1.1) 1) Jaccard similarity between {1,2,3,4}, {2,3,5,7} and {2,4,6}

$J(\{1,2,3,4\}, \{2,3,5,7\}) = \frac{2}{6} = \frac{1}{3}$

$J(\{2,3,5,7\}, \{2,4,6\}) = \frac{1}{6}$

$J(\{1,2,3,4\}, \{2,4,6\}) = \frac{2}{5}$

2) First 3-shingles in the first sentence of Section 3.2? $k=3$

Sentence: The most effective way to represent documents as sets;

shingles = { The, he_, e_m, _mo, mos, ost, st_, t_e, _ef, eff }

3) $h_1(x) = 2x+1 \mod 6$ $\quad h_2(x) = 3x+2 \mod 6$ $\quad h_3(x) = 5x+2 \mod 6$

a)

Element	S_1	S_2	S_3	S_4	$(2x+1) \mod 6$	$(3x+2) \mod 6$	$(5x+2) \mod 6$
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

1st iteration:

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

2nd iteration:

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

3rd iteration:

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

4th iteration:

	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

5th iteration:

	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

6th iteration:

	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

Our final signature matrix is

	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

c>

Jaccard similarity according to characteristic matrix

$S(S_1, S_2) = 0$ $\quad$ $S(S_2, S_3) = 0$

$S(S_1, S_3) = 0$ $\quad$ $S(S_2, S_4) = 1/4$

$S(S_1, S_4) = 1/4$ $\quad$ $S(S_3, S_4) = 1/4$

Jaccard similarity according to signature matrix

$S(S_1, S_2) = 1/3$ $\quad$ $S(S_2, S_3) = 2/3$

$S(S_1, S_3) = 1/3$ $\quad$ $S(S_2, S_4) = 2/3$

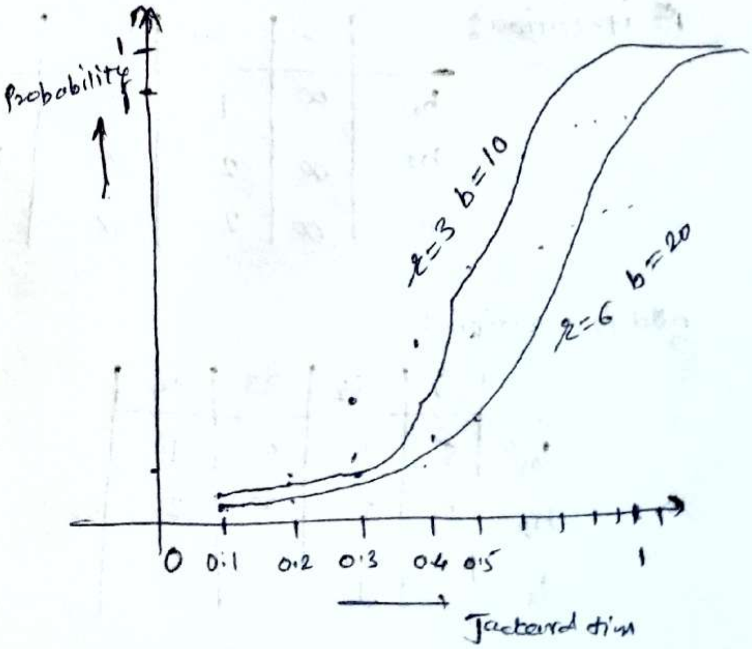
$S(S_1, S_4) = 2/3$ $\quad$ $S(S_3, S_4) = 2/3$

b> h_3 is a true permutation

4> Evaluate the S-curve $1-(1-s^r)^b$ for $s = 0.1, 0.2, 0.3, \ldots 0.9$,

a> $r=3$ & $b=10$ $\quad$ b> $r=6$ $b=20$

s	$r=3$ $b=10$	$r=6$ $b=20$
0.1	0.001	0.0001
0.2	0.078	0.0013
0.3	0.22	0.0145
0.4	0.48	0.0789
0.5	0.74	0.2702
0.6	0.91	0.6155
0.7	0.98	0.919
0.8	0.99	0.9978
0.9	1.04	0.9999



1.2

1) herd of Asian elephants:

All given columuse are numerical. but the range of columns is varying, also the attributes are not similar, it asymmetric which removes the options cosine and correlation distance.

