Spark Assignment 21.1

This assignment is about **Counting popular hashtags using Spark sql**

Objective: Let us find out the count of various hashtags used during the demonetization by analysing the tweets from twitter. Here is the dataset where twitter tweets are gathered in CSV format. The data set is given in the below link. So download the dataset from the below link

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

This assignment is done in the spark shell of Acadgild VM. I've used Pig(to convert CSV to JSON) and spark shell for this assignment

Steps Followed To convert to JSON:

1) Copied the dataset file (CSV) in the path /home/acadgild/pig/demonetization.csv. Loaded the data using PigStorage as below.

```
Load_Data = LOAD '/home/acadgild/pig/demonetization.csv' USING PigStorage(',') AS(text:chararray,favorited:chararray,favoriteCount:int,replyToSN:chararray,created:chararray,
truncated:chararray,replyToSID:chararray,id:long,replyToUID:chararray,statusSource:chararray,screenName:chararray,retweetCount:int,isRetweet:chararray,retweeted:chararray);
```

2) Parsed thru the Load_data 'text' field using Tokenize, flatten the text and stored as hashtag.

```
tweetwords=FOREACH Load_Data GENERATE text,favorited,favoriteCount,replyToSN,created,truncated,replyToSID,id,replyToUID,statusSource,screenName,retweetCount,isRetweet,retweeted, FLATTEN(TOKENIZE(text)) AS hashtag;
```

3) Then Filter the hashtag having character # .So hashtags will contain filtered hashtag, along with all the items of Load_Data.

```
hashtags = FILTER tweetwords BY UPPER(hashtag) MATCHES '#\\s*(\\w+)';
```

4) Finally converted the hashtags data into json and stored in the below path.To be used as a input in Spark SQL.

```
STORE hashtags INTO '/home/acadgild/spark/assignment_21_1' USING org.apache.pig.builtin.JsonStorage();
```

Analysis Using Spark SQL:

1) First, read the json file by using spark.read.json and registered a Temp table as below.

```
val tweets = spark.read.json("/home/acadgild/spark/assignment_21_1/part-m-00000").registerTempTable("tweets")
```

2) Now using select function, selected id, hashtag, converted the hashtag into Lowercase as below and registered as a temporary table.

```
val hashtags = spark.sql("select id,LOWER(hashtag) as hashtag  from tweets").registerTempTable("hashtags")
```

```
scala> val tweets = spark.read.json("/home/acadgild/spark/assignment_21_1/part-m-00000").registerTempTable("tweets")
warning: there was one deprecation warning; re-run with -deprecation for details
tweets: Unit = ()

scala> val hashtags = spark.sql("select id,LOWER(hashtag) as hashtag  from tweets").registerTempTable("hashtags")
warning: there was one deprecation warning; re-run with -deprecation for details
hashtags: Unit = ()
```

3) Finally, selecting the hashtag keywords and count of the hashtag and grouping by hashtag and order by hashtag count of highest order.

```
val popular_hashtags = spark.sql("select hashtag, count(hashtag) as hashtag_count from hashtags group by hashtag order by hashtag_count desc").show
```

Spark Shell Output:

```
scala> val popular_hashtags = spark.sql("select hashtag, count(hashtag) as hashtag_count from hashtags group by hashtag order by has
htag_count desc").show
+-------------------+-------------+
|            hashtag|hashtag_count|
+-------------------+-------------+
|     #demonetization|         5689|
|        #nitishkumar|          258|
|#corruptionfreeindia|          103|
|         #blackmoney|          102|
|#indiafightscorru...|           79|
|               #modi|           71|
|        #nomoneyyaar|           53|
|     #demonetisation|           52|
|              #india|           47|
|                #bjp|           44|
|           #insights|           44|
|           #bulletin|           44|
|          #ratantata|           44|
|                #rbi|           34|
|              #nmapp|           29|
|        #netascashin|           26|
|             #pmmodi|           23|
|             #hitler|           20|
|             #survey|           19|
|   #jaichandkejriwal|           18|
+-------------------+-------------+
only showing top 20 rows

popular_hashtags: Unit = ()
```

Output:

hashtag	hashtag_count
#demonetization	5689
#nitishkumar	258
#corruptionfreeindia	103
#blackmoney	102
#indiafightscorru...	79
#modi	71
#nomoneyyaar	53
#demonetisation	52
#india	47
#bjp	44
#insights	44
#bulletin	44
#ratantata	44
#rbi	34
#nmapp	29
#netascashin	26
#pmmodi	23
#hitler	20
#survey	19
#jaichandkejriwal	18