IDS 561 Homework 4

Due date: 04/27/2021 Tuesday 3:30pm (before class)

In this homework, you have two options: Recommender System and K-means. Please select **ONE** option and complete the required tasks on Spark.

Option 1: Recommender System

Data

MovieLens is a dataset that is collected by the GroupLens Research Project at the University of Minnesota and made available rating data sets from the MovieLens web site. Download and unzip the MovieLens 100K Dataset (ml-100k.zip). http://grouplens.org/datasets/movielens/

u.data is the dataset for this assignment

The full dataset contains 100,000 ratings by 943 users on 1682 items. Each user has rated at least 20 movies. Users and items are numbered consecutively from 1. The data is randomly ordered. The format is:

user_id<tab>item_id<tab>rating<tab>timestamp.

The time stamps are unix seconds since 1/1/1970 UTC

TODO

1. Import the MovieLens dataset.
2. Build a recommendation model using Alternating Least Squares.
3. Report the original performance (Mean Squared Error)
4. Try to improve the performance of the original model using 10-fold cross validation and solve the cold-start problem.
5. Optimize the model based on step 4 and report the improvement of performance.
6. Output top 10 movies for all the users with the following format:

userID<\tab>itemID1,itemID2,itemID3 ...,itemID10

...

A tutorial of spark can be found at:

https://spark.apache.org/docs/latest/mllib-collaborative-filtering.html

Submission

1. Python program
2. A .txt file of recommendation output with the required format.

Option 2: K-means

Data

The Iris data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters.

Download here: https://archive.ics.uci.edu/ml/datasets/iris

TODO

1. Import the Iris dataset.
2. Build a K-means model to classify the species of Iris. You can choose a k value randomly at this step.
3. Report the original performance using Silhouette score.
4. Try to improve the performance of the original model by trying at least 10 different k values.
5. Select the best k based on step 4 and print out the following sentence in your code:

   "k=xx gives the best performance, Silhouette =xx "

   (replace xx with your own numbers)

A tutorial of spark can be found at:

https://spark.apache.org/docs/latest/ml-clustering.html

Submission

Python program