PM
training@localhost:~/Downloads
File Edit View Search Terminal Help
grunt> dump Demonit_Tweets;
training@localhost:~/...
[Assignments – BUAN ...
Downloads - File Brow...

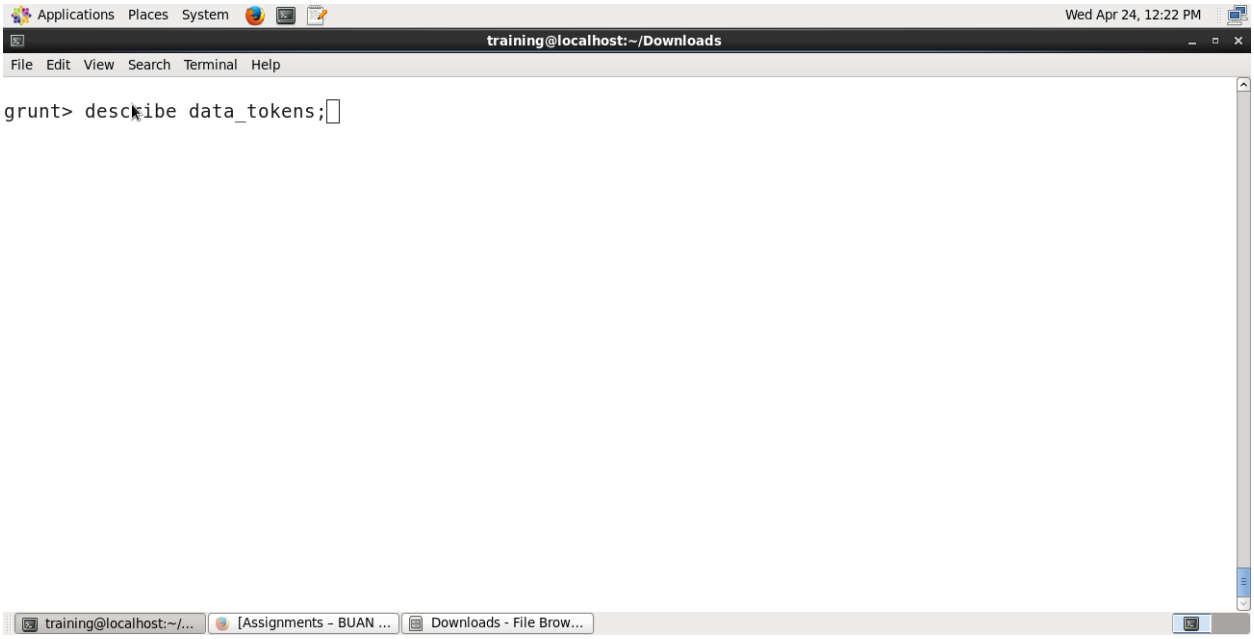

Applications Places System
Wed Apr 24, 12:15 PM
training@localhost:~/Downloads
File Edit View Search Terminal Help
("7995","@baliramsingh2 So many restrictions. Not easy to avail the facility by anyone. Multiple U-turns by GOI on the issue. #DeMonetization #RBI",FALSE,0,"baliramsingh2","2016-11-22 10:58:03",FALSE,"801011262805012480","801016889082163200","2275972058","<a href=""http://twitter.com"" rel=""nofollow"">Twitter Web Client</a>","RahulCipher",0,FALSE,FALSE)
("7996","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,"801016871856140289",NA,"<a href=""http://www.twitter.com"" rel=""nofollow"">Twitter for Windows Phone</a>","Rishimathur18",0,FALSE,FALSE)
("7997","RT @sukanyaiyer2: #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,"801016867473227776",NA,"<a href=""http://twitter.com/download/android"" rel=""nofollow"">Twitter for Android</a>","asitawasthi_",2,TRUE,FALSE)
("7998","#demonetization will help combat terror because Pak won't be able to print new notes! And now, this. *slow claps* https://t.co/p0p93sXiMn",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 @deepparihar @Tabassumzia @AshSharma1102",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>
training@localhost:~/...
[Assignments – BUAN ...
Downloads - File Brow...

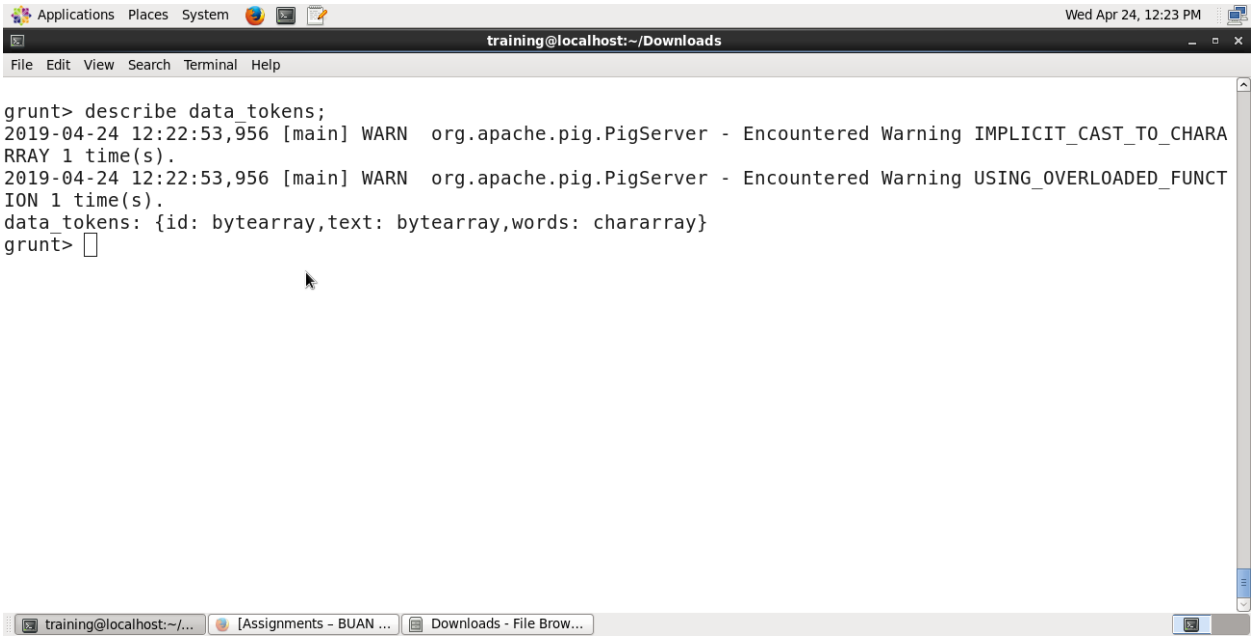
Step-8: We are defining and collecting the first two data elements in the csv file and dumping the data at the grunt shell.



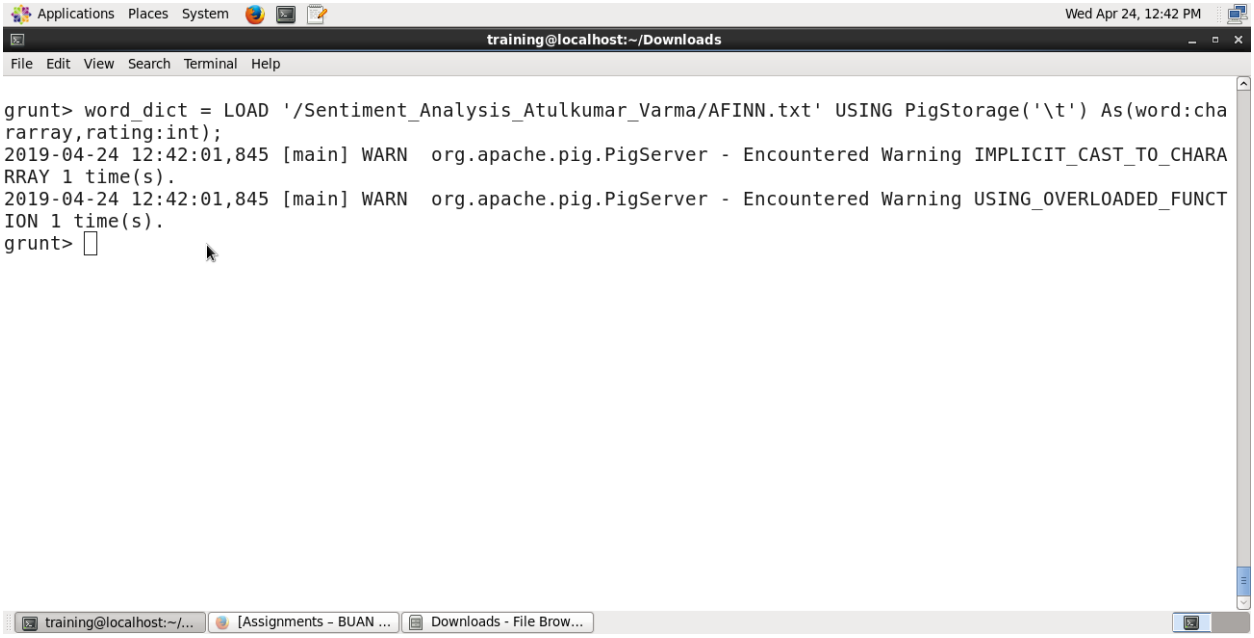


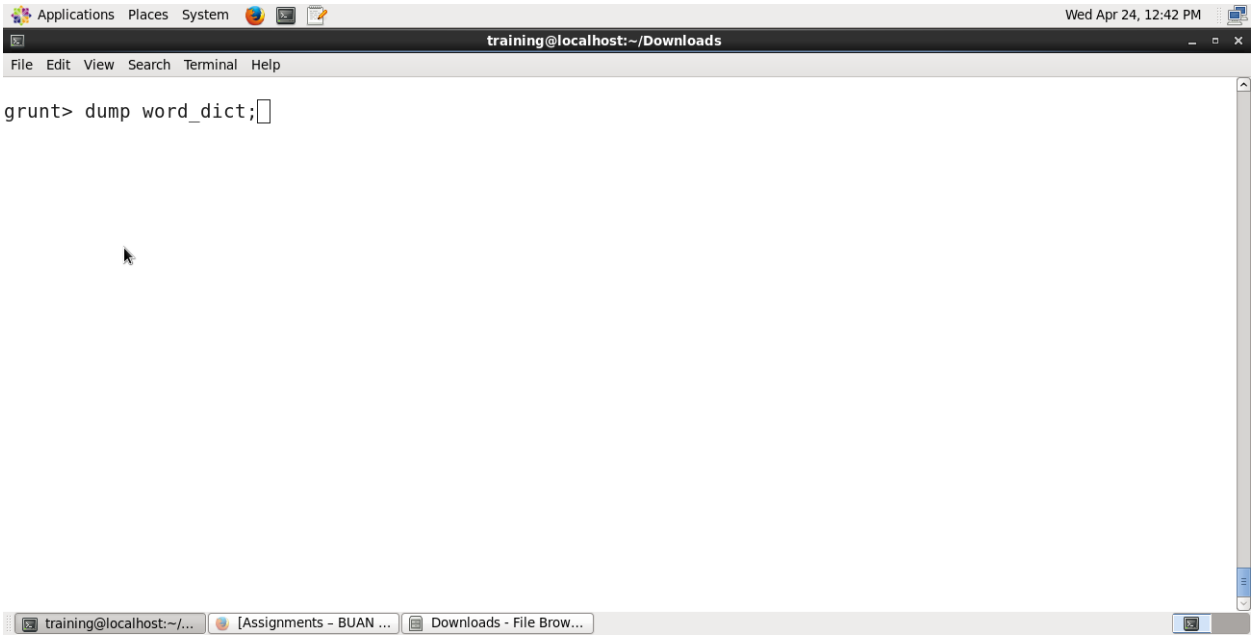

Applications Places System
Wed Apr 24, 12:20 PM
training@localhost:~/Downloads
File Edit View Search Terminal Help
early fishy and requires full disclosure &

