PES University, Bengaluru

(Established under Karnataka Act 16 of 2013)

END SEMESTER ASSESSMENT (ESA) - JULY - 2023

UE20CS332 - Algorithms for Information Retrieval and Intelligence Web

Total Marks : 100.0

1.a. Define
1. Information Retrieval
2. Boolean Retrieval Model

(4.0 Marks)

1.b. Suppose you want to retrieve documents for the wildcard query “ho*l” . Explain how would you use permuterm index for the retrieval process. Take example of term “hotel” to explain. (4.0 Marks)

1.c. Write an algorithm to intersect posting lists with the help of skip pointers (4.0 Marks)

1.d. Match the list of surnames with their corresponding SOUNDEX codes and restore the missing code characters.
Surnames: Allaway, Buckingham, Colquhoun, Kingscott, Whytehead, Lewis, Littlejohns, Tocher.
Soundex Codes: _252, _252, _330, L2_0, A400, L_42, T_6_, C42_.
You are also provided the required letter to digit mappings
B, F, P, V --> 1
C, G, J, K, Q, S, X, Z --> 2
D, T --> 3
L --> 4
M, N --> 5
R --> 6 (8.0 Marks)

2.a. Explain Single Pass In-Memory Indexing with algorithm (6.0 Marks)

2.b. Suppose you want to index corpus from a new language, for which average word size(number of characters) per token is 8, and average word size per term is 12. Also, it is not very uncommon to have words having 24 characters in the language, so you assign 24 bytes per term for the dictionary storage.

1. Assume that in your corpus, you have 3 million unique terms. Estimate the size of the dictionary while using standard array of fixed width entries. (consider term frequency, pointer to posting list) (2 marks)
2. How much compression can you achieve on this, if you store dictionary as a (long) string, with pointer to the next word showing end of the current word? [Indicate the final size of the dictionary] (6 marks)

(8.0 Marks)

2.c. Which of the two novels are more similar among Sense and Sensibility(SaS), Pride and Prejudice (PaP), and Wuthering Heights (WH)? The term frequencies in all three novels is listed below.

term	**SaS**	**PaP**	**WH**
affection	115	58	20
jealous	10	7	11
gossip	2	0	6
wuthering	0	0	38

(6.0 Marks)

3.a. Suppose the table given below lists all the documents retrieved by an algorithm. If total number of relevant documents is 6, calculate the value of recall, precision, and F-score.

Sn	Doc ID	relevant
1	D1	no
2	D2	no
3	D3	yes
4	D4	no
5	D5	yes
6	D6	yes
7	D7	no
8	D8	no
9	D9	yes

(4.0 Marks)

3.b. Explain the crawler functionality with a neat diagram of crawler architecture.
(6.0 Marks)

3.c. What is kappa statistic? What does it measure? What do the different values of Kappa indicate?
(6.0 Marks)

3.d. How does Rocchio algorithm incorporate relevance feedback into the vector space model?
(4.0 Marks)

4.a. Discuss the goals of Recommender System (4.0 Marks)

4.b. Consider a matrix that shows ratings of Alice, Bob, Carol and Dave on different items. The ratings range from 1 to 5. The ? indicates that the user has not rated the item. Complete the user item rating matrix. Use weighted average of similarity to compute the not rated item. Use mean centered ratings for computation of Pearson correlation coefficient.

User/Item	Item1	Item2	Item3	Item4	Item5
Alice	5	4	1	4	?
Bob	3	1	2	3	3
Carol	4	3	4	3	5
Dave	3	3	1	5	4

(8.0 Marks)

4.c. Discuss the advantages of model-based recommender systems over neighborhood-based recommender systems. (3.0 Marks)

4.d. What are knowledge based recommender systems? Explain its types. (5.0 Marks)

5.a. Explain the architecture of Semantic Web (8.0 Marks)

5.b. What are the ontology classes based on the information it represents? (6.0 Marks)

5.c. Write down the syntax of RDF statement (3.0 Marks)

5.d. Define a class called ConferencePaper which is a sub class of Publication using owl construct. Make the class ConferencePaper disjoint from the other two classes namely JournalPaper and Book. (3.0 Marks)