DS5603 Information Retrieval
Dr. Koninika Pal
IIT Palakkad
Test 1: Date 18.09.2025, Time 8:00 - 8:50

INDIAN INSTITUTE OF TECHNOLOGY PALAKKAD

Question 1: Text Processing (2 points)

(i) What will be the Levenshtein distance between *planet and paint* if an insert operation costs twice a delete operation? Present the dynamic table.

Question 2: IR Effectiveness (2+2 points)

Suppose we have a corpus of 100 documents. 20 of them are relevant for the query. A model retrieves 20 documents and 10 of them are correct. The relevant documents are retrieved in the following positions – 1, 2, 5, 6, 8, 9, 10, 12, 14, 15. Rest of the relevant documents are retrieved at position 51 to 60.

(i) Compute precision, recall, precision@5, and F_1 for the methods.

(ii) Compute average precision of the method?

Question 3: Ranking Model (1.5+ 2.5 + 1 points)

(i) How could you implement ranking in Boolean retrieval model for documents containing a title and different subsections?

(ii) How could you support the tf-idf scoring with probabilistic ranking principal ? [Must derive the relation between them]

(iii) What is the main difference between tf-idf scoring and BM-25?

Question 4: Language model ((1+2)+ 1 points)

(i) What is the probabilistic representation of query-likelihood approach? which parameter do you estimate from the corpus and how?

(ii) Does this approach support synonymy ? State one plausible solution for it.

DS5603 Information Retrieval
Dr. Koninika Pal
IIT Palakkad
Test II: Date 23.10.2025, Time 8:00 - 8:50

INDIAN INSTITUTE OF TECHNOLOGY PALAKKAD

Question 1: Topic Modelling(2+2)

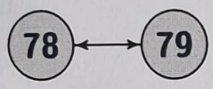
(a) Given a word as a query, how do you find similar words using LSI?

(b) Compare LSI and PLSI.

Question 2: Link Analysis(2+3)

(a) What is a random surfer model? How do you calculate pageRank of a page p ?

(b) Assume that a graph consisting of $|V| = 100$ vertices (all of which with non-zero outdegree) has the following isolated component consisting of two vertices:



What is the PageRank score of the two vertices assuming $\epsilon = 0.2$?

Question 3: Indexing and Compression (2 + (1+3))

(a) What kind of index needs to be created to support the following search scenarios, and what information needs to be stored in the postings?

(i) boolean search

(ii) proximity search

(b) What encoding methods do you use to encode the following index entries $\langle 1004, 1009, 1021, 1042, 1047, 1135, 1234 \rangle$? What is the final encoded result? Show each step of the encoding process. [Hint: you have to use two compression methods]

DS5603 Information Retrieval
Dr. Koninika Pal
IIT Palakkad
Final: Date 28.12.2022, Time 9:30 - 11:30

INDIAN INSTITUTE OF TECHNOLOGY PALAKKAD

Question 1: Web Graph ((2+2)+ 1 =5)

(a) Consider the Markov chain described by the following transition probability matrix

$$P = \begin{bmatrix} 0.0 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.5 & 0.5 \\ 0.0 & 0.0 & 0.0 & 1.0 \\ 0.5 & 0.5 & 0.0 & 0.0 \end{bmatrix}$$

(i) Show that the Markov chain is ergodic.

(ii) Determine its stationary state probabilities for the graph.

(b) How can you use PageRank in retrieval models?

Question 2: Query processing (4 + 2 + 2= 8)

Consider a top-k query with $m = 3$ terms, the user is interested in $k = 2$ results, and (non-weighted) summation as score aggregation. The underlying three index lists have the following (document identifier, score) entries:

$term_1$
d_1 0.9
d_7 0.7
d_3 0.3
d_2 0.3
d_4 0.3
d_5 0.2

$term_2$
d_3 0.8
d_4 0.8
d_7 0.5
d_1 0.3
d_6 0.2
d_5 0.2
d_2 0.2

$term_3$
d_1 0.7
d_6 0.6
d_7 0.5
d_4 0.4
d_2 0.3

(a) Apply the NRA method (without random accesses) to this setting. Document all index accessing steps and the top-k after each of them. How many sorted accesses (SA) and random accesses (RA) does the method need? Show the table for current score and upper bound for candidate documents and top-k documents for each round.

(b) Can you apply the aggregate function $spread = max - min$ to NRA method. Justify your answer.

(c) If you have the above index sorted by the document id, how do you process the conjunctive query $term_1 \wedge term_2 \wedge term_3$?

Question 3: Indexing and Compression (2 + 4 + 2 + 3+ 2= 13)

(a) Write Ziv-Lempel Compression for the text: *she sells seashells*

(b) What is the optimal ways to encode the following index entries $1004, 1009, 1021, 1042, 1047, 1135, 1234$? What is the final encoded result? Show each step of the encoding process.

(c) How much space (bits or byte-level) will you save if the entries are encoded using variable byte encoding?

(d) Decode the following compressed sequence that was created using Golomb encoding with M = 9.

0011111011111111110101

(e) How does the near-duplicate-detection method improve the efficiency and effectivity of a retrieval model?

Question 4: Advanced IR(2+(1+1) = 4)

(a) If you know the static embedding for the words of two sentences, how do you find the similarity between two sentences?

(b) If you know the dynamic embedding for the words, what is the best way to find the similarity between two sentences? what is the advantage of this approach compared to the approach considered in the prevision question?