Both matrices are symetric square and real.

Symetric: $(M^T M)^T = M^T (M^T)^T = M^T M$. And $(MM^T) = (M^T)^T M^T = MM^T$

Square: $M^T M$ is size $q$ x $q$. $MM^T$ is size $p$ x $p$. If we multiply matrix size $p$ x $q$ by its transponse, which is size $q$ x $p$ we get square matrix size $p$ x $p$.

Real: If M is real, the product of M and its transponse will also be real.

For each eigenvalue $MM^T$, we can write $MM^Tv = v\lambda$. By multypling both sides by prefix of $M^T$, we get $M^TM(M^Tv) = (M^Tv)\lambda$, where are the parentheses can be added because of assiciativity rule. $\lambda$ is also an eigenvalue of $M^TM$, with a corresponding eigenvector being $M^Tv$. The eigenvector are ussualy not the same. An exception is $M^T = I$.

We can write $M^T M = Q\lambda Q^T$, because $M^T M$ is symetric, square and real by 1(a).

$M^T M = V \sum U^T U \sum V^T = V \sum^2 V^T$, which matches 1(c).

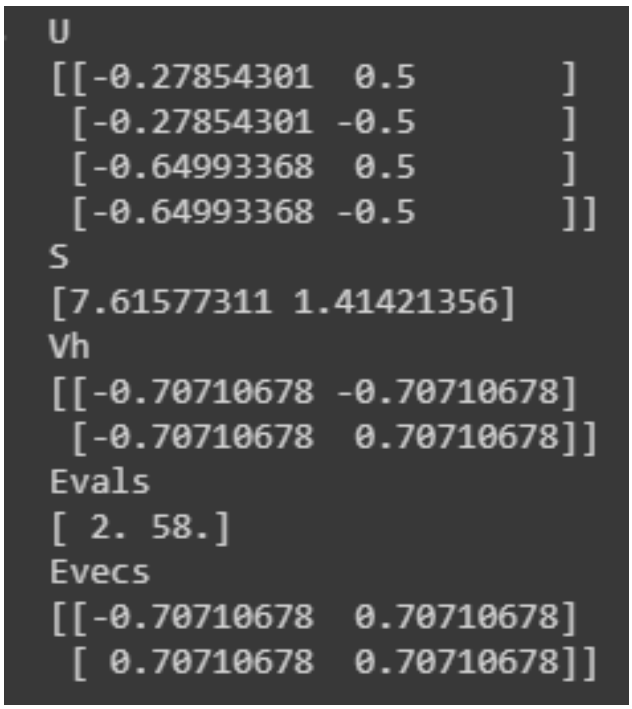


Figure 1: Output

On figure 1 we can see output of assignment.

V is equvivalent to the matrix of the eigenvectors if we reorder the colums based on ordering of their singular values.

Each and every singular value of M is square root of an eigenvalues of $M^T M$.

| Initialization | Percentage change after 10 iterations |
|---|---|
| c1.txt | 26.5% |
| c2.txt | 76.7% |

Table 1: Euclidian distance

C2 is better than c1, because it distributes the initial clusters far apart. Because there is less overlap true clusters are split less often which leads to a better final set of clusters.

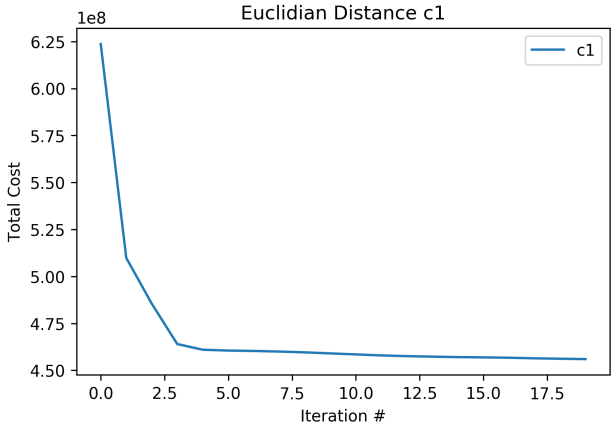


Figure 2: Euclidian graph c1

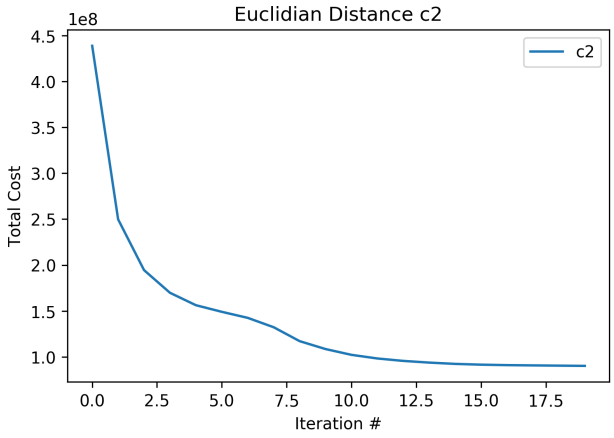


Figure 3: Euclidian graph c2

| Initialization | Percentage change after 10 iterations |
|---|---|
| c1.txt | 18.7% |
| c2.txt | 51.6% |

