ASTMA

1. BOW - BOW
2. TF - IDF
3. ACP Problem
4. prop. of core text mining ope
5. NLP tech in text category & explain process of text mining *
6. Text categorization - theory *
7. POS tagging
8. Entropy & purity
9. outlier detection (IQR, Z-Score)
10. CRF in social NW analysis *
11. Page Rank
12. K-core graph.
13. NW Analysis.
14. whole NW Anaysis.
15. Two mode NW Analysis (Egocentric)
16. Viterbi and forward (HMM)
17. SEO *
18. core text mining operations *
19. problems models for info extraction *
20. pre processing tech. of NLP *
21. Sentiment analysis *
22. clustering & topic detection. *
23. G-test
24. centrality measures
25. ~~HMM~~

1. BOW - BOW , TF - IDF.

DOC 1: cat sat on the mat
DOC 2: Dog sat on the mat
DOC 3: cat chased the dog.

Soln:

Step 1: BOW (Bag of words)

a) create voculalary
["cat", "sat", "on", "the", "mat", "dog", "chased"]

b) count word frequencies

word	cat	sat	on	the	mat	dog	chased
DOC1	1	1	1	1	1	0	0
DOC2	0	1	1	1	1	1	0
DOC3	1	0	0	1	0	1	1

c) Bow output:
The matrix is;
[[1,1,1,1,1,1,1,0,0], [0,1,1,1,1,1,0], [1,0,0,1,0,1,1]]

Step 2: TF-IDF.

a) compute Term Frequency (TF)
TF = word F in DOC / total words in DOC.

TF(cat, DOC1) = 1/5 = 0.2
TF(sat, DOC1) = 1/5 = 0.2
TF(on, DOC1) = 1/5 = 0.2
TF(the, DOC1) = 1/5 = 0.2
TF(mat, DOC1) = 1/5 = 0.2

TF(dog, DOC2) = 1/5 = 0.2
TF(sat, DOC2) = 1/5 = 0.2
TF(on, DOC2) = 1/5 = 0.2
TF(the, DOC2) = 1/5 = 0.2
TF(mat, DOC2) = 1/5 = 0.2

TF(cat, DOC3) = 1/4 = 0.25
TF(chased, DOC3) = 0.25
TF(the, DOC3) = 0.25
TF(dog, DOC3) = 0.25

b) IDF = log(Total Docs / Docs containing the word)

eg: cat appears 2 times in doc.

IDF(cat) = log (3/2) = 0.176

IDF(sat) = log (3/2) = 0.176

IDF(on) = log (3/2) = 0.176

IDF(the) = log(3/3) = 0

IDF(mat) = log(3/2) = 0.176

IDF(dog) = log(3/2) = 0.176

IDF(chased) = log(3/1) = 0.477

c) compute TF - IDF.

TF-IDF = TF * IDF

Eg: TF-IDF (cat, DOC1) = 0.2 * 0.176

TF-IDF (sat, DOC1) = 0.2 * 0.176

(on, DOC1) = 0.2 * 0.176

(the, DOC1) = 0.2 * 0

(mat, DOC1) = 0.2 * 0.176

(dog, DOC2) = 0.2 * 0.176

(chased, DOC3) = 0.25 * 0.477

d) TF-IDF output

word	cat	sat	on	the	mat	dog	chased.
DOC 1	0.0352	0.0352	0.0352	0	0.0352	0	0
DOC 2	0	0.0352	0.0352	0	0.0352	0.0352	0
DOC 3	~~0.0352~~ 0.044	0	0	0	0	~~0.0352~~ 0.044	0 ~~0.0352~~ 0.1192

Teacher's Signature:

2. ACP problem (Average concept proportion)

H/W : Laptop, Desktop, tab
S/W : OS, Applications
NW: LAN, WAN, MAN, VPN

Soln:

Step 1: count the parent nodes - HW, SW, NW

Step 2: count the child node for each parent

HW → 3
SW → 2
NW → 4

< 0.5 unevenly distributed
> 0.5 toward evenly distributed
1 → evenly distributed (no need any modification)

Step 3: HW (3 children) $= \frac{1}{3} = 0.33$

SW (2 children) $= \frac{1}{2} = 0.5$

NW (4 children) $= \frac{1}{4} = 0.25$

Step 4: Calculate ACP

$\Rightarrow \frac{0.33 + 0.5 + 0.25}{3} = 0.36$

$\Rightarrow 0.36 < 0.5$

4. Distribution properties of / core / text mining operations.

Expt. No. :
Page No. :
Date :

3. calculate entropy and purity.

