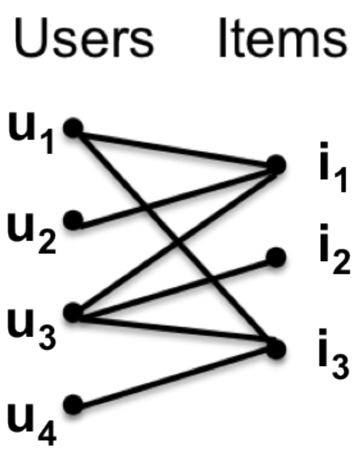
Problem Set 2

1 Recommendation Systems

Consider a user-item bipartite graph where each edge in the graph between user U to item I, indicates that user U likes item I. We also represent the ratings matrix for this set of users and items as R, where each row in R corresponds to a user and each column corresponds to an item. If user i likes item j, then $R_{i,j} = 1$, otherwise $R_{i,j} = 0$. Also assume we have m users and n items, so matrix R is $m \times n$.

Let's define a matrix P, $m \times m$, as a diagonal matrix whose i-th diagonal element is the degree of user node i, *i.e.* the number of items that user i likes. Similarly, a matrix Q, $n \times n$, is a diagonal matrix whose i-th diagonal element is the degree of item node i or the number of users that liked item i. See figure below for an example.



$$R = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$$P = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad Q = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

Figure 1: User-Item bipartite graph.

(a) [4 points]

Define the non-normalized user similarity matrix $T = R * R^T$. Explain the meaning of T_{ii} and T_{ij} $(i \neq j)$, in terms of bipartite graph structures (See Figure 1) (e.g. node degrees, path between nodes, etc.).

Cosine Similarity: Recall that the cosine similarity of two vectors u and v is defined as:

$$\textit{cos-sim(u,v)} = \frac{u \cdot v}{\|u\|\|v\|}$$

(b) [8 points]

Let's define the *item similarity matrix*, S_I, $n \times n$, such that the element in row i and column j is the cosine similarity of *item* i and *item* j which correspond to column i and column j

of the matrix R. Express S_I in terms of R, P and Q. Your answer should be an operation on the matrices, in particular you should not define each coefficient of S_I individually.

Repeat the same question for *user similarity matrix*, S_U where the element in row i and column j is the cosine similarity of *user* i and *user* j which correspond to row i and row j of the matrix R.

Your answer should show how you derived the expressions.

(Note: To make the element-wise square root of a matrix, you may write it as matrix to the power of $\frac{1}{2}$.)

(c) [8 points]

The recommendation method using user-user collaborative filtering for user u, can be described as follows: for all items s, compute $r_{u,s} = \Sigma_{x \in users} \text{cos-sim}(x, u) * R_{xs}$ and recommend the k items for which $r_{u,s}$ is the largest.

Similarly, the recommendation method using item-item collaborative filtering for user u can be described as follows: for all items s, compute $r_{u,s} = \Sigma_{x \in items} R_{ux} * \text{cos-sim}(x, s)$ and recommend the k items for which $r_{u,s}$ is the largest.

Let's define the recommendation matrix, Γ, $m \times n$, such that $\Gamma(i, j) = r_{i,j}$. Find Γ for both item-item and user-user collaborative filtering approaches, in terms of R, P and Q.

Your answer should show how you derived the expressions.

(d) [15 points]

In this question you will apply these methods to a real dataset. The data contains information about TV shows. More precisely, for 9985 users and 563 popular TV shows, we know if a given user watched a given show over a 3 month period.

Download the dataset[1] on http://snap.stanford.edu/class/cs246-data/hw2-q1-dataset.zip.

The ZIP file contains:

- `user-shows.txt` This is the ratings matrix R, where each row corresponds to a user and each column corresponds to a TV show. $R_{ij} = 1$ if user i watched the show j over a period of three months. The columns are separated by a space.
- `shows.txt` This is a file containing the titles of the TV shows, in the same order as the columns of R.

We will compare the user-user and item-item collaborative filtering recommendations for the 500^{th} user of the dataset. Let's call him Alex.

In order to do so, we have erased the first 100 entries of Alex's row in the matrix, and replaced them by 0s. This means that we don't know which of the first 100 shows Alex has watched. Based on Alex's behaviour on the other shows, we will give Alex recommendations on the first 100 shows. We will then see if our recommendations match what Alex had in fact watched.

- Compute the matrices P and Q.
- Using the formulas found in part (c), compute Γ for the user-user collaborative filtering. Let S denote the set of the first 100 shows (the first 100 columns of the matrix). From all the TV shows in S, which are the five that have the highest similarity scores for Alex? What are their similarity scores? In case of ties between two shows, choose the one with smaller index. Do not write the index of the TV shows, write their names using the file `shows.txt`.
- Compute the matrix Γ for the movie-movie collaborative filtering. From all the TV shows in S, which are the five that have the highest similarity scores for Alex? In case of ties between two shows, choose the one with smaller index. Again, hand in the names of the shows and their similarity score.

[1]The original data is from Chris Volinksy's website at http://www2.research.att.com/~volinsky/DataMining/Columbia2011/HW/HW6.html.

Alex's original row is given in the file `alex.txt`. For a given number k, **the true positive rate at top-k** is defined as follows: using the matrix Γ computed previously, compute the top-k TV shows in S that are most similar to Alex (break ties as before). The true positive rate is the number of top-k TV shows that were watched by Alex in reality, divided by the total number of shows he watched in the held-out 100 shows.

- Plot the true positive rate at top-k (defined above) as a function of k, for $k \in [1, 19]$, with predictions obtained by the user-user collaborative filtering.
- On the same figure, plot the true positive rate at top-k as a function of k, for $k \in [1, 19]$, with predictions obtained by the item-item collaborative filtering.

What to submit:

(i) Interpretation of T_{ii} and T_{ij} [for 1(a)]

(ii) Expression of S_I and S_U in terms of R, P and Q and accompanying explanation [for 1(b)]

(iii) Expression of Γ in terms of R, P and Q and accompanying explanation [for 1(c)]

(iv) The answer to this question should include the followings: [for 1(d)]

- The five TV shows that have the highest similarity scores for Alex for the user-user collaborative filtering
- The five TV shows that have the highest similarity scores for Alex for the item-item collaborative filtering
- The graph of the true positive rate at top-k for both the user-user and the item-item collaborative filtering
- Upload the source code