MIS -NO: 11 2.00 3 61 2

College of Engineering Pune

Test1

e : 27/2/2023

Subject: Data Science (Div1 +Div 2)

TY:Computer Engineering

ructions:

cientific calculators are only permissible

ver written answers would not be considered for evaluation

Time: 10 to 11 a.m

Which of the following is/are instance(s) of lemmatization? 1

a) eat -> ate
b) conflation->conflate
c) Studies -> Studi
d) walked -> walk

How many different lexemes are there in the following list? 1
man, men, girls, girl, mouse

a) 5
b) 4
c) 3
d) 2

Select categorical type of data attributes: 1

a) Nominal
b) Ordinal
c) Numeric
d) Discrete

[Fill in the blanks: 4-8]

IQR is used to identify the outlier 1

Transformation of a continuous attribute into ordinal categorical attribute is called as 1

_____

In case of many outliers in the dataset , the most representative value is maximum

Matching strings that should not have matched _____ [which error]

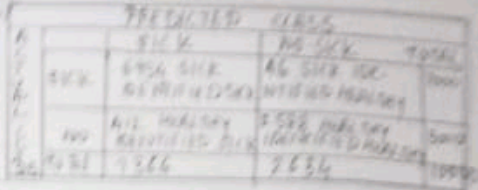
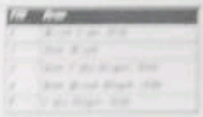
Low value for chi-square means there is a high similarity between two sets of data.

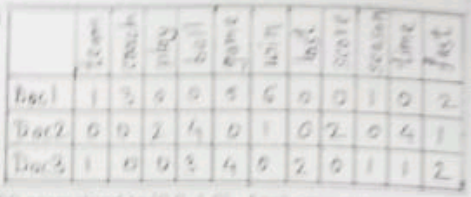
Reasons of noisy data [Select the appropriate] :

a) faulty data collection instruments
b) data entry problems
c) data transmission problems
d) inconsistent with other recorded data
e) technology limitation

All

10 Match the following

1. Term frequency matrix	a) 
2. Transactional data	(a) b) 
3. Ordered data	(b)
4. Cross tabs	(c) c)



11 Let data be in the range 10,000 to 80,000 normalized to [0.0, 1.0]. Apply min-max normalization to determine the normalized value of 56,500

Sol $v_i^* = \frac{v_i - min_A}{max_A - min_A} [new\_max_A - new\_min_A]$

$= \frac{56,500 - 10000}{80000 - 10000} [1 - 0]$

$v_i^* = 0.6571$

12 Let there be two Cricket players: Rohit and Hardik, and you have to select one for the cricket world cup. The score of both the players in the last five one-day matches are as follows:

Rohit	Hardik
28	92
38	59
45	2
59	17
63	83

Sol Find the player who is more consistent

finding mean, Q_1, Q_3, min, max for both of them

$mean(Rohit) = \frac{28+38+45+59+63}{5} = 46.6$

$mean(Hardik) = \frac{92+59+2+17+83}{5} = 58.6$

from taking diff. bet^n.

$min_R = 28$ $max_R = 63$

$min_H = 2$ $max_H = 83$

By taking diff. bet^n mean & min also mean & max it is found that difference is more in case of Hardik. while for Rohit diff. is quite same.

∴ Rohit is consistent.

$(mean_R - min_R) = 14.6 = R_1$

$(-(mean_R) + max_R) = 16.4 = R_2$

less diff. in R_1 & R_2.

$mean_H - min_H = 36.6 = H_1$

$max_H - mean_R = 44.4 = H_2$

more diff in H_1 & H_2 also more diff. bet^n mean & max

Also plotting boxplot can be find out i.e. Rohit is more consistent.

13 Consider the corpus C of sentences

1. there is a big garden
2. children play in a garden
3. they play inside beautiful garden

Calculate P(Children play in a big house) assuming a bi-gram language model using

i) Maximum likelihood estimation (MLE), and

ii) Add one smoothing

[Hint: Augment each sentence with sentence beginning and end markers]

Ma;?

Sol: from above corpus, using bigram lang model

$P(w_i | w_1, w_2, \ldots w_{i-1}) = \frac{P(w_{i-1}, w_i)}{P(w_{i-1})} = \frac{C(w_{i-1}, w_i)}{C(w_{i-1})}$

$P(play/children) = \frac{1}{1} = 1$

$P(in/play) = \frac{1}{2} = 0.5$

$P(a/in) = 1$, $P(big/a) = \frac{1}{2} = 0.5$,

$P(house/big) = 0$ — $P(\langle/s\rangle | house) = ?$

Here considering only one before word and bigram model. By using MLE, the probability of house follows big is 0, ∴

P(children play in a big house)

$= P(play/children) \times P(in/play) \times P(a/in) \times P(big/a)\ P(house/big)$

$= 1 \times 0.5 \times 1 \times 0.5 \times 0$

$= 0$

Add-one smoothing,

$P(w_i | w_1, w_2, \cdots w_{i-1}) = \frac{C(w_{i-1}, w_i) + 1}{C(w_{i-1}) + V}$

$P(in/play) = \frac{1+1}{1+11} = \frac{2}{12} = \frac{1}{6}$

$P(a/in) = \frac{1+1}{2+11} = \frac{2}{13}$

$P(a/in) = \frac{1+1}{1+11} = \frac{1}{6}$

$P(big/a) = \frac{1+1}{2+11} = \frac{2}{13}$

$P(house/big) = \frac{0+1}{1+11} = \frac{1}{12}$

∴ P(children play in a big house) $= \frac{1}{6} \times \frac{2}{13} \times \frac{1}{6} \times \frac{2}{13} \times \frac{1}{12}$

$= \frac{1}{18252}$

$= 0.000054788$

$1\frac{1}{2}$

incomplete

14 Suppose you have the following data of time in seconds for different cyclothon runners 2
52, 65, 62, 52, 53, 55, 60, 70, 54, 56, 58, 58, 62, 62, 54, 65, 56, 59, 52, 61, 67. Find median
and estimated median values.

Sol Arranging in increasing order = 52, 53, 54, 55, 56, 56, 58, 58, 59, 59, 60, 61, 61, 62, 62, 62, 64, 65, 65, 67, 70.

Here $N = 21$,

$\text{median} = x\left[\frac{N+1}{2}\right]$

$= x\left[\frac{21+1}{2}\right]$

$= x[11]$

$\therefore$ median = 60

1/2

Hai!

15 Suppose you are fishing in a lake with 8 species (bass, carp, catfish, eel, perch, salmon, trout, whitefish) and you have seen 6 species with the following counts: 10 carp, 3 perch, 2 whitefish, 1 trout, 1 salmon, and 1 eel (so you haven't yet seen the catfish or bass). What are the Good Turing probabilities (P^*_{GT}(trout)) for catching the next fish as "trout" or the Good Turing probabilities (P^*_{GT}(bass)) for catching the next fish as "bass"? 2

Sol N_c – frequency of frequency

Here $N_1 = 3$, $N_2 = 1$, $N_3 = 1$,

$N_4 = N_5 = N_6 = N_7 = N_8 = N_9 = 0$, $N_{10} = 1$

Total no. of fish found = 18 = N

So the probability that next fish is trough trout = 1/18

(since term freq^n. of trout = 1)

Now for bass, as in 18 already found fishes, # of bass = 0.

But $N_1 = 3$,

$\therefore P_{bass} = \frac{N_1}{N} = \frac{3}{18}$

But now finding next trout i.e. P^*_{GT}(trout) will change

It is less than 1/18.

$P^*_{GT} = \frac{N_1}{N}$

$c^* = \frac{(c+1)N_{c+1}}{N_c}$

for bass (which is unseen)

$c = 0$

$\therefore \boxed{P^*_{GT}(bass) = \frac{N_1}{N} = \frac{3}{18}}$

for seen trout,

$c = 1$ (1 time seen)

$\therefore c^* = \frac{(c+1)N_{c+1}}{N_c} = \frac{2N_2}{N_1}$

$c^* = \frac{2 \times 1}{3}$

2

$\therefore P^*_{GT}(trout) = \frac{\frac{2}{3}}{\frac{1}{18}}$

$\therefore P^*_{GT}(trout) = \frac{2}{3+18}$

$\therefore \boxed{P^*_{GT}(trout) = \frac{1}{27}}$

$$\cos(d_2, d_4) = \frac{1\times1 + 1\times1 + 0\times1}{\sqrt{1^2+1^2+0^2}\,\sqrt{1^2+1^2+1^2}}$$

$$= \frac{2}{\sqrt{4}\cdot\sqrt{3}} = \sqrt{\frac{2}{4}} = 0.86$$

$$(d_2, d_4) = \cos^{-1}(0.86) = 30.31$$

$$\cos(d_3, d_4) = \frac{1\times1}{\sqrt{1^2+1^2+1^2}\,\sqrt{1^2+1^2+1^2}} = \frac{1}{\sqrt{3}\,\sqrt{3}} = \frac{1}{3} = 0.33$$

$$(d_3, d_4) = \cos^{-1}(0.33) = 71$$

2 1/2

3 Given the following data. Fill the matrix by computing the minhash signature for each column using the given hash functions. 5

Element	S1	S2	S3	S4	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

Determine signatures matrix? [Note: Show steps while updating signature matrix starting from initialization]

Justify, how close are the estimated Jaccard similarities for all pairs of columns to the true Jaccard similarities?

signatures matrix:- for 2x+1(mod 6).

1st step:- We select 0th row, in which S_2 & S_4 are 1.

	1		1

∴ 1 1

2nd step:- Now, select 2nd row, in which S1 & S4 are 1.

5	1		1

3rd step:- Now, select 4th row, S_3 & S_4 are 1.

signature matrix:- For 3x+2(mod 6).

1st step:- Select 4th row.

		2	2

2nd step:- Select 5th row

5		2	2

3rd step:- Select 2nd row.

5	5	2	2

signature matrix:- For 5x+2(mod 6).

1st step:- Select 0th row

	1		

2nd step:- Select 4th row

	1	4	4

3rd step:- Select 5th row

3	1	4	4

Final

3	1	4	4

2023.04.20 14:59

Column 1 & 2.
Jsim = 0

Column 2 & 3.
Jsim = 0.

Column 3 & 4
Jsim = $\frac{1}{4}$ = 0.25.

∴ Columns 1 & 2 and Columns 2 & 3 have the Jaccard similarity. i.e. zero.

[Fill in the blanks: 4-7]

4 Anti-monotone property is: Support of an itemset does not exceed the support of its subset, is called anti-monotone property. 1

5 State Apriori principle: Support of an itemset is frequent, then the support of its subset is frequent. Apriori principle holds the property :- ∀ X, Y : X ⊆ Y then S(X) ≥ S(Y) 1

6 The Prune step of Apriori algorithm is : There are 3 steps :- 1) Brute-force 2) $F_{k-1} \times F_1$ merge 2) Merge $F_{k-1} \times F_1$ itemsets & 3) $F_{k-1} \times F_{k-1}$ itemsets 1

7 If minsup is set too low then the frequent itemset will be more

8 Dataset given below shows weight of students in a class. Draw box plot with whiskers for the same. Analyze the box plot. 2

Weight: 13, 12, 9, 11, 14, 12, 10, 20, 11, 10, 7

Sorted data: 7, 9, 10, 10, 11, 11, 12, 12, 13, 14, 20

median = 11

Q_1 (1^{st} quartile) = median(7, 9, 10, 10, 11)

$Q_1 = 10$

Q_3 (3^{rd} quartile) = median(12, 12, 13, 14, 20)

$Q_3 = 13$

$IQR = Q_3 - Q_1 = 13 - 10$

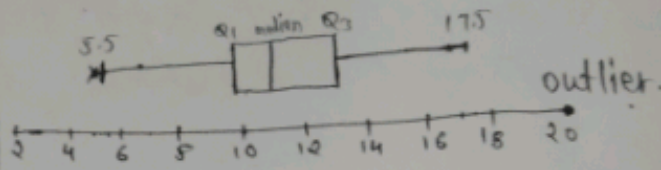
$IQR = 3$

Lower limit = $Q_1 - 1.5 \times IQR = 5.5$

Upper limit = $Q_3 + 1.5 \times IQR = 17.5$

There is outlier, which is greater than 17.5 is 20.

∴ Box plot:-



9 Investigate whether there's a relationship between the number of students and the marks scored. Which plot will help and how? Assume the data below : 2

No of students	Marks Obtained	Percentage Of Students
5	40	2.5
6	60	3
25	70	12.5
11	65	5.5
30	80	15
4	50	2
6	55	3
10	75	5
14	90	7
18	45	9
20	40	10
22	95	11
2	100	1
11	35	5.5
16	25	8

For the given relationship, the line plot will helpful.

2023.04.20 14:59

MIS -NO: 112003012

College of Engineering Pune

Test2

Subject: Data Science (Div1 +Div 2)

Date : 20/3/2023 TY-Computer Engineering Time: 10 to 11 a.m.

Instructions:

1. Scientific calculators are only permissible
2. Over written answers would not be considered for evaluation

1 Find the edit distances **using only insertions (cost 1) and deletions(cost 1)** between the following pairs of strings.

Source: abccdabc , Target: acbdcab

$D[i,j] = \min \begin{cases} D[i-1,j]+1 \\ D[i,j-1]+1 \\ D[i-1,j-1]+0 \text{ if } x=y \end{cases}$

Here only considering insertions & deletion (cost 1)

	#	a	b	c	c	d	a	b	c
#	0	1	2	3	4	5	6	7	8
a	1	0	1	2	3	4	5	6	7
c	2	1	2	3	2	3	4	5	6
b	3	2	1	2	3	4	5	4	5
d	4	3	2	3	4	3	4	5	6
c	5	4	3	2	3	4	5	6	5
a	6	5	4	3	4	5	4	5	6
b	7	6	5	4	5	6	5	4	5

(3)