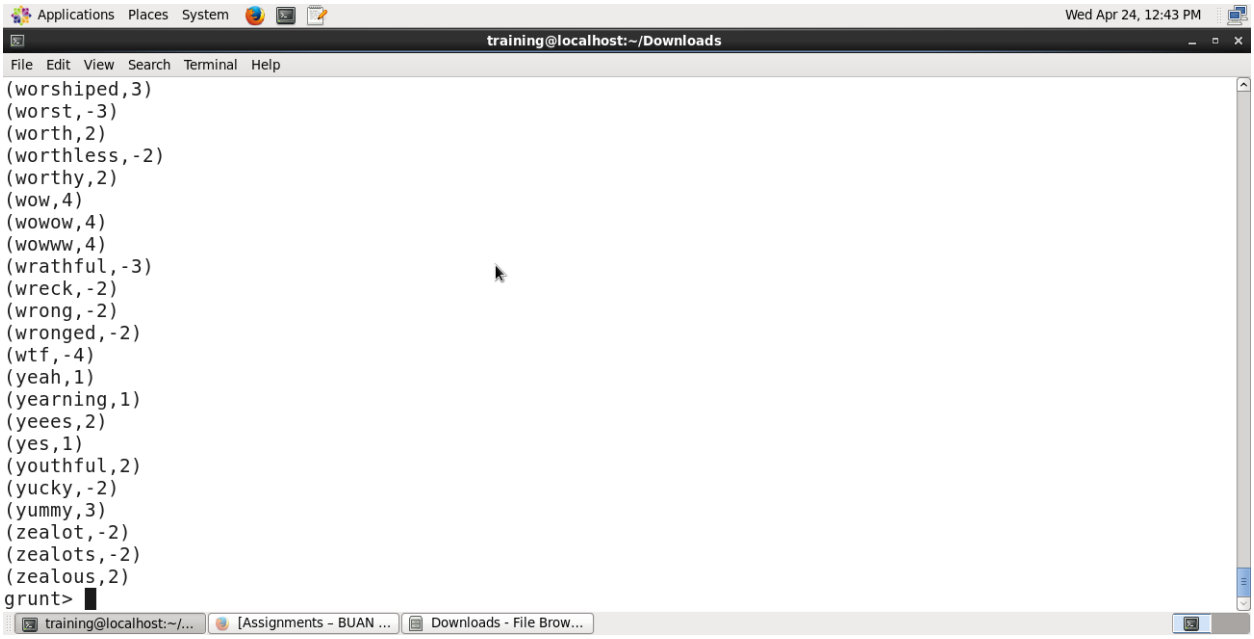

Applications Places System
Wed Apr 24, 12:22 PM
training@localhost:~/Downloads
File Edit View Search Terminal Help
grunt> describe data_tokens;
training@localhost:~/...
[Assignments – BUAN ...
Downloads - File Brow...


Applications Places System
Wed Apr 24, 12:23 PM
training@localhost:~/Downloads
File Edit View Search Terminal Help
grunt> describe data_tokens;
2019-04-24 12:22:53,956 [main] WARN org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_CHARARRAY 1 time(s).
2019-04-24 12:22:53,956 [main] WARN org.apache.pig.PigServer - Encountered Warning USING_OVERLOADED_FUNCTION 1 time(s).
data_tokens: {id: bytearray,text: bytearray,words: chararray}
grunt>
training@localhost:~/...
[Assignments – BUAN ...
Downloads - File Brow...

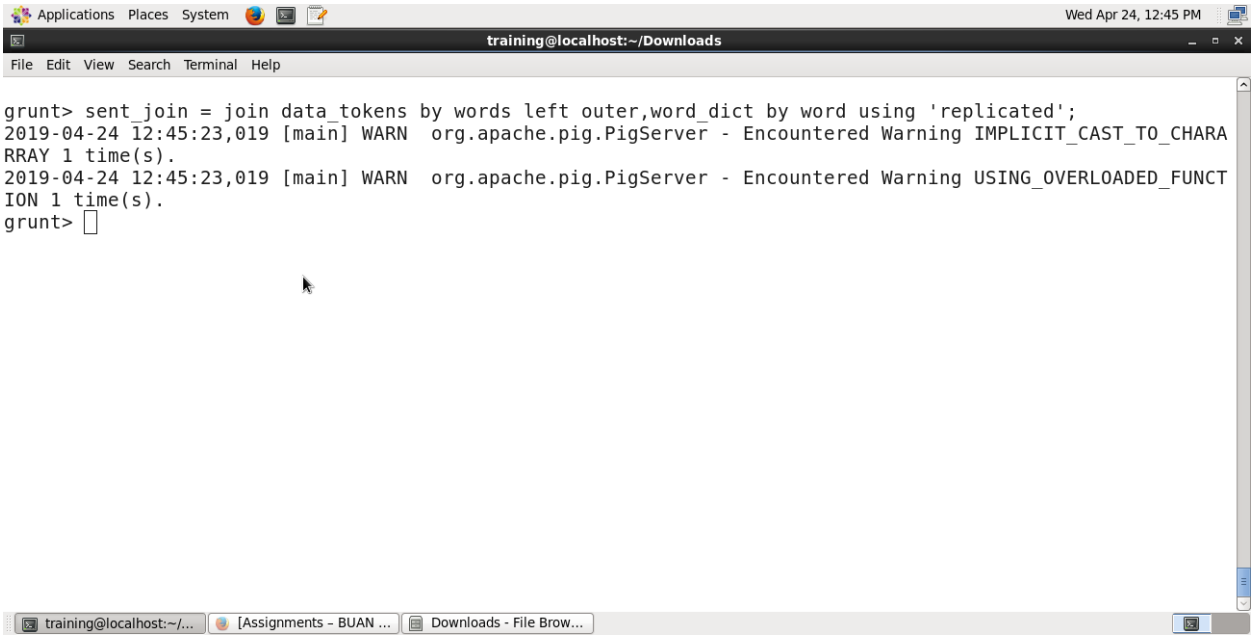
Step-10: We will be using AFINN dictionary to perform the sentiment analysis on our tweet data.

