C1 Assessment (BTech) / Midsem (MTech) **Information Retrieval**

Total Marks: 25
Time: 1 Hour

Section A:

Attempt all 5 questions in this section [2 marks each]

1. Give an example for a situation, when for a given query two documents have the same cosine similarity score but different TF-IDF scores for specific terms?

2. Explain the steps of any stemming algorithm for reducing the word "connections" to its stem.

3. If you increase the number of terms in a query, how will it impact precision and recall? Justify your answer.

4. Explain Zipf's Law in the context of natural language processing and how it impacts term weighting in information retrieval.

5. What is the difference between a 'stem' and a 'root word; in the context of search queries?

Section B:

Attempt any 2 questions (out of 3) in this section. Each question carries 4 marks.

Ques 1. Consider a document collection with the following three documents:

1. **Doc1:** "Information retrieval techniques can be complex."
2. **Doc2:** "Complex retrieval methods can improve information processing."
3. **Doc3:** "Techniques for processing information are crucial."

You are given a Boolean query: **"information AND (retrieval OR techniques)"**.

a) Using the Boolean model, determine which documents satisfy the query. Explain your reasoning. [1 mark]

b) Convert this Boolean query into a Vector Space model query. Represent this query as a vector in a term space where terms are "information," "retrieval," "techniques," and "processing." Assume TF-IDF weights are provided. Show the query vector in this space. [2 marks]

c) Suppose a new document is added to the collection: **"Advanced techniques in information retrieval are explored."** How would this new document affect the Boolean and Vector Space models for the query? Discuss any changes in ranking for the existing documents. [1 mark]

Ques2. Given the sentence:

"Natural language processing improves machine learning models."

a) How many unique **unigrams**, **bigrams**, and **trigrams** can be formed from this sentence? Show the counts for each. [1.5 mark]

b) What is the probability of the word **"machine"** in a **unigram model** using Maximum Likelihood Estimation (MLE)? [1.5 mark]

c) If the word **"improves"** did not appear in the training data, how would a **trigram model** handle this case? Briefly mention the most common solution. [1 mark]

Ques3. Given the following data:

1. **Relevant Documents (R):** 8
2. **Retrieved Documents (Retrieved):** 12
3. **Relevant & Retrieved Documents (R ∩ Retrieved):** 5

a. Calculate the precision and recall for this retrieval. [1 mark]

b. Calculate the F1-measure. [1 mark]

c. If $\beta = 2$, calculate the E-measure**.** [2 marks]

Section C:

Attempt any one question from this section [7 marks]

Ques 1 **.** Consider a small corpus consisting of 3 documents:

I. **Doc1:** "Machine learning is a fascinating field."

II. **Doc2:** "Deep learning is a subset of machine learning."

III. **Doc3:** "Data science and machine learning are closely related."

You are interested in the term **"learning"**.

a) Calculate the TF-IDF score for the term **"learning"** in **Doc2**. [3 mark]
Use the following formulae:

- **TF (Term Frequency)** = Number of times term occurs in document/ Total number of terms in document
- **IDF (Inverse Document Frequency)** = log(Total number of documents/ Number of documents containing the term)

- **TF-IDF** = TF × IDF

b) If **Doc2** is expanded with the term **"neural networks"** and the new document becomes: "Deep learning and neural networks are a subset of machine learning," how would the TF-IDF score of the term **"learning"** in **Doc2** change? [2 marks]

c) Suppose a new document, **Doc4**, is added to the corpus: "Learning is essential in artificial intelligence." How would the addition of **Doc4** affect the IDF of the term **"learning"** and consequently its TF-IDF score in the existing documents (Doc1, Doc2, Doc3)? [2 marks]

Ques 2. Explain the impact on the overall relevance scoring.In a scenario where the search query is vague or ambiguous, such as "apple," the information retrieval system retrieves a set of documents. You are given the following details:

- **Total documents in the collection**: 100
- **Documents retrieved for the query "apple"**: 20
- **Documents relevant to "apple (fruit)"**: 15
- **Documents retrieved and relevant to "apple (fruit)"**: 10

Based on this information, answer the following:

a. Discuss the limitations of basic keyword-based retrieval for the query "apple" in distinguishing between **"apple (fruit)"** and **"Apple (company)"**. [3 marks]

b. Explain how document clustering and 2- Nearest Neighbour algorithm can be applied to enhance the effectiveness of the system. [4 marks]