**PRICIPLES OF BIG DATA MANAGEMENT (CS-5540)**

**PROJECT REPORT**

****

Recreational Activities Trends across World

Tweets Analysis

**TEAM NUMBER 18**

**TEAM MEMBERS:**

RASHMI TRIPATHI (rtrh6)

RITIKA CHOWDRI (rcc92)

NITIN BHANDARI (nbgm3)

HARMEET SINGH (hsp79)

**Table of Contents**

1. Introduction……………………………………………………………………………...3
2. Architecture……………………………………………………………………………...4
3. Process Flow…………………………………………………………………………….5
4. System Setup…………………………………………………………………………….6
5. Queries…………………………………………………………………………………..7

5.1 Query1 Dataframe ………………………………………………………………….8

5.2 Query 2 Dataframe ………………………………………………………………….9

5.3 Query 3 Dataframe…………………………………………………………………10

5.4 Query 4 RDD ……………………………………………………………………...12

5.5 Query 5 RDD ……………………………………………………………………...14

1. Runtime Measurement………………………………………………………………....15
2. References……………………………………………………………………………...15

Introduction

**Theme:** Recreational Activities Trends across World

This project is a powerful web application to help you comparing you to others on many fronts in terms of recreation activities followed by people around the world. Recreation is part of any human life. All human beings need some break from their hectic schedule to spend time on activities they enjoy like music, sports, dance, watching television etc. This boost their energy multifold.

As part of this project we have collected 130000+ instances of tweets across the globe using Twitter APIs in form of JSON file. The JSON(JavaScript Object Notation) data is one of lightweight data-interchange format. It is very easy to interpret by human beings as well as machines.

Now on top of this tweets we are analyzing various trends across the globe using Spark SQL.

We have designed meaningful queries to analyze various recreational activities people of various age groups choose and talk about, across the world by using twitter data

and designing meaningful queries.

Some of the recreational activities we analyzed are as, we included:

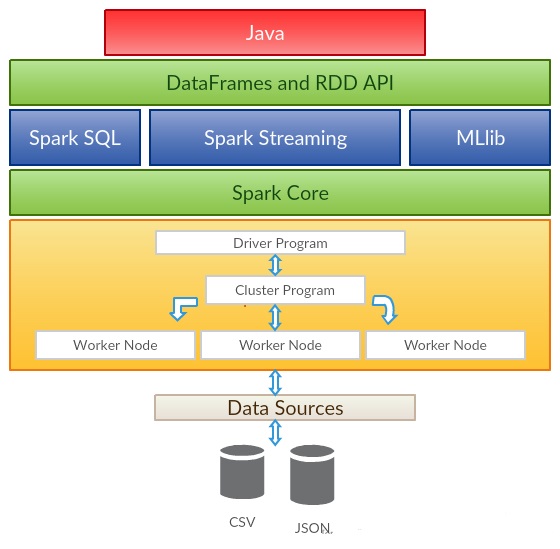
* Sports, Extreme Sports
* Travel, Trips, Adventure, Theme Parks, National Parks
* Watching, Listen, Eating activities
* Dancing, Yoga, Gym ,Relaxation activities
* Hobbies such as poetry, collection, DIY etc.

We particularly designed queries to analyze what all countries people involves in recreational activities and in turn revealing their work life balance. Also what all are the most popular activities, what king of age group they belong to, popular trending hashtags are part of this project.

GitHub URL: <https://github.com/rashmitripathi/TwitterTrendAnalysis_BigData>

Architecture

Following is the software architecture for our project.



Process Flow

Following is the process flow for the software architecture:

Twitter Public API

Keywords filter

Tweets

Input Datastream (JSON)

Spark Streaming

Interactively

Query with SQL

Process with

Java

DataFrames &RDD

Spark SQL

Output Datastream (JSON,CSV)

