Jay Patel
CS422

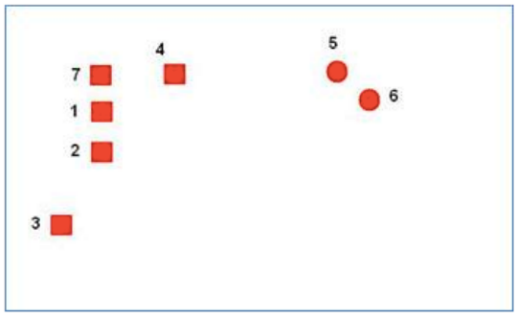
Assignment 6

CLUSTERING | RECOMMENDER SYSTEM | PAGE RANK

I Clustering

Empty cluster can form due to many reasons. It can be ranging from, bad selection of initial centroid to inappropriate number of k. The Kmeans algorithm is not made to avoid such situation. Hence it is possible that due to placement of data points or other situation, the result can lead to an empty cluster.

In our case ,we want k = 3.



Analyzing this graph, we see that there are two naturally occurring clusters. We can assume the we dont know this number, and we can keep our k = 3.

Initially, we chose 3, 5, and 6 as our starting cluster centers. Based on the graph provided in the assignment, we can say that after first iteration, 3, 2, 1, & 7 will be in one cluster, 4 & 5 will in second, and 6 will be in third. This is calculated based on the close distance of points from initial centers.

Important thing to note here is that, after the first iteration, even though it seems that 4 and 3 are grouped together, the distance between 4 and 5 is fairly smaller than 4 and 3. As the algorithm continue, the center points are updated. The new center points are near 2 (RED), between 4 and 5 (BLUE), and on 6 (GREEN). Now, the 4 will move towards RED due to 7, 1, 2 are closer now, and 5 is closer to 6(GREEN), and at the end of second iteration we are left if an empty cluster BLUE, in the middle.

$$\min_{x} \sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij} ||p_i - c_j||^2$$

subject to:

$$\sum_{j=1}^{k} x_{ij} = 1 \quad \forall i$$

c_j *is the centroid of cluster* j

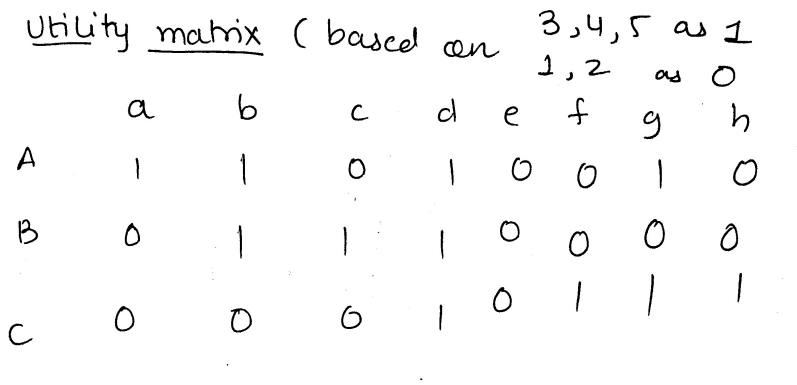
$$x_{ij} \in \{0, 1\} \quad \forall i, j$$

The variable x in the picture denotes if the value is assigned to cluster j or not, symbol p and c denotes the coordinates of x and centroid, and the formula try to minimize that distance. The rules that are put to the formula is that, each point should be assigned to exactly one cluster, and second is that, the centroid coordinate j depends on the variable x assigned to that particular cluster. In our case, due to the placement of the points, we try to optimize the

distance to just any set of points, rather than minimizing the distance to the assigned cluster. Now Since we end up with an empty cluster, thats due the the fact that at 2nd iteration, we see that the middle (Blue) is not closer to any of the set of points nearby hence, resulting into NULL set, and we could achieve a global minimum value by iterating to every possible k value, however, the placement of the data points leads the convergence and the BLUE set end up with null set, empty set, and can be accounted as global minimun.

II Recommender System

In this section we are given a Utility matrix, and we have to calculate the Jaccard, and Cosine scores between each pair and user. Also provide an analysis on which one is better. For the simplicity of assignment, I decided to do calculation part of the assignment on paper.



