SVKM

Mithibai College (Arts, Sci & Comm)

Programme: B.Sc (Computer Science) - (CBCGS)

Year: III/Semester VI (Exam Year: 2023-2024)

Subject: INFORMATION RETRIEVAL

Date: 23 Mar 2024 **Time:** 10:30 am - 01:00 pm (02:30 Hrs.)

Max Marks: 75

FINAL EXAMINATION(2023-2024)

Instructions:

1. This question paper contains 3 pages
2. Answer to each new question to be started on a fresh page.
3. Figure in right hand side indicates full marks
4. Draw neat and well labelled diagrams wherever necessary.
5. Use of scientific calculator is permitted

1. (Attempt Any 3 Questions) **15**

5

A. