CS550: Massive Data Mining and Learning
Homework 2

Due 11:59pm Friday, March 12, 2021
Only one late period is allowed for this homework (11:59pm Saturday 3/13)

Submission Instructions

Assignment Submission Include a signed agreement to the Honor Code with this assignment. Assignments are due at 11:59pm. All students must submit their homework via Canvas. Students can typeset or scan their homework. Students also need to include their code in the final submission zip file. Put all the code for a single question into a single file.

Late Day Policy Each student will have a total of *two* free late days, and for each homework only one late day can be used. If a late day is used, the due date is 11:59pm on the next day.

Honor Code Students may discuss and work on homework problems in groups. This is encouraged. However, each student must write down their solutions independently to show they understand the solution well enough in order to reconstruct it by themselves. 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 is considered an honor code violation. We check all the submissions for plagiarism. We take the honor code seriously and expect students to do the same.

Discussion Group (People with whom you discussed ideas used in your answers):
Rohan Shah (ras513)

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

I acknowledge and accept the Honor Code.

(Signed) RG

If you are not printing this document out, please type your initials above.

Answer to Question 1(a)

MM^T is square: M has dimensions $p \times q$, and M^T has dimensions $q \times p$, so MM^T has dimensions $p \times p$. MM^T is real, because the values of M are all real. MM^T is symmetric, as shown below.

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

M^TM is square: M^T has dimensions $q \times p$, and M has dimensions $p \times q$, so M^TM has dimensions $q \times q$. M^TM is real, because the values of M are all real. M^TM is symmetric, as shown below.

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

Answer to Question 1(b)

Let v be an eigenvector of MM^T, with λ as the corresponding eigenvalue. Thus, $MM^Tv = \lambda v$. We can multiply both sides of this equation by M^T: $M^TMM^Tv = M^T\lambda v = \lambda M^Tv$. Since M^T has dimensions $q \times p$, and v has dimensions $p * 1$, M^Tv is a vector with dimensions $q * 1$. Therefore, M^Tv can be treated as an eigenvector for M^TM with corresponding eigenvalue λ. This means that MM^T and M^TM have the same eigenvalues, although the corresponding eigenvectors will be different as long as M is not the identity matrix.

Answer to Question 1(c)

As proven in 1.a, M^TM is a real, symmetric, and square matrix. So, using the definition provided, M^TM can be expressed with Q, Λ, and Q^T as:

$$M^TM = Q\Lambda Q^T$$

Answer to Question 1(d)

If $M = U\Sigma V^T$ as per SVD, then $M^T = (U\Sigma V^T)^T = (V^T)^T\Sigma^T U^T = V\Sigma^T U^T$. Since Σ is a square diagonal matrix, $\Sigma^T = \Sigma$. Therefore, in terms of V, V^T, and Σ, M^TM can be expressed as:

$$M^TM = V\Sigma^T U^T U\Sigma V^T = V\Sigma^2 V^T$$

Answer to Question 1(e)(a)

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

$$\Sigma = \begin{bmatrix} 7.61577311 & 1.41421356 \end{bmatrix}$$

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

Answer to Question 1(e)(b)

$$Evals = \begin{bmatrix} 58.0 & 2.0 \end{bmatrix}$$

$$Evacs = \begin{bmatrix} 0.70710678 & -0.70710678 \\ 0.70710678 & 0.70710678 \end{bmatrix}$$

Answer to Question 1(e)(c)

Transposing V^T shows that $V^T = V$. The first column of $Evecs$ is equal to -1 times the first column of V, and the second column of $Evecs$ is exactly equal to the second column of V.

Answer to Question 1(e)(d)

The singular values as shown in Σ are the square roots of the eigenvalues as shown in $Evals$.

Answer to Question 2(a)

$$r'_i = \sum_{j=1}^{n} M_{ij} r_j$$

$$w(r') = \sum_{i=1}^{n} r'_i = \sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} r_j = \sum_{i=1}^{n} (M_{i1} r_1 + \cdots + M_{in} r_n) = r_1 \sum_{i=1}^{n} M_{i1} + \cdots + r_n \sum_{i=1}^{n} M_{in}$$

Since there are no dead ends in the Web, then all columns of M have k values of $1/k$ that sum to 1. Therefore, the last expression above can be simplified to $r_1 \cdot 1 + \cdots + r_n \cdot 1$. The proof then continues as follows:

$$w(r') = r_1 + \cdots + r_n = \sum_{i=1}^{n} r = w(r)$$

$$\therefore w(r') = w(r)$$

Answer to Question 2(b)

If a teleportation probability is introduced, then $w(r) = w(r')$ if $w(r) = 1$.

$$w(r') = \sum_{i=1}^{n} r'_i = \sum_{i=1}^{n} \left(\beta \sum_{j=1}^{n} M_{ij} r_j + \frac{1-\beta}{n} \right) = \frac{1-\beta}{n} \cdot n + \beta \sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} r_j$$

As proven in **2.a**, $\sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} r_j = w(r)$. So, we can simplify the above expression as follows:

$$w(r') = 1 - \beta + \beta w(r)$$

If we were to substitute in $w(r)$ for $w(r')$ in this equation (or vice versa), the only solution for the variable would be 1.

