IR Midsem

Max marks -
Duration - 1 hr

Description:
1. The exam contains 17 MCQs.
2. There may be more than one option correct for each question.
3. Most questions are worth 1 point, the rest are for 2 points.
4. There is no partial marking. Full marks will be awarded for a question if and only if all correct and no wrong options are selected.
5. No negative marking.

Important Guidelines:
1. You may use a calculator (**do not use mobile phone calculator)
2. Kindly ensure your videos are on.
3. No extension will be given.

shubham21091@iiitd.ac.in Switch account

Your email will be recorded when you submit this form

What is the time complexity of the BSBI index creation algorithm, where n is the number of termID-docID pairs?

[✓ □ O(nlogn)]

□ O(n^2logn)

□ O(n)

□ None of the above

Postings list should be sorted by

- ☑ Term Frequency
- □ DocID
- □ TermID
- □ Document Frequency

Intersection operation of posting lists is always optimal when merged in increasing order of list size.

- □ True
- ☑ False

When using bigram indexes for processing the query mon*day, mark ALL words that will be matched while processing the query:

- □ monsundaay
- □ moonday
- □ daymon
- □ monday

Which of the following are true

□ SPIMI is more time-efficient than BSBI

□ SPIMI is more memory-consuming than BSBI

□ SPIMI is more suitable than block sort-based indexing when working with very large collections

□ Block sort-based indexing is more suitable than SPIMI when working with very large collections

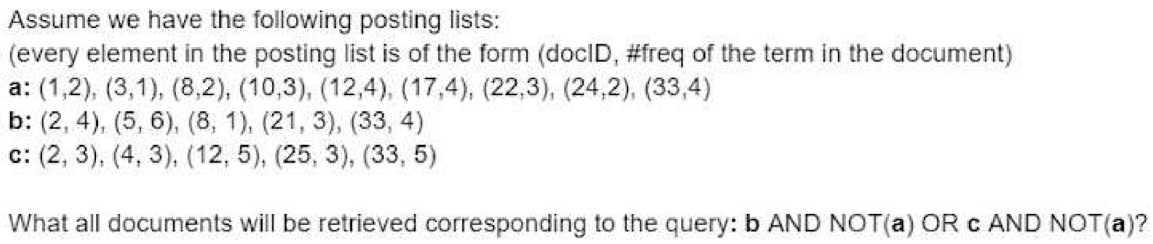
What is the edit distance between the words “mississippi” and “isssssippe” (where the allowed edits are insertion, deletion, and substitution)

□ 1

□ 2

□ 3

□ 4



Assume we have the following posting lists:
(every element in the posting list is of the form (docID, #freq of the term in the document)
a: (1,2), (3,1), (8,2), (10,3), (12,4), (17,4), (22,3), (24,2), (33,4)
b: (2, 4), (5, 6), (8, 1), (21, 3), (33, 4)
c: (2, 3), (4, 3), (12, 5), (25, 3), (33, 5)

What all documents will be retrieved corresponding to the query**: b** AND NOT(**a**) OR **c** AND NOT(**a**)?

□ 2, 5, 21, 33

□ 2, 4, 25, 33

□ 2, 4, 5, 21, 25

□ 2, 8, 12, 33

Mark ALL words that would be matched corresponding to the query: se*ate AND fil*er

□ senate

□ senate filler

□ filter

□ seagate filter

What will be the query corresponding to the query *tion to search in a permuterm index

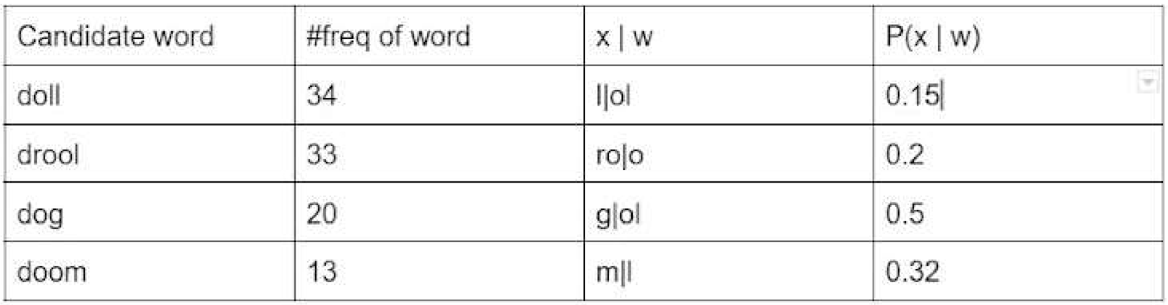
□ $*tion

□ *tion$

□ tion$*

□ tion*$*

Given the following information, what will be the most suitable spelling correction for the word “dool”



Candidate word	#freq of word	x | w	P(x | w)
doll	34	l|ol	0.15
drool	33	ro|o	0.2
dog	20	g|ol	0.5
doom	13	m|l	0.32

