**Problem 1**

* **Similarity measure:** In our approach, we plan to use Cosine similarity as our similarity measure to compute user to user similarities because using, although, normalization in Pearson correlation removes bias from the data, it would be a huge overhead in terms of space and computation time because we need to compute a average when every new rating arrives. Pearson correlation would give a more accurate result but we won’t be able to compute much results if we use it. Cosine, on the other hand, would give a little less accurate data but we would be able to compute on larger number of data records.
* **Number of similar users k:** Currently set to 3. Will compare with other k’s to get the best possible prediction
* **Prediction method (weighted or un-weighted average):** We would be using Weighted mean because unweighted does not take into consideration, the difference in the number of users.
* **Thresholding the ratings:** We are not going to threshold the ratings because that would make more sense while using Jaccard similarity.
* **Normalization of the ratings:** We are not going to normalize the ratings for the same reason we are using Pearson Correlation; it would ensure the results are not affected by overly enthusiastic or overly critical users.
* **User-user model or Item-item model:** We’ll use User-user model, since the memory consumed in User-User model would be much less than that in item-item model.

**Problem 2**

**(a)**

Training:

User number: **28978**

Item number: **1821**

Testing:

User number: **27555**

Item number: **1701**

**(b)**

For user-user overlap, there are **13549535** pairs in total

For item-item overlap, there are **1002208278** pairs in total.

So we should try to use the user-user model since the memory needs for user-user model is smaller.

**(c)** For the **similarity measure**, we are planning to use Cosine similarity for the reasons stated above, that although it would be a little less accurate than Pearson correlation, we would be able to predict much larger number of user ratings. And since our purpose is to get better accuracy but, at the same time, computing as much number of predictions, Cosine similarity would be a better choice.

For the **data model** choice, we would be using **User-user** model because it would consume less space and time than Item-item model.

**(d)** Code committed.