So using Euclidean distance to have mean of 0 and standard deviation of 1.

2) a)

Hamming distance = 3

Jaccard similarity = $\frac{2}{5} = 0.4$

b) Formulae for simple matching coefficient = Hamming dist / no. of bits

so hamming distance is similar to SMC. As Jaccard & cosine ignores 0-0 so they are both similar.

c) As we need to compare only the genes & how many genes these two organisms share so Jaccard is more appropriate

d) In these case we have to focus on differences so hamming distance will be more appropriate.

3)

a) $\cos(x,y) = 1$ Euclidean $(x,y) = 2$

$\text{corr}(x,y) = 0$

b) $\cos(x,y) = 0$ Euclidean $(x,y) = 2$

$\text{Corr}(x,y) = -1$ Jaccard $(x,y) = 0$

c) $\cos(x,y) = 0$ $\text{corr}(x,y) = 0$

Euclidean = 2

d) $\cos(x,y) = 0.75$ $\text{corr}(x,y) = 0.25$

Jaccard $(x,y) = 0.6$

e) $\cos(x,y) = 0$ $\text{corr}(x,y) = 0$

5> Find the Jaccard distances between following pairs of sets

a) $\{1,2,3,4\}$ and $\{2,3,4,5\}$

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

$\longrightarrow$

Jaccard dist for $\{1,2,3,4\}$ & $\{2,3,4,5\} = 1 - sim(A,B)$

$$= 1 - \frac{3}{5} = \frac{2}{5}$$

Jaccard dist for $\{1,2,3\}$ & $\{4,5,6\} = 1 - sim(A,B)$

$$= 1 - 0 = 1$$

6> a) $\{3,-1,2\}$ & $\{-2,3,1\}$

$$= \frac{-6-3+2}{\sqrt{9+1+4}\ \sqrt{4+9+1}} = \frac{-7}{\sqrt{14}\ \sqrt{14}} = -0.5 = \cos^{-1}(0.5) = 120^\circ$$

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

$$= \frac{2+8+18}{\sqrt{1+4+9}\ \sqrt{4+16+36}} = 1 = \cos^{-1}(1) = 0^\circ$$

c) $(5,0,-4)$ & $(-1,-6,2)$

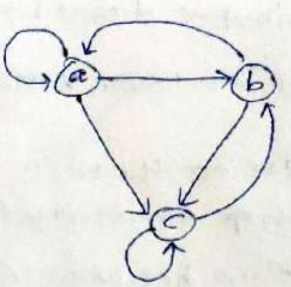
$$= \frac{-5+0-8}{\sqrt{25+16}\ \sqrt{1+36+4}} = \frac{-13}{\sqrt{41}\ \sqrt{41}} = \frac{-1}{3} = \cos^{-1}\left(\frac{-1}{3}\right) \approx 108^\circ$$

d) $(0,1,1,0,1,1)$ $(0,0,1,0,0,0)$

$$= \frac{0+1+0}{\sqrt{4}} = \frac{1}{2} = \cos^{-1}\left(\frac{1}{2}\right) = 60^\circ$$

1.3

1) Compute the PageRank assuming no taxation



→ Transition Matrix $M = \begin{array}{c} \begin{matrix} A & B & C \end{matrix} \\ \begin{bmatrix} 1/3 & 1/2 & 0 \\ 1/3 & 0 & 1/2 \\ 1/3 & 1/2 & 1/2 \end{bmatrix} \end{array}$

Iterations :-

$$\begin{bmatrix} 1/3 & 1/2 & 0 \\ 1/3 & 0 & 1/2 \\ 1/3 & 1/2 & 1/2 \end{bmatrix} \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} = \begin{bmatrix} \frac{1}{9}+\frac{1}{6} \\ \frac{1}{9}+\frac{1}{6} \\ \frac{1}{9}+\frac{1}{6}+\frac{1}{6} \end{bmatrix} = \begin{bmatrix} 5/18 \\ 5/18 \\ \frac{8}{18} \end{bmatrix} = \begin{bmatrix} 5/18 \\ 5/18 \\ 2/9 \end{bmatrix}$$

$$\begin{bmatrix} \frac{1}{3} & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 5/18 \\ 5/18 \\ 2/9 \end{bmatrix} = \begin{bmatrix} \frac{5}{54}+\frac{5}{36} \\ \frac{5}{54}+\frac{2}{18} \\ \frac{5}{54}+\frac{5}{36}+\frac{2}{18} \end{bmatrix} = \begin{bmatrix} 25/108 \\ 22/108 \\ 37/108 \end{bmatrix}$$