Output File

System Setup

**Operating System:** Windows 10

**IDE:** Eclipse Neon 4.6.1

**Browser:** Google Chrome

**Language:** Java

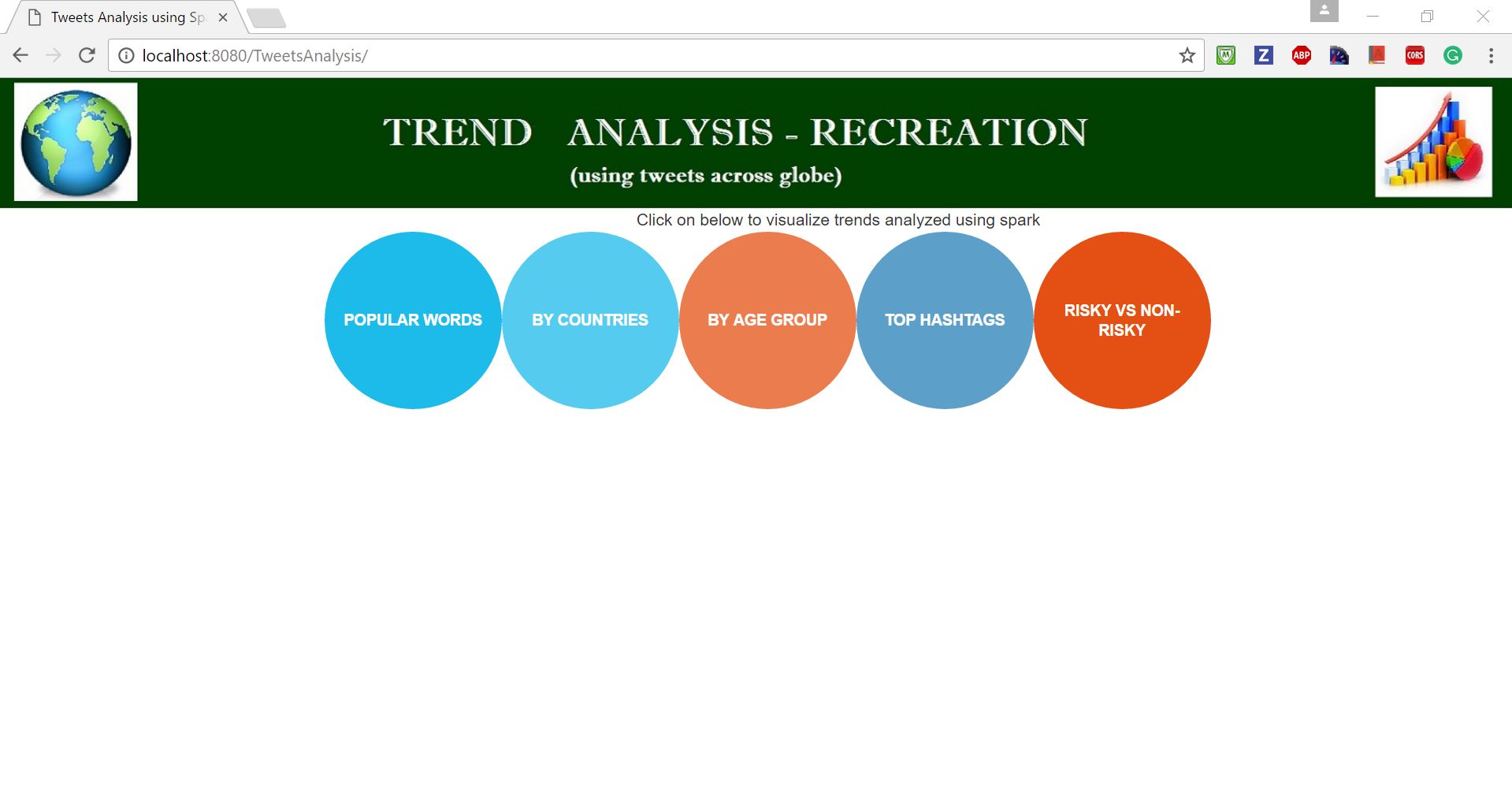
**Database**: Apache Spark SQL

**Library:** Apache Spark Core, Spark Ml lib, Spark SQL, Twitter4j, Sun Jersey

**Public API**: uClassify

**Visualization Charts**: HighCharts.js, Google Geo Charts.js, d3.js

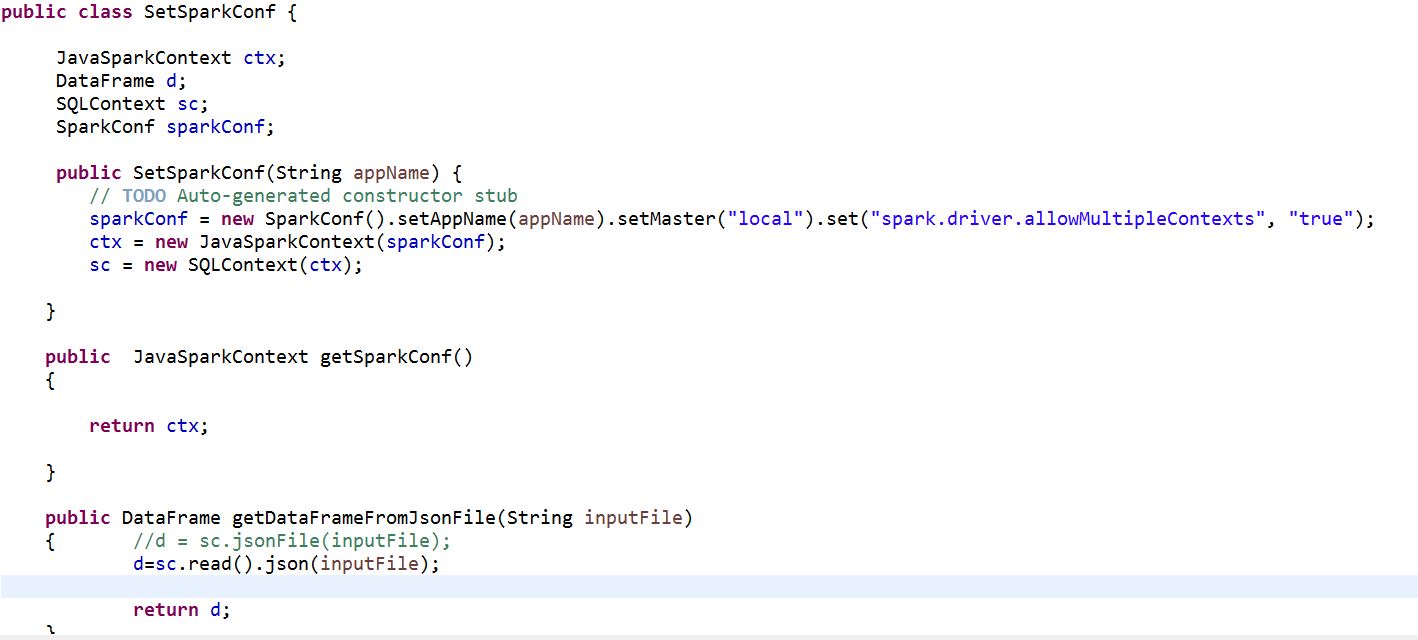
**User Interface**

Snapshot:

Queries

As part of this project, I have created one Java class to setup Spark Configuration and SQLContext. This class will be utilized in all further classes.

