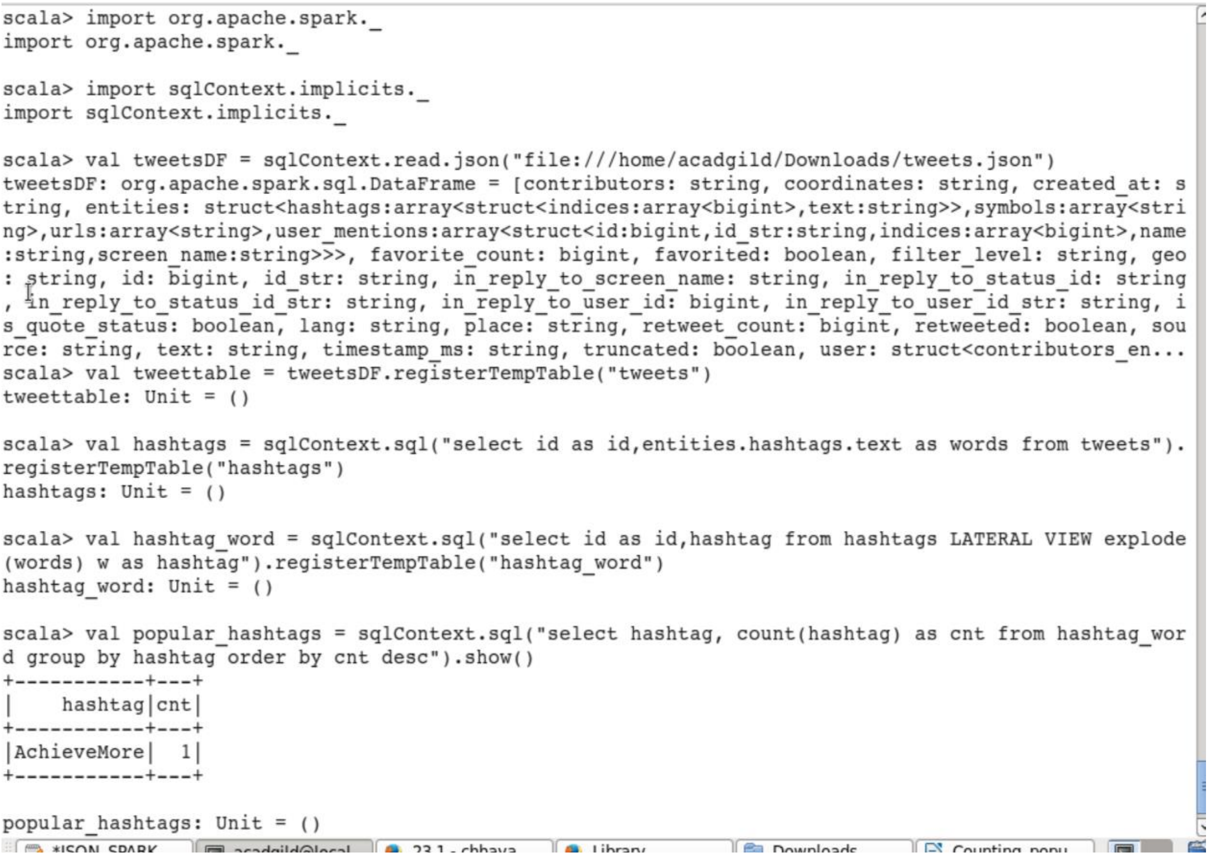
SPARK ASSIGNMENT 23.1

Problem Statement :Counting popular hashtags using Spark sql

~~~~~~

```
//Importing the SPARK SQL Packages

import org.apache.spark._

import sqlContext.implicits._

//Reading JSON file with sqlContext object

val tweetsDF =
sqlContext.read.json("file:///home/acadgild/Downloads
/tweets.json")

//Conversion of Dataframe and Creating a temporary
table

val tweettable = tweetsDF.registerTempTable("tweets")

//Running SPARK SQL and save the output as Temporary
Table

val hashtags = sqlContext.sql("select id as
id,entities.hashtags.text as words from
tweets").registerTempTable("hashtags")

//Running SPARK SQL and save the output as Temporary
Table

val hashtag_word = sqlContext.sql("select id as
id,hashtag from hashtags LATERAL VIEW explode(words)
w as hashtag").registerTempTable("hashtag_word")

//Running SPARK SQL and showing the result

val popular_hashtags = sqlContext.sql("select
hashtag, count(hashtag) as cnt from hashtag_word
group by hashtag order by cnt desc").show()
```
~~~~~~

Screenshots:



```
scala> import org.apache.spark._
import org.apache.spark._

scala> import sqlContext.implicits._
import sqlContext.implicits._

scala> val tweetsDF = sqlContext.read.json("file:///home/acadgild/Downloads/tweets.json")
tweetsDF: org.apache.spark.sql.DataFrame = [contributors: string, coordinates: string, created_at: s
tring, entities: struct<hashtags:array<struct<indices:array<bigint>,text:string>>,symbols:array<stri
ng>,urls:array<string>,user_mentions:array<struct<id:bigint,id_str:string,indices:array<bigint>,name
:string,screen_name:string>>>, favorite_count: bigint, favorited: boolean, filter_level: string, geo
: string, id: bigint, id_str: string, in_reply_to_screen_name: string, in_reply_to_status_id: string
, in_reply_to_status_id_str: string, in_reply_to_user_id: bigint, in_reply_to_user_id_str: string, i
s_quote_status: boolean, lang: string, place: string, retweet_count: bigint, retweeted: boolean, sou
rce: string, text: string, timestamp_ms: string, truncated: boolean, user: struct<contributors_en...
scala> val tweettable = tweetsDF.registerTempTable("tweets")
tweettable: Unit = ()

scala> val hashtags = sqlContext.sql("select id as id,entities.hashtags.text as words from tweets").
registerTempTable("hashtags")
hashtags: Unit = ()

scala> val hashtag_word = sqlContext.sql("select id as id,hashtag from hashtags LATERAL VIEW explode
(words) w as hashtag").registerTempTable("hashtag_word")
hashtag_word: Unit = ()

scala> val popular_hashtags = sqlContext.sql("select hashtag, count(hashtag) as cnt from hashtag_wor
d group by hashtag order by cnt desc").show()
+-----------+---+
|    hashtag|cnt|
+-----------+---+
|AchieveMore|  1|
+-----------+---+

popular_hashtags: Unit = ()
```