□ doll

□ drool

□ dog

□ doom

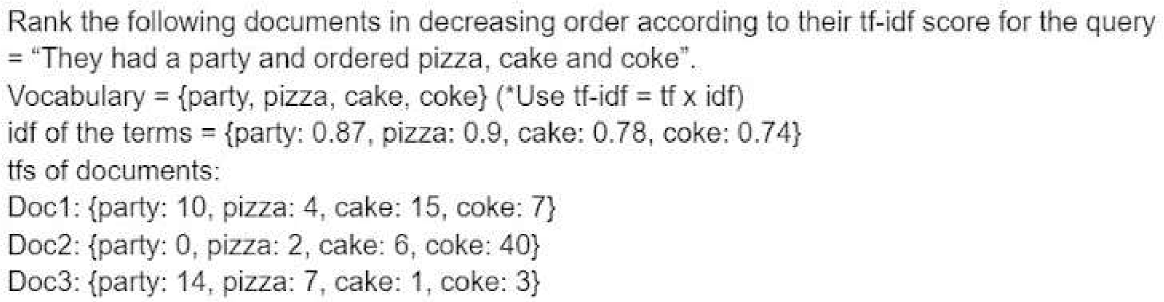
Choose the Correct statements for the Poisson model: i) It is a reasonable fit for general words. ii) It is a poor fit for general words. iii) It is a reasonable fit for topic-specific words. iv) It is a poor fit for topic-specific words.

□ i) and iii)

□ ii) and iv)

□ i) and iv)

□ ii) and iii)



Rank the following documents in decreasing order according to their tf-idf score for the query = "They had a party and ordered pizza, cake and coke".
Vocabulary = {party, pizza, cake, coke} (*Use tf-idf = tf x idf)
idf of the terms = {party: 0.87, pizza: 0.9, cake: 0.78, coke: 0.74}
tfs of documents:
Doc1: {party: 10, pizza: 4, cake: 15, coke: 7}
Doc2: {party: 0, pizza: 2, cake: 6, coke: 40}
Doc3: {party: 14, pizza: 7, cake: 1, coke: 3}

□ Doc3, Doc1, Doc2

□ Doc2, Doc3, Doc1

□ Doc2, Doc1, Doc3

□ Doc1, Doc2, Doc3

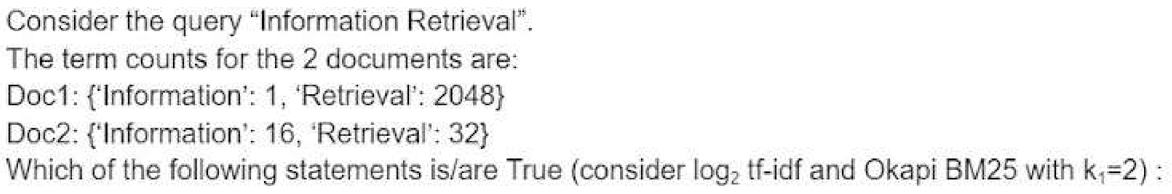
Which of the following statements is/are False:

□ Probabilistic Retrieval leads to Famine (too few(=0) results)

□ Boolean Search may lead to feast (too many results)

□ Boolean Search provides ranking of documents

□ Boolean search models may lead to Famine (too few(=0) results)



Consider the query "Information Retrieval".
The term counts for the 2 documents are:
Doc1: {'Information': 1, 'Retrieval': 2048}
Doc2: {'Information': 16, 'Retrieval': 32}
Which of the following statements is/are True (consider $\log_2$ tf-idf and Okapi BM25 with k_1=2) :

□ Tf-idf will return the ranking: Doc2, Doc1

□ Tf-idf will return the ranking: Doc1, Doc2

□ BM25 will return the ranking: Doc1, Doc2

□ BM25 will return the ranking: Doc2, Doc1

Paragraph (For next 3 questions):

Query - b a d c e
Doc1 - a b d
Doc2 - a c d e b e
Doc3 - c d e f b c
Doc4 - c a b a d
Individual Terms: a b c d e f
While ranking the documents using Binary Independence Model (BIM), in a particular iteration, we get user feedback which tells us that -
(i) All documents are relevant
(ii) A term/token is relevant to a document if the document contains that specific term/token.
Now for this particular iteration, answer the following:
(Note:
Use log10 wherever log is required in Q11-13.
For smoothing, add 0.5 to every count.)

Paragraph Q1

Calculate the log-odds ratio for term ‘a’ (Upto 4 decimal points, No round-off) (Note: Use the contingency table. For smoothing, add 0.5 to every count in the table)

○ 0.3679

○ 0.9542

○ 0.6020

○ -0.4771

Paragraph Q2

Calculate the log-odds ratio for term ‘e’ (Upto 4 decimal points, No round-off) (Note: Use the contingency table. For smoothing, add 0.5 to every count in the table)

○ 0.9542

○ -0.6020

○ 0

○ 0.4771

Paragraph Q3

Which of the following is/are true? (RSV(D) denotes the Retrieval Status Value for document D)

□ RSV(Doc4) = 3.0121

□ RSV(Doc1) = 1.7236

□ RSV(Doc3) = 2.6442

□ RSV(Doc2) = RSV(Doc3)

Submit

Clear form