Source code for this is shown below:

****

Keywords to collect Tweets:

****

**Query1: Dataframe to return top countries**

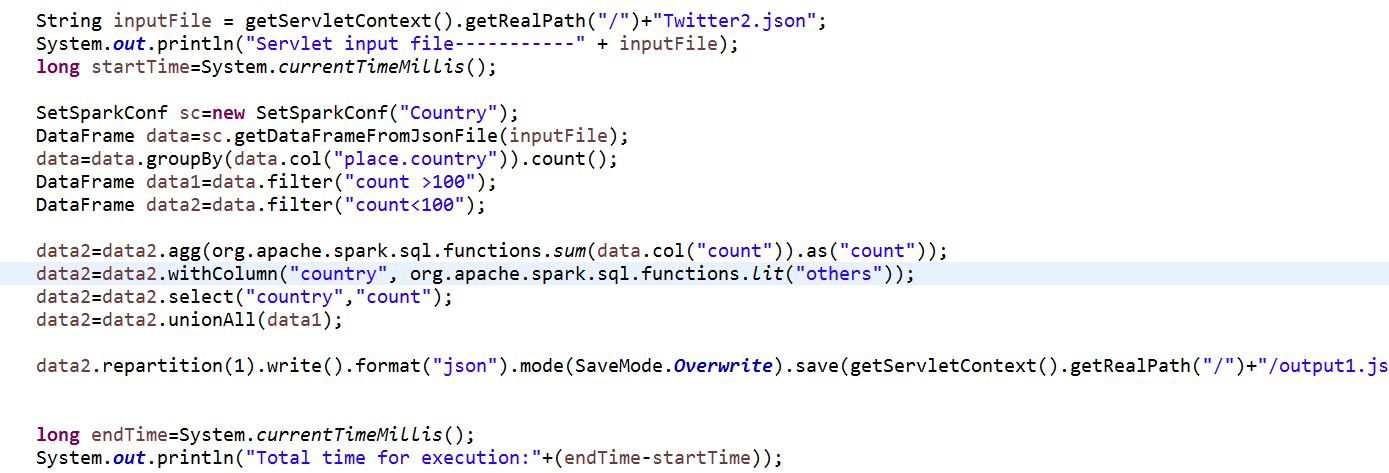
**Description**: It returns the top countries where people talk about recreation.

Numerous keywords related to sports, adventure, and activities were included to extract a large number of tweets.

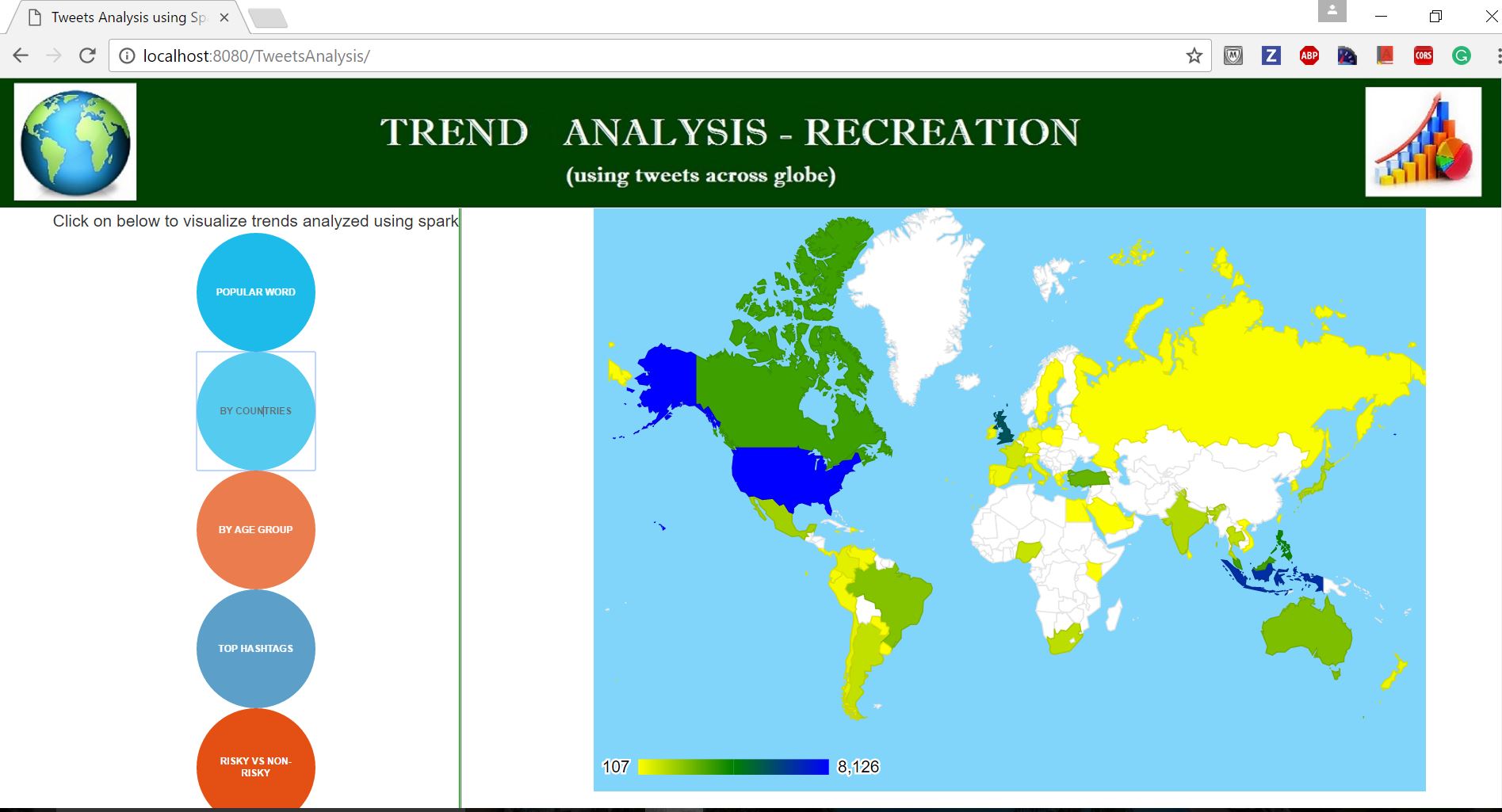
Here the countries whose tweets are less than 100 will be considered as others.

And the countries whose tweets are more than 100 will be respective countries and union operation is performed with the former dataframe.