Table 2: Manhattan distance

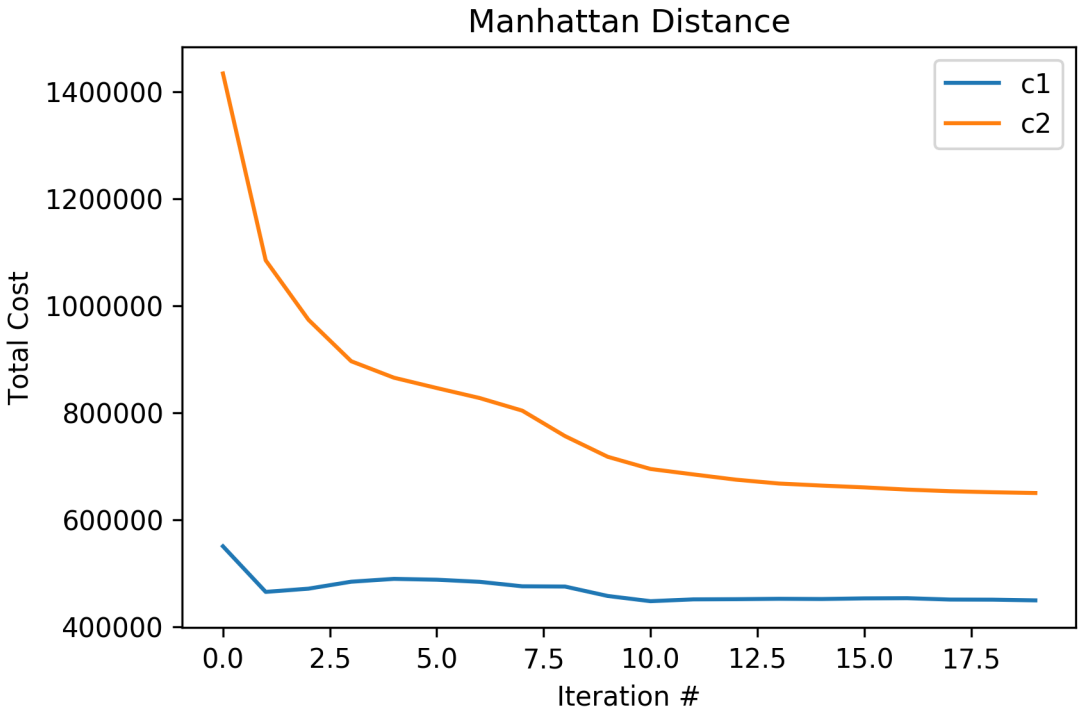


Figure 4: Manhattan graph

For the Manhattan distance, random initialitiazation doesn't apper to be better, because points in c2 are far apart in Euclidian distance, they are not necessarily apart in the Manhattan distance.

$$\epsilon_{iu} = 2 * (R_{iu} - q_i * p_u^T)$$

For SGD we then denote:

$$\frac{\vartheta E}{\vartheta q_i} = \left[ \sum_{u_{(iu)\epsilon ratings}} 2(R_{iu} - q_i p_u^T)(-p_u) \right] + 2\lambda q_i$$

$$\frac{\vartheta E}{\vartheta p_u} = \left[ \sum_{u_{(iu)\epsilon ratings}} 2(R_{iu} - q_i p_u^T)(-q_i) \right] + 2\lambda p_u$$

$$q_i := q_i + \eta * (\epsilon_{iu} * p_u - 2 * \lambda * q_i)$$

$$p_u := p_u + \eta * (\epsilon_{iu} * q_i - 2 * \lambda * p_u)$$

We got best result for $\eta = 0.015$.

