Data Mining
Assignment 1
CS422
A20422490

1 Exercises

1.1 Leskovec Ch 3.3

Q.1 Exercise 3.1.1

A = {1, 2, 3, 4}
B = {2, 3, 5, 7}
C = {2, 4, 6}

Sim (A, B) = 2/6 = 1/3 = 0.33

Sim (B, C) = 1/6 = 0.166

Sim (A, C) = 2/5 = 0.4

Q.2 Exercise 3.2.1

- first 10 - 3 shingles
{"The", "he_", "e m", " mo", "mos", "ost", "st ", "t e", " ef", "eff"}

- If we consider as words
{"The most effective", "most effective way", "effective way to", "way to represent", "to represent documents", "represent documents as", "documents as sets", "as sets for", "sets for purpose", "for purpose of"}

Q.3 Exercise 3.3.3

Element	S_1	S_2	S_3	S_4	$h_1(x)$ $2x+1 \bmod 6$	$h_2(x)$ $3x+2 \bmod 6$	$h_3(x)$ $5x+2 \bmod 6$
0	0	1	0	1	1	2	2
1	0	1	0	0	3	5	1
2	1	0	0	1	5	2	0
3	0	0	1	0	1	5	5
4	0	0	1	1	3	2	4
5	1	0	0	0	5	5	3

