



Session 20: MLIB AND GRAPHX

Assignment 21.1

Student Name: Abarajithan SA
Course: Big Data Hadoop & Spark Training
Start Date: 2017-09-09
End Date: 2017-11-26

Assignment 21.1– MLIB AND GRAPHX.

Contents

Introduction	1
Problem Statement.....	1
Counting popular hashtags using Spark sql	2
Source codes	3
Expected output.....	3

Introduction

In this assignment, we are going to Counting popular hashtags using Spark sql.

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

Counting popular hashtags using Spark sql



Counting popular hashtags using Spark sql

Below is the sample tweet, we uploaded the below tweets as a .json file to the local file system,

```
{
  "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 @acadgild 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": "ACADGILD",
        "indices": [42, 51],
        "screen_name": "acadgild",
        "id_str": "2649945906"
      }
    ]
  },
  "is_quote_status": false,
  "source": "<a href='\"https://about.twitter.com/products/tweetdeck\"' rel='\"nofollow\"'>TweetDeck</a>",
  "favorited": 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": "9AD487",
    "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, Namo Fan",
    "contributors_enabled": false,
    "profile_sidebar_border_color": "000000",
    "name": "Bahubali",
    "profile_background_color": "000000",
    "created_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/themes/theme1/bg.png",
    "profile_background_image_url_https": "https://abs.twimg.com/images/themes/theme1/bg.png",
    "follow_request_sent": null,
    "url": null,
    "utc_offset": 19800,
    "time_zone": "Chennai",
    "notifications": null,
    "profile_use_background_image": false,
    "friends_count": 810,
    "profile_sidebar_fill_color": "000000",
    "screen_name": "Ashok_Uppuluri",
    "id_str": "197865769",
    "profile_image_url": "http://pbs.twimg.com/profile_images/664486535040000000/GOjDUiuK_normal.jpg",
    "listed_count": 50,
    "is_translator": false
  }
}
```

The file is uploaded as Counting_popular_hashtags_using_spark.json into the below location,

/home/acadgild/hadoop/Counting_popular_hashtags_using_spark.json



Source codes

```
val tweets =  
spark.read.json("file:///home/acadgild/hadoop/Counting_popular_hashtags_using_spark.json").registerTempTable("tweets")
```

```
scala> val tweets = spark.read.json("file:///home/acadgild/hadoop/Counting_popular_hashtags_using_spark.json").registerTempTable("tweets")  
warning: there was one deprecation warning; re-run with -deprecation for details  
18/01/15 17:59:03 WARN ObjectStore: Failed to get database global_temp, returning NoSuchObjectException  
18/01/15 17:59:04 WARN SizeEstimator: Failed to check whether UseCompressedOups is set; assuming yes  
18/01/15 17:59:12 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 = ()
```

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

```
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 = ()
```

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

```
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 = ()
```

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

Expected output

```
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| 1|  
+-----+  
-----
```