CS246: Mining Massive Datasets
Homework 2

Answer to Question 1(a)

Below is the proof, this proof use the properties of M: $\sum_i M_{ij} = \mathbf{1}$

$$\begin{aligned}
\omega(r') &= \omega(Mr) \\
&= \sum_i \sum_j M_{ij} r_j \\
&= \sum_j r_j \sum_i M_{ij} \\
&= \sum_j r_j 1 \\
&= \sum_j r_j \\
&= \omega(r)
\end{aligned}$$

Answer to Question 1(b)

$$\begin{aligned}
\omega(r') &= \omega(Mr) \\
&= \sum_i (\sum_j \beta M_{ij} r_j + (1-\beta)/n) \\
&= \sum_j r_j \sum_i \beta M_{ij} + \sum_i (1-\beta)/n \\
&= \beta \sum_j r_j + (1-\beta)
\end{aligned}$$

In order to make $\omega(r') = \omega(r)$:

$$\begin{aligned}
\beta \sum_j r_j + (1-\beta) &= \sum_j r_j \\
\omega(r) &= \sum_j r_j = 1
\end{aligned} \tag{1}$$

Answer to Question 1(c)

We can divide the node into two sets one for dead nodes and the other for live nodes

$$r'_i = \beta \sum_j M_{ij} r_j + [(1-\beta) \sum_{j \in live} r_j + \sum_{j \in dead} r_j]/n$$

As we know for dead node: $\sum_{j \in dead} M_{ij} = \mathbf{0}$. Therefore,

$$\begin{aligned} r'_i &= \beta \sum_{j \in live} M_{ij} r_j + (1-\beta) \sum_{j \in live} r_j/n + \sum_{j \in dead} r_j/n \\ &= \sum_{j \in live} (\beta M_{ij} + (1-\beta)/n) r_j + \sum_{j \in dead} r_j/n \end{aligned} \tag{2}$$

Then use above equation to prove $\omega(r') = \omega(r)$:

$$\begin{aligned} \omega(r') &= \omega(Mr) \\ &= \sum_i \sum_{j \in live} (\beta M_{ij} + (1-\beta)/n) r_j + \sum_i \sum_{j \in dead} r_j/n \\ &= \sum_{j \in live} r_j \sum_i (\beta M_{ij} + (1-\beta)/n) + \sum_{j \in dead} r_j 1 \\ &= \sum_{j \in live} r_j + \sum_{j \in dead} r_j \\ &= \sum_j r_j \\ &= \omega(r) \end{aligned} \tag{3}$$

Answer to Question 2(a)

MM^T and M^TM are symmetric.

$$(MM^T)^T = MM^T$$

$$(M^TM)^T = M^TM$$

It is obvious they are square matrix, MM^T has dimension of $p \times p$ and M^TM has dimenstion of $q \times q$. Given most of the data are real number, then these matrix are real as well.

Answer to Question 2(b)

Suppose **e** is an eigenvector of $M^T M$, that is:

$$M^T M e = \lambda e$$

Multiply both sides by M:

$$M M^T (Me) = M\lambda e = \lambda (Me)$$

Where $Me \neq \mathbf{0}$, becuase if $Me = \mathbf{0}$ then $M^T Me = \mathbf{0}$ thus e cannot be eigenvector. Therefore, we find Me is a eigenvector of MM^T, and λ is also the eigenvalue if MM^T.

In conclusion, MM^T and $M^T M$ has the same eigenvalue. Their eigenvectors are not necessarily the same.

Answer to Question 2(c)

Since already approved in part a that $M^T M$ is a real, symmetric and square matrix. Then we can apply eigenvalue decomposition of a real,symmetric and square matrix:

$$M^T M = Q \Lambda Q^T \tag{4}$$

Answer to Question 2(d)

Apply the SVD decomposition for matrix M, also it is important to note that $U^T U = \mathbf{I}$

$$\begin{aligned} M^T M &= (U\Sigma V^T)^T (U\Sigma V^T) \\ &= V\Sigma^T U^T U \Sigma V^T \\ &= V\Sigma^T \Sigma V^T \\ &= V\Sigma^2 V^T \end{aligned} \tag{5}$$

Answer to Question 2(e)

1. From SVD we obtain:

$$M = U\Sigma V^T$$

Where,

$$U = \begin{pmatrix} -0.27854301 & 0.5 \\ -0.27854301 & -0.5 \\ -0.64993368 & 0.5 \\ -0.64993368 & -0.5 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 7.61577311 & 0 \\ 0 & 1.41421356 \end{pmatrix}$$

$$V^T = \begin{pmatrix} -0.70710678 & -0.70710678 \\ -0.70710678 & 0.70710678 \end{pmatrix}$$

2. The eigenvalue and eigenvector for $M^T M$:

$$Evectors = \begin{pmatrix} 0.70710678 & -0.70710678 \\ 0.70710678 & 0.70710678 \end{pmatrix}$$

$$Evals = 58, 2$$

3. Based on the experiments, it is found that the eigenvector of $M^T M$ is the same for the V in the SVD. This can be proved using the result in 2d:

$$M^T M = V\Sigma^T \Sigma V^T$$

Therefore, V is the matrix composed by its eigenvectors.