**Special features**: Dataframe , Union

**Source code:**

Output:

Visualization:

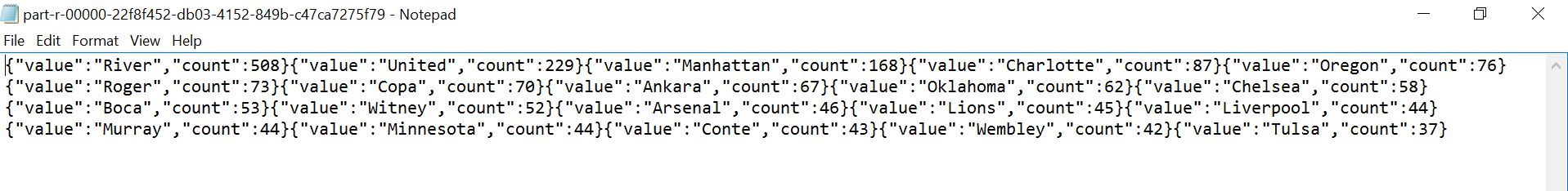
**Query2: Dataframe to return top hashtags**

**Description :** This query returns the top hashtags people used in their tweets.

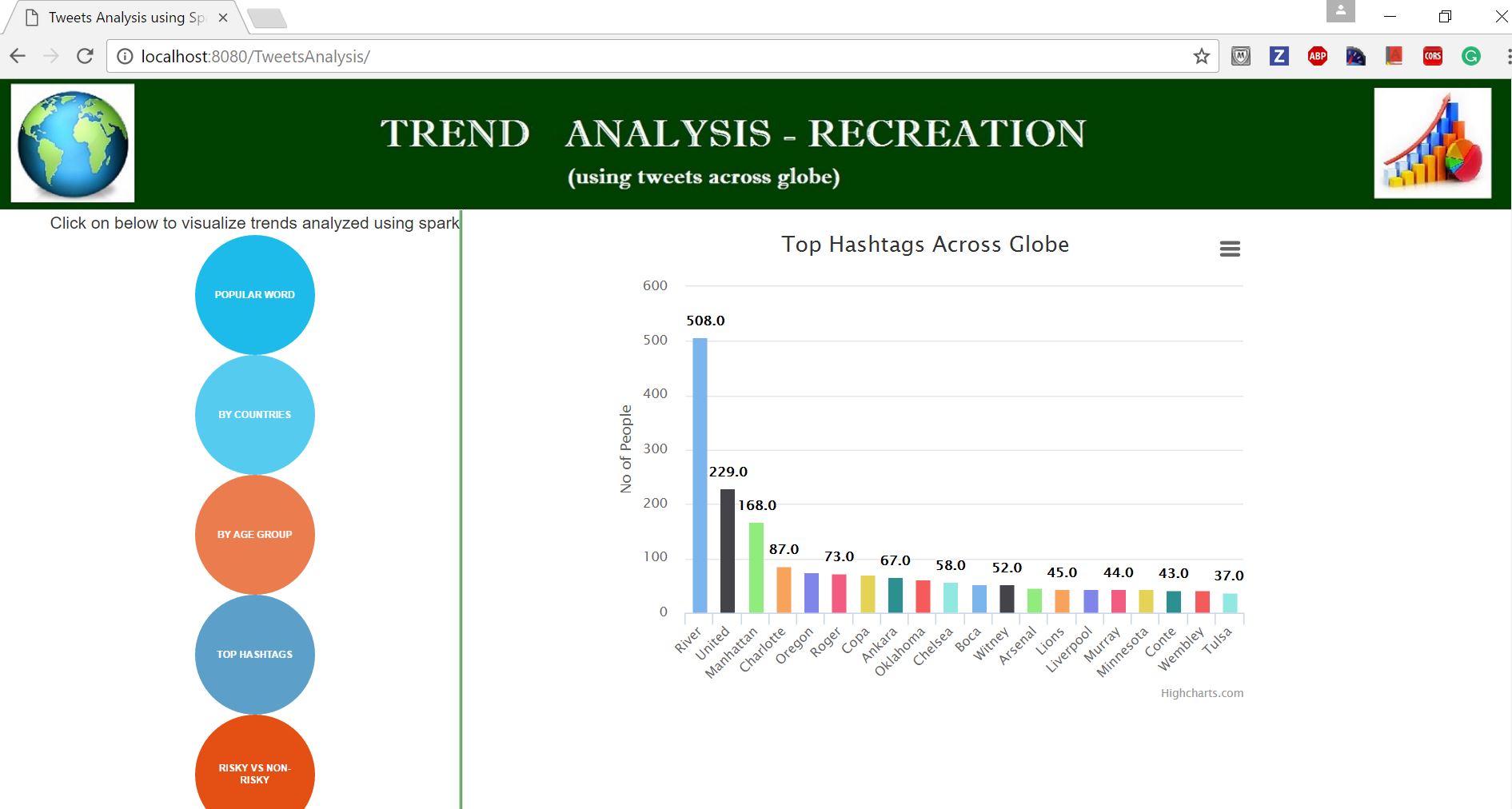
This query makes use of the hashtag file provided. It is basically join of hashtags from hashtag file plus tweets.

**Special features**: Dataframe , Join with another hashtags text file ,groupBy,orderBy

**Source code:**

Output:

Visualization:

****

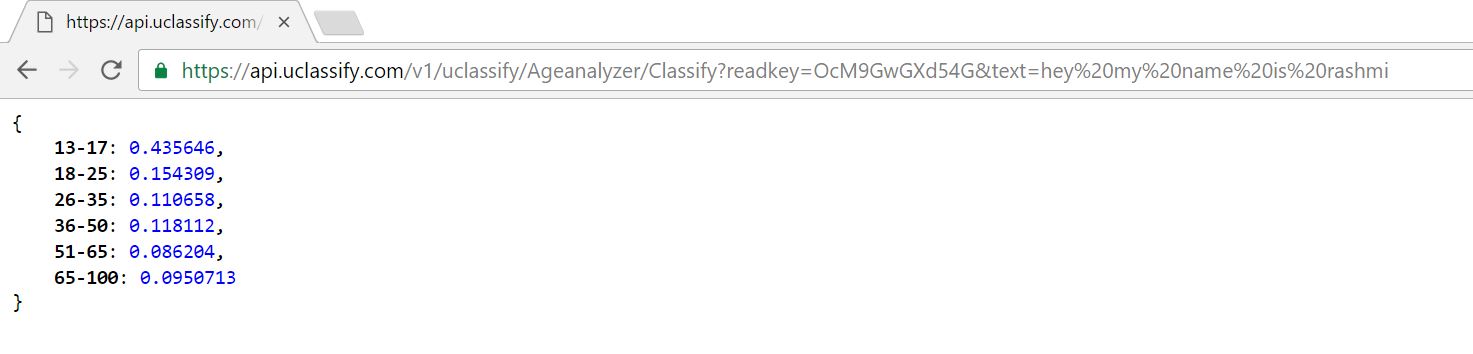
**Query 3: Dataframe to call Public API to classify tweets based on age group**