Minimum Cost= 5

2 Give the formula for Cosine Similarity and determine the Cosine Similarity between the given document collection D and test document D4 specifying the document closer to test document. 4

[Apply stop-word removal, lemmatization] [Stop words: to, from, is]

Document Collection D:

D1: Jack travelled to London

D2: Jack travelled from Oakland to London

D3: Travel to Oakland is wonderful

Test Document D4: Jack travelled to Oakland

Cosine similarity

$$\cos(a,b) = \frac{|a \cap b|}{|a \cup b|}$$

$\cos(D_1, D_2) = \frac{4}{6} = 0.667$

$\cos(D_2, D_3) = \frac{2}{9} = 0.22$

$\cos(D_1, D_3) = \frac{1}{8} = 0.125$

Now: $\cos(D_1, D_4) = \frac{4}{5} = 0.8$

$\cos(D_2, D_4) = \frac{4}{6} = 0.667$

$\cos(D_3, D_4) = \frac{2}{7} = 0.285$

$0.8 > 0.667 > 0.285$

Now ~~for~~ $\cos(D_1, D_4) > \cos(D_2, D_4) \cos(D_3, D_4)$

$\therefore D_4$ is document closer to D_1

By Using Stop word removal

D_1 : Jack travelled London

D_2 : Jack travelled oakland london

D_3 : Travel oakland wonderful

D_4 : Jack travelled oakland

$\cos(D_1, D_2) = \frac{3}{4} = 0.75$

$\cos(D_1, D_3) = \frac{0}{6} = 0$

$\cos(D_2, D_3) = \frac{1}{6} = 0.167$

$\cos(D_1, D_4) = \frac{3}{4} = 0.75$

$\cos(D_2, D_4) = \frac{3}{4} = 0.75$

$\cos(D_3, D_4) = \frac{1}{5} = 0.2$

(0)

2023.04.20 14:59

3 Given the following data. Fill the matrix by computing the minhash signature for each column using the given hash functions. 5

Element	S1	S2	S3	S4	(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	4

Determine signatures matrix? [Note: Show steps while updating signature matrix starting from initialization]

Justify, how close are the estimated Jaccard similarities for all pairs of columns to the true Jaccard similarities?

2023.04.20 14:59

(x)

[Fill in the blanks: 4-7] $s(x) \geq s(y)$ $x \subset y$

4 Anti-monotone property is: candidate does not have more support than its subset. 1

5 State Apriori principle: If any set is frequent then its subset is also frequent itemset. 1

6 The Prune step of Apriori algorithm is : candidate elimination used to infrequent itemsets & prevents searching for association in infrequent item 1

7 If **minsup** is set **too low** then neraly all candiate itemsets becomes frequent itemsets. 1

all candi... are frequ...

2023.04.20 14:59

8 Dataset given below shows weight of students in a class. Draw box plot with whiskers for the same. Analyze the box plot. 2

Weight: 13, 12, 9, 11, 14, 12, 10, 20, 11, 10, 7

Arranging weights in increasing order

weight (w): 7, 9, 10, 10, 11, 11, 12, 12, 13, 14, 20

$n = 11$

min = 7, max = 20,

median = $w\left[\frac{n+1}{2}\right]$

$= w[6]$

$= 11$

$Q_1 = w\left[\frac{n+1}{2}\right]$, (n below median) $n = 5$

$= w[3]$

$Q_1 = 10$

$Q_3 = w\left[\frac{n+1}{2}\right]$, n above median

$= w[9]$

$Q_3 = 13$

∴ Using 5 no. summary.

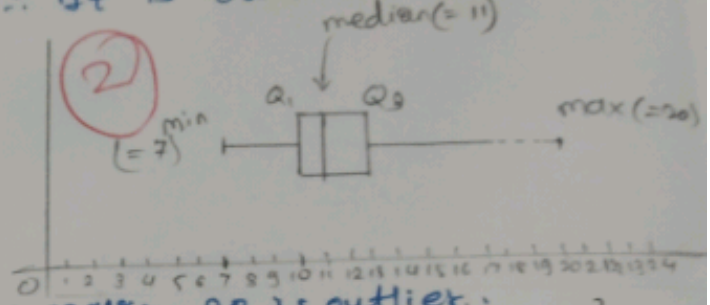
Box plot, IRQ = 13 − 10 = 3

outlier is less than $Q_1 - 1.5IRQ = 5.5$

more than $Q_3 + 1.5IRQ = 17.5$

now max (20) is more than 17.5

∴ It is outlier.