Jaccard Similarity $= \dfrac{M_{11} + M_{00}}{M_{10} + M_{01} + M_{00} + M_{11}}$

Jaccard distance = 1 — jacard similarity

∴ for pairs

$J(A,B) = \dfrac{2+3}{2+1+3+2} = 5/8$ ∴ $J_D = 1 - 0.625 = 0.375$

$J(A,C) = \dfrac{2+2}{2+2+2+2} = 1/2$ ∴ $J_D = 1 - 1/2 = 0.5$

$J(B,C) = \dfrac{1+2}{3+2+1+2} = 3/8 = 0.375$ $J_D = 1 - 0.375 = 0.625$

Cosine Distance

$\text{Cosine}(A,B) = \dfrac{A \cdot B}{|A||B|}$

$= \dfrac{0+1+0+1+0+0+0+0}{2\sqrt{3}}$

$= \dfrac{2}{2\sqrt{3}} = 1/\sqrt{3}$

$\cos(A,C) = \dfrac{2}{4} = 1/2$

$\cos(B,C) = \dfrac{1}{2\sqrt{3}} = 1/2\sqrt{3}$

We cant say for sure which one is better over the other since Cosine uses two real valued vectors, while Jaccard uses sets. In cosine similarities, the number of common attributes are divided with total attributes, while in Jaccard, the common attributes are divided by the number of attributes that occur it at-least on of the set. In our case, we aim to get our score as close to 1 as possible, and hence, we will discuss separately for each set.

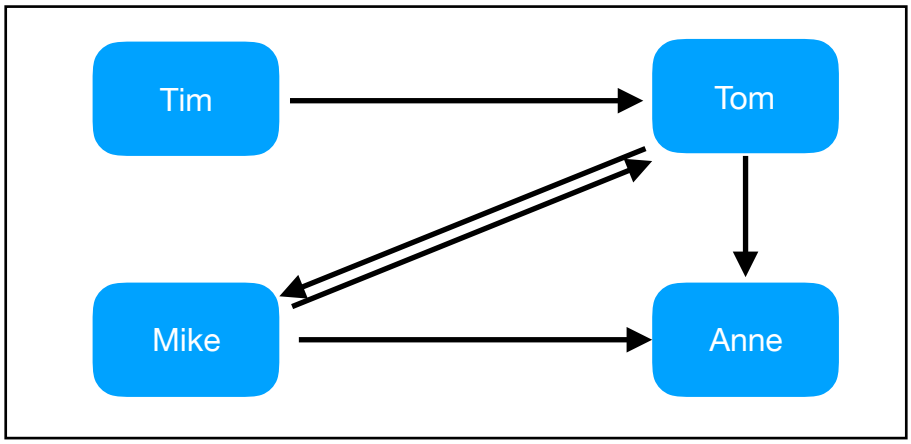
(A,B) ==> Cosine Proves to be closer to 1 then Jaccard
(A,C)==> Both are same
(B, C) ==> Jaccard is better than Cosine since its value is higher than Cosine's

III Page Rank

In this section, we will be performing some calculation using Page Rank. In the first part, we have to find influential twitter users.

The "*@user"* mentions can be formed into a directed graph.

user: Tim, tweet: "@Tom Howdy!"
user: Mike, tweet: "Welcome @Tom and @Anne!"
user: Tom, tweet: "Hi @Mike and @Anne!"
user: Anne, tweet: "Howdy!"



In this model, we can represent each user as a node, and each mentions as links to which they are directed.

Further Calculations were done on paper as follow:

- Transition Matrix

	Tim	Mike	Tom	Anne
Tim	0	0	0	0
Mike	0	0	1/2	0
Tom	1	1/2	0	0
Anne	0	1/2	1/2	0

⇒ Looking it as matrix, we make sure all column add up to 1.

⇒ "M"

$$V_0 = \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}$$

we represent each iteration as $V = MV_0$

$$V_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 \\ 1 & 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \end{bmatrix} \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix} = \begin{bmatrix} 0 \\ 1/8 \\ 3/8 \\ 1/4 \end{bmatrix}$$

$$V_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 \\ 1 & 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1/8 \\ 3/8 \\ 1/4 \end{bmatrix} = \begin{bmatrix} 0 \\ 3/16 \\ 1/16 \\ 1/4 \end{bmatrix}$$

$$V_3 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 \\ 1 & 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 3/16 \\ 1/16 \\ 1/4 \end{bmatrix} = \begin{bmatrix} 0 \\ 1/32 \\ 3/32 \\ 1/8 \end{bmatrix}$$

