

# Assignment 21.1:

## Problem Statement:

Implement the below blog at your end and send the complete documentation.

[https://drive.google.com/file/d/0B\\_Qjau8wv1KobUlaOEtFNetQNkU/view?usp=sharing](https://drive.google.com/file/d/0B_Qjau8wv1KobUlaOEtFNetQNkU/view?usp=sharing)

## Solution:

1. Data set is places at below location:

```
[acadmild@localhost spark]$ pwd
/home/acadmild/spark
[acadmild@localhost spark]$ cat tweets.txt
{"filter_level":"low","retweeted":false,"in_reply_to_screen_name":"FilmFan","truncated":false,"lang":"en","in_reply_to_status_id_str":null,"id":"689085590822891521","in_reply_to_user_id_str":"6048122","timestamp_ms":"1453125782100","in_reply_to_status_id":null,"created_at":"Mon Jan 18 14:03:02 +0000 2016","favorite_count":0,"place":null,"coordinates":null,"text":"@filmfan hey its time for you guys follow @acadmild To #AchieveMore and participate in contest Win Rs.500 worth vouchers","contributors":null,"geo":null,"entities":{"symbols":[],"urls":[]},"hashtags":[{"text":"AchieveMore","indices":[56,68]}],"user_mentions":[{"id":"6048122","name":"Tanya","indices":[0,8],"screen_name":"FilmFan","id_str":"6048122"}],"id":"2649945906","name":"ACADMILD","indices":[42,51],"screen_name":"acadmild","id_str":"2649945906"},"is_quote_status":false,"source":"<a href='\"https://about.twitter.com/products/tweetdeck\"' rel='\"nofollow\"'>TweetDeck</a>","retweeted":false,"in_reply_to_user_id":"6048122","retweet_count":0,"id_str":"689085590822891521","user":{"location":"India ","default_profile":false,"profile_background_tile":false,"statuses_count":86548,"lang":"en","profile_link_color":"940487","profile_banner_url":"https://pbs.twimg.com/profile_banners/197865769/1436198000","id":"197865769","following":null,"protected":false,"favourites_count":1002,"profile_text_color":"000000","verified":false,"description":"Proud Indian, Digital Marketing Consultant,Traveler, Foodie, Adventurer, Data Architect, Movie Lover, Nam o Fan","contributors_enabled":false,"profile_sidebar_border_color":"000000","name":"Bahubali","profile_background_color":"000000","create_at":"Sat Oct 02 17:41:02 +0000 2010","default_profile_image":false,"followers_count":4467,"profile_image_url_https":"https://pbs.twimg.com/profile_images/664486535040000000/GojDUiuk_normal.jpg","geo_enabled":true,"profile_background_image_url":"http://abs.twimg.com/images"/>
```

2. First we will read the JSON file stored in the local file system and create a temporary table tweets

```
val tweets =
```

```
spark.read.json("/home/acadmild/spark/tweets.txt").registerTempTable("tweets")
```

```
scala> val tweets = spark.read.json("/home/acadmild/spark/tweets.txt").registerTempTable("tweets")
warning: there was one deprecation warning; re-run with -deprecation for details
18/01/16 08:30:10 WARN Utils: Truncated the string representation of a plan since it was too large. This behavior can be adjusted by setting 'spark.debug.maxToStringFields' in SparkEnv.conf.
tweets: Unit = ()
scala> █
```

3. Now, from the above temporary table we will select the ID's, hashtag and create another temporary table hashtags.

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

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

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

scala> val hashtag_word = spark.sql("select id as id,hashtag from hashtags LATERAL VIEW explode(words) w as hashtag").registerTempTable("hashtag_word")
warning: there was one deprecation warning; re-run with -deprecation for details
hashtag_word: Unit = ()
scala> █
```

4. Finally, we will get the popular hashtags used in twitter and its count with the following command:

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

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

hashtag	cnt
AchieveMore	11
Hadoop	5
bigdata	2
WhitePaper	1
GartnerEIM	1
masterdata	1
BigData	1
contest	1
data	1
chiefdataofficer	1
HDFS	1
informationgovern...	1
Virtualization	1
OReilly	1
dataquality	1
Spark	1
Infonomics	1

```
popular_hashtags: Unit = ()
scala> █
```