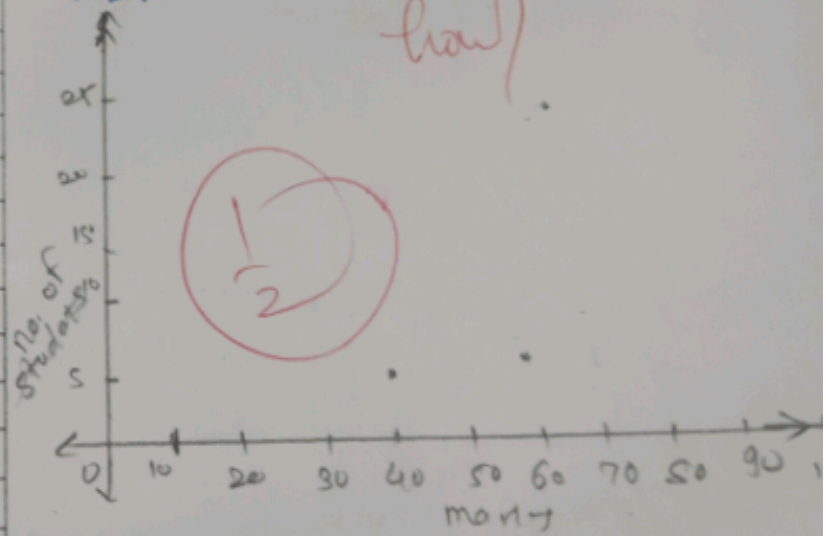


By angles, 20 is outlier. 2

9 Investigate whether there's a relationship between the number of students and the marks scored. Which plot will help and how? Assume the data below :

No of students	Marks Obtained	Percentage Of Students
5	40	2.5
6	60	3
25	70	12.5
11	65	5.5
30	80	15
4	50	2
6	55	3
10	75	5
14	90	7
18	45	9
20	40	10
22	95	11
2	100	1
11	35	5.5
16	25	8

We need to find out relationship between no. of students and marks. → Scatter plot.



2023.04.20 14:59

Add-k -

$$P_{Add\text{-}k}(w_{i-1}, w_i) = \frac{C(w_{i-1}, w_i) + k}{+V}$$

$$= \frac{C(w_{i-1}, w_i) + m\left(\frac{1}{V}\right)}{C(w_{i-1}) + m}$$

5] * Good - Turing smoothing -

N_c - frequency of frequency c

N_c - count of things we've seen c times

Sam I am I am Sam I do not eat

I	3
sam	2
am	2
do	1
no	1
eat	1

how many words are occuring once

$N_1 = 3$ (three words)

$N_2 = 2$

$N_3 = 1$

how many words occuring thrice (only one - I)

IMP

eg. suppose find 18 fish.

10 carp, 3 perch, 2 eel, 1 trout, 1 salmo = 18

next for trout = 1/18 → now new species (catfish) finding probability.

so here $N_1 = 3$, (no. of fishes occured only once)

SO $N_1/18 = 3/18$

Now probability of next trout is not 1/18 (less than it)

How to find new estimate?

2023.04.22 18:52

p^*_{GT} (thing with zero freq^n) $= \frac{N_1}{N}$

$c^* = \frac{(c+1)\,N_{c+1}}{N_c}$

Unseen (bass or catfish)

$c = 0$

MLE $= p = 0/18 = 0$

$p^*_{GT} = \frac{3}{18}$

Seen once (trout)

$c = 1$

MLE $= p = 1/18$

$c^*(\text{trout}) = \frac{2 \times N_2}{N_1}$

$= 2 \times 1/3$

$= 2/3$

So $p^*_{GT}(\text{trout}) = \frac{2}{3} \times \frac{1}{18} = \frac{1}{27}$

2023.04.22 18:53

MIS –NO: 111920052

College of Engineering Pune

Test1

Subject: Data Science (Div1 +Div 2)

Date : 27/2/2023 TY-Computer Engineering Time: 10 to 11 a.m.

Instructions:

1. Scientific calculators are only permissible
2. Over written answers would not be considered for evaluation

1 Which of the following is/are instance(s) of lemmatization? 1

a) eat -> ate
b) conflation->conflate
c) Studies -> Studi
d) walked -> walk

2 How many different lexemes are there in the following list? 1
man, men, girls, girl, mouse

a) 5
b) 4
c) 3
d) 2

3 Select categorical type of data attributes: 1

a) Nominal
b) Ordinal
c) Numeric
d) Discrete

[Fill in the blanks: 4-8]

4 IQR is used to identify the outlier 1

5 Transformation of a continuous attribute into ordinal categorical attribute is called as 1
discritisation

6 In case of many outliers in the dataset , the most representative value is mean 1

7 Matching strings that should not have matched Type II [which error] 1

8 Low value for chi-square means there is a high dissimilarity between two sets of data. 1

9 Reasons of noisy data [Select the appropriate] :

a) faulty data collection instruments
b) data entry problems
c) data transmission problems
d) inconsistent with other recorded data
e) technology limitation