cluster	science	sports	politics	
1	250	20	10	|280|
2	20	180	80	|280|
3	30	100	210	|340|
	300	300	300	$\boxed{900}$

soln:

$$\text{entropy}(D_i) = -\sum_{j=1}^{k} P(c_j)\log_2 P(c_j)$$

$$\text{entropy}_{\text{total}}(D) = \sum_{i=1}^{k} \frac{|D_i|}{|D|} \times \text{entropy}(D_i)$$

$$\text{Purity}(D_i) = \max_j (P(c_j))$$

$$\text{Purity}_{\text{total}}(D) = \sum_{i=1}^{k} \frac{|D_i|}{|D|} * \text{purity}(D_i)$$

Probabilities;

cluster	science	sports	politics
1	250/280	20/280	10/280
2	20/280	180/280	80/280
3	30/340	100/340	210/340

→

cluster	S	S	P
1	0.893	0.071	0.036
2	0.071	0.643	0.286
3	0.088	0.294	0.618

∵ purity (D_i) = max (P_i)

cluster 1 = 0.893

cluster 2 = 0.643

cluster 3 = 0.618

Teacher's Signature:

For cluster 1: (ci)

Entropy (c1) = $-\sum p \log_2 p(c_i)$.

$\Rightarrow -\frac{250}{280}\boxed{\log_2 \frac{250}{280}} - \frac{20}{280}\log_2\frac{20}{280} - \frac{10}{280}\log_2\frac{10}{280}$

$\Rightarrow \cancel{-0.893(-0.163)}$

$\Rightarrow 0.146 + 0.272 + 0.172$

$\Rightarrow 0.590$

Entropy (c2) = $-\frac{20}{280}\log_2\frac{20}{280} - \frac{180}{280}\log_2\frac{180}{280} - \frac{80}{280}\log_2\frac{80}{280}$

$\Rightarrow 0.272 + 0.410 + 0.516$

$\Rightarrow 1.198$

Entropy (c3) = $-\frac{30}{340}\log_2\frac{30}{340} - \frac{100}{340}\log_2\frac{100}{340} - \frac{210}{340}\log_2\frac{210}{340}$

$\Rightarrow 0.309 + 0.519 + 0.429$

$\Rightarrow 1.257$.

Entropy total $= |D| = 900$.

Entropy total (D) = $\frac{280}{900} \times 0.590 + \frac{280}{900} \times 1.198 + \frac{340}{900} \times 1.257$

$\Rightarrow 0.184 + 0.373 + 0.475$

$\Rightarrow 1.032$

4- outlier detection

Z-score & IQR method.

A data pt is considered an outlier of its z-score exceeds a threshold (Eg: $|z| > 3$)

Data pts: 4, 8, 10, 14, 16, 18, 20, 22, 24, 28.

$z = \frac{x-\mu}{\sigma}$, $\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$, $\mu = \frac{\text{All add}}{N}$

$$\mu = \frac{4+8+10+14+16+18+20+22+24+28}{10}$$

$$\mu = \frac{164}{10} = 16.4.$$

$$\sigma = \sqrt{\frac{(4-16.4)^2+(8-16.4)^2+(10-16.4)^2+(14-16.4)^2+(18-16.4)^2+(20-16.4)^2+(22-16.4)^2+(24-16.4)^2+(28-16.4)^2}{10}}$$

$$\sigma = \sqrt{\frac{153.76+70.56+40.96+5.76+2.56+12.96+31.36+57.76+134.56}{10}}$$

$$\sigma = \sqrt{\frac{510.240}{10}} = \sqrt{51.024} = 7.143$$

IQR (Inter Quartile range)

Q1 = lower quartile = median of lower half of data

$Q1 = 10 \Rightarrow 25^{th}$ percitble $= \frac{(N+1)\times 25}{100} = \frac{11 \times 25}{100} = 3.$

Q3 = upper quartile = median of upper half of data.

$Q3 = 22. \Rightarrow 75^{th}$ percentile $= \frac{(N+1)\times 75}{100} = \simeq 8$

$IQR = Q3 - Q1 = 22 - 10 = 12$

Any value below 10 and above 22 is considered an outlier.

Teacher's Signature:

lower boundary $= Q1 - 1.5 \times IQR$

$= 10 - 1.5 \times 12$

$= 8$

upper boundary $= Q3 + 1.5 \times IQR$

$= 22 + 1.5 \times 12$

$= 40$

Z-score : Data pts: [10, 12, 14, 18, 100]

Threshold = 3

$$\mu = \frac{10+12+14+18+100}{5} = \frac{154}{5} = 30.8$$

$$\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$$

$$= \sqrt{\frac{(10-30.8)^2 + (12-30.8)^2 + (14-30.8)^2 + (18-30.8)^2 + (100-30.8)^2}{5}}$$

$= \sqrt{\;}$ => 34.70.

