Project

The purpose of this project is to practice the concepts of data mining and recommender systems that were discussed during the lectures/workshops and to implement a recommender system using “MovieLens” dataset based on a collaborative filtering system. In this project, we will be using package “recommenderlab” in R. The details of this package can be found at:

https://cran.r-project.org/web/packages/recommenderlab/recommenderlab.pdf

Please download “MovieLens” dataset - Latest Datasets (***ml-latest-small.zip*** size: 1 MB and **ml-latest.zip** 224 MB):

https://grouplens.org/datasets/movielens/latest/

You can use **ml-latest-small.zip** (1 MB) for testing your R scripts. After testing your scripts please use **ml-latest.zip** (224 MB) for the final evaluation and analysis of the results. This can save you a lot of time, as **ml-latest.zip** (224 MB) is quite large and requires a lot of time and system resources.

The datasets include the following CSV files:

- *ratings.csv*
- *movies.csv*
- tags.csv
- links.csv
- genome-tags.csv
- genome-scores.csv

Please look at the README.html for more details on each file:

http://files.grouplens.org/datasets/movielens/ml-latest-small-README.html

Please note that **ratings.csv** is main file you will be using in this project.

The following tutorial provides a sample project using “recommenderlab” using similar datasets.

https://ashokharnal.wordpress.com/2014/12/18/using-recommenderlab-for-predicting-ratings-for-movielens-data/

The following script cab be used to load “MovieLens” dataset in R:

```
> install.packages("recommenderlab")
> library(recommenderlab)
> library(ggplot2)
> library(data.table)
> library(reshape2)
> setwd("C:/MyFiles/.../Project")
> r <- read.csv("ratings.csv", header=TRUE)
> …
```

Tasks:

1. Please divide ratings.csv in two sets: Training (80%) and Test (20%) based on a random sampling method.
2. Apply User-Based Collaborative Filtering (UBCF) based on different similarity/dissimilarity methods including “cosine”, “pearson” and “jaccard” and compare the quality of recommendations based on each method.
3. You can use Normalized Mean Absolute Error (NMAE) for the comparison:

   NMAE=∑(|predicted rating – real rating|) / n(max rate – min rate)

4. Please submit your project for evaluation and feedback. Your submission should include two files:
   a. Your R script
   b. Your Report in MS Word format, that describes your project and presents the results in the form of graph(s) or table(s) and analysis of the results.