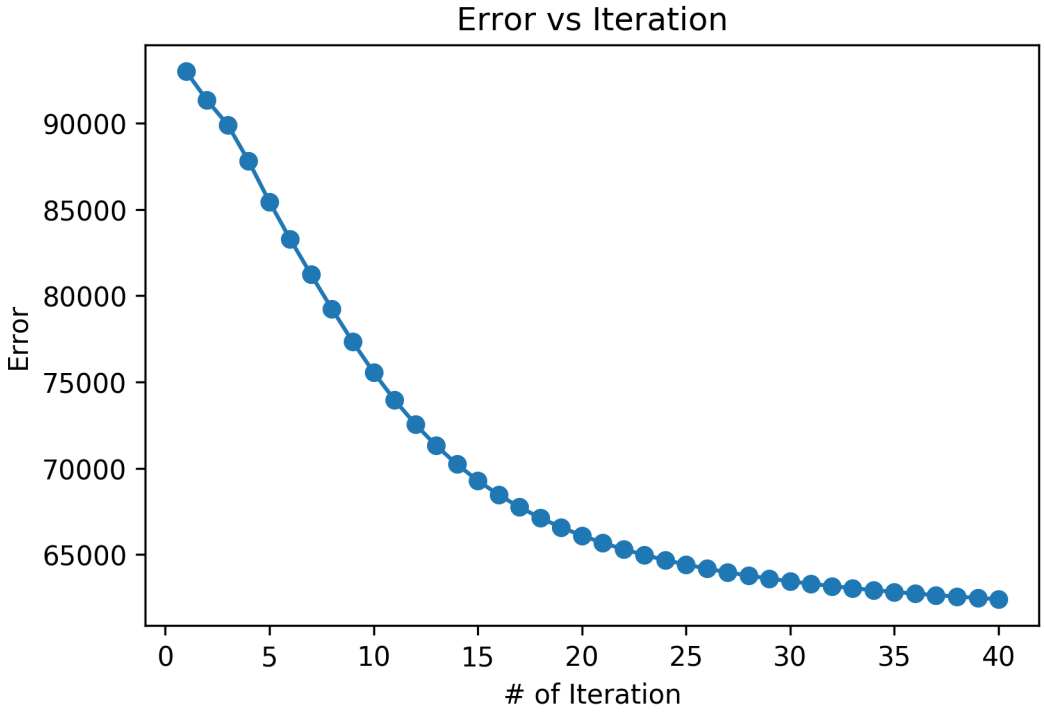


Figure 5: Error vs Iterations

$T_{ii}$ is the number or items that user i likes, so out-degree of user node i.

Some math behind this: $T_{ii} = \sum_{j=1}^{n} R_{ij} * (R^T)_{ji} = \sum_{j=1}^{n} R_{ij}^2 = \sum_{j=1}^{n} R_{ij}$. $R_{ij} = 0$ or 1, so $T_{ii} =$ degree(user).

$T_{ij}$ is the number of paths between user i and user j with the length of 2. It also represents the number of items they both like. I.e. Number of common neighbours uf user i and j.

Again some math behind this: $T_{ij} = \sum_{k=1}^{n} R_{ik} * R_{jk}, R_{ik} * R_{jk} (\neq 0)$ this means it exists a 2 step path started from user i ended in user j via $\text{item}_k$.

A diagonal matrix can scale the columns of its left multiplier, so to normlaize items we use $RQ^{-\frac{1}{2}}$. So, similirar to 4(a) the item similiratiy matrix $S_I = (RQ^{-\frac{1}{2}})^T RQ^{-\frac{1}{2}} = Q^{-\frac{1}{2}} R^T RQ^{-\frac{1}{2}}$.

Simililar to $S_I$, $S_U = (R^T P^{-\frac{1}{2}})^T R^T P^{-\frac{1}{2}} = P^{-\frac{1}{2}} RR^T P^{-\frac{1}{2}}$.

User-user collaborative filtering: $\Gamma_{us} = \sum_{x \epsilon users} cos - sim(x, u) * R_{x,s}$

So $\Gamma = S_U R = P^{-\frac{1}{2}} R R^T P^{-\frac{1}{2}} R$.

Item-item collaborative filtering: $\Gamma_{us} = \sum_{x \epsilon items} R_{ux} * cos - sim(x, s)$

So $\Gamma = R S_I = R Q^{-\frac{1}{2}} R^T R Q^{-\frac{1}{2}}$.

Note: We take $i_{alex} = 499$ due to 0-indexing.

User_user shows:

- FOX 28 News at 10 pm
- Family Guy
- 2009 NCAA Basketball Tournament
- NBC 4 at Eleven
- Two and a Half Men

item_item shows:

- FOX 28 News at 10 pm
- Family Guy
- NBC 4 at Eleven
- 2009 NCAA Basketball Tournament
- Access Hollywood

There is no significant advantage to any of used mehtods. Precision also decreases for both datasets as k increases.

# Information sheet
# CS246: Mining Massive Data Sets

**Assignment Submission** Fill in and include this information sheet with each of your assignments. This page should be the last page of your submission. Assignments are due at 11:59pm and are always due on a Thursday. All students (SCPD and non-SCPD) must submit their homework via Gradescope (http://www.gradescope.com). Students can typeset or scan their homework. Make sure that you answer each (sub-)question on a separate page. That is, one answer per page regardless of the answer length. Students also need to upload their code on Gradescope. Put all the code for a single question into a single file and upload it.

**Late Homework Policy** Each student will have a total of *two* late periods. *Homework are due on Thursdays at 11:59pm PT and one late period expires on the following Monday at 11:59pm PT*. Only one late period may be used for an assignment. Any homework received after 11:59pm PT on the Monday following the homework due date will receive no credit. Once these late periods are exhausted, any assignments turned in late will receive no credit.

**Honor Code** We strongly encourage students to form study groups. Students may discuss and work on homework problems in groups. However, each student must write down their solutions independently, i.e., each student must understand the solution well enough in order to reconstruct it by him/herself. Students should clearly mention the names of all the other students who were part of their discussion group. Using code or solutions obtained from the web (GitHub/Google/previous year's solutions etc.) is considered an honor code violation. We check all the submissions for plagiarism. We take the honor code very seriously and expect students to do the same.

**Your name:** _____

**Email:** _____ **SUID:** _____

Discussion Group: _____

I acknowledge and accept the Honor Code.

*(Signed)* _____