INF 553 Homework 2

My three scala codes, were saved in a package called Georgios.HW2.
As a result in front of each scala file name, I have to specify the package.
This is the code required to run it successfully:

```
spark-submit --class Georgios.HW2.ModelBasedCF Georgios_Iliadis_hw2.jar
spark-submit --class Georgios.HW2.ItemBasedCF --executor-memory 8G --driver-memory 8G Georgios_Iliadis_hw2.jar
```

I had to use –executor memory 8G –driver memory 8G on my machine for it to work – I am not entirelly sure you will need to do the same on your machine, but in case the normal:
spark-submit --class Georgios.HW2.ItemBasedCF Georgios_Iliadis_hw2.jar
does not work, the one I provided works for me.

First you have to cd where the jar file is and then use the three commands provided to run Task 1 and 2.
I have the data needed in the directory where the .jar file is.
I did not include any arguments for the command line, so the data to be loaded has to be where the .jar file is, as I have it in my submission.

By using the commands provided, it will load the file necessary and it will create the output file with the name required.

Arguments:

I have the args as comments, in case you want to load different files for training and testing.
You simply uncomment lines 22 and 23, and replace where train_review.csv with readTrainFile
and replace where test_review.csv with readTestFile
Lines 22-25 Task1
You can find it in Task2 at the start of the code too.

Task 1:

Here I implement the ModelBased CF, where I use the ALS model to predict the ratings. In case there are missing ratings, I calculate the average rating of each user and fill the missing ratings with the average. Finally I join them to create my final ratings and predictions rdd.

Accuracy Information:

```
>= 0 and < 1: 29133
>= 1 and < 2: 12905
>= 2 and < 3: 2328
>= 3 and < 4: 315
>= 4: 42
RMSE: 1.0713477776519622
Time: 80 sec
```

The text file saved is Georgios_Iliadis_ModelBasedCF.txt. It is in the format as in the example with the three columns: UserID, ProductID and Prediction rating.

Task 2:

For Task2, I implement the ItemBasedCF. I use the equations provided in the notes to calculate the weight and the prediction. Finally I do the same thing as in Task1, when there are missing ratings, I fill them with the average rating of the user and I join the two predictions to create my final ratings and predictions rdd.

Accuracy information:
>= 0 and < 1: 28465
>= 1 and < 2: 13460
>= 2 and < 3: 2462
>= 3 and < 4: 297
>= 4: 39
RMSE: 1.099870773462598
Time: 357 sec

The text file saved is Georgios_Iliadis_ItemBasedCF.txt. It is in the format as in the example with the three columns: UserID, ProductID and Prediction rating.

Georgios Iliadis
3668057286