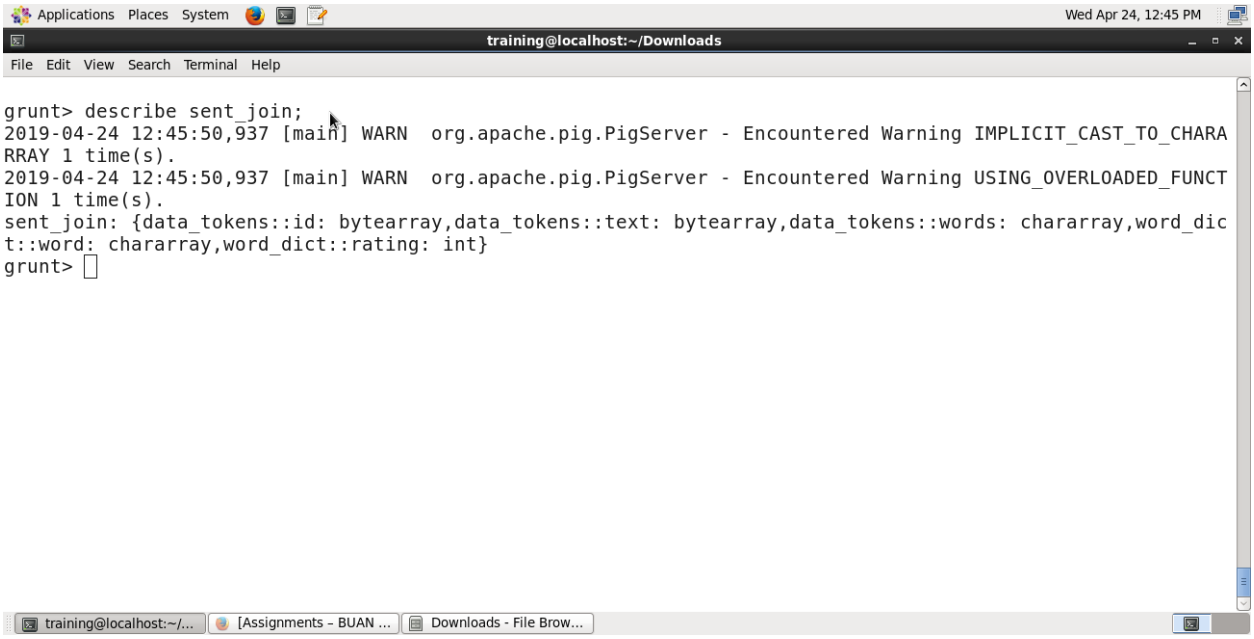




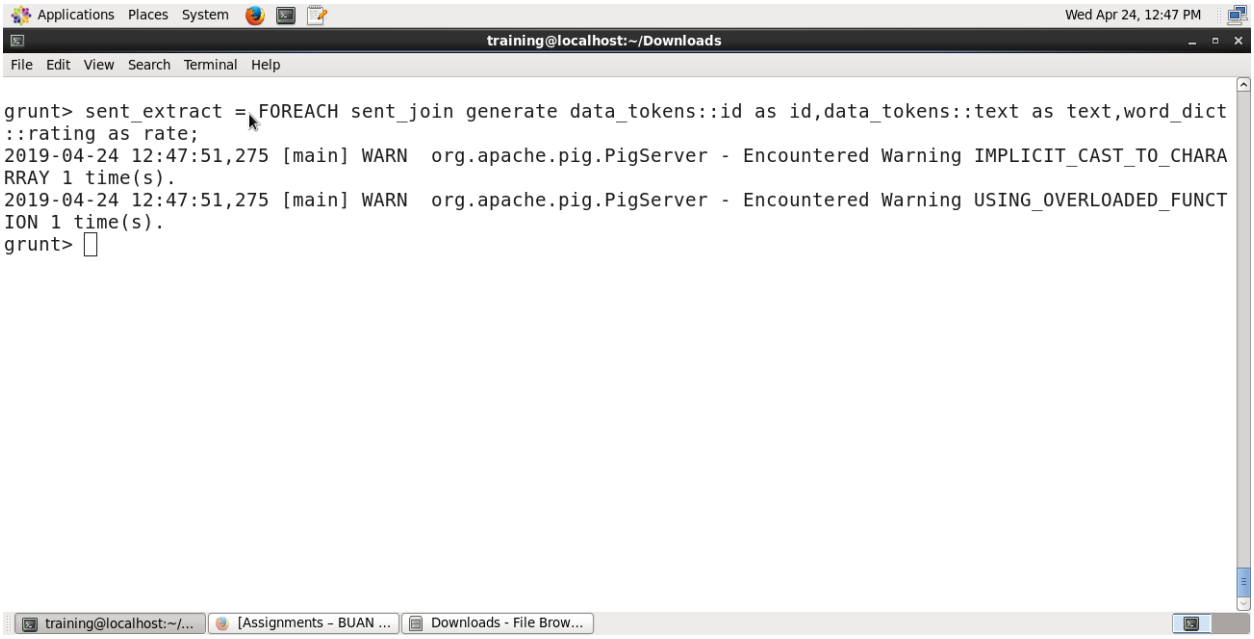


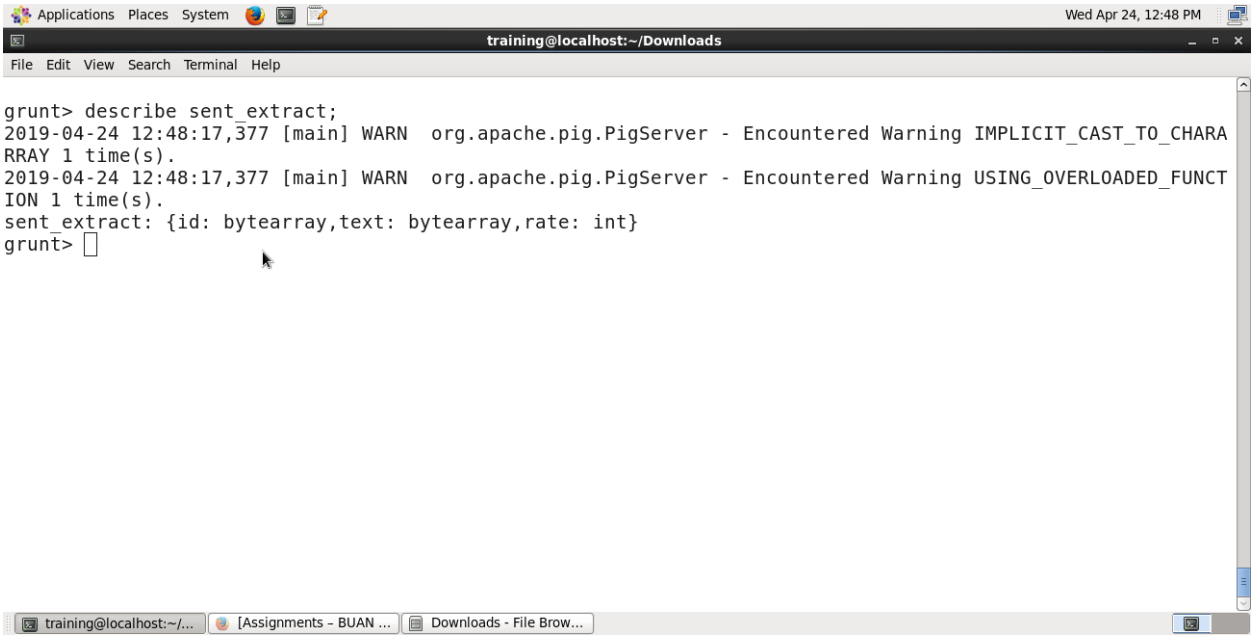
Step-11: We are joining the sentiments values from the AFINN dictionary with the words in the text from each tweet.



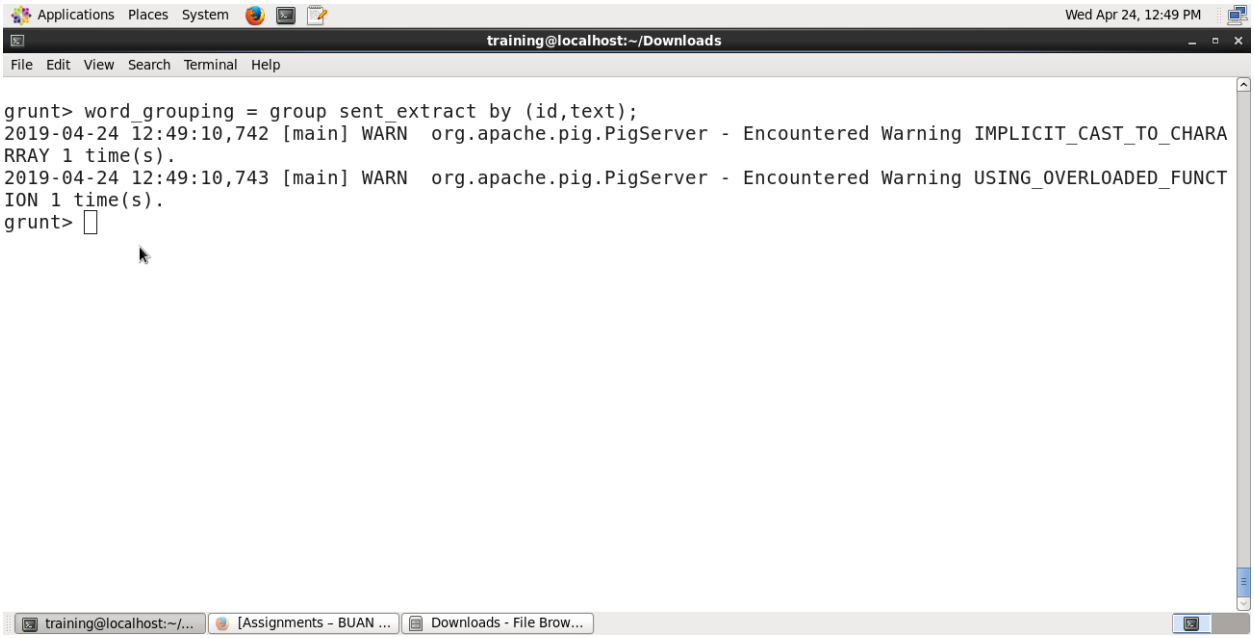


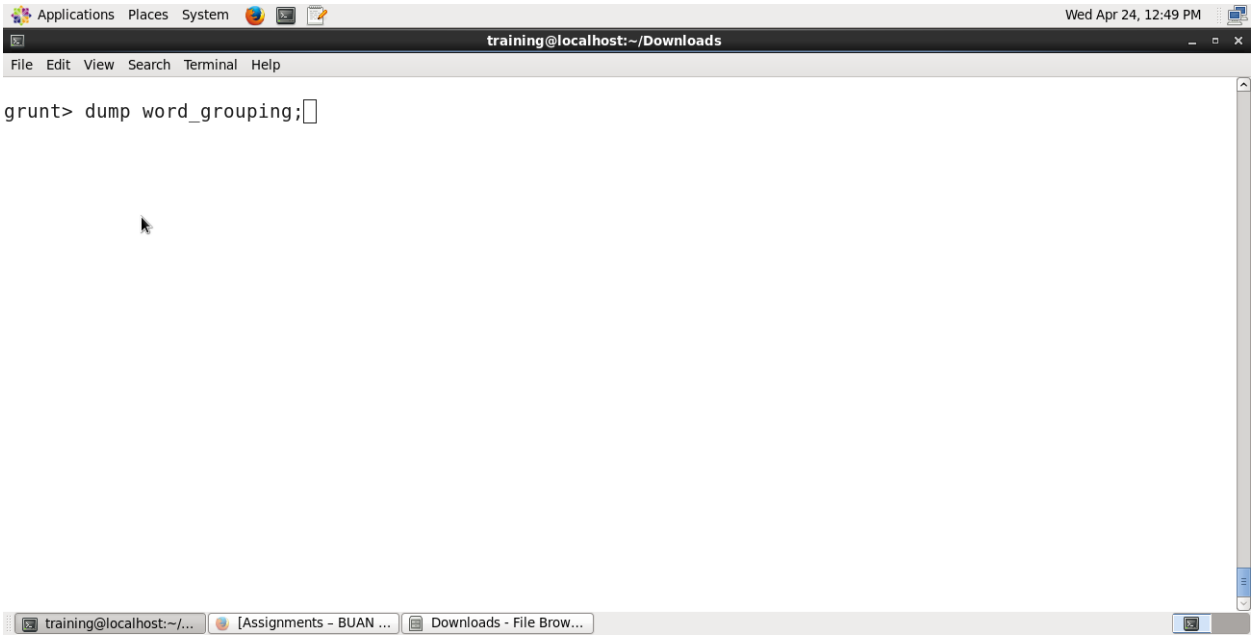
Step-12: : We are now extracting the data from the variable in step 11 that we need to complete our sentiment analysis.





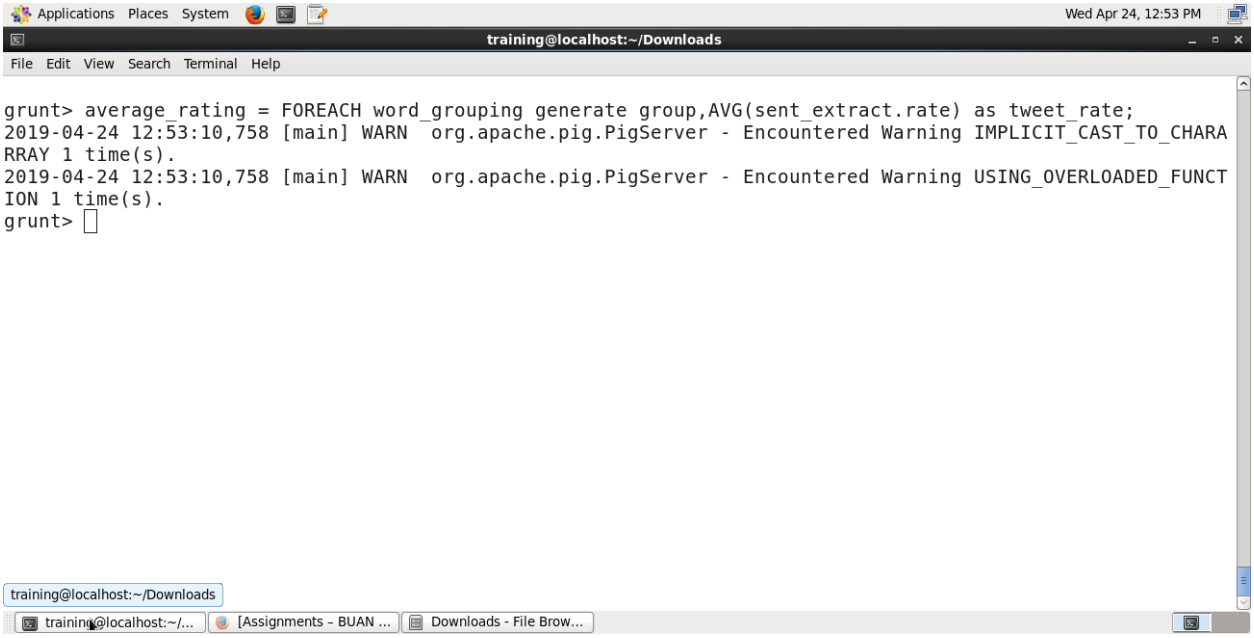
Step-13: We are grouping each record by its id and text.