4. Based on the experiments, it is found that the singular value of M (denote as s) and eigenvalue of $M^T M$ (denote as λ):

$$s^2 = \lambda \tag{6}$$

It is obvious to prove: the singular value of M are on the diagonal of Σ while the eigenvalue of $M^T M$ are on the diagonal of $\Sigma^T \Sigma$.

Answer to Question 3(a)

5 nodes with top page rank: [53, 14, 40, 1, 27]
5 nodes with lowest page rank: [85, 59, 81, 37, 89]

Answer to Question 3(b)

5 nodes with top hubbiness: [59, 39, 22, 11, 58]
5 nodes with top hubbiness: [9, 35, 15, 95, 53]
5 nodes with top authority: [66, 40, 27, 53, 1]
5 nodes with top authority: [54, 33, 24, 67, 50]

Answer to Question 4(a)

Percentage change:
Euclidean c1.txt 0.26398863292043157
Euclidean c2.txt 0.7525973243724743

It is found that by initializing with c2.txt the cost is smaller than random initialization. Random initialization seems get trapped in its local optimal.

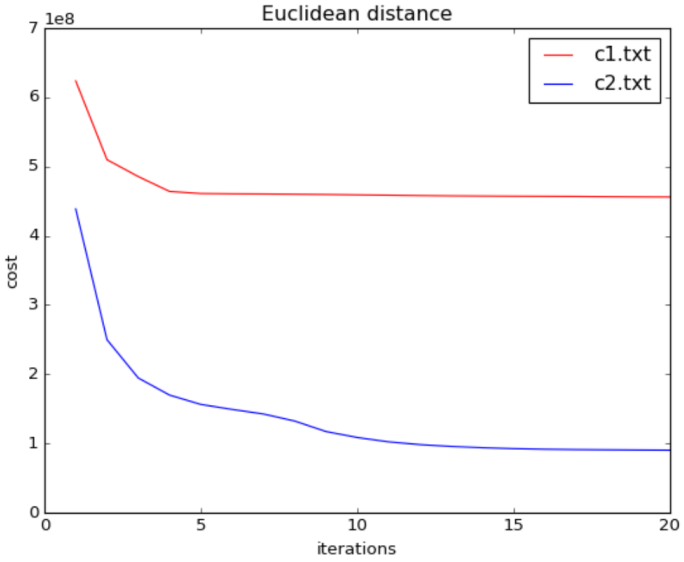

Figure 1: Euclidean Distance

Answer to Question 4(b)

Percentage change:
Manhattan c1.txt 0.16884444201690121
Manhattan c2.txt 0.49967689555556216

From the figure below we can find the result is different from previous part. When using different distance measure, the ideal clustering is changed. Therefore c2.txt may not be closer to the ideal clustering than c1.txt.

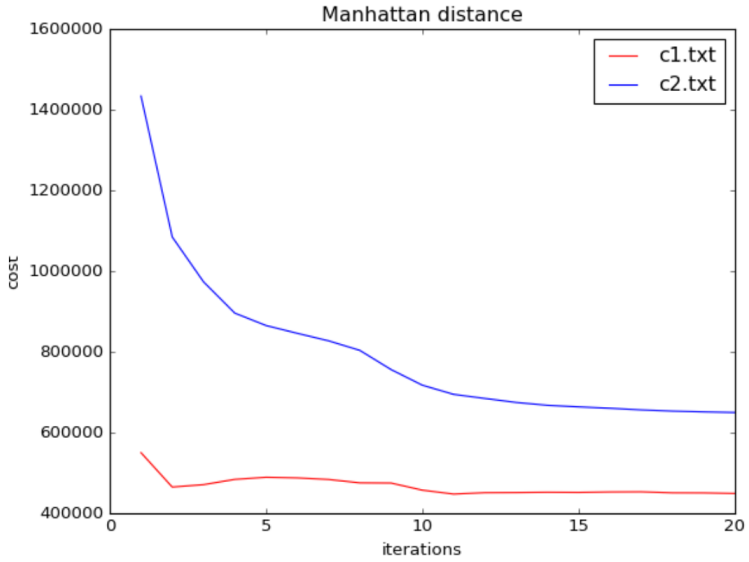

Figure 2: Manhattan Distance

Cover Sheet

Assignment Submission Fill in and include this cover sheet with each of your assignments. Assignments are due at 11:59pm. All students (SCPD and non-SCPD) must submit their homeworks via GradeScope (`http://www.gradescope.com`). Students can typeset or scan their homeworks. Make sure that you answer each question on a separate page. Students also need to upload their code at `http://snap.stanford.edu/submit`. Put all the code for a single question into a single file and upload it. Please do not put any code in your GradeScope submissions.

Late Day Policy Each student will have a total of *two* free late periods. *One late period expires at the start of each class.* (Homeworks are usually due on Thursdays, which means the first late periods expires on the following Tuesday.) Once these late periods are exhausted, any assignments turned in late will be penalized 50% per late period. However, no assignment will be accepted more than *one* late period after its due date.

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 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: Jiajun Sun

Email: jiajuns@stanford.edu **SUID:** jiajuns

Discussion Group (People with whom you discussed ideas used in your answers):

On-line or hardcopy documents used as part of your answers:

Use the answer for this on my kmeans question.
http://stackoverflow.com/questions/19844649/java-read-file-and-store-text-in-an-array

I acknowledge and accept the Honor Code.

(Signed) 孙嘉缘