Seifallah Ben Mustapha Hanfi
274033
seifallah.benmustaphahanfi@epfl.ch

1 Similarity.predictor

1.1 Question 1

The prediction accuracy (MAE on ml-100k/u1.test) of (Eq. 3) using Adjusted Cosine similarity is 0.747765. The prediction accuracy is better than the baseline and the difference is 0.01913.

1.2 Question 2

The mathematical formulation used for the Jaccard Coefficient is $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$. The prediction accuracy obtained with this similarity metric is 0.76225. The Jaccard Coefficient is slightly worst than the Adjusted Cosine similarity and the difference is 0.014484.

1.3 Question 3

Let n be the size of U (the set of users). The number of $s_{u,v}$ computations is $C(n+1,2)$ which is 445096 for the 'ml-100k' dataset.

1.4 Question 4

The min, max, average, and standard deviation of the number of multiplication required for each possible $s_{u,v}$ for all similarities are reported in the following table:

	number of multiplications required for each $s_{u,v}$
Min	0.0
Max	685.0
Average	12.28202
Standard deviation	18.52030

1.5 Question 5

Let n be the size of U (the set of users). The amount of memory required to store all $s_{u,v}$, both zero and non-zero values is 8*$C(n+1,2)$. The number of bytes needed to store only the non-zero $s_{u,v}$ on the 'ml-100k' dataset, assuming each non-zero $s_{u,v}$ is stored as a double is 3273776.

1.6 Question 6

	The time required for computing predictions (with 1 Spark ex- ecutor), including computing the similarities $s_{u,v}$ over five measurements
Min	1.0466394591E7
Max	1.2328250616E7
Average	1.1087389305799998E7
Standard deviation	649500.28192

We notice that the average time is higher than the one for the previous methods of Milestone 1 because this method is more computational expensive due to the computation of similarities between the users and then for getting the similarities values needed for the prediction.

1.7 Question 7

	The time required for computing the similarities $s_{u,v}$ over five measurements
Min	4291478.66
Max	4806925.522
Average	4568126.661
Standard deviation	209365.10080

On average, the average time per $s_{u,v}$ in microseconds is 11.155921. The ratio between the computation of similarities and the total time required to make predictions is 0.4120110. The computation of similarities is significant for the predictions since a big part of the time needed for the predictions is taken by computing the similarities.

2 KNN.Predictor

2.1 Question 1

N° of neighboors k	**Prediction accuracy (MAE)**
10	0.840703
30	0.791422
50	0.774940
100	0.756135
200	0.748456
300	0.746931
400	0.74739
800	0.747239
943	0.74776

The lowest k such that the MAE is lower than for the baseline method is 100. It's lowest by 0.01076.

2.2 Question 2

Let n be the size of U (the set of users). The minimum number of bytes required to store only the k nearest similarity values for all possible users u for all previous values of k (with the ml-100k dataset) is given by n*k*8 and are represented in the following table:

N° of neighboors k	Minimum number of bytes
10	75440
30	226320
50	377200
100	754400
200	1508800
300	2263200
400	3017600
800	6035200
943	7106448

2.3 Question 3

The RAM available in my computer is 4GB=4e9Bytes.Given the lowest k we have provided in Q.3.1.1, the maximum number of users we could store in RAM is 4e9/100*3*8=9.8e8 assuming we were storing values in a simple sparse matrix implementation that used 3x the memory than what we have computed in the previous section.

2.4 Question 4

Varying k doesn't have an impact on the number of similarity values to compute since we have to compute all the similarities in order to choose the k best.

3 Recommendation

3.1 Question 5

My personal top 5 recommendations using the neighbourhood predictor with k=30 are :

- 11, Seven (Se7en) (1995), 5.0
- 42, Clerks (1994), 5.0
- 47, Ed Wood (1994), 5.0
- 48, Hoop Dreams (1994), 5.0
- 57, Priest (1994), 5.0

With k=300:

- 884, Year of the Horse (1997), 5.0
- 1260, Total Eclipse (1995), 5.0
- 1293, Star Kid (1997), 5.0
- 1347, Ballad of Narayama, 5.0
- 1358, The Deadly Cure (1996), 5.0

In this case, changing k results in totally different recommendations. Only one recommendation is similar to the previous Milestone for k=300.