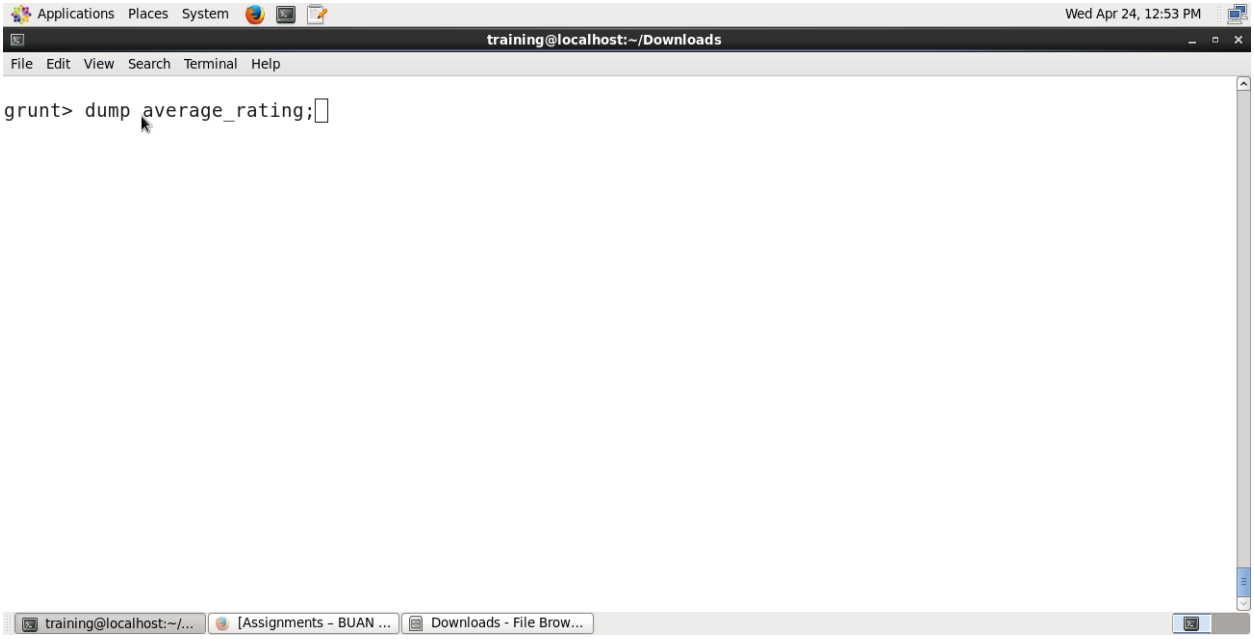


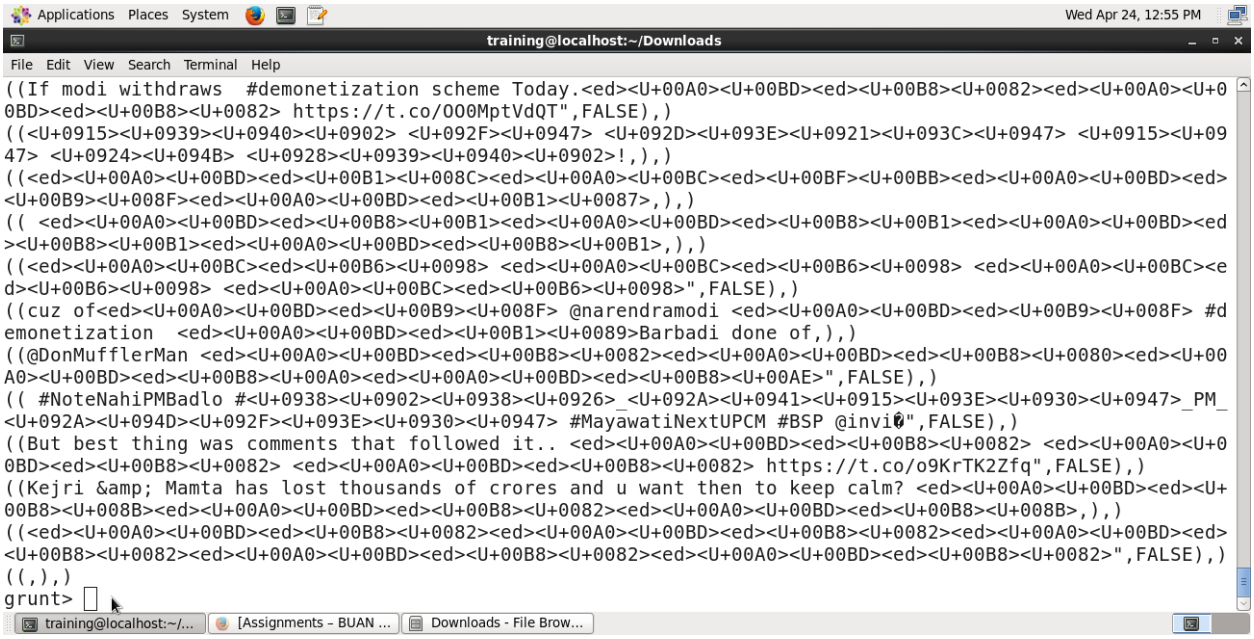

Applications Places System
Wed Apr 24, 12:49 PM
training@localhost:~/Downloads
File Edit View Search Terminal Help
grunt> dump word_grouping;
training@localhost:~/...
[Assignments – BUAN ...
Downloads - File Brow...

Applications Places System
Wed Apr 24, 12:51 PM
training@localhost:~/Downloads
File Edit View Search Terminal Help
grunt>
training@localhost:~/...
[Assignments – BUAN ...
Downloads - File Brow...

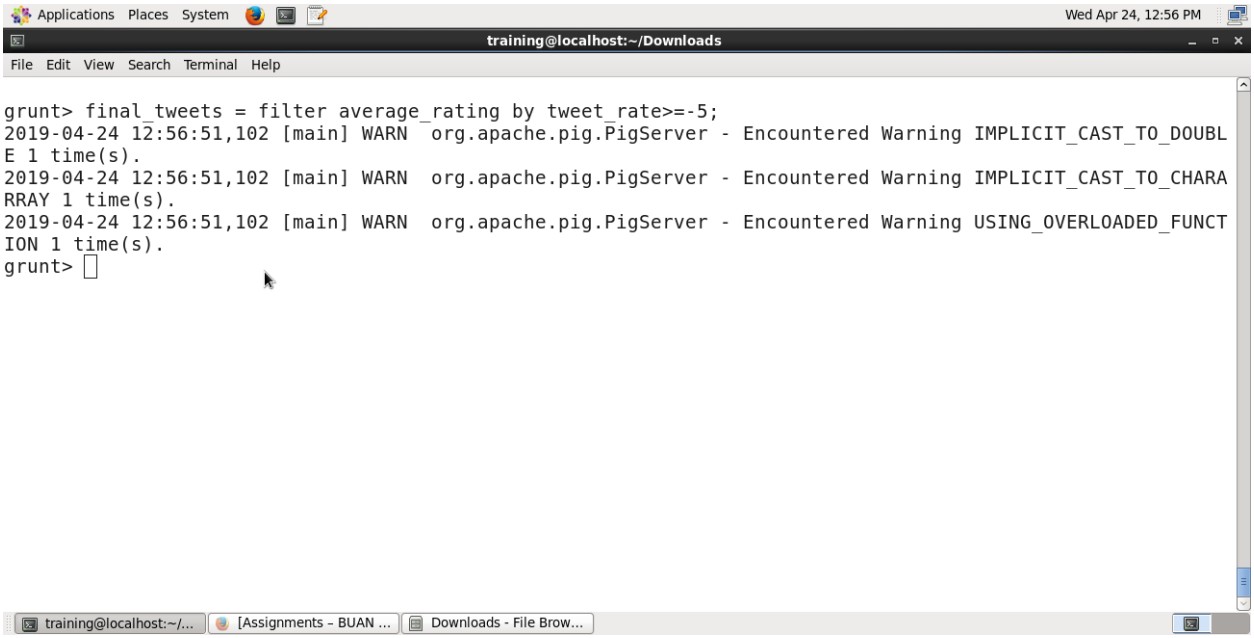
Step-14: We are averaging the rating for all the words in each tweet and storing the result in average_rating.

