Session 22 : SPARK

Assignment 2

**Problem Statement**

Go through below blog and reiterate the same at your end.

https://docs.google.com/document/d/14YUd\_wi-KJTBEqqtyoNtPMFeORjFU-

\_yId4hJ11q90o/edit

**Sentiment analysis on demonetization**

Let us find out the views of different people on the demonetization by analysing the tweets from twitter. Here is the dataset where twitter tweets are gathered in CSV format. You can download the dataset from the below link

<https://drive.google.com/open?id=0ByJLBTmJojjzNkRsZWJiY1VGc28>

Dataset -  AFINN.txt - 

**Solution:-**

Starting Spark-shell

Command – spark-shell



Loading data:-

**scala>** val tweets = sc.textFile("demonetization-tweets.csv").map(x => x.split(",")).filter(x=>x.length>=2).map(x => (x(0).replaceAll("\"",""),x(1).replaceAll("\"","").toLowerCase)).map(x => (x.\_1,x.\_2.split(" "))).toDF("id","words");



**scala>** tweets.registerTempTable("tweets");

**scala>** val explode = spark.sql("select id as id,explode(words) as word from tweets").registerTempTable("tweet\_word");

**scala>** val afinn = sc.textFile("AFINN.txt").map(x => x.split("\t")).map(x => (x(0),x(1))).toDF("word","rating").registerTempTable("afinn");

**scala>** val join = spark.sql("select t.id,AVG(a.rating) as rating from tweet\_word t join afinn a on t.word=a.word group by t.id order by rating desc");



**scala>** join.show

