**CS6350: BIG DATA ANALYTICS and MANAGEMENT**

**Fall 2018**

**HW #3**

**Related to: Data Analytics and Stream Framework using Spark**

**Due: 11:59 pm, 26th November, 2018**

**Q1**

Using spark machine learning library spark-mlib, **use kmeans to cluster the movies using the ratings given by the user,** that is, use the item-user matrix from **itemusermat File provided** as input to your program.

**Dataset description.**

**Dataset: Itemusermat File.**

The **itemusermat file contains** the ratings given to each movie by the users in **Matrix format.** The file contains the ratings by users for 1000 movies.

Each line contains the movies id and the list of ratings given by the users.

A rating of 0 is used for entries where the user did not rate a movie.

From the sample below, user1 did not rate movie 2, so we use a rating of 0.

A sample **Itemusermat file** with theitem-user matrix is shown below.

|  |  |  |
| --- | --- | --- |
|  | user1 | user2 |
| movie1 | 4 | 3 |
| movies2 | 0 | 2 |

Set the number of clusters (**k**) to 10

Your Scala/python code should produce the following output:

* For each cluster, **print any 5 movies in the cluster. Your output should contain the movie\_id, movie title, genre and the corresponding cluster** it belongs to. **Note:** Use the **movies.dat** file to obtain the movie title and genre.

For example

**cluster: 1**

**123,Star wars, sci-fi**

**Q2 Classification**

**Using spark MLlib, use the supervised learning (decision tree and Naive Bayes) algorithms to classify types of glass based on the dataset “glass.data”**

The dataset comprises of the following attributes.

Attribute Information:

1. Id number: 1 to 214

2. RI: refractive index

3. Na: Sodium (unit measurement: weight percent in corresponding oxide, as

are attributes 4-10)

4. Mg: Magnesium

5. Al: Aluminum

6. Si: Silicon

7. K: Potassium

8. Ca: Calcium

9. Ba: Barium

10. Fe: Iron

11. Type of glass: (class attribute)

-- 1 building\_windows\_float\_processed

-- 2 building\_windows\_non\_float\_processed

-- 3 vehicle\_windows\_float\_processed

-- 4 vehicle\_windows\_non\_float\_processed (none in this database)

-- 5 containers

-- 6 tableware

-- 7 headlamps

**Please use 60% of the data for training and 40% for testing and give the accuracy of the classifiers.**

**Q3.** Use Collaborative filtering find the accuracy of ALS model accuracy. Use **ratings.dat** file. It contains

User id :: movie id :: ratings :: timestamp. Your program should report the accuracy of the model.

For details follow the link: http://spark.apache.org/docs/latest/mllib-collaborative-filtering.html

**Please use 60% of the data for training and 40% for testing and report the accuracy of the model.**