x	$z = \frac{x-\mu}{\sigma}$
10	−0.599
12	−0.542
14	−0.484
18	−0.369
100	1.99

$|z| > 3$ → External high value

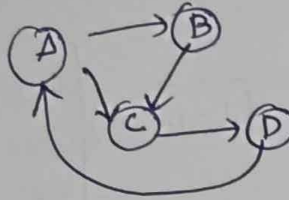
$|z| < 3$ → External low value

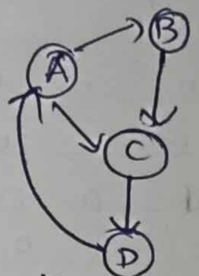
$|z| \leq 3$ → not outlier (within normal range)

5. Page rank.

To calculate the page rank for the directed graph with nodes A, B, C, D and edges A → B, A → C, B → C, C → D, D → A.

soln:





a) Represent the graph as transition matrix probability of transitions from one-to-other node,

$$M = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1/2 & 0 & 0 & 0 \\ 1/2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \end{array}$$

col-wise vanthu 1 mattu tha varanum

Max prob. is 1, No. of edges = 2, so 1/2

b) Initilize the page rank vector.

total node = 4 ⇒ 1/4

$$p^0 = \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix} \Rightarrow \begin{bmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{bmatrix}$$

c) Applying damping factor (d)

α (or) $d = 0.85$ (constant)

$\frac{1}{2} \times 0.25$

$\frac{0.25}{2}$ Teacher's Signature:

vector i.e. uniform distributions

Page rank formula becomes $p^{(k+1)} = d \cdot M \cdot p^{(k)} + (1-d) \cdot v$

$$v = \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{bmatrix} \Rightarrow \begin{bmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{bmatrix}$$

$$p^{(k+1)} = 0.85 \times \begin{bmatrix} 0 & 0 & 0 & 1 \\ 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \times \overset{p(k)}{\begin{bmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{bmatrix}} + (1-0.85) \begin{bmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{bmatrix}$$

$P(1) =$

$$\Rightarrow \cancel{0.85 \begin{bmatrix} 0.25 \\ 0.125 \\ 0.375 \\ 0.25 \end{bmatrix}} \Rightarrow 0.85 \begin{bmatrix} 0.25 \\ 0.125 \\ 0.375 \\ 0.25 \end{bmatrix} + (0.15) \begin{bmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} 0.213 \\ 0.106 \\ 0.319 \\ 0.213 \end{bmatrix} + \begin{bmatrix} 0.038 \\ 0.038 \\ 0.038 \\ 0.038 \end{bmatrix} \Rightarrow \begin{bmatrix} 0.251 \\ 0.144 \\ 0.357 \\ 0.251 \end{bmatrix}$$

$$P(1) = \begin{bmatrix} 0.25 \\ 0.144 \\ 0.35 \\ 0.25 \end{bmatrix} \qquad P(7) = \begin{bmatrix} 0.25 \\ 0.14 \\ 0.26 \\ 0.34 \end{bmatrix}$$

$P(K+1) \Rightarrow P(2)$

$$= 0.85 \times \begin{bmatrix} & & \\ & M & \\ & & \end{bmatrix} \times \begin{bmatrix} 0.25 \\ 0.14 \\ 0.35 \\ 0.25 \end{bmatrix} + (1-0.85) \begin{bmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{bmatrix}$$

Expt. No. :

Page No. :

upto $P(5)$ ie, $k = 4$

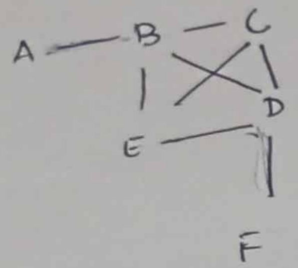
Date :

6. K-core graph

A, B, C, D, E, F

edges: AB, BC, BE, BD, CD, CE, DE, DF

a) compute K-core collapse sequence for k=1, 2, 3, 4.

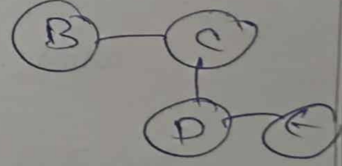


k=1
k=2
k=3
k=4

B, C, D, E.

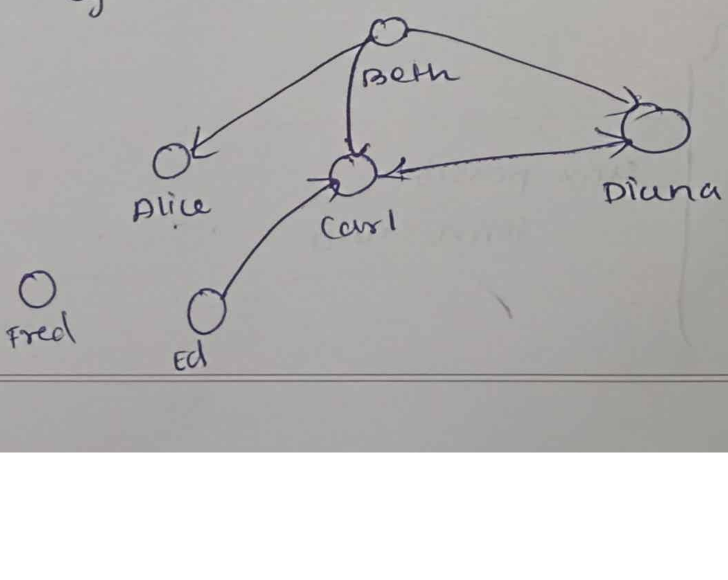
B — C — D — E.