① 1st step

	S_1	S_2	S_3	S_4
h_1	∞	∞	∞	∞
h_2	∞	∞	∞	∞
h_3	∞	∞	∞	∞

② 2nd step

	S_1	S_2	S_3	S_4
h_1	∞	1	∞	1
h_2	∞	2	∞	2
h_3	∞	2	∞	2

③ 3rd step

	S_1	S_2	S_3	S_4
h_1	∞	1	∞	1
h_2	∞	2	∞	2
h_3	∞	1	∞	2

④ 4th step

	S_1	S_2	S_3	S_4
h_1	5	1	∞	1
h_2	2	2	∞	2
h_3	0	1	∞	0

⑤ 5th step

	S_1	S_2	S_3	S_4
h_1	5	1	1	1
h_2	2	2	5	2
h_3	0	1	5	0

⑥ 6th step

	S_1	S_2	S_3	S_4
h_1	5	1	1	1
h_2	2	2	2	2
h_3	0	1	4	0

⑦ 7th step

	S_1	S_2	S_3	S_4
h_1	5	1	1	1
h_2	2	2	2	2
h_3	0	1	4	0

∴ Minhashing ~~Matrix~~ Signature Matrix

S_1	S_2	S_3	S_4
5	1	1	1
2	2	2	2
0	1	4	0

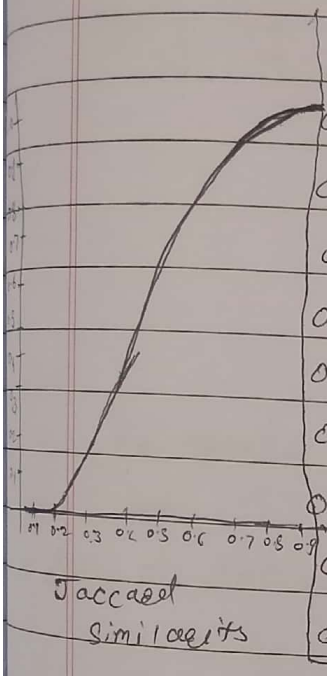
04 Exercise 3.4.1

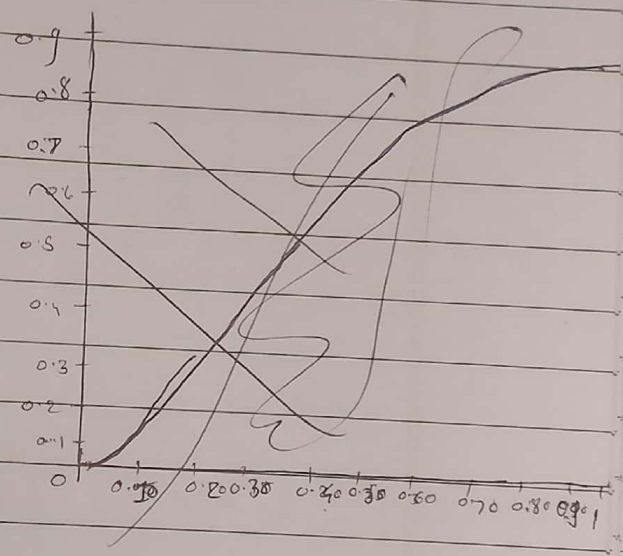
S curve for $1-(1-s^r)^b$, for

$s = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$

a) $r = 3$ & $b = 10$

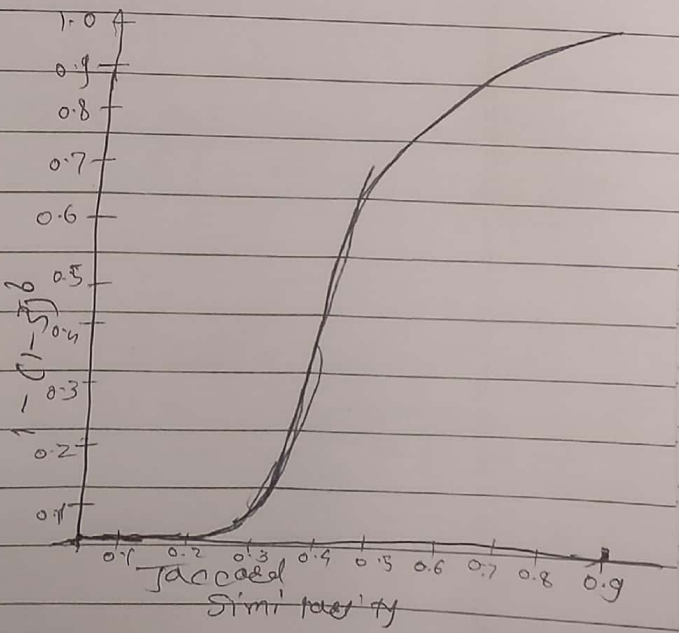
s	$1-(1-s^r)^b$
0.1	0.00995
0.2	0.07718
0.3	0.23945
0.4	0.48387
0.5	0.73692
0.6	0.91227
0.7	0.98502
0.8	0.99923
0.9	1.0000





b) $r = 6$ $b = 20$

s	$(1-(1-s^r)^b)$
0.1	0.0002
0.2	0.00128
0.3	0.01448
0.4	0.07881
0.5	0.27019
0.6	0.61541
0.7	0.91819
0.8	0.99771
0.9	1.000



Q.5 Exercise 3.5.4

Ans: a) $\{1,2,3,4\}$ & $\{2,3,4,5\}$

$S(\{1,2,3,4\},\{2,3,4,5\}) = 3/5$

Distance $= 1 - S = 1 - \frac{3}{5} = 2/5 = 0.4$

b) $\{1,2,3\}$ & $\{4,5,6\}$

$S(\{1,2,3\},\{4,5,6\}) = 0/6 = 0$

Distance $= 1 - S = 1 - 0 = 0$

Q6

Q6 Exercise 3.5.5

Ans:

Ans: a) cosine $((-3,-1,2),(-2,3,1)) = -0.5 = \cos(120)$

b) cosine $((1,2,3),(2,4,6)) = \frac{28}{\sqrt{14} * \sqrt{56}} = \frac{28}{28} = 1$

$= \cos(0)$

c) cosine $((5,0,-4),(-7,-6,2)) = \frac{-13}{\sqrt{41}\sqrt{41}} = -0.317$

$= \cos(108)$

d) cosine $((0,1,1,0,1,1),(0,0,1,0,0,0)) = 0.50$

$= \cos(60)$

12 Leskovac Ch 9

1) Exercise 9.2.1

Ans: a)

$$\cos(A,B) = \frac{8.2008 + 160000\alpha^2 + 24\beta^2}{\sqrt{9.3136 + 250000\alpha^2 + 36\beta^2}\sqrt{7.1824 + 102400\alpha^2 + 16\beta^2}}$$

$$\cos(B,C) = \frac{7.8256 + 204800\alpha^2 + 24\beta^2}{\sqrt{7.1824 + 102400\alpha^2 + 16\beta^2}\sqrt{8.5264 + 409600\alpha^2 + 36\beta^2}}$$

$$\cos(A,C) = \frac{8.9352 + 320000\alpha^2 + 36\beta^2}{\sqrt{9.3136 + 250000\alpha^2 + 36\beta^2}\sqrt{8.5264 + 409600\alpha^2 + 36\beta^2}}$$

b) if $\alpha = \beta$

a) $\cos(A,B) = \alpha = \beta = 0.9999973$

b) $\cos(B,C) = \alpha = \beta = 0.9999879$

c) $\cos(A,C) = \alpha = \beta = 0.9999953$

c) a) $\cos(A,B) = \alpha = 0.1$, $\beta = 0.5$ $\Big\}$ $= 0.9908815$

b) $\cos(B,C) = \alpha = 0.1$, $\beta = 0.5$ $\Big\}$ 0.9691779

c) $\cos(A,C) = \alpha = 0.1$, $\beta = 0.5$ $\Big\}$ 0.9915547

d) $\mu(\alpha) = (500+320+640)/3 = 1460/3 = 486.6666$

$\mu(B) = (6+4+6)/3 = 16/3 = 5.3333$

we want to make its inversly proportional,

SO

$\alpha = 1/486.6666 = 3/1460$

$B = 1/5.333 = 3/16$

Now put value of α & B in (a)

