THEORY ASSIGNMENTS

Answer all questions

01. For a 2 word query, the posting lists are as mentioned below:
[2, 5, 7, 13, 16, 18, 23, 32, 39, 58, 59, 63, 68, 76, 80] and [12, 56].
Show how many comparison (step by step) would be done to find out the intersection of the above two posting lists using skip pointer with a skip length of $\sqrt{P}$

Answer -

L1 = 2, 5, 7, 13, 16, 18, 23, 32, 39, 56, 59, 63, 68, 76, 80

P_1

L2 = 12, 56

P_2

For L1, skip span = $\sqrt{15} = 3.84 \approx 4$

L2, skip span = $\sqrt{2} = 1.41 \approx 1$

P1		P2	Remark
2	<	12	Not Found Match
16	>	12	Not Found Match
16	<	56	Not Found Match
39	<	56	Not Found Match
68	>	56	Not Found Match
56	=	56	Found Match

Total Comparisons needed are 6.

Q2. For a given document sorted in a data warehouse, compress the words by applying following preprocessing technique separately.

i) Normalization

ii) Stemming (Use Porter Stemmer)

iii) Stopwords removal

Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on full text or other content based indexing. Automated information retrieval systems are used to reduce what has been called "information overload". Many universities and public libraries use IR systems to provide access to books, journals and other documents. Web search engines are the most visible IR applications.

Answer-

i) Normalization: information retrieval is the activity of obtaining information resources relevant to an information need from a full text collection of information resources. Searches can be based on full text or other context based indexing. Automated information retrieval systems are used to reduce what has been called "information overload". Many universities and public libraries use IR systems to provide access to books, journals and other documents. Web search engines are the most visible IR applications.

ii) Stemming (Use Porter Stemmer): Information retriev is the activ of obtain inform resoure relev to an inform need from a collect of inform resoure. search can be base on full text system are or other content base index. autom inform retriev system are used to reduc what ha been call "inform overload". Mani univers and public librari use ir system to provid access to book, journal and other document. Web search engin are the most visible ir applic.

iii) Stopwords Removal : Information retrieval activity obtaining information resources relevant information collection information resources. Searches based full text content based indexing. Automated information retrieval systems reduce called "Informatio overload". universities public libraries IR systems provide access books, journals documents. Web search engines visible IR applications

03. Edit distance can be used for spelling correction in search queries.

i) Define edit distance

ii) As an example of how to calculate edit distance efficiently, show how dynamic programming can be used to calculate the edit distance between "system" and "item".

Answer-

i) Edit Distance :

→ Edit Distance calculate distance between mispelled term and correct form of term by using minimum number of operations

→ Let S1 and S2 be two character strings, then the edit distance between them is the minimum number of edit operations require to transform S1 to S2.

→ Most common edit operations allowed here are :
- insert
- delete
- replace

ii) S1 = system
S2 = item

		i		t		e		m	
	0	1	1	2	2	3	3	4	4
s	1	1	2	2	3	3	4	4	5
	1	2	1	2	2	3	3	4	4
y	2	2	2	2	3	3	4	4	5
	2	3	2	3	2	3	3	4	4
s	3	3	3	3	3	3	4	4	5
	3	4	3	4	3	4	3	4	4
t	4	4	4	3	4	4	4	4	5
	4	5	4	5	3	4	4	5	4
e	5	5	5	5	4	3	5	5	5
	5	6	5	6	4	5	3	4	4
m	6	6	6	6	5	5	4	3	5
	6	7	6	7	5	6	4	5	③

04. Suppose the vocabulary for your inverted index consists of the following 6 terms.

elite, elope, ellipse, eloquent, eligible, elongate

Assume that the dictionary data structure used for this index stores the actual terms using Dictionary-as-a string storage with front coding and a block size of 3. Show the resulting storage of the above vocabulary of 6 terms.

Answer-

Terms :- elite, elope, ellipse, eloquent, eligible, elongate

Given, Block size = 3

1st Block = elite, elope, ellipse

2nd Block = eloquent, eligible, elongate

Using Front coding -

5el*ite 3◇ope 5◇lipse → 1st Block

8 el *oquent 6◇igible 6◇ongate → 2nd Block

05. Consider a collection made of the following 4 documents (one document per line):

D1 : John gives a book to Mary.

D2 : John who reads a book loves Mary.

D3 : Who does John think Mary love?

D4 : John thinks a book is a good gift.

- These documents are preprocessed using a stop-list and a stemmer The resulting index is built ~~to~~ show by applying vector-based queries Give a (graphical or textual representation of this index)

- We now focus on 3 terms belonging to the dictionary, namely book love and Mary. Compute the tf-idf based vector representation for the 4 documents in the collection.

- Consider the query " love Mary". Give the results of a ranked retrieval for this query. What document is (are) considered to be the most relevant ?