After few iterations we get $\begin{bmatrix} 3/13 \\ 4/13 \\ 6/13 \end{bmatrix}$ Page rank $A = 3/13$, $B = 4/13$, $C = 6/13$

2) $r = \beta M + (1-\beta)\frac{1}{n}$

$$r = \begin{bmatrix} 4/15 & 2/5 & 0 \\ 4/15 & 0 & 2/5 \\ 4/15 & 2/5 & 2/5 \end{bmatrix} + \begin{bmatrix} 1/15 & 1/15 & 1/15 \\ 1/15 & 1/15 & 1/15 \\ 1/15 & 1/15 & 1/15 \end{bmatrix}$$

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

$$r = A = \begin{bmatrix} 1/3 & 7/15 & 1/15 \\ 1/3 & 1/15 & 7/15 \\ 1/3 & 7/15 & 7/15 \end{bmatrix} \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix} = \begin{bmatrix} 0.288 \\ 0.288 \\ 0.422 \end{bmatrix}$$

After iterations $r = \begin{bmatrix} 0.25 \\ 0.308 \\ 0.432 \end{bmatrix}$

Page rank

$A = 7/21$, $B = 25/81$, $C = 35/81$

3)

O → O → O ...

Only first node has a self loop thus page rank of first node will be 1 & Page rank of remaining nodes will be 1/2

4) a) Range of possible values for cosine measure is $[-1,1]$

b) might be. because their values of attributes differ by constant factor

c) If both are same then we have to look at mean & standard deviation

d) As very big amount of data is on curve so there is relationship between Euclidean distance & cosine similarity for normalized data. There is inverse relationship between cosine similarity & Euclidean distance

e) There is relationship between Euclidean distance & correlation similarity & There is inverse relationship between correlation & Euclidean distance.

f) Consider x & y are two vectors where L_2 length is 1

variance is n times the sum of its squared attribute values

& correlation between two vectors is

$$d(x,y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} = \sqrt{2(1-\cos(x,y))}$$

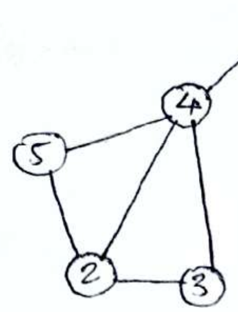
g) Consider x, & y be two vectors & mean is 0 standard deviation is 1 . so variance is just n times the sum of its squared attribute

$$d(x,y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} = \sqrt{n - 2n\,corr(x,y) + n}$$

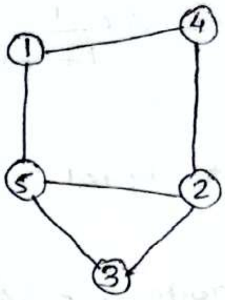
$$= \sqrt{2n(1-corr(x,y))}$$

1.4 Centrality Measures

Graphs:



a>



b>

a> Normalized degree centrality

Graph a: For node 1 $= \frac{1}{5-1} = \frac{1}{4}$

For node 4 $= \frac{4}{4} = 1$

For node 5 $= \frac{2}{4} = 0.5$

For node 2 $= \frac{3}{4}$

For node 3 $= \frac{2}{4} = 0.5$

Graph b: For node 1 $= \frac{2}{4}$

For node 4 $= \frac{2}{4}$

For node 2 $= \frac{3}{4}$

For node 3 $= \frac{3}{4}$

For node 5 $= \frac{3}{4}$

b> Normalized closeness centrality

Graph a: node ① $= \frac{1}{\text{closest path distance to all nodes}} = \frac{1}{d(1,4)+d(1,3)+d(1,5)+(1,2)}$