$\cos(A,B) = 0.994$ $\cos(A,C) = 0.994$

$\cos(B,C) = 0.982$

e) Excercise 9.2.3

Ans: →

a) $A = 4$, $B = 2$, $C = 5$

$avg = (4+2+5)/3 = 11/3$

after applying normalization

$A = 4 - \frac{11}{3} = 1/3$

$B = 2 - \frac{11}{3} = -5/3$

$C = 5 - \frac{11}{3} = 4/3$

b)

Processor speed = 3.06 * (1/3) + 2.68 * (-5/3) + 2.93 * (4/3)

= 0.4667

Disk Size = 500 * (1/3) + 320 * (-5/3) + 640 * (4/3)

= 486.6667

Main-Memory Size = 6 * (1/3) + 4 * (-5/3) + 6 * (4/3)

= 3.3330

3) Exercise 9.3.1

a) Jaccard (A, B) = 4/8 = 1/2

Jaccard (B, C) = 4/8 = 1/2

Jaccard (A, C) = 4/8 = 1/2

b) cos (A, B) = 2/3

cos (B, C) = 2/3

cos (A, C) = 2/3

c) Jac (A, B) = 1 - 2/5 = 3/5

Jac (B, C) = 1 - 1/6 = 5/6

Jac (A, C) = 1 - 2/6 = 4/6 = 2/3

d) cos (A, B) = 0.57735

cos (B, C) = 0.28868

cos (A, C) = 0.50

cos dist (A, B) = 1 − cos (A, B) = 0.42265

cos dist (B, C) = 0.71132

cos dist (A, C) = 0.5

e) Normalization

Avg (A) = 3.33

Avg (B) = 2.33 Avg (C) = 3

	a	b	c	d	e	f	g	h
A	0.667	1.667	0.000	1.667	−2.333	0.000	−0.33	−1.333
B	0.000	0.667	−1.667	0.667	−1.333	−0.3333	−1.333	0.000
C	−1	0	−2	0	0	1	2	0

f) cos (A, B) = 0.58408

cos (B, C) = −0.73918

cos (A, C) = −0.11518

cos dist (A, B) = 1 − cos (A, B) = 0.41592

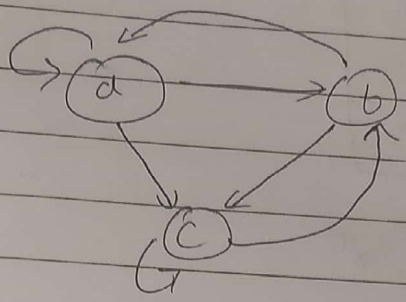
cos dist (B, C) = 1 − cos (B, C) = 1.73918

cos dist (A, C) = 1 − cos (A, C) = 1.11518

1.3 Leskovec Link Analysis

5.1.1

Ans:



$r_a = r_a/3 + r_b/2$

$r_b = r_a/3 + r_c/2$

$r_c = r_a/3 + r_b/2 + r_c/2$

Transition matrix $= \begin{bmatrix} 1/3 & 1/2 & 0 \\ 1/3 & 0 & 1/2 \\ 1/3 & 1/2 & 1/2 \end{bmatrix}$

by eqn

$$\begin{bmatrix} 2/3 & -1/2 & 0 \\ -1/3 & 1 & -1/2 \\ 1/3 & 1/2 & -1/2 \end{bmatrix} \begin{bmatrix} r_a \\ r_b \\ r_c \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

A Constraint to force uniqueness soln

$r_a + r_b + r_c = 1$

by solving eqn

$$r = \begin{bmatrix} 3/13 \\ 4/13 \\ 6/13 \end{bmatrix}$$

by solving with power method:

$$r^0 = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

$$r^{(t+1)} = M \cdot r^t$$

$$\begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} \begin{bmatrix} 0.2777 \\ 0.2777 \\ 0.4444 \end{bmatrix} \begin{bmatrix} 0.2314 \\ 0.3148 \\ 0.4537 \end{bmatrix} \begin{bmatrix} 0.2345 \\ 0.304 \\ 0.4614 \end{bmatrix} \begin{bmatrix} 0.2301 \\ 0.3088 \\ 0.4609 \end{bmatrix} \begin{bmatrix} 0.2311 \\ 0.3071 \\ 0.4616 \end{bmatrix}$$

$$\begin{bmatrix} 0.2306 \\ 0.3078 \\ 0.4614 \end{bmatrix} \begin{bmatrix} 0.2308 \\ 0.3076 \\ 0.4615 \end{bmatrix} \begin{bmatrix} 0.2307 \\ 0.3077 \\ 0.4615 \end{bmatrix} \begin{bmatrix} 0.2307 \\ 0.3076 \\ 0.4615 \end{bmatrix}$$