Answer-

Terms : book, love, Mary

N = No. of Documents
= 4

Terms	tf				df	idf	tf * idf			
	D1	D2	D3	D4		$= \log \frac{N}{df}$	D1	D2	D3	D4
book	1	1	0	1	3	0.123	0.123	0.123	0	0.123
love	0	1	1	0	2	0.301	0	0.301	0.301	0
Mary	1	1	1	0	3	0.123	0.123	0.123	0.123	0

- Ranking the documents :

query : "love Mary"

We can rank the 4 documents by using the Jaccard co-efficient Similarity :

For D1 :

$$JC = \frac{q \cap D1}{q \cup D1} = \frac{1}{7} = 0.142$$

For D2 :

$$JC = \frac{q \cap D2}{q \cup D2} = \frac{2}{7} = 0.285$$

For D3 :

$$JC = \frac{q \cap D3}{q \cup D3} = \frac{2}{6} = 0.333$$

For D4 :

$$JC = \frac{q \cap D4}{q \cup D4} = \frac{0}{10} = 0$$

So, $D3 > D2 > D1 > D4$

Thus, D3 is considered to be the most rekevant document.

06. Compute the Jaccard matching score and the tf matching score for the following query-document pairs.

q: [information on cars] d: "all you've ever wanted to know about cars"

q: [information on cars] d: "information on trucks, information on planes, information on trains"

q: [red cars and red trucks] d: "cops stop red cars more often"

-Answer-

Jaccard Matching Score:

i) $JC = \frac{q \cap d}{q \cup d} = \frac{1}{10}$

ii) $JC = \frac{q \cap d}{q \cup d} = \frac{2}{6} = \frac{1}{3}$

iii) $JC = \frac{q \cap d}{q \cup d} = \frac{2}{8} = \frac{1}{4}$

tf matching Score:

i)

terms	tf	
	query	document
information	1	0
on	1	0
cars	1	1
all	0	1
you've	0	1
ever	0	1
wanted	0	1
to	0	1
know	0	1
about	0	1

ii)

terms	tf	
	query	document
information	1	3
on	1	3
cars	1	0
trucks	0	1
planes	0	1
trains	0	1

iii)

terms	tf	
	query	document
red	2	1
cars	1	1
and	1	0
trucks	1	0
cops	0	1
stop	0	1
more	0	1
often	0	1

07. The figure below shows the output of an information retrieval system on two queries. Crosses correspond to the relevant documents, dashes to non-relevant documents. Let the two documents contain 3 and 6 relevant documents respectively, but those only shown in the figure are retrieved by the system, not the others.

Rank	Q1	Q2
1	X	–
2	–	X
3	–	–
4	X	X
5	X	–
6	–	X
7	–	X
8	–	–
9	–	X
10	–	X

a) Draw the precision - recall curve

b) Compute the mean-average position

c) Compute the R-precision

Answer -

a) Precision - Recall Curve :

Rank	Precision		Recall	
	Q1	Q2	Q1	Q2
1	1/1 ✓	0/1	1/3	0/6
2	1/2	1/2 ✓	1/3	1/6
3	1/3	1/3	1/3	1/6
4	2/4 ✓	2/4 ✓	2/3	2/6
5	3/5	2/5	3/3	2/6
6	3/6 ✓	3/6 ✓	3/3	3/6
7	3/7	4/7 ✓	3/3	4/6
8	3/8	4/8	3/3	4/6
9	3/9	5/9 ✓	3/3	5/6
10	3/10	6/10	3/3	6/6

b) Average Precision of Q1 $= \frac{1}{1} + \frac{2}{4} + \frac{3}{6}$

$= 2$

Average Precision of Q2

$= \frac{1}{2} + \frac{2}{4} + \frac{3}{6} + \frac{4}{7} + \frac{5}{9} + \frac{6}{10}$

$= 3.22$

Mean Average Precision

$= \frac{2 + 3.22}{2} = 2.61$

c) R - Precision of Q1 $= \frac{1}{3}$

R - Precision of Q2 $= \frac{3}{6}$

08. Rank the documents in collection {d1, d2} for query q using the language model approach to IR introduced in class with Jelinek-Mercer smoothing. Use the mixture coefficient $\lambda = 0.4$

d_1: Scottish sheep getting smaller due to climate change study says

d_2: The analysis has shown a dramatic shift in the natural ranges for US Bird species in response to climate change

Query q : climate change

Answer-

According to Jelinek-Mercer smoothing:

$$P(q|d) = \prod_{1<k\leq|q|} [\lambda P(t_k|M_d) + (1-\lambda) P(t_k|M_c)]$$

For D1 =

$$P(q|D1) = \left[0.4 * \frac{1}{15} + 0.6 * \frac{1}{25}\right] * \left[0.4 * \frac{1}{15} + 0.6 * \frac{1}{25}\right] \approx 0.051^2 = 0.0026$$

$$P(q|D2) = \left[0.4 * \frac{0}{10} + 0.6 * \frac{1}{25}\right] * \left[0.4 * \frac{0}{10} + 0.6 * \frac{1}{25}\right] = 0.024^2 = 0.000576$$

