National University of Computer and Emerging Sciences, Lahore Campus

	Course:	Information Retrieval	Course Code:	CS4051
	Program:	BS (Computer Science)	Semester:	Spring 2022
			Total Marks:	25
	Deadline:	11-March-2022		
	Section:	BCS-8A		
	Assessment	QUIZ - 1		

Instruction/Notes:

25/25

1. **Please mark the correct answer.** (5)

5

i) A corpus in a directory has 50 articles. If you perform 1-query based search to find the closest article to this query article, how many times you must compute the similarity between articles and query.

A. 1

B. 49

[✓ C. 50] ✓

D. 0

ii) What is the upper bound for BM25 transformation?

[✓ A. k+1] ✓

B. k-1

C. k

D. 0

iii) Which is true about pivoted length normalization?

A. It always rewards.

B. It always penalizes.

[✓ C. It has both a penalization and reward effect.] ✓

D. None

iv) Consider the instantiation of the vector space model where documents and queries are represented as bit vectors. Assume we have the following query and two documents.

Q = "healthy diet plans"

D1 = "healthy plans for weight loss. Check out other healthy plans"

D2 = "the presidential candidate plans to change the educational system."

Let V(X) = [b1 b2 b3] represent a part of the bit vector for document or query X, where b1, b2, and b3 are the bits corresponding to "healthy," "diet," and "plans," respectively.

Which of the following is true?

A. V(Q) = [1 1 1] V(D1) = [1 1 1] V(D2) = [0 0 1]
B. V(Q) = [1 1 1] V(D1) = [1 1 1] V(D2) = [0 0 0]
C. V(Q) = [1 1 1] V(D1) = [2 0 2] V(D2) = [0 0 1]
[✓ D.] V(Q) = [1 1 1] V(D1) = [1 0 1] V(D2) = [0 0 1]

v) consider the same scenario as in Question iv, with dot product as the similarity measure. Which of the following is true?

[✓ A.] Sim (Q,D1) = 2 Sim(Q,D2) = 1
B. Sim (Q,D1) = 3 Sim (Q,D2) = 0
C. Sim (Q,D1) = 3 Sim(Q,D2) = 1
D. Sim (Q,D1) = 4 Sim (Q,D2) = 1

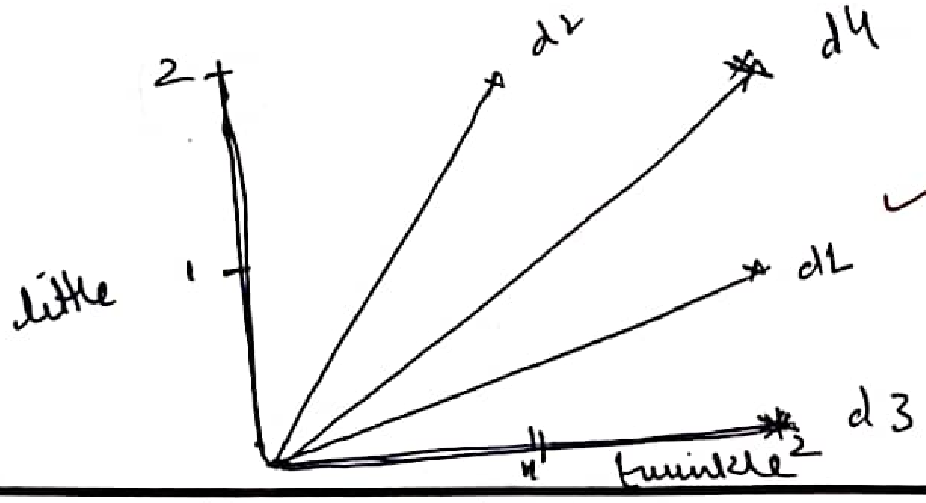
2. **These documents as vectors in space with dimension R^n. calculate the value of N and draw these documents as vectors in R^n.** (5)

Corpus= {d1, d2, d3, d4}

N=2

5

d1= twinkle twinkle little.
d2= little little twinkle.
d3= twinkle twinkle.
d4= little little twinkle twinkle.



5

3. **What is BM25 Model? which problems it addresses and how?** (5)

BM25 model is a modification/enhancement ~~vsm~~ for vector space model which is used for finding similiarity between a query and a document. The original version tackles the problem of repeated terms in the document (Term Frequency) by putting an upper band on the TF. It ~~so~~ achieves this by transforming TF with a "k" parameter, and this K1 (K+1) denotes the upper band of TF.

The modified BM25 accounts for document length. It introduces document length normalization in the denominator of Term Frequency, thus penalizing long documents while rewarding short ones.

4. **Consider the table of term frequencies for 3 documents denoted Doc1, Doc2, Doc3 in following table. Compute the idf weights for the terms car, auto and also compute tf-idf for both terms in for each document. Total number of documents are 806,791.** **(10)**

	Doc1	Doc2	Doc3
car	27	4	24
auto	3	33	0
insurance	0	33	29
best	14	0	17

$$\left(\text{IDF} \rightarrow \log_{10}\left(\frac{M+1}{k+1}\right) \right)$$

10

Write Idf values in the following table, where N=806,791

	df (no of docs containing a term)	Idf
Car	18,165	1.648
Auto	6723	2.080

Write the tf-idf values in the following table:

		Doc1	Docs2	Doc3
Car	1.648 × 27	44.496	6.592	39.55
Auto	2.08 × 3	6.24	68.64	0

Using TF_2 ~~$\frac{(K+1)x}{K+x}$~~ n

So TF-IDF ⇒ $\frac{(K+1)x}{K+x}$ × IDF