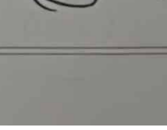
		Removed nodes	Remaining nodes	subgraph
		–	A, B, C, D, E, F	
k=1	A, F	A, F	B, C, D, E	
k=2	–	–	B, C, D, E	
k=3	C, E	C, E	B, D	
k=4	B, D			



7. NW Analysis

Ego analysis (Directed graph)





Ego NW Analysis for Alice.

Alice	Alice	Beth	Rowsum
Alice	0	1	1
Beth	1	0	1

Ego NW Analysis for Beth

Beth	Beth	Alice	carl	Diana	Rowsum
Beth	0	1	1	1	3
Alice	1	0	0	0	1
carl	0	0	0	1	1
Diana	0	0	1	0	1

Ego for all.

Fred has no connections → so independent

8. whole NW analysis

consider a NW with 4 nodes.

Node A - 2 neighbors with 1 connection
Node B - 3 " " 3 "
Node C - 1 " " 0 "
Node D - 3 " " 2 "

Soln:

Total possible $\frac{n(n-1)}{2}$ n→neighbor

$A \to \frac{2(2-1)}{2} = \frac{2(1)}{2} = 1$

$B \to \frac{3(3-1)}{2} = \frac{3(2)}{2} = 3$

$C \to \frac{1(1-1)}{2} = 0$

$D \to \frac{3(3-1)}{2} = 3$

Max possible connections

Cn : neighbors   conn - among neighbor

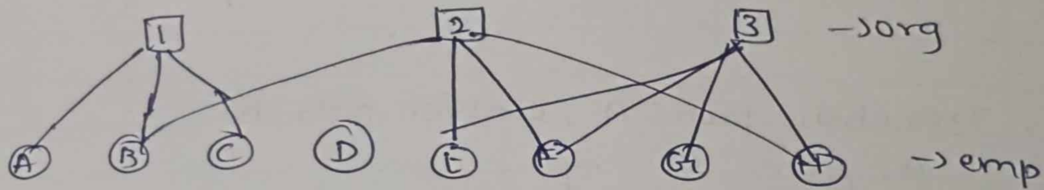
A -> 2   A -> 1
B -> 3   B -> 3
C -> 1   C -> 0
D -> 3   D -> 2

$\frac{\text{Total connection}}{\text{Max conn.}}$

CC(A) = 1/1 = 1
CC(B) = 3/3 = 1
CC(C) = 0/0 = 0
CC(D) = 2/3 = 0.67.

Avg clustering : $\frac{1+1+0+0.67}{4}$
= 0.667

9. two mode NW analysis.



Adj matrix.

	1	2	3
A	1	0	0
B	1	1	0
C	1	0	0
D	0	0	0
E	0	1	1
F	0	1	1
G	0	0	1
H	0	1	1

density for 2 mode NW analysis

$\Rightarrow \frac{L}{M \times N}$

m -> no. of nodes

n -> no. of org.

$\Rightarrow \frac{11}{3 \times 8} = \frac{11}{24}$.

11. G-test.

$$G\ test = 2 \sum_{i=1}^{n} O_i \ln\left(\frac{O_i}{E_i}\right)$$

$O = [8, 10, 12, 15, 7, 8]$
$E = [10, 10, 10, 10, 10, 10]$
$\alpha = 0.05$

Step 1 > Cal $\frac{O_i}{E_i}$ & $O_i \ln\left(\frac{O_i}{E_i}\right)$

category (i)	O_i	E_i	$\left(\frac{O_i}{E_i}\right)$	$O_i \ln\left(\frac{O_i}{E_i}\right)$
1	8	10	0.8	-1.785
2	10	10	1	0
3	12	10	1.2	2.188
4	15	10	1.5	6.082
5	7	10	0.7	-2.497
6	8	10	0.8	-1.785

$$G = 2 \sum_{i=1}^{n} O_i \ln\left(\frac{O_i}{E_i}\right)$$

$G = 2 \times (-1.785 + 2.188 + 6.082 + (-2.497) + (-1.785))$
$= 2 \times 2.20$
$= 4.49$

Degree of freedom = df => $k-1 = 6-1 = 5$
$\therefore p = 0.48$
$4.49 < 11.09$
accept (or) else reject