0.25

14 Suppose you have the following data of time in seconds for different cyclothon runners: 59, 65, 61, 62, 53, 55, 60, 70, 64, 56, 58, 58, 62, 62, 54, 65, 56, 59, 52, 61, 67. Find median and estimated median values. 2

Sol: data : 52 53 54 55 56 56 58 58 59 59 60 61 61 62
64 65 65 67 70 $n = 21$

since n is odd median $= x_{n/2+1} = 60$

mean $= \frac{\Sigma x_i}{n} = \frac{1259}{21} = 59.95$

$\frac{1}{2}$

estimated median = ~~median - mean~~ ~~x_i - m~~

= 60 - x_i - mean

15 Suppose you are fishing in a lake with 8 species (bass, carp, catfish, eel, perch, salmon, trou whitefish) and you have seen 6 species with the following counts: 10 carp, 3 perch, whitefish, 1 trout, 1 salmon, and 1 eel (so you haven't yet seen the catfish or bass). What a the Good Turing probabilities (P^*_{GT}(trout)) for catching the next fish as "trout" or the Go Turing probabilities (P^*_{GT}(bass)) for catching the next fish as "bass"?

Sol:

0

11 Consider the corpus C of sentences

1. there is a big garden
2. children play in a garden
3. they play inside beautiful garden

Calculate P(children play in a big house) assuming a bi-gram language model using
i) Maximum likelihood estimation (MLE) , and
ii) Add one smoothing
[Hint: Augment each sentence with sentence beginning and end markers]

Sol:
1 <s> there is a big garden </s>
2 <s> children play in a garden </s>
3 <s> they play inside beautiful garden </s>

P(<s> children play in a big house </s>)

$P(w_i \mid w_{i-1}) = \dfrac{count(w_{i-1}, w_i)}{count(w_i)}$

no of variable = 6 + 2 = 8 (<s>, </s>)

$P = P(children \mid \langle s \rangle) \times P(play \mid children) \times P(in \mid play) \times P(a \mid in) \times P(big \mid a) \times P(house \mid big) \times P(\langle /s \rangle \mid house)$

$= \dfrac{c(\langle s \rangle, children)}{c(children)} \times \dfrac{c(children, play)}{c(play)} \times \dfrac{c(play, in)}{c(in)} \times \dfrac{c(in\ a)}{c(a)} \times \dfrac{c(a, big)}{c(big)} \times \dfrac{c(big\ house)}{c(house)} \times \dfrac{c(house\ \langle /s \rangle)}{c(\langle /s \rangle)}$

$P = \dfrac{1}{1} \times \dfrac{1}{2} \times \dfrac{1}{1} \times \dfrac{1}{2} \times \dfrac{1}{1} \times 0 \times 0 = 0$

Probability by adding one smoothing $= \dfrac{c(w_i \mid w_{i-1}) + 1}{c(w_{i-1}) + V}$

$P = \dfrac{1+1}{9} \times \dfrac{1+1}{2+8} \times \dfrac{1+1}{1+8} \times \dfrac{1+1}{2+8} \times \dfrac{1+1}{1+8} \times \dfrac{0+1}{0+8} \times \dfrac{0+1}{1+8}$

one smoothing $\dfrac{2}{9} \times \dfrac{2}{10} \times \dfrac{2}{9} \times \dfrac{2}{10} \times \dfrac{2}{9} \times \dfrac{1}{8} \times \dfrac{1}{9}$

$= \dfrac{32}{5248800} = \dfrac{1}{164025} = 6.0966 \times 10^{-6}$

one smoothing

10 Match the following:

1. Term Document matrix	d ✓	a)
2. Transactional data	c ✓	b)
3. Ordered data	b ✓	c)
4. Cross tabs	a ✓	d)

a)

ACTUAL CLASS	PREDICTED CLASS: SICK	PREDICTED CLASS: NO SICK	TOTAL
SICK	6954 SICK IDENTIFIED SICK	46 SICK IDENTIFIED HEALTHY	1000
NO	412 HEALTHY IDENTIFIED SICK	2588 HEALTHY IDENTIFIED HEALTHY	3000
TOTAL	7366	2634	10000

b)

Healthy person ...GTGCTGGCCCAT...

