Sr. No.311508

December 2022

BCA- V SEMESTER

Web Technology and Information Retrieval (GEC-4)

Time: 3 Hours **Max. Marks:75**

Instructions:
1. *It is compulsory to answer all the questions (1.5 marks each) of Part -A in short.*
2. *Answer any four questions from Part -B in detail.*
3. *Different sub-parts of a question are to be attempted adjacent to each other.*

PART -A

Q1 (a) State Zipf's Law (1.5)

(b) What do you mean by freshness of a page? (1.5)

(c) Differentiate between stemming and lemmatization (1.5)

(d) What do you mean by Secure Socket Layer? (1.5)

(e) List any five Web Servers (1.5)

(f) What is the task of DNS resolver? (1.5)

(g) What are Bi-word indices? (1.5)

(h) What is a web graph? (1.5)

(i) What do you mean by ranked retrieval system? (1.5)

(j) What is the need of index construction? (1.5)

PART -B

Q2 (a) Consider the postings list (4, 10, 11, 12, 15, 62, 63, 265, 268, 270, 400) with a corresponding list of gaps (4, 6, 1, 1, 3, 47, 1, 202, 3, 2, 130). Assume that the length of the postings list is stored separately, so the system knows when a postings list is complete. Using variable byte encoding: (10)
(i) What is the largest gap you can encode in 1 byte?
(ii) What is the largest gap you can encode in 2 bytes?
(iii) How many bytes will the above postings list require under this encoding?

(b) What do you mean by Information Retrieval? (5)

Q3 (a) Compute the vocabulary size M for this scenario: (5)
In a collection of web pages, there are 4500 different terms in the first 45,000 tokens and 50,000 different terms in the first 5,000,000 tokens.
Assume a search engine indexes a total of 70,000,000,000 pages, containing 300 tokens on average. What is the size of the vocabulary of the indexed collection as predicted by Heap's Law

(b) How SPIMI differs from BSBI? Discuss BSBI in detail. (10)

Q4 How do Search Engines work, technically? Discuss the general architecture of (15)

Search Engine with the detailed working of following components:
(a) Indexer Module
(b) Ranking Module

Q5 (a) Consider the following fragment of a positional index with the format: (10)
word: document: hposition, position, ...i; document: hposition, ...i ...
Gates: 1: h3i; 2: h6i; 3: h2,17i; 4: h1i;
IBM: 4: h3i; 7: h14i;
Microsoft: 1: h1i; 2: h1,21i; 3: h3i; 5: h16,22,51i;
The /k operator, word1 /k word2 finds occurrences of word1 within k words of word2 (on either side), where k is a positive integer argument. Thus k=1 demands that word1 be adjacent to word2.
(a) Describe the set of documents that satisfy the query Gates /2 Microsoft.
(b) Describe each set of values for k for which the query Gates /k Microsoft returns a different set of documents as the answer.

(b) List various limitations of term document incident matrix. (5)

Q6 (a) How distributed index is created? Explain in detail the working of map and reduce phase with suitable example. (10)
(b) Explain with suitable example how skip pointers can help in reducing the number of comparisons ? (5)

Q7 Write short notes on (any three): (15)
a) IIS
b) Incremental Web Crawlers
c) Dynamic Indexing
d) Posting List Compression

****