5.1.2

Ans:

$$B = 0.8$$

$$A = BM + (1-B)\left[\frac{1}{N}\right]_{N \times N}$$

$$= 0.8 \begin{bmatrix} 1/3 & 1/2 & 0 \\ 1/3 & 0 & 1/2 \\ 1/3 & 1/2 & 1/2 \end{bmatrix} + 0.2 * \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$$

$$A = \begin{bmatrix} 1/3 & 7/15 & 1/15 \\ 1/3 & 1/15 & 7/15 \\ 1/3 & 7/15 & 7/15 \end{bmatrix}$$

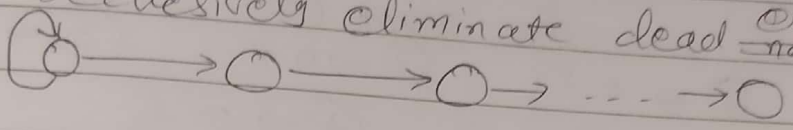
$$r^0 = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

$r^{(t+1)} = M \cdot r^{t}$

$$\begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} \begin{bmatrix} 0.2888 \\ 0.2888 \\ 0.4222 \end{bmatrix} \begin{bmatrix} 0.2592 \\ 0.3125 \\ 0.4281 \end{bmatrix} \begin{bmatrix} 0.2608 \\ 0.307 \\ 0.432 \end{bmatrix} \begin{bmatrix} 0.2590 \\ 0.3090 \\ 0.4318 \end{bmatrix} \begin{bmatrix} 0.2593 \\ 0.3085 \\ 0.4321 \end{bmatrix} \begin{bmatrix} 0.2592 \\ 0.3086 \\ 0.4320 \end{bmatrix}$$

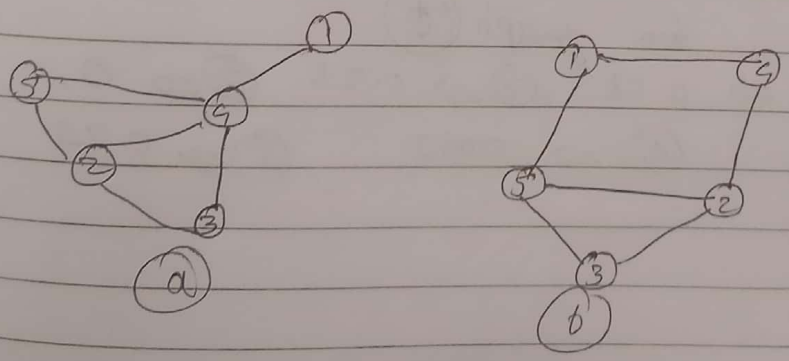
5.1.6

1.3 We recursively eliminate dead ~~nod~~ ends of the graph.



The head node with a self-direction will survive.
The pagerank for that node is 1
& if we calculate page rank for all remaining nodes then it comes 1/2

1.4 Centrality Measures.



a) Normalize degree centrality for each node

~~1 ⇒ 0.25~~ For Graph (a)

1 ⇒ 0.25 2 ⇒ 0.75 3 ⇒ 0.5

4 ⇒ ~~0.1~~ 5 ⇒ 0.5

b) for graph (b)

(1) ⇒ 0.5 (2) ⇒ 0.75 (3) ⇒ 0.5

(4) ⇒ 0.5 (5) ⇒ 0.75

c) Normalize closeness centrality of each node

for graph (a)

(1) ⇒ 0.57 (2) ⇒ 0.8 (3) ⇒ 0.66

(4) ⇒ 1 (5) ⇒ 0.5

for graph (b)

(1) ⇒ 0.66 (2) ⇒ 0.80 (3) ⇒ 0.6

(4) ⇒ 0.66 (5) ⇒ 0.8

d) Normalize betweenness centrality of each node

for graph (a)

(1) ⇒ 0 (2) ⇒ 0.08 (3) ⇒ 0

(4) ⇒ 0.58 (5) ⇒ 0

for graph (b)

(1) ⇒ 0.083 (2) ⇒ 0.25 (3) ⇒ 0

(4) ⇒ 0.083 (5) ⇒ 0.25