Person with β-thalassemia ...GTGCCGGCCCAT...

c)

TID	Items
1	Bread, Cake, Milk
2	Beer, Bread
3	Beer, Cake, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Cake, Diaper, Milk

d)

	team	coach	play	ball	game	win	lost	score	season	time	fast
Doc1	1	3	0	0	5	6	0	0	1	0	2
Doc2	0	0	2	4	0	1	0	2	0	4	1
Doc3	1	0	0	3	4	0	2	0	1	1	2

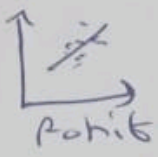
1 Let data be in the range 10,000 to 80,000 normalized to [0.0, 1.0]. Apply min-max normalization to determine the normalized value of 56,500. 1

min-max normalization $= \frac{v + min}{max - min} \times (newmax - newmin) + newmin$ (0.5)

normalized value of 56,500 $= \frac{56500 + 810000}{80000 - 10000} \times (1 - 0) + 0 = 0.95$

Let there be two Cricket players: Rohit and Hardik, and you have to select one for the cricket world cup. The score of both the players in the last five one-day matches are as follows: 2

Rohit X	Hardik y
28	32
38	79
45	2
59	17
63	83



Find the player who is more consistent

$\frac{28+38+45+59+63}{5} = 46.6$ ✓

$\frac{32+79+2+17+83}{5} = 42.6$ ✓

$\frac{\sum (x_i - \bar{x})^2}{n} = 169.04$ ✓

$\frac{(y_i - \bar{y})^2}{n} = 1064.112$ ~

from the 6 values s.d of hardik is more than the s.d of Rohit (1 1/2)

so Rohit is more consistant ✓

MIS –NO: 112003053

College of Engineering Pune

Test2

Subject: Data Science (Div1 +Div 2)

Date : 20/3/2023 TY-Computer Engineering Time: 10 to 11 a.m.

Instructions:

1. Scientific calculators are only permissible
2. Over written answers would not be considered for evaluation.

1 Find the edit distances **using only insertions (cost 1) and deletions(cost 1)** between the following pairs of strings. 3

Source: **abccdabc** , Target: **acbdcab**

	#	a	b	c	c	d	a	b	c
#	0	01	02	03	04	05	06	07	08
a	01	0	1	2	3	4	5	6	7
c	2	1	2	31	42	53	64	75	6
b	3	2	1	2	3	4	5	4	5
d	4	3	2	3	4	3	4	5	6
c	5	4	3	2	3	4	5	6	5
a	6	5	4	3	4	5	4	5	6
b	7	6	5	4	5	6	5	4	5

3

Minimum Cost= 5

2 Give the formula for Cosine Similarity and determine the Cosine Similarity between the given document collection D and test document D4 specifying the document closer to test document. 4

[Apply stop-word removal, lemmatization] [Stop words: to, from, is]

Document Collection D:

D1: Jack travelled to London

D2: Jack travelled from Oakland to London

D3: Travel to Oakland is wonderful

Test Document D4: Jack travelled to Oakland

D

	support count
Jack	2
travel	3
London	2
Oakland	2
wonderful	1

D4

	support count
Jack	1
travel	1
london	0
Oakland	1
wonderful	0

$[2,3,2,2,1] \cdot [1,1,0,1,0] = |D1||D4| \cos\theta$

D1: $[1,1,1,0,0] \cdot [1,1,0,1,0] = |D1||D4| \cos\theta_1$

$\therefore \cos\theta_1 = \frac{1+1}{\sqrt{3}\cdot\sqrt{3}} = \frac{2}{3}$

$\theta_1 = 48.16°$

Shot on OnePlus
By Jay 2023.04.20 14:54

8 Dataset given below shows weight of students in a class. Draw box plot with whiskers for the same. Analyze the box plot. 2

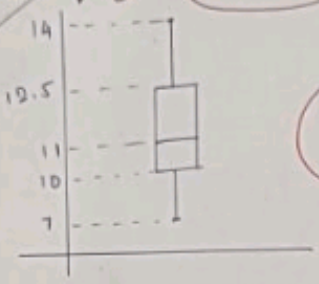
Weight: 13, 12, 9, 11, 14, 12, 10, 20, 11, 10, 7

7, 9, 10, 10, 11, 11, 12, 12, 13, 14, 20

$Q_2 = 11$, $Q_1 = 10$, $Q_3 = 12.5$

IQR = 2.5

1.5 IQR = 3.75

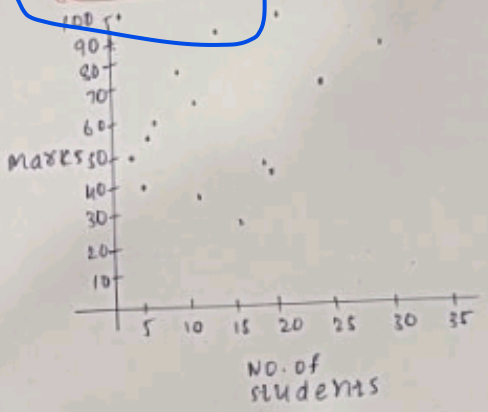


½

9 Investigate whether there's a relationship between the number of students and the marks scored. Which plot will help and how? Assume the data below : 2

No of students	Marks Obtained	Percentage Of Students
5	40	2.5
6	60	3
25	70	12.5
11	65	5.5
30	80	15
4	50	2
6	55	3
10	75	5
14	90	7
18	45	9
20	40	10
22	95	11
2	100	1
11	35	5.5
16	25	8

Scatter plot will help using Regression.