**Description :**

For query 3, a public API “UClassify” is used in which tweet text is fed to API and it returns the age group to which the user belongs who tweeted that particular text. Also this query is dynamically collecting tweets. As the uClassify calls have limitation, we have performed this query with 100 tweets.

**Special features:** UDF, calling Public API

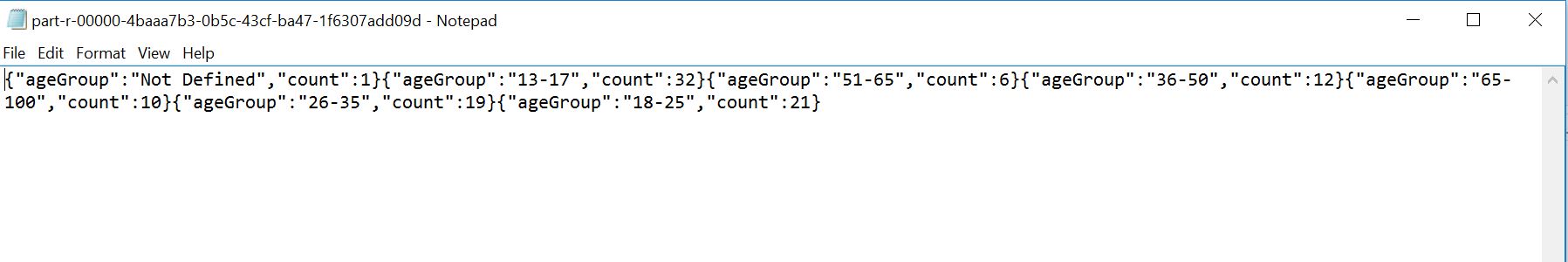
UClassify JSON Response:



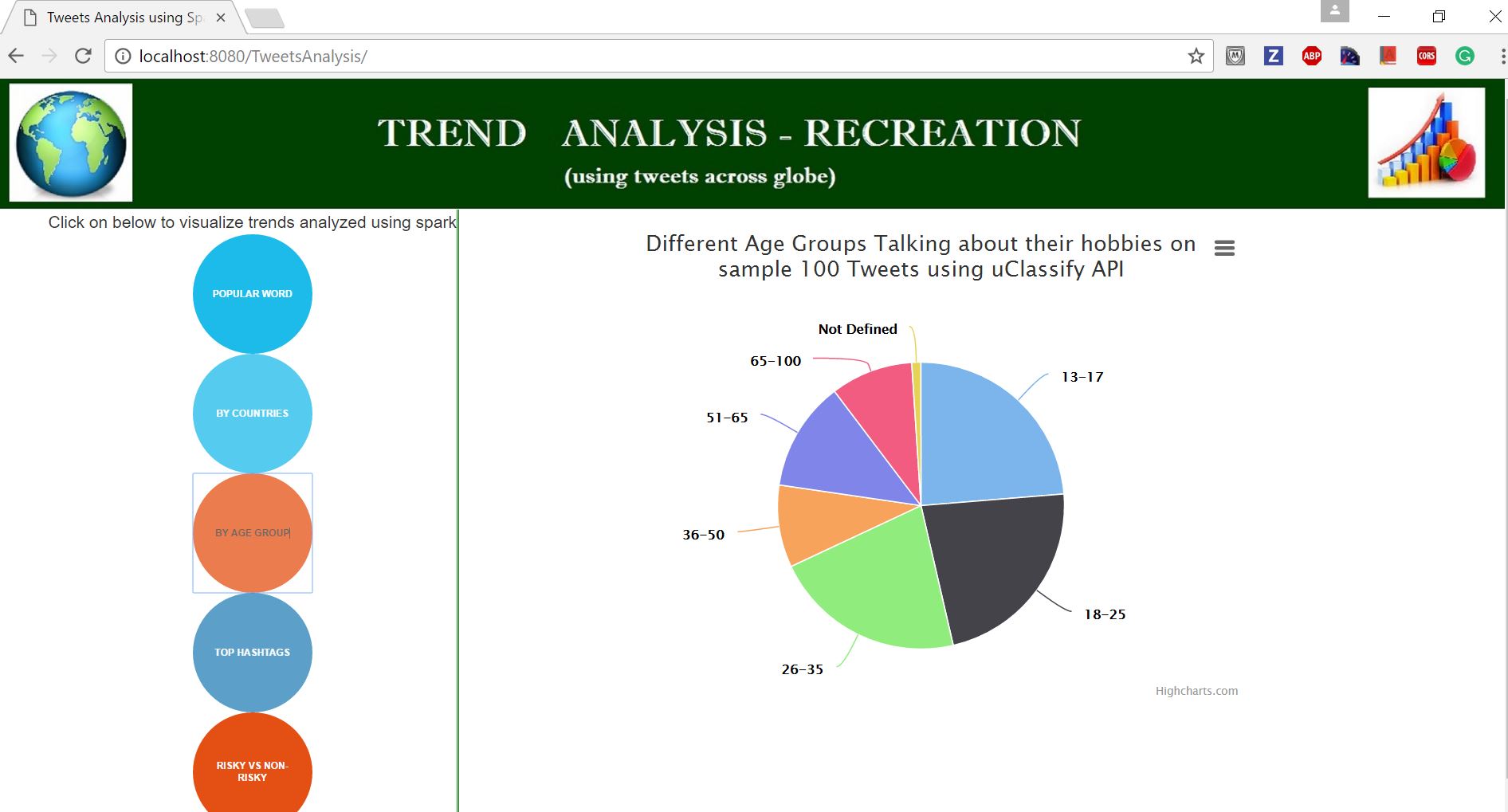
**Source code: UDF source code:**



**Output:**

****

Visualization:

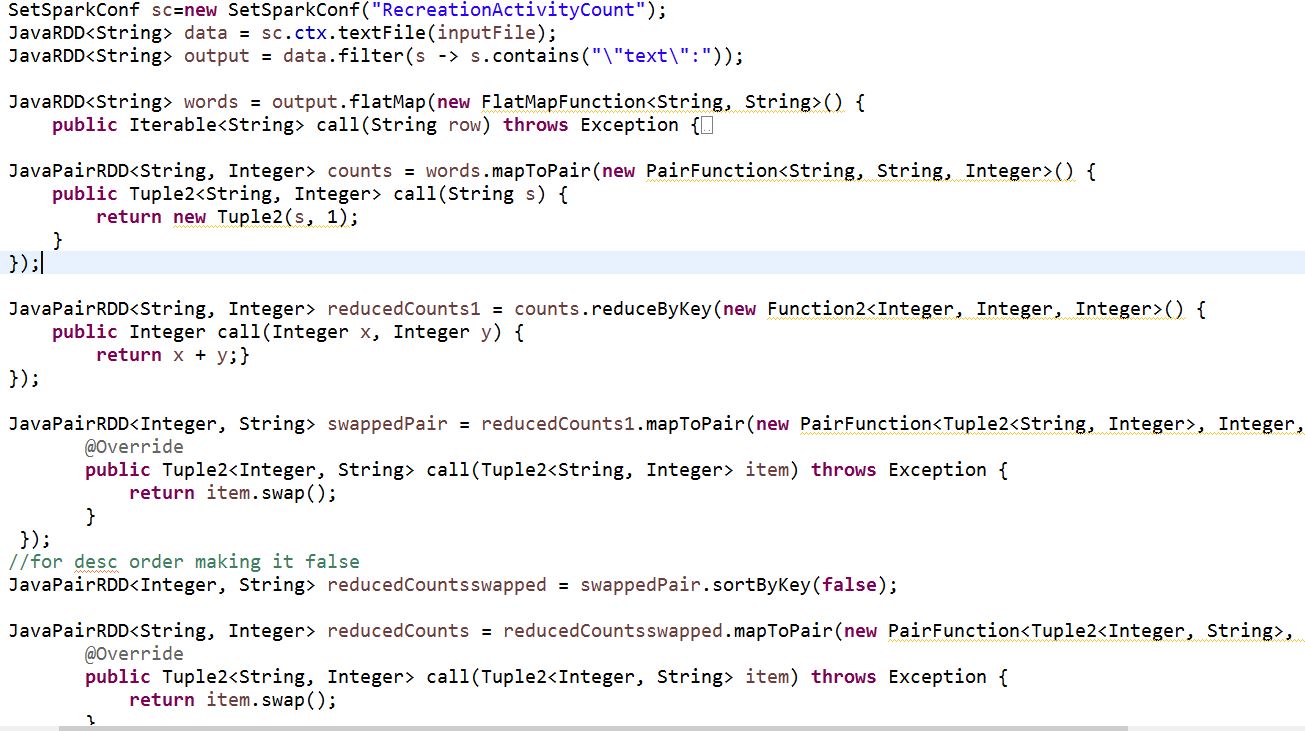


**Query 4: JAVA RDD query to get Top activities**

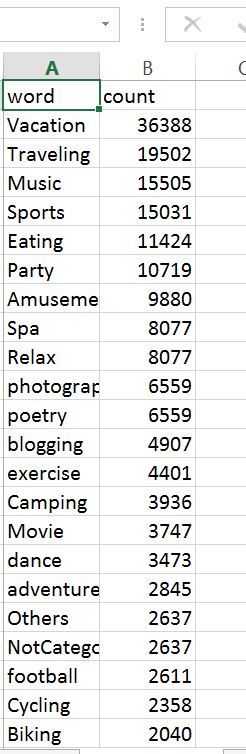
**Description :** This query will perform the map and reduce on tweets . It will get the top activities among all countries.