The values of the resultant matrix at each iteration are moving to 0, there for its gonna lead to dead end, hence we will have to change the ranking matrix such that

$$V' = \beta \cdot M \cdot r + (1-\beta)\frac{e}{n}$$

$1-\beta = 0.1 \quad \therefore \beta = 0.9$

here β = Teleportation Prob.
n = num. user
e = element matrix

$$e = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad \therefore e/n = \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix}$$

$$\therefore \beta \cdot M = 0.9 \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 \\ 1 & 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0.45 & 0 \\ 0.9 & 0.45 & 0 & 0 \\ 0 & 0.45 & 0.45 & 0 \end{bmatrix}$$

$$(1-\beta)e/n = 0.1 \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix} = \begin{bmatrix} 0.025 \\ 0.025 \\ 0.025 \\ 0.025 \end{bmatrix}$$

- Using the same method,
we calculate each iteration

$$V_1' = \beta \cdot M \cdot V_0 + (1-\beta)e/n$$

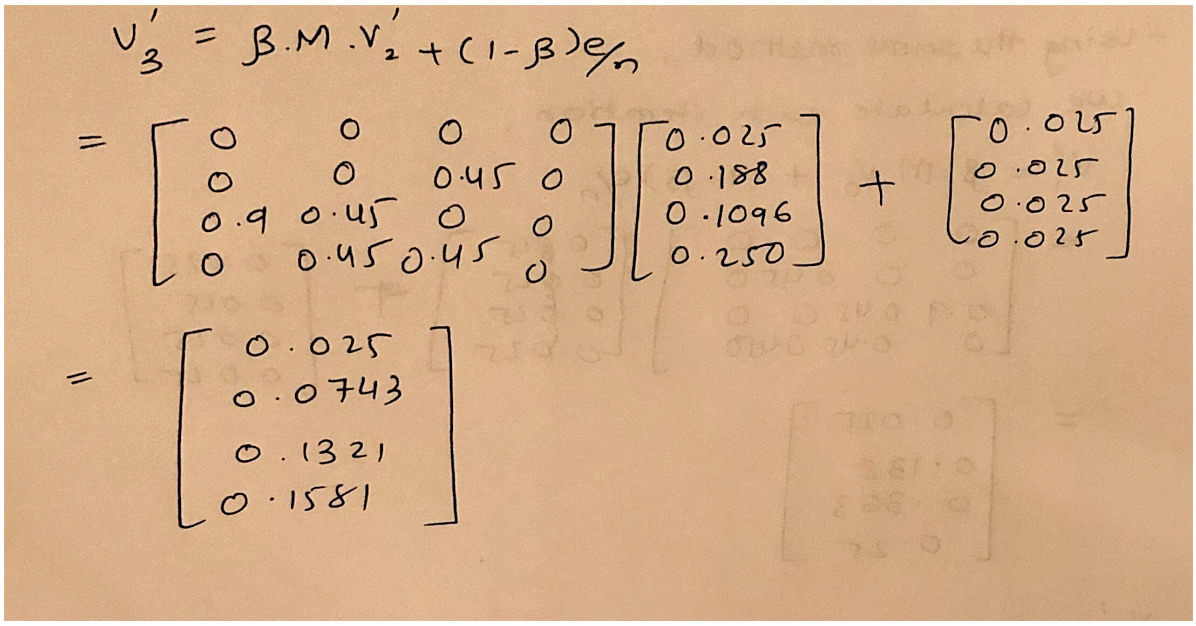
$$= \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0.45 & 0 \\ 0.9 & 0.45 & 0 & 0 \\ 0 & 0.45 & 0.45 & 0 \end{bmatrix} \begin{bmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{bmatrix} + \begin{bmatrix} 0.025 \\ 0.025 \\ 0.025 \\ 0.025 \end{bmatrix}$$

$$= \begin{bmatrix} 0.025 \\ 0.138 \\ 0.363 \\ 0.25 \end{bmatrix}$$

$$V_2' = \beta \cdot M \cdot V_1 + (1-\beta)e/n$$

$$= \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0.45 & 0 \\ 0.9 & 0.45 & 0 & 0 \\ 0 & 0.45 & 0.45 & 0 \end{bmatrix} \begin{bmatrix} 0.025 \\ 0.138 \\ 0.363 \\ 0.25 \end{bmatrix} + \begin{bmatrix} 0.025 \\ 0.025 \\ 0.025 \\ 0.025 \end{bmatrix}$$

$$= \begin{bmatrix} 0.025 \\ 0.188 \\ 0.1096 \\ 0.250 \end{bmatrix}$$



Hence, we can build a final table based on the values in the third iteration.

Tim = 0.025
Mike = 0.0743
Tom = 0.1321
Anne = 0.1589

According to which, we can say, Anne is the most influential because she has the highest score. Then Tom then Mike, and time at last.