∴ There's no relation between them.

2

Shot on OnePlus
By Jay 2023.04.20 14:55

	Estimated Jaccard similarity	True Jaccard similarity
S_1, S_2	1/3	0
S_1, S_3	1/3	0
S_1, S_4	2/3	1/4
S_2, S_3	2/3	0
S_2, S_4	2/3	1/4
S_3, S_4	2/3	1/4

4 1/2

[Fill in the blanks: 4-7]

4 Anti-monotone property is: Support count of itemset is ≤ support count of its subsets 1

5 State Apriori principle: All the subsets of frequent itemset are also frequent itemsets. 1

6 The Prune step of Apriori algorithm is: If support count of itemset < minsup or an itemset is a superset of an itemset whose support count < minsup prune its too 1

7 If minsup is set too low then too many values and calculations, large memory needed. 1

4

Shot on OnePlus
By Jay 2023.04.20 14:54

D2 : $[1,1,1,1,0] \cdot [1,1,0,1,0] = |D2||D4| \cos\theta_2$

$1+1+1 = \sqrt{4} \cdot \sqrt{3} \cos\theta_2$

$\cos\theta_2 = \frac{3}{2\sqrt{3}} = \frac{\sqrt{3}}{2}$ $\quad \theta_2 = 30°$

D3 : $[0,1,0,1,1] \cdot [1,1,0,1,0] = |D3||D4| \cos\theta_3$

$1+1 = \sqrt{3} \cdot \sqrt{3} \cdot \cos\theta_3$

$\cos\theta_3 = \frac{2}{3}$ $\quad \theta_3 = 48.16°$

As $\theta_2 < \theta_1, \theta_3$

D2 is closer to D4.

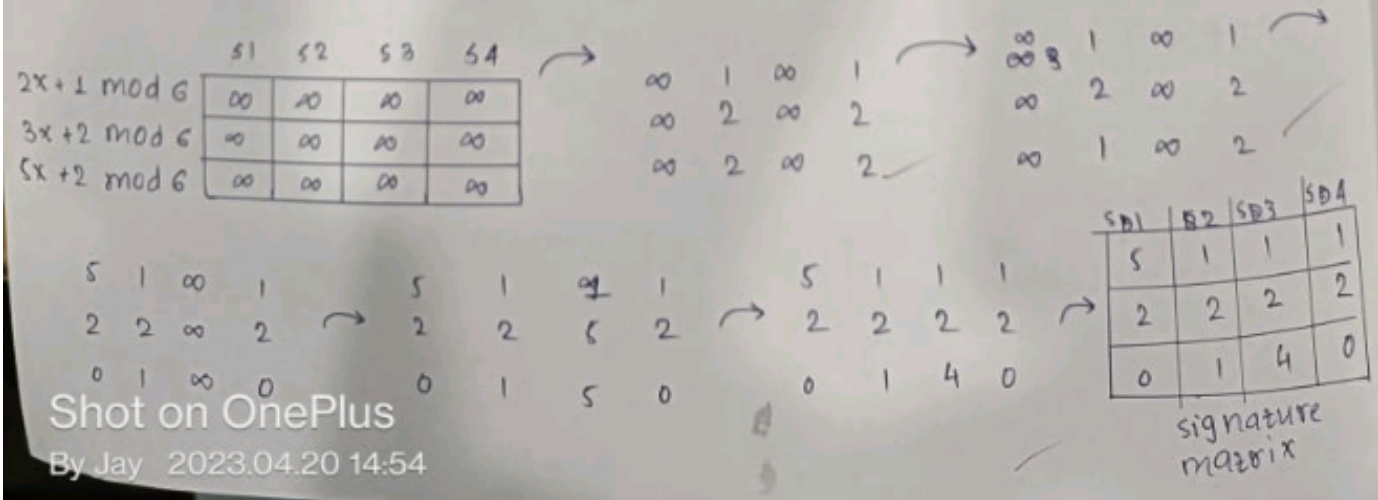
(4)

3 Given the following data. Fill the matrix by computing the minhash signature for each column using the given hash functions. 5

Element	S1	S2	S3	S4	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

Determine signatures matrix? [Note: Show steps while updating signature matrix starting from initialization]

Justify, how close are the estimated Jaccard similarities for all pairs of columns to the true Jaccard similarities?



Shot on OnePlus
By Jay 2023.04.20 14:54