**Special features:** Java RDD, mapToPair, reduceByKey, flatMap, sortBy, swap

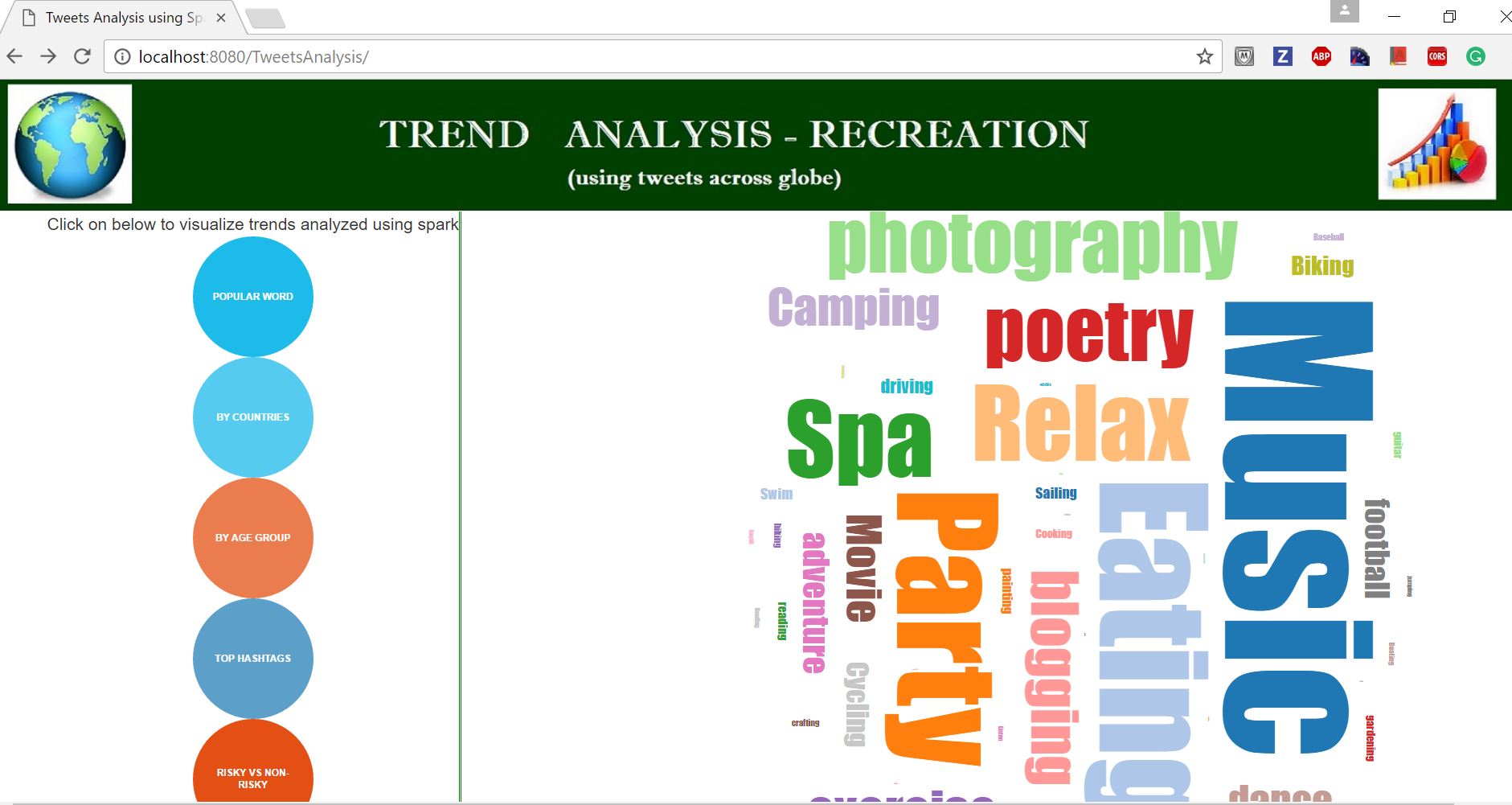
**Source code:**

****

**Output:**



Visualization:

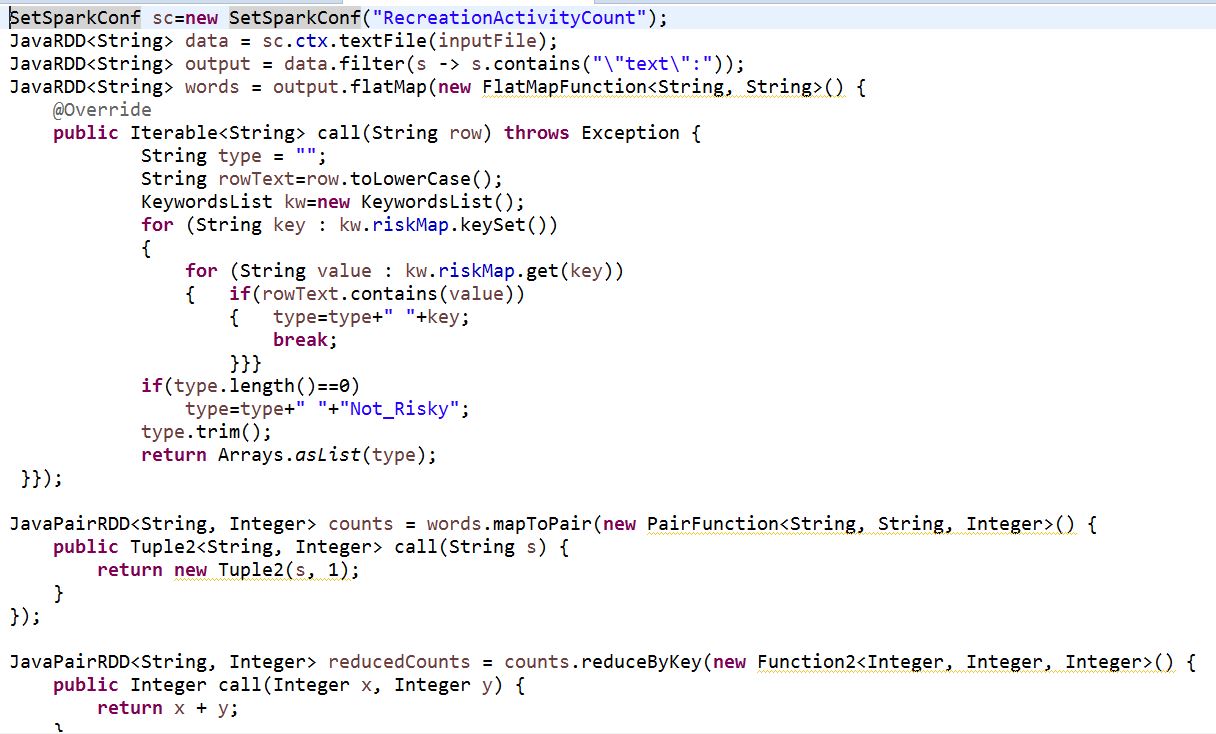


**Query 5: JAVA RDD query to classify Risky vs Non Risky Activities**

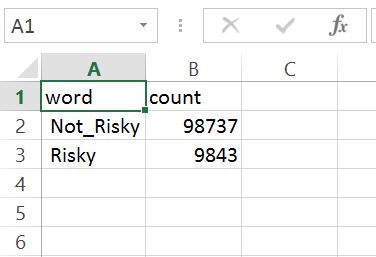
**Description:** This query will perform the map and reduce on tweets . It will classify the tweets based on people following risky activities like sky diving, jumping, kayak etc.

