

Indian Statistical Institute

End-Semester Examination

Second Semester: 2024-25

Information Retrieval

Time: 3 Hours
Full Marks: 100

Date: 2 May 2025

Answer Question 1, and any 6 questions from the rest.

1. (a) Suppose the TF.IDF score for a term t in a document d is defined as

$$\text{score}(t, d) = \frac{\text{freq}(t, d)}{\text{df}(t)},$$

where $\text{freq}(t, d)$ denotes the number of times the term t occurs in document d , and $\text{df}(t)$ is the number of documents in which the term t appears. Using this formula, construct an inverted index for the corpus consisting of the following three documents:

- D1: data science is about fun with data
- D2: science needs data
- D3: models in data science are powerful

- (b) Consider the following alternative formula for TF.IDF, given as:

$$\text{score}(t, d) = (1 + \log(\text{freq}(t, d))) \cdot (\log N - \log(\text{df}(t))),$$

where N denotes the total number of documents in the corpus. For large collections, which of these two formulae is expected to result in better retrieval? Explain why.

$$[6 + (1 + 3) = 10]$$

2. (a) Define non-interpolated precision and interpolated precision at a recall point $r \in [0, 1]$ and explain the advantage of one over another briefly, if any.
- (b) Prove or disprove: For a query and two lists of ranked results for the query, the interpolated precisions are same at recall 0.5 if and only if the non-interpolated precisions are also same at that recall point.

$$[(3 + 3 + 3) + 6 = 15]$$

3. (a) Does query expansion help in improving precision, recall, or both? Explain with example scenarios.
- (b) Briefly describe how pseudo relevance feedback works.
- (c) Describe one scenario when pseudo relevance feedback hurts retrieval performance.

[5 + 5 + 5 = 15]

4. (a) Explain why dimension reduction using truncated SVD can be interpreted as an autoencoder with no non-linear activation function.
- (b) Suppose, in a min-hash signature with b bands of r rows each, two documents (columns) are classified as a candidate pair if and only if their signatures agree on at least one row of each band. If the Jaccard similarity of two particular documents is s , then derive the probability p that those two documents are classified as a candidate pair. Based on your intuition, draw a rough plot of p vs s , for $b = r = 5$.

[4 + 8 + 3 = 15]

5. (a) The Word2Vec embeddings are trained based on the primary assumption that words that appear close to each other should be represented by similar vectors. Very frequent words appear close to many other words. How is this problem mitigated while training Word2Vec?
- (b) Explain the advantages of contextual word embeddings over context-independent embeddings, and vice-versa, if any.
- (c) Give one example each of a contextual and a context-independent word embedding model. Briefly explain why each example fits its category.

[6 + 3 + (2 + 4) = 15]

6. (a) Consider the question:

Who accidentally discovered penicillin?

Show how the askMSR system would rewrite this question into multiple search queries as part of its QA pipeline.

- (b) What is the primary assumption based on which query suggestion using query flow graph works? When are two nodes of the query flow graph connected by an edge?
- (c) Briefly explain how TrustRank modifies PageRank. Why is TrustRank effective against link-spam?

[5 + 2 + 2 + (3 + 3) = 15]

7. (a) Can self-attention in transformer models be used without multi-head attention? Justify your answer.
- (b) Describe what modification or addition of layers, with suitable activation functions, are required to fine-tune BERT for the following tasks. Also describe how the input will be fed into BERT.
- (i) Sentiment analysis: takes a small paragraph as input and outputs one of the two classes, namely positive or negative.
 - (ii) Paraphrase detection: takes two sentences and determines whether they have the same meaning.

$$[5 + (5 + 5) = 15]$$

8. (a) In the context of Retrieval-Augmented Generation (RAG), why is it challenging to evaluate the retrieval and generation components independently?
- (b) What is Knowledge-F1? Mention one of its limitations.
- (c) Briefly explain the strategy of training a ColBERT model. Design and justify an appropriate loss function for it.

$$[3 + (2 + 2) + (4 + 4) = 15]$$