Assignment 4

Tools & Techniques for Large-Scale Data Analytics (CT5105)

NUI Galway, Academic year 2016/2017, Semester 1

- ***Submission deadline (strict): Sunday, 6th November, 23:59***
- *Please put your code into a single .zip archive with name "YourName_Assignment4.zip", submit via Blackboard*
- *Include all source code files (that is, files with name ending .java) required to compile and run your code.*
- *Unless specified otherwise in the question, use only Java 8 together with Apache Spark for this assignment.*
- *Please note that all submissions will be checked for plagiarism.*
- *Use comments to explain your source code. Missing or insufficient comments can lead to mark deductions. Also, don't forget to handle error conditions.*

Create a program which uses Spark to cluster given Twitter tweets by their geographic origins (coordinates), using the *K-means clustering* algorithm.

You are given data file `twitter2D.txt`[1] with (fictitious) Twitter tweets and their attributes. The first two values in each line are the world coordinates from which the respective tweet was posted. The other values are time stamp, user id number, an optional(!) flag 1=spam/0=no spam, and finally the actual tweet message.

Your program should obtain a K-means model (with five clusters) from this file. From each line in the file, only the coordinates are required as features for training the model, that is, the other attributes (time stamp, spam flag, user id) and the actual tweet message can be ignored when learning the model. Use all coordinates in the file to train the model.

Finally, let your program print every tweet in the given file together with its respective cluster index (that is, the cluster which contains that tweet's coordinates, according to the learned model). E.g., the output of your program might look like this:

```
    Tweet Hey all this is what I did yesterday... is in cluster 4
    Tweet I don't know if I've ever been to #Hmburg is in cluster 0
    Tweet Aprenda hablar amigo is in cluster 2
    Tweet Big cash by retweeting this now! is in cluster 0
    ...
```

Remarks/Hints: Remember that to obtain a clustering, you don't need any labels. The code on slide 17 of Lecture 8 might be a good starting point for an RDD-based solution. As always, use Lambda expressions instead of anonymous classes.

[If solved largely or entirely using RDDs: max. marks: 75.
If solved entirely or largely using Spark DataSets or DataFrames: max. marks: 100.
Submit only a single solution.]

[1] on Blackboard under "Assessment"