Ranking - D1 > D2

09. Classify whether a given person is a male or a female based on a measured features. The features include height, weight and foot size Dataset is shown below

sex	height (feet)	weight (lbs)	foot size
male	6	180	12
male	5.92 (5'11")	190	11
male	5.58 (5'7")	170	12
male	5.92 (5'11")	165	10
female	5	100	06
female	5.5 (5'6")	150	08
female	5.42 (5'5")	130	07
female	5.75 (5'9")	150	09

Below is a sample to be classified using Neive Bayes algorithm as a male or female.

sex	height	weight	foot size
sample	6	130	8

Answer-

Classification of training data set :

sex	mean			variance		
	height	weight	foot size	height	weight	foot size
male	5.855	176.25	11.25	3.5033×10^{-2}	1.2292×10^{2}	9.1667×10^{-1}
female	5.4175	132.5	7.5	9.7225×10^{-2}	5.5833×10^{2}	1.6667

Classification of testing data set :

For male -

P(male) = 0.5

$$P(\text{height}\,|\,\text{male}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(6-\mu)^2}{2\sigma^2}\right) = 1.5789$$

Similarly,

$P(\text{weight}\,|\,\text{male}) = 5.9881\times10^{-6}$

$P(\text{foot size}\,|\,\text{male}) = 1.3112\times10^{-3}$

Thus,

$$\text{Posterior Numerator (male)} = \frac{P(\text{male})\,P(\text{height}/\text{male})\,P(\text{weight}/\text{male})\,P(\text{foot size}/\text{male})}{\text{evidence}}$$

$$= \frac{0.5\times1.5789\times5.9881\times10^{-6}\times1.3112\times10^{-3}}{1}$$

$$= 6.1984\times10^{-9}$$

For Female -

P(Female) = 0.5

$P(\text{height}\,|\,\text{female}) = 2.2346\times10^{-1}$

$P(\text{weight}\,|\,\text{female}) = 1.6789\times10^{-2}$

$P(\text{footsize}\,|\,\text{female}) = 2.8669\times10^{-1}$

$$\text{Posterior Numerator (female)} = \frac{P(\text{female})\,P(\text{height}|\text{female})\,P(\text{weight}\,|\,\text{female})\,P(\text{footsize}|\text{female})}{\text{evidence}}$$

$$= \frac{0.5 \times 2.2346 \times 10^{-1} \times 1.6789 \times 10^{-2} \times 2.8669 \times 10^{-1}}{1}$$

$$= 5.3778 \times 10^{-4}$$

Since, posterior numerator is greater in the female case, so we predict the sample is female.

10. Based on the data below, estimate a multinomial Naive Bayes Classifier and apply the classifier to the test document. Calculate the probability that the classifier assigns the test document to c = china or.

	dOCID	words in document	In c = China ?
training set	1	Taipei Taiwan	Yes
	2	Macao Taiwan Shanghai	Yes
	3	Japan Sapporo	no
	4	Sapporo Osaka Taiwan	no
	5	london	no
test set	6	Taiwan Taiwan Sapporo	?

-Answer-

Priors =

$\hat{P}(c) = \frac{2}{5}$ and $\hat{P}(\bar{c}) = \frac{3}{5}$

Conditional Probabilities =

$$\hat{P}(t|c) = \frac{T_{ct}+1}{\left(\sum_{t' \in v} T_{ct'}\right)+B}$$

where, B is the number different words

Here, B = 8

$$\hat{P}(\text{Sapporo}|c) = \frac{0+1}{5+8} = \frac{1}{13}$$

$$\hat{P}(\text{Taiwan}|c) = \frac{2+1}{5+8} = \frac{3}{13}$$

$$\hat{P}(\text{Sapporo}|\bar{c}) = \frac{2+1}{6+8} = \frac{3}{14}$$

$$\hat{P}(\text{Taiwan}|\bar{c}) = \frac{1+1}{6+8} = \frac{2}{14} = \frac{1}{7}$$

Now, $\hat{P}(c|d_5) \propto \frac{2}{5} \cdot \left(\frac{1}{13}\right)^3 \cdot \frac{3}{13} \cdot \frac{3}{13} \approx 0.000008$

$\hat{P}(\bar{c}|d_5) \propto \frac{3}{5} \cdot \left(\frac{3}{14}\right)^3 \cdot \frac{1}{7} \cdot \frac{1}{7} \approx 0.001$

Thus, the classifier assigns the test document to not China.