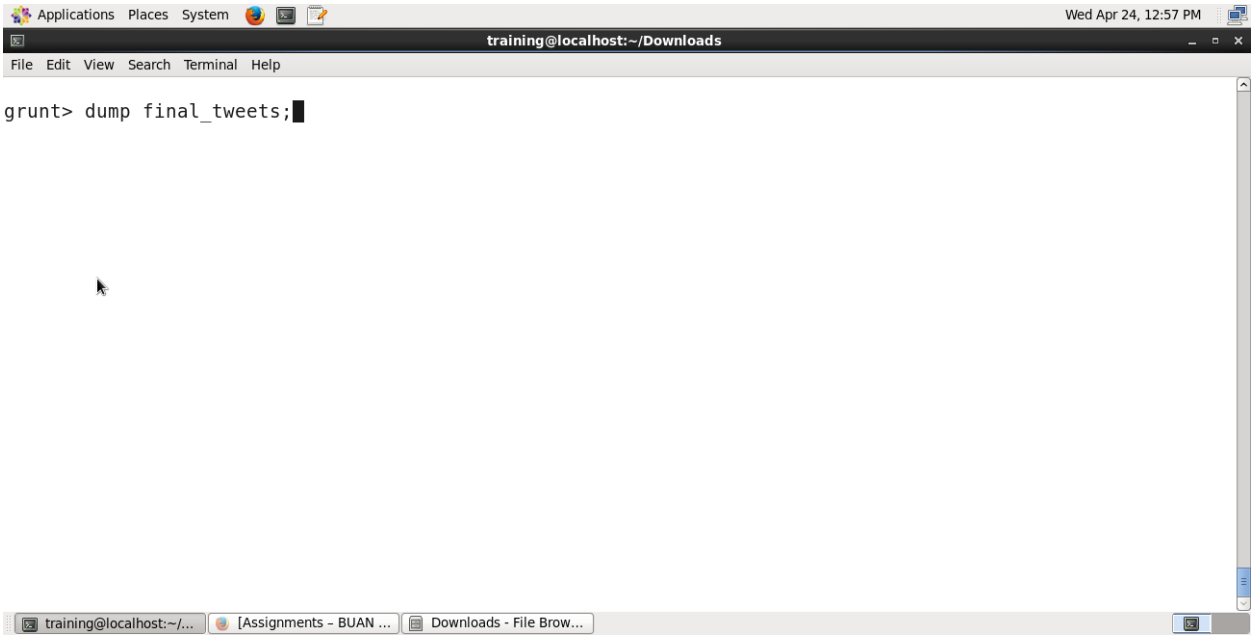


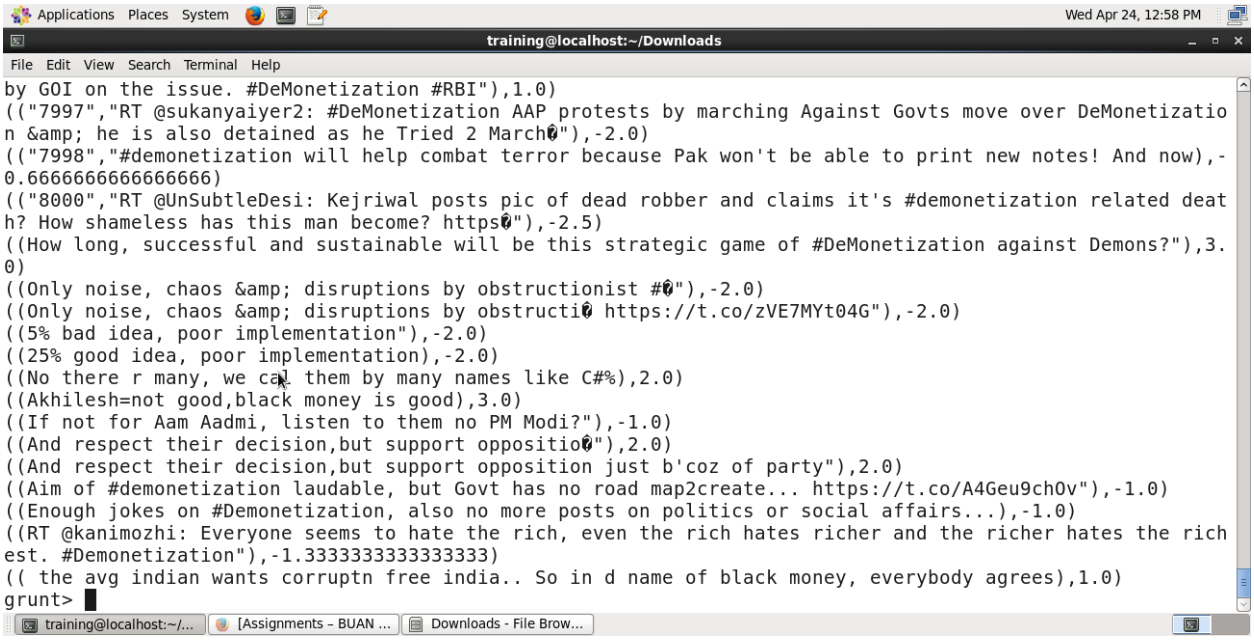




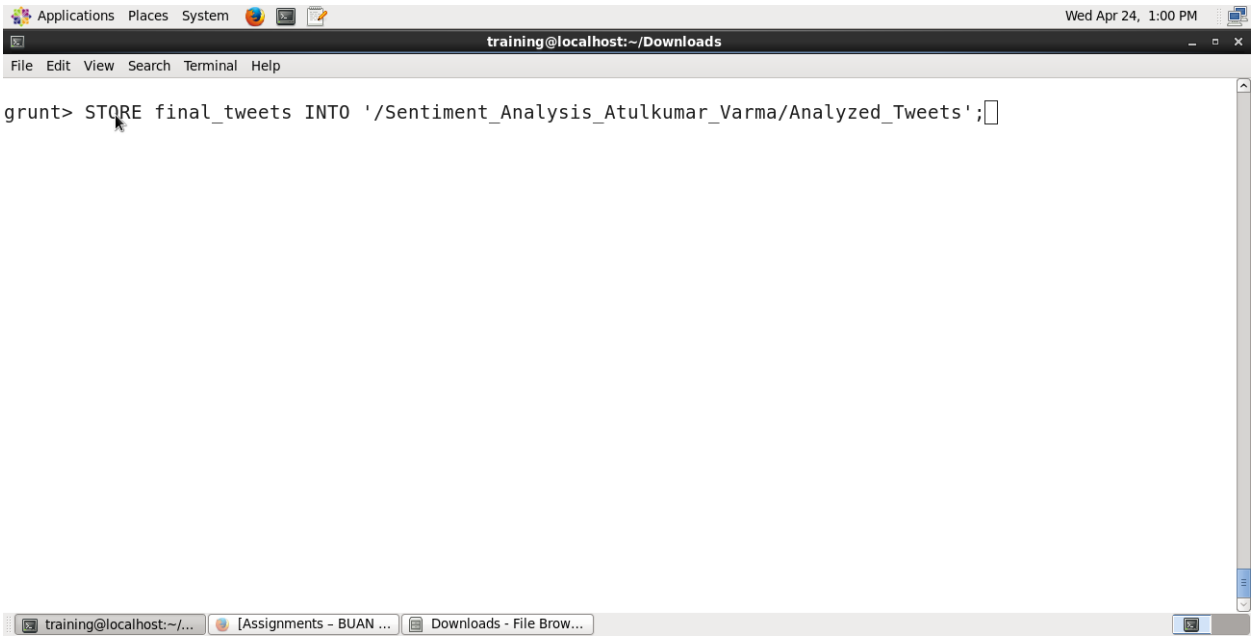
Step-15: Filtered all the tweets for those tweets that have an average rating >= -5

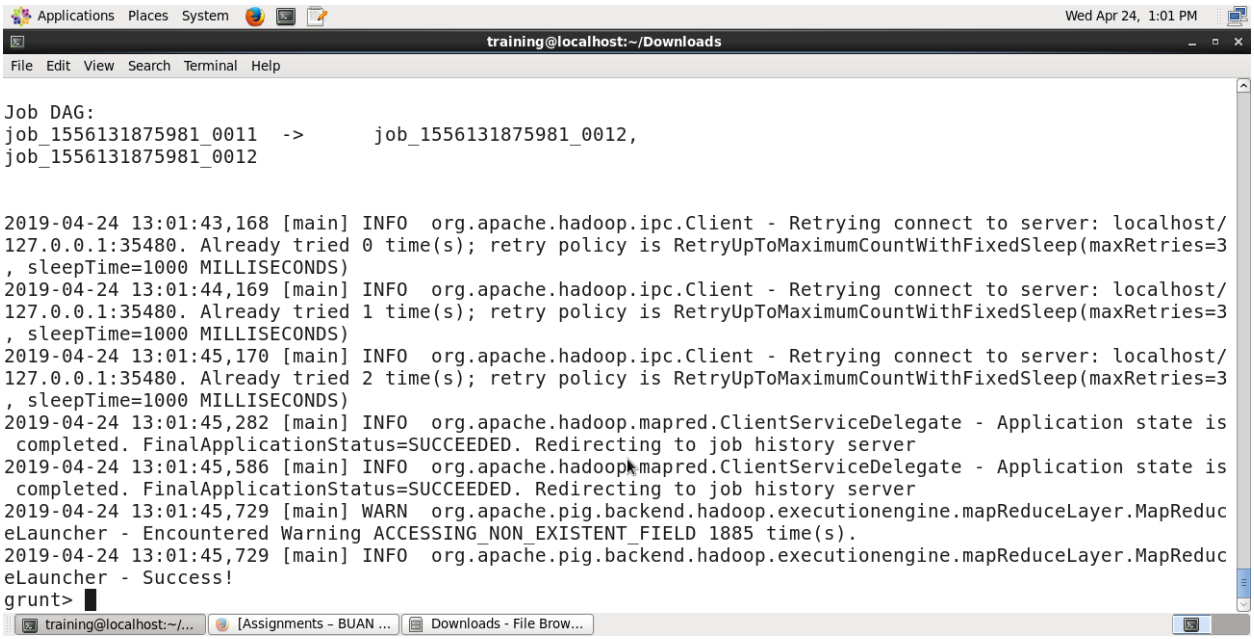


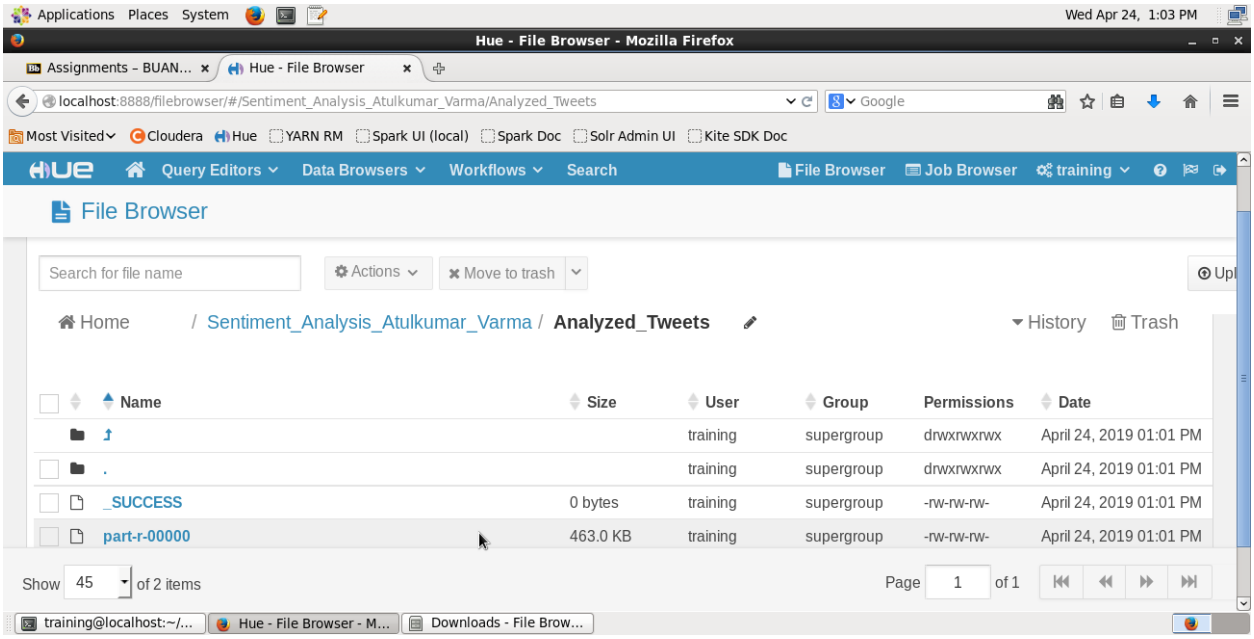

Applications Places System
Wed Apr 24, 12:57 PM
training@localhost:~/Downloads
File Edit View Search Terminal Help
grunt> dump final_tweets;
training@localhost:~/...
[Assignments – BUAN ...
Downloads - File Brow...


Applications Places System
Wed Apr 24, 12:58 PM
training@localhost:~/Downloads
File Edit View Search Terminal Help
by GOI on the issue. #DeMonetization #RBI"),1.0)
(("7997","RT @sukanyaiyer2: #DeMonetization AAP protests by marching Against Govts move over DeMonetizatio
n & 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),-
0.6666666666666666)
(("8000","RT @UnSubtleDesi: Kejriwal posts pic of dead robber and claims it's #demonetization related deat
h? 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 obstructi https://t.co/zVE7MYt04G"),-2.0)
((5% bad idea, poor implementation"),-2.0)
((25% good idea, poor implementation),-2.0)
((No there r many, we call 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/A4Geu9chOv"),-1.0)
((Enough jokes on #Demonetization, also no more posts on politics or social affairs...),-1.0)
((RT @kanimozhi: Everyone seems to hate the rich, even the rich hates richer and the richer hates the rich
est. #Demonetization"),-1.3333333333333333)
((the avg indian wants corruptn free india.. So in d name of black money, everybody agrees),1.0)
grunt>
training@localhost:~/...
[Assignments – BUAN ...
Downloads - File Brow...

