Bashar Alhafni **INF-553**
USC ID: 9065856238 **Fall 2018 Assignment 2**

How to run the code?

The code should be executed from the command line in the following manner (Assuming the PATH is pointing to spark and Java):

spark-submit --class <className> <JarFileName.jar> <training_file> <testing_file>

Model Based CF:

To get the result for task one (Model Based CF), the following command must be executed:

```
spark-submit --class Bashar_Alhafni_ModelBasedCF
Bashar_Alhafni_hw2.jar
/Users/alhafni/Desktop/repos/inf553_data_mining/hw2/Data/train_review.
csv
/Users/alhafni/Desktop/repos/inf553_data_mining/hw2/Data/test_review.c
sv
```

Results:

```
>=0 and <1: 28781
>=1 and <2: 13108
>=2 and <3: 2237
>=3 and <4: 286
>=4: 48
RMSE = 1.0763489126861916
```

Item Based CF:

To get the result for task two (Item Based CF), the following command must be executed:

```
spark-submit --class Bashar_Alhafni_ItemBasedCF Bashar_Alhafni_hw2.jar
/Users/alhafni/Desktop/repos/inf553_data_mining/hw2/Data/train_review.
csv
/Users/alhafni/Desktop/repos/inf553_data_mining/hw2/Data/test_review.c
sv
```

Results:

```
>=0 and <1: 25188
>=1 and <2: 12896
>=2 and <3: 3699
>=3 and <4: 1330
>=4: 2123
RMSE: 1.8578234583899136
TIME: 74 sec
```

NOTE:
I found a way to reduce the RMSE of the Item Based CF method by not using the pearson correlations at all. I was able to reduce it to 1.08 and the main reason for this significant reduce is because there are so many users in the Yelp dataset who rated one and only one business. Therefore, the pearson correlation would be zero and the average rating for that user will give very good results.

Results using the above methods:

>=0 and <1: 28681
>=1 and <2: 13119
>=2 and <3: 2923
>=3 and <4: 469
>=4: 44
RMSE: 1.0883543731575855
TIME: 65 sec

Versions used:
Scala: 2.11.0
Spark: 2.3.1

References:

https://spark.apache.org/docs/2.2.0/mllib-collaborative-filtering.html