Answer to Question 2(c)(a)

Let L be the set of "live" nodes and D be the set of "dead" nodes.

$$r'_i = \beta \sum_{j=1}^{n} M_{ij} r_j + \frac{1-\beta}{n} + \frac{\beta}{n} \sum_{d \in D} r_d$$

Answer to Question 2(c)(b)

Again, let L be the set of "live" nodes and D be the set of "dead" nodes.

$$w(r') = \sum_{i=1}^{n} r'_i = \sum_{i=1}^{n} \left(\beta \sum_{j=1}^{n} M_{ij} r_j + \frac{1-\beta}{n} + \frac{\beta}{n} \sum_{d \in D} r_d \right)$$

$$= n \cdot \left(\frac{1-\beta}{n} + \frac{\beta}{n} \sum_{d \in D} r_d \right) + \beta \sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} r_j = 1 - \beta + \beta \sum_{d \in D} r_d + \beta \sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} r_j$$

As proven in **2.a**, $\sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} r_j = w(r)$. Furthermore, for any dead nodes j, the value of $\sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} r_j$ is 0. Therefore, $w(r) = \sum_{i=1}^{n} \sum_{j=1}^{n} M_{ij} r_j$ can be simplified to $\sum_{l \in L} r_l$. The final expression above can then be simplified:

$$w(r') = 1 - \beta + \beta \sum_{d \in D} r_d + \beta \sum_{l \in L} r_l = 1 - \beta + \beta \left(\sum_{d \in D} r_d + \sum_{l \in L} r_l \right)$$

Since the Web is composed entirely of dead nodes and live nodes, $\sum_{d \in D} r_d$ and $\sum_{l \in L} r_l$ sum to $\sum_{j=1}^{n} r_j$, or $w(r)$.

$$w(r') = 1 - \beta + \beta w(r)$$

Substituting in the assumed value of 1 for $w(r)$, $w(r')$ has a value of 1.

Answer to Question 3(a)

The 5 node IDs with the highest PageRank scores are:

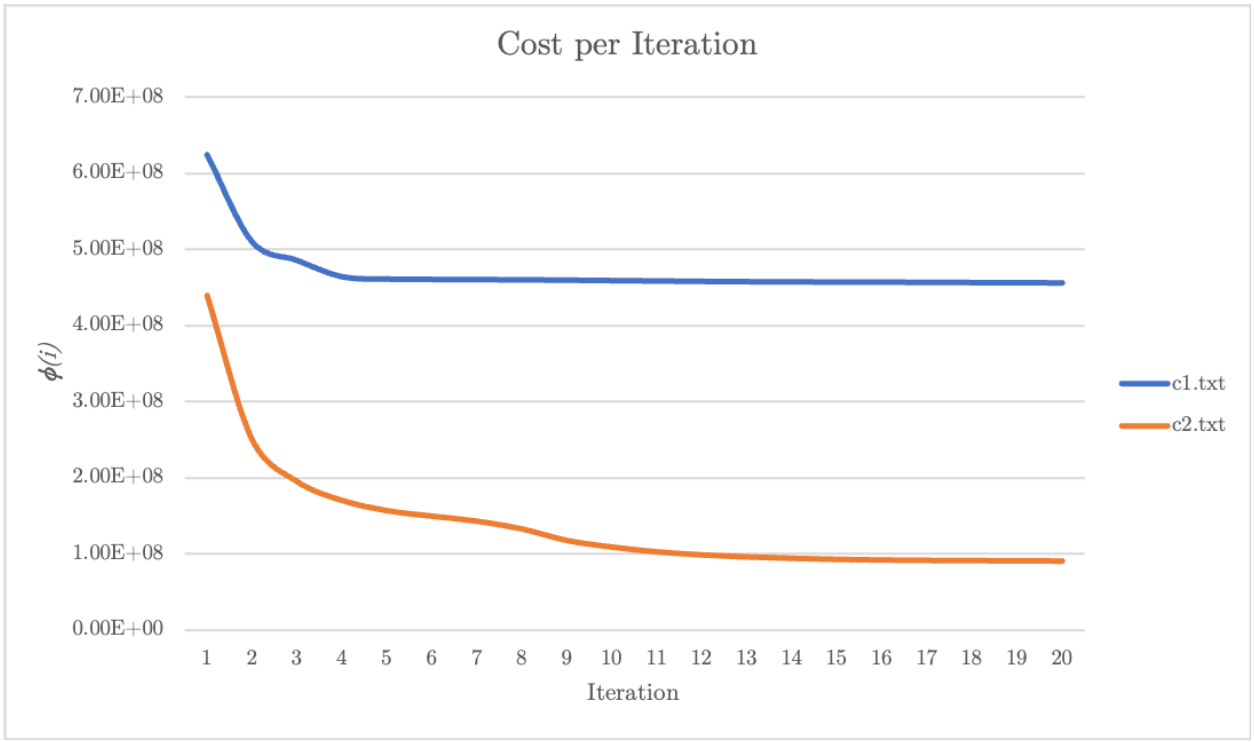
Node ID	PageRank Score
53	0.03786861
14	0.03586677
1	0.03514138
40	0.03383064
27	0.0331302

Answer to Question 3(b)

The 5 node IDs with the lowest PageRank scores are:

Node ID	PageRank Score
85	0.00323481
59	0.00344425
81	0.00358043
37	0.00371428
89	0.00383985

Answer to Question 4(a)



Answer to Question 4(b)

The percentage improvement after 10 iterations for `c1.txt` is **26.40%**. The percentage improvement after 10 iterations for `c2.txt` is **75.26%**.

Random initialization of k-means using `c1.txt` is not better than initialization using `c2.txt` in terms of cost $\phi(i)$. As can be seen in **4.a**, the cost function using random initialization converges to its lower bound much quicker, and this lower bound is significantly greater than it would be if the centroids from `c2.txt` were used. Optimal clusters do not have a high likelihood of being achieved if random initialization is used, resulting in high error values. By choosing centroids that are as far away from each other as possible, we place data points in more defining clusters and thus choose more meaningful centroids.