$= \frac{1}{1+2+2+2} = \frac{1}{7}$

normalized C_c for node 1 $= (n-1)\frac{1}{7} = \frac{4}{7}$

normalized C_c for node 4 $= 4 \times \frac{1}{(1+1+1+1)} = \frac{4}{4} = 1$

normalized C_c for node 5 $= 4 \times \frac{1}{2+1+1+2} = \frac{4}{6} = \frac{2}{3}$

normalized C_c for 2 = $4 \times \frac{1}{1+1+1+2} = \frac{4}{5}$

normalized C_c for 3 = $4 \times \frac{1}{1+1+2+2} = \frac{4}{6}$

Graph b:

normalized C_c for node 1 = $4 \times \frac{1}{1+1+2+2} = \frac{4}{6}$

normalized C_c for node 4 = $4 \times \frac{1}{1+1+2+2} = \frac{4}{6}$

normalized C_c for node 5 = $4 \times \frac{1}{1+1+1+2} = \frac{4}{5}$

normalized C_c for node 2 = $4 \times \frac{1}{1+1+1+2} = \frac{4}{5}$

normalized C_c for node 3 = $4 \times \frac{1}{1+1+2+2} = \frac{4}{6}$

c) Normalized between centrality:

Normalized between centrality = $\frac{C_B}{2 \times {}^4C_2} = \frac{C_B}{12}$

Graph a

For node 1 = 0

For node 4 = $\frac{2 \times (1+1+1+0)}{12} = \frac{6}{12} = \frac{1}{8}$

For node 5 = 2×(

Node 2

$1 \to 4 = 0$
$1 \to 5 = 0$
$1 \to 3 = 0$
$4 \to 5 = 0$
$4 \to 3 = 0$
$3 \to 5 = \frac{1}{2}$

Normalised = $\frac{2 \times \frac{1}{2}}{12}$

$= \frac{1}{12}$

Node 3

$1 \to 5 = 0$
$1 \to 2 = 0$
$1 \to 4 = 0$
$4 \to 5 = 0$
$4 \to 2 = 0$

Normalized = 0

Node 4

$1 \to 5 = 1$
$1 \to 2 = 1$
$1 \to 3 = 1$
$3 \to 5 = \frac{1}{2}$
$3 \to 2 = 0$

$= 3 + \frac{1}{2} = \frac{7}{2}$

Normalized = $\frac{2 \times \frac{7}{2}}{12}$

$= \frac{7}{12}$

Node 5

$1 \to 4 = 0$
$1 \to 2 = 0$
$1 \to 3 = 0$
$3 \to 2 = 0$

Normalized = 0

Graph 2

For node 1

$4 \rightarrow 2 = 0$

$4 \rightarrow 3 = 0$

$4 \rightarrow 5 = \frac{1}{2}$

$2 \rightarrow 5 = 0$

$2 \rightarrow 3 = 0$

$$\text{Norm } C_B = \frac{2 \times \frac{1}{2}}{12} = \frac{1}{12}$$

node 2

$1 \rightarrow 4 = 0$

$1 \rightarrow 5 = 0$

$1 \rightarrow 3 = 0$

$3 \rightarrow 5 = 0$

$3 \rightarrow 4 = 1$

$4 \rightarrow 5 = \frac{1}{2}$

$$\text{Norma } C_B = \frac{3 \times 2}{12 \times 2} = \frac{1 \times 3}{6 \times 2} = \frac{3}{12}$$

node 3

$1 \rightarrow 2 = 0$

$1 \rightarrow 4 = 0$

$1 \rightarrow 5 = 0$

$2 \rightarrow 4 = 0$

$2 \rightarrow 5 = 0$

$4 \rightarrow 5 = 0$

$\text{Norm } C_B = 0$

node 4

$1 \rightarrow 2 = \frac{1}{2}$

$1 \rightarrow 3 = 0$

$1 \rightarrow 5 = 0$

$2 \rightarrow 3 = 0$

$2 \rightarrow 5 = 0$

$3 \rightarrow 5 = 0$

$$\text{Norma } C_B = \frac{0.5 \times 2}{12} = \frac{1}{12}$$

node 5

$1 \rightarrow 2 = \frac{1}{2}$

$1 \rightarrow 3 = 1$

$1 \rightarrow 4 = 0$

$2 \rightarrow 3 = 0$

$2 \rightarrow 4 = 0$

$3 \rightarrow 4 = 0$

Total = 1.5

$$\text{Norm } C_B = \frac{2 \times 1.5}{12} = \frac{3}{12}$$