**Special features:** Java RDD, mapToPair, reduceByKey, flatMap

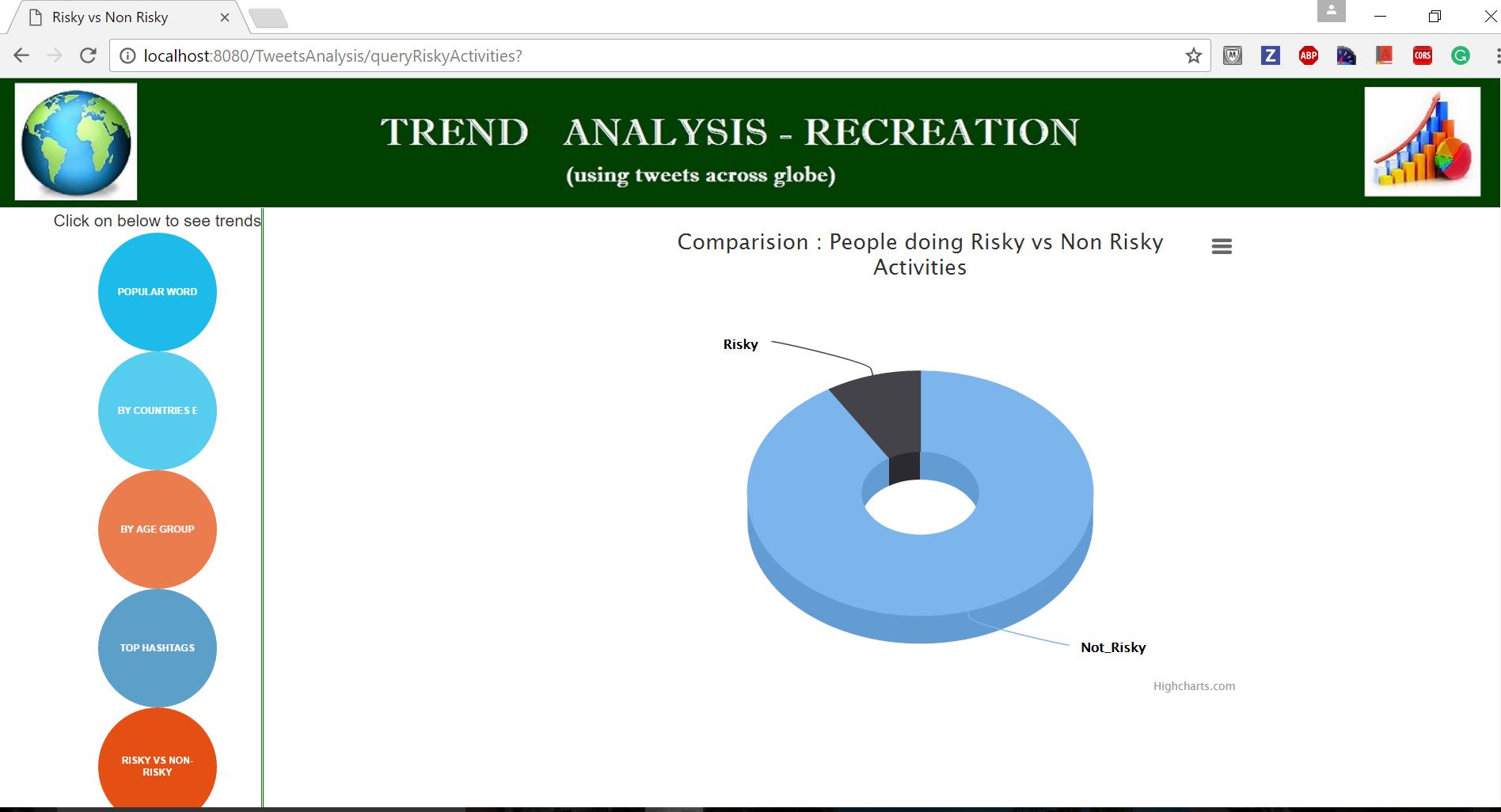
**Source code:**

****

**Output:**

****

Visualization:



Runtime Measurements

Below is total run time to execute above queries:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Q\_No | Description | Type | Features | Running time  (in ms) | (seconds) |
| 1 | By Countries | Dataframe | Union | 37722 | 38 |
| 2 | Top hashtags | Dataframe | join with hashtags file | 285737 | 286 |
| 3 | Age classification of tweets user | Dataframe | UDF with public API+ automatically collecting tweets | 62811 | 63 |
| 4 | Top activities | RDD | flatMap+ mapReduce+sortBy | 42955 | 43 |
| 5 | Risky vs non Risky | RDD | flatMap+ mapReduce | 18667 | 19 |

Extra Credits

* Used google Map in query 1 to visualize. Please refer the code for the same in “show.differnt.graphs.js” file.
* For query 5 we are dynamically executed one of the queries upon the user’s click using form submit and Java servlet execution.
* W.r.t Phase 2 we have collected 100 tweets at runtime in query 3 and then executed out query

Github URL-Code Repository

https://github.com/rashmitripathi/TwitterTrendAnalysis\_BigData

References

* <https://spark.apache.org/docs/1.6.1/sql-programming-guide.html>
* <http://twitter4j.org/en/code-examples.html>
* <https://www.uclassify.com/docs>
* <http://stackoverflow.com/questions/35211993/how-to-join-mutiple-columns-in-spark-sql-using-java-for-filtering-in-dataframe>
* <http://stackoverflow.com/questions/18471043/twitter-stream-api-to-get-users-tweets>
* <http://getbootstrap.com/css/#grid-offsetting>
* <http://www.knowstack.com/webtech/charts_demo/highchartsdemo4.html>
* <https://developers.google.com/chart/interactive/docs/gallery/geochart>