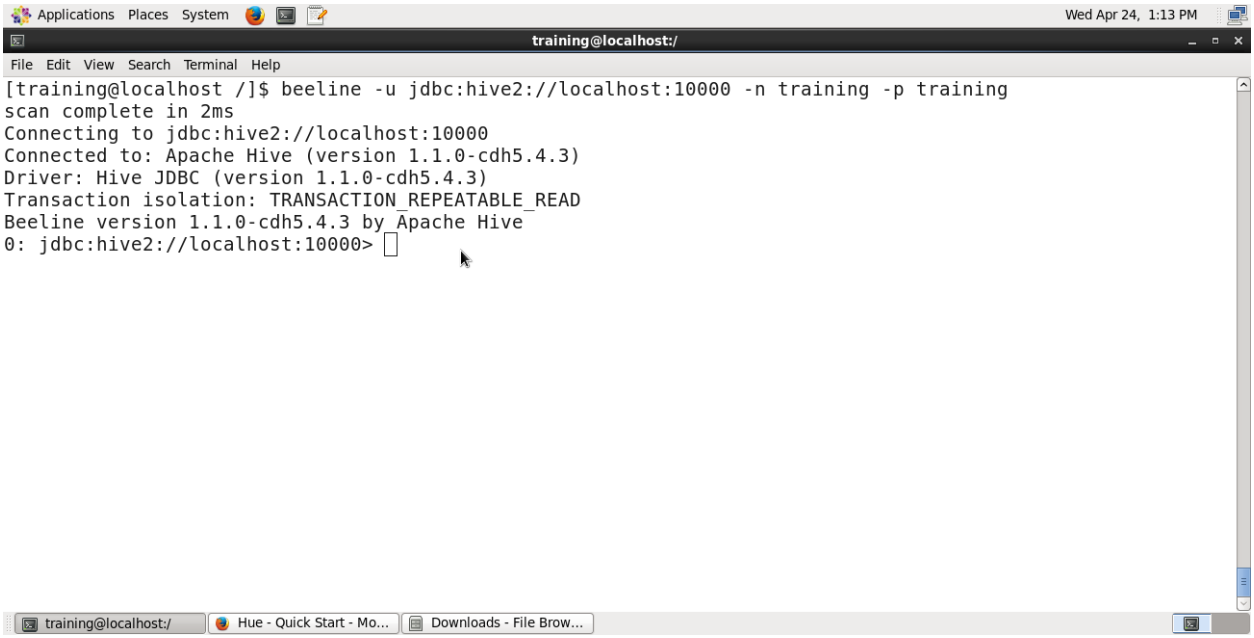
Step-16: We are viewing the output of the Analyzed_Tweets file in HUE



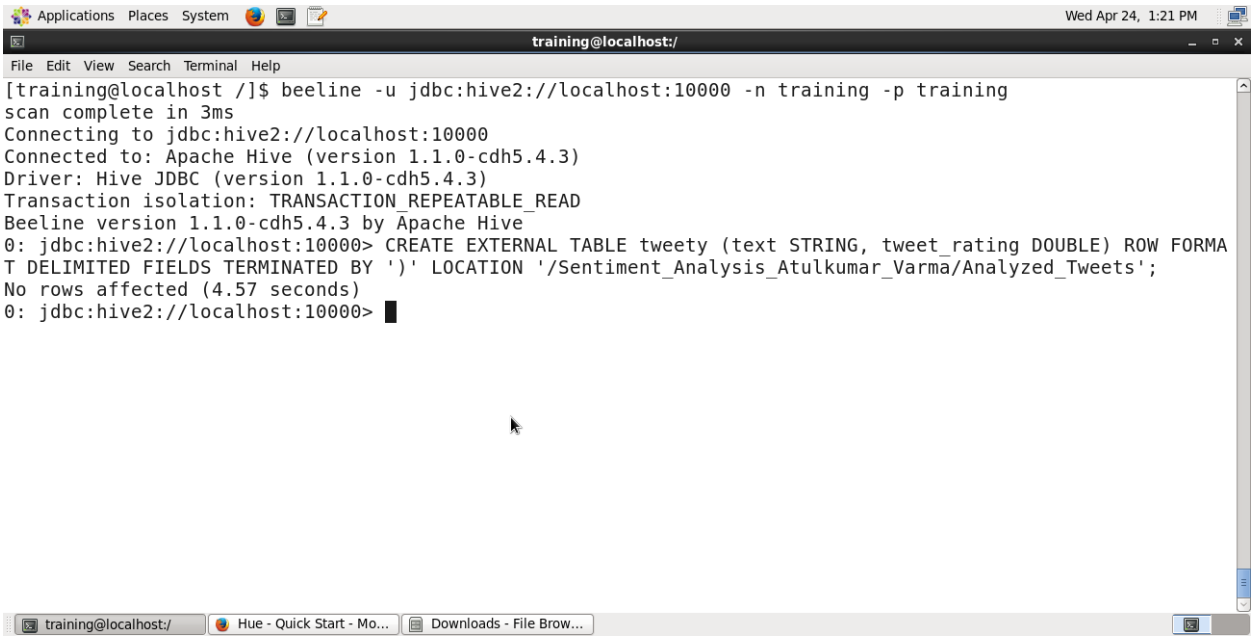


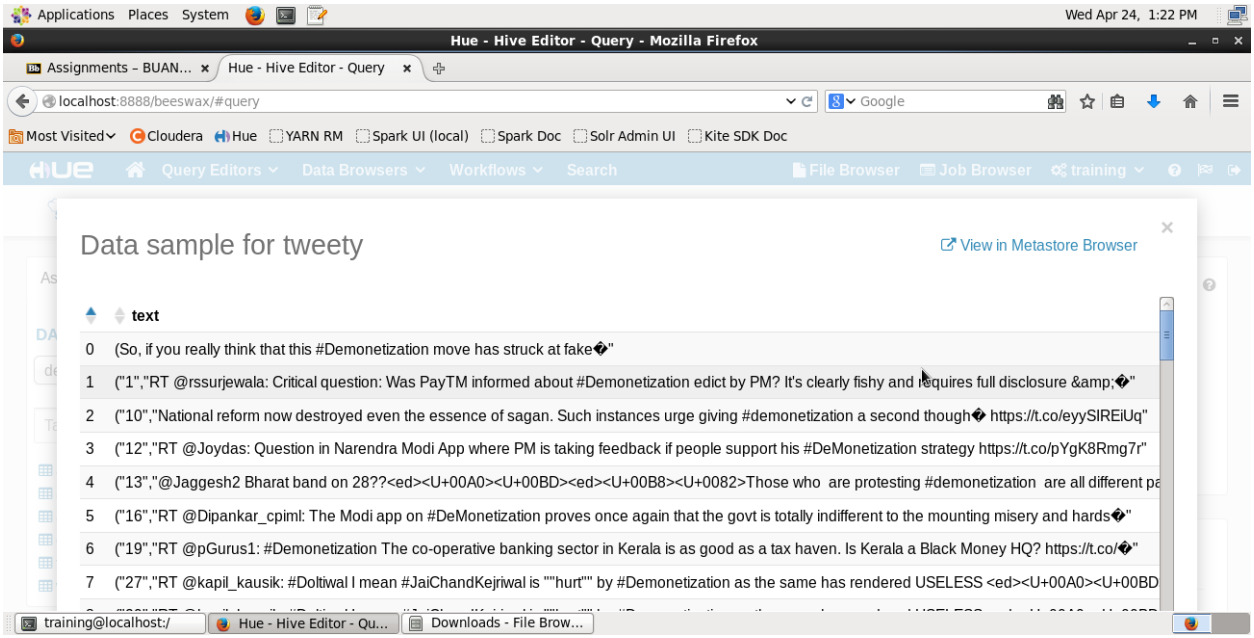


Step-17: We are starting up beeline, to execute the hive commands

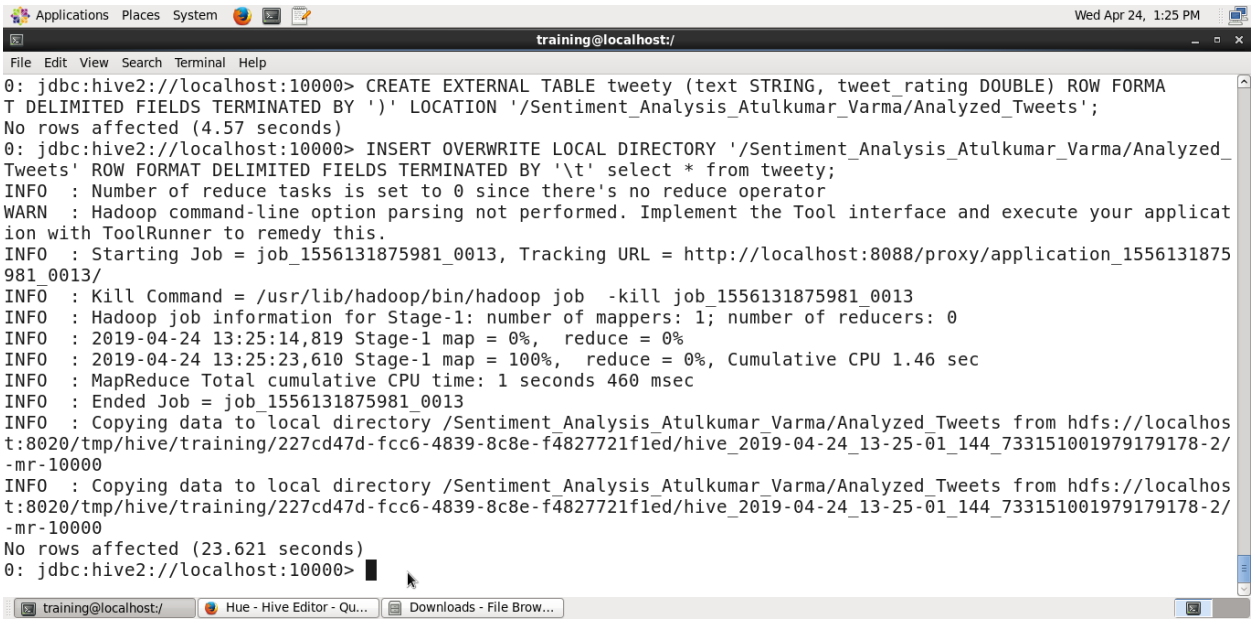


Step-18: Creating a hive table to structure the unstructured data

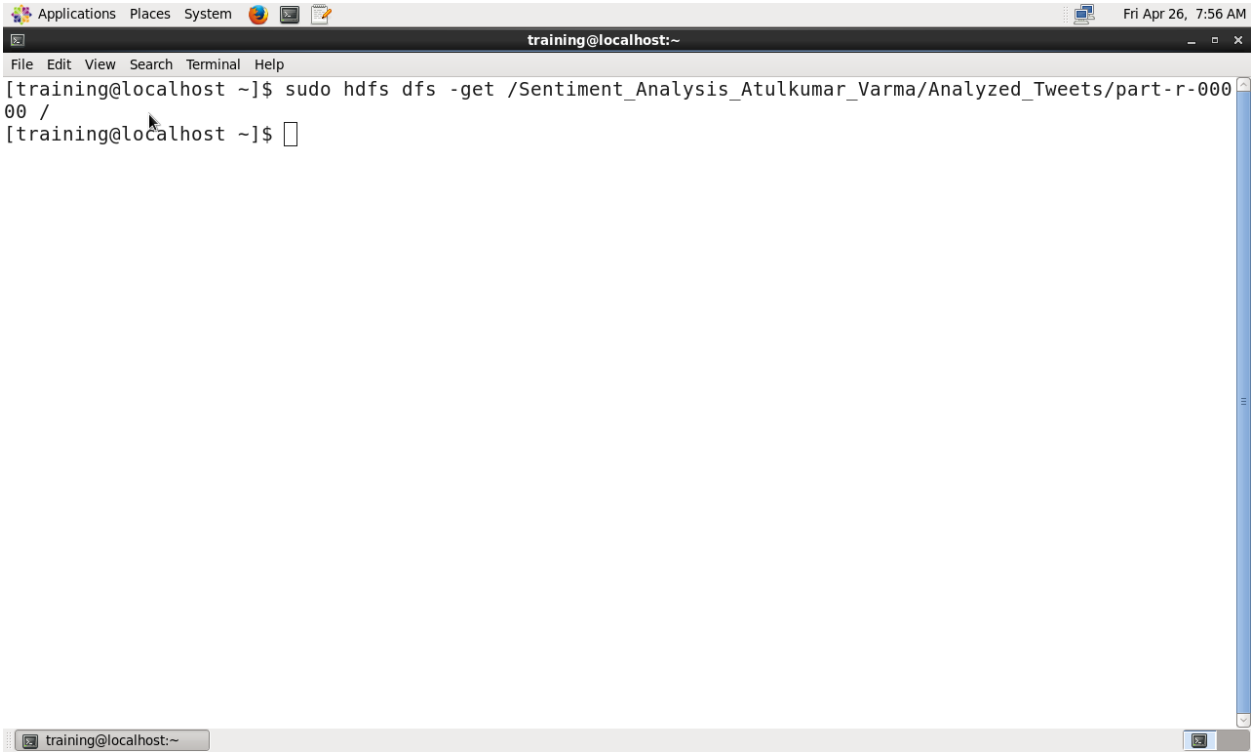


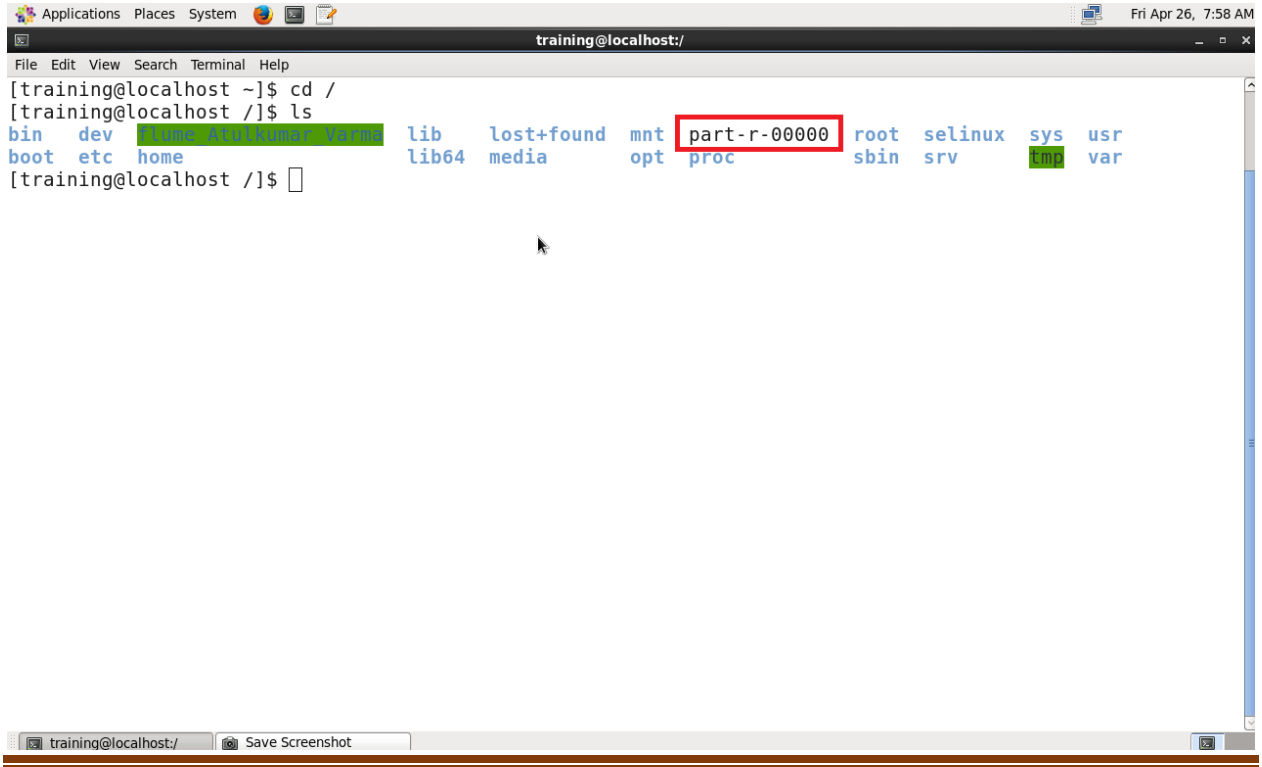


Step-19: Creating a tsv file from the table that we created in Step-18

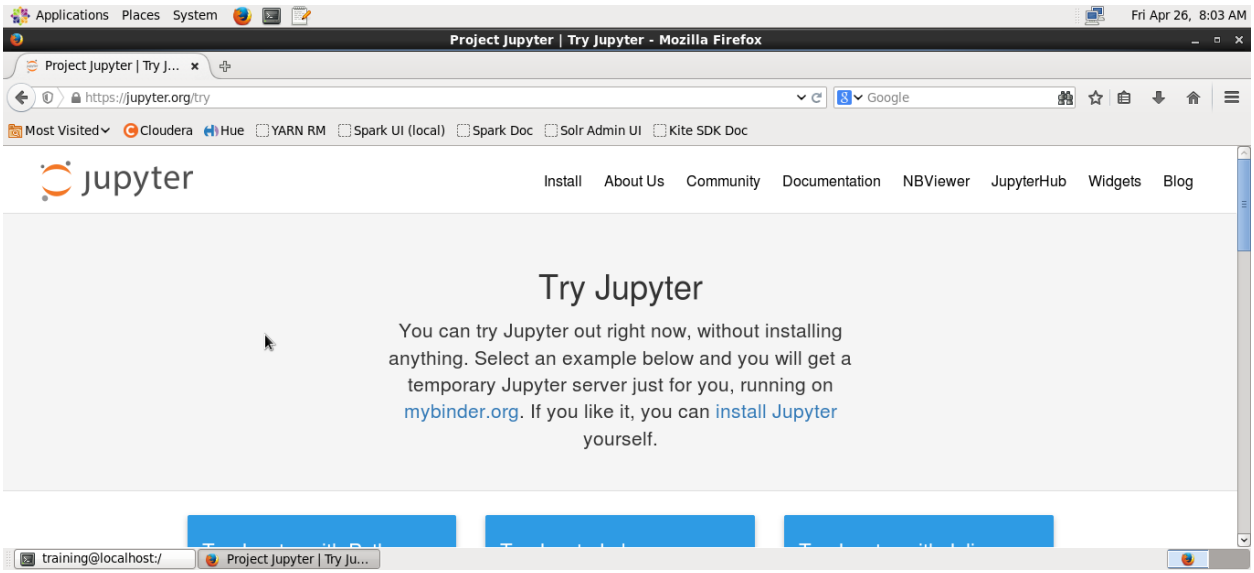


Step-20: Copying the results of our sentiment analysis back on the local file system to visualize.

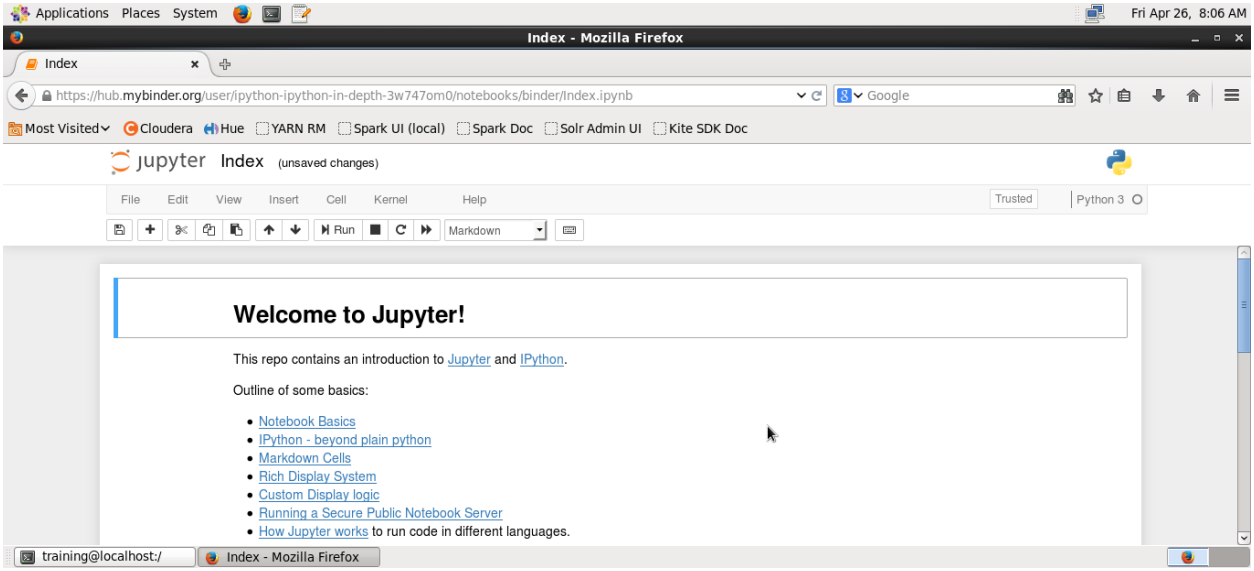




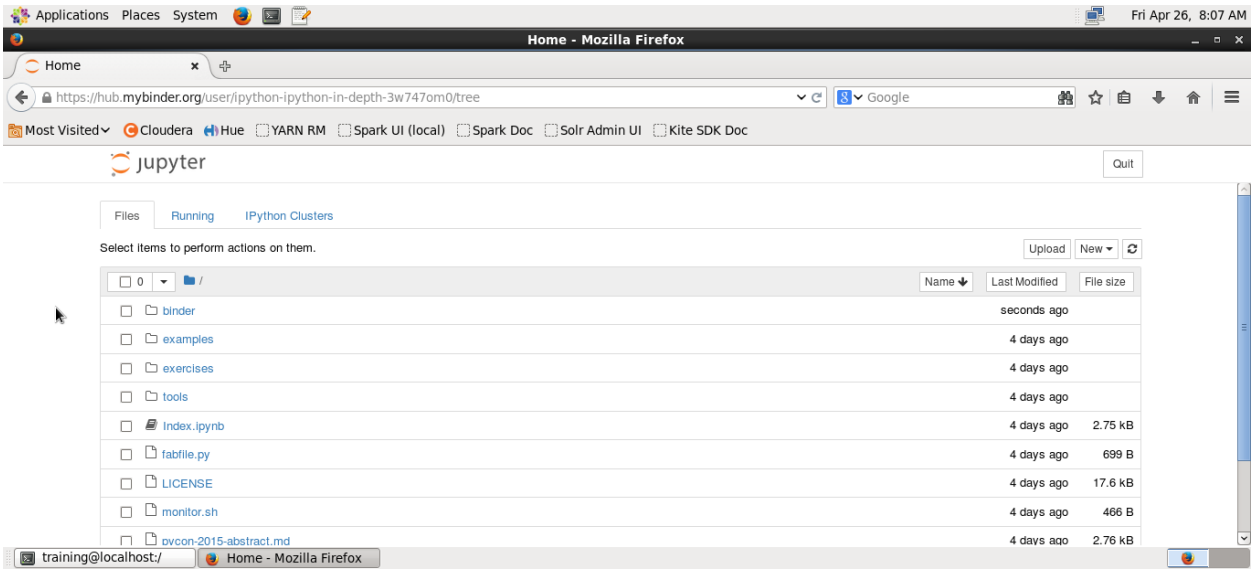
Step-21:



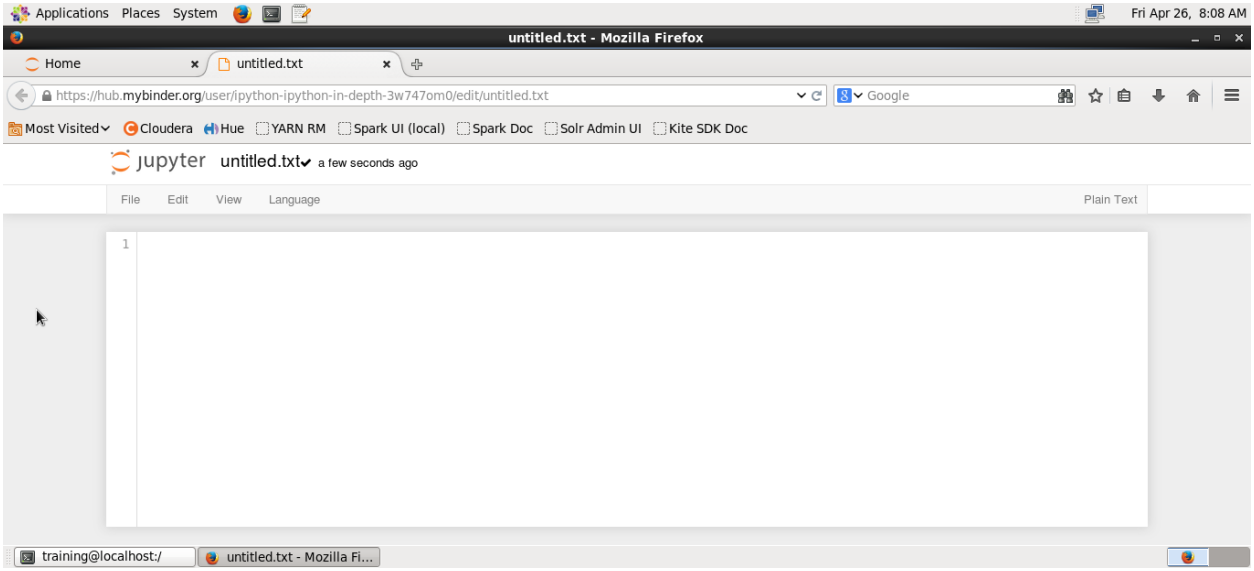
Step-22:



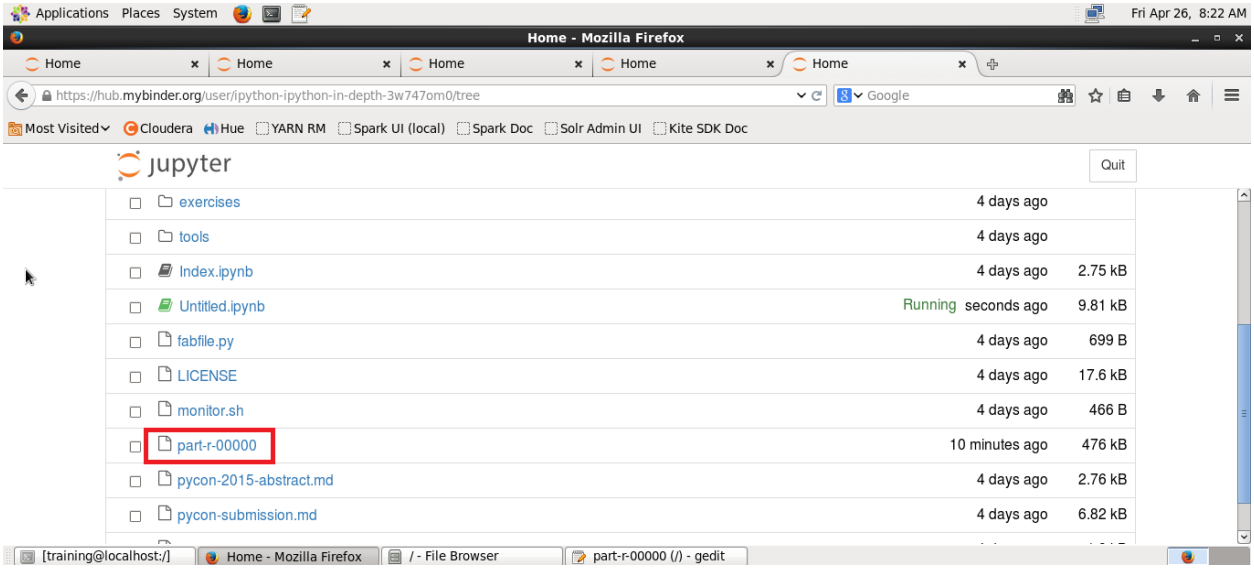
Step-23:



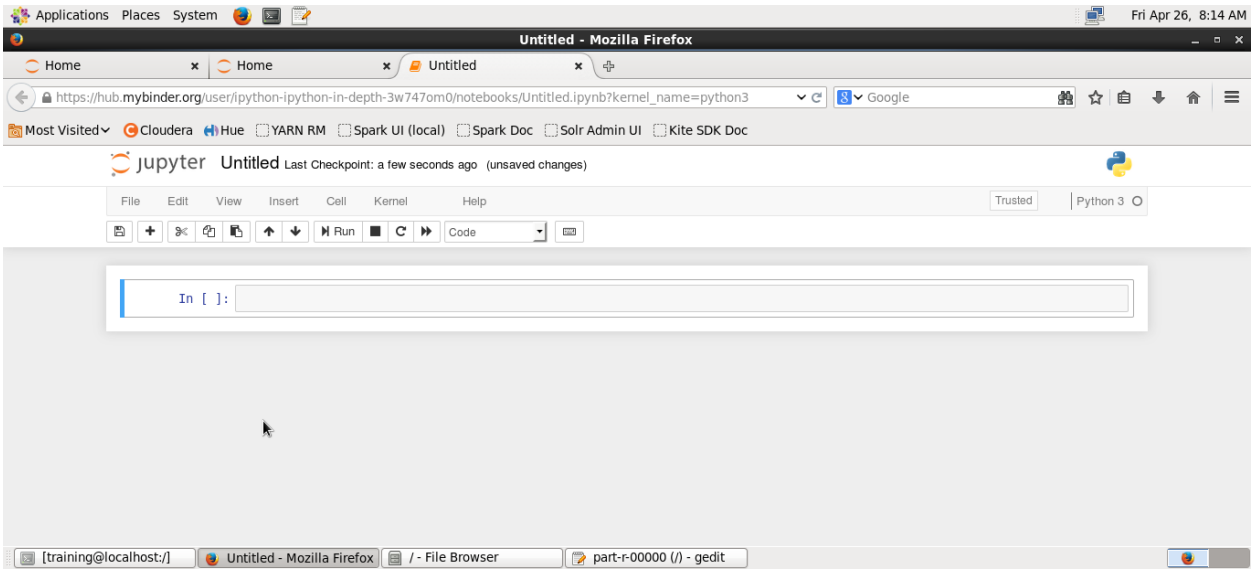
Step-24:



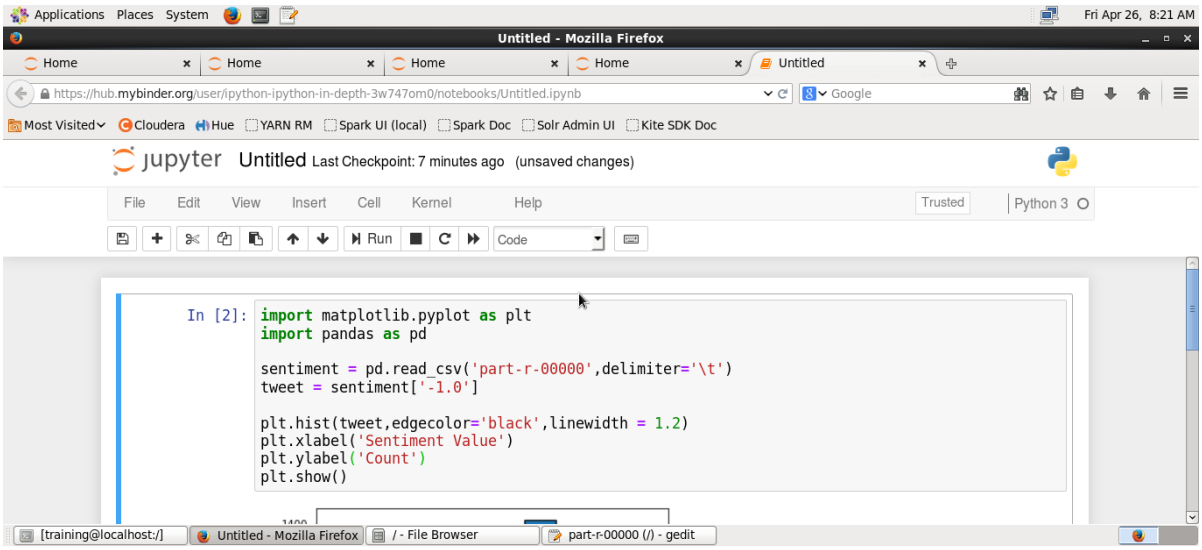
Step-25:

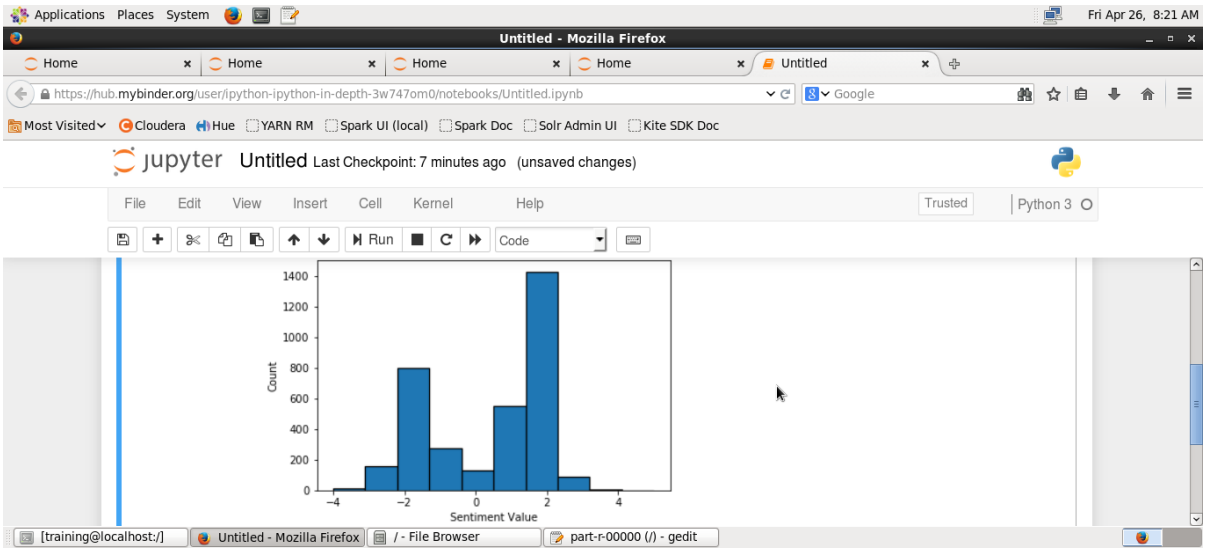


Step-26:



Step-27:





Y axis represents the count of sentiments

X axis whether the sentiment tweeted by people is negative or positive

Based on the AFINN Dictionary, few people(around 20-30) tweeted -4th level of negative sentiment

Around 800 tweeted -2nd level of negative sentiment

Around 100 tweeted neutral sentiment

Around 1400 tweeted 2nd level of positive sentiment

And finally, around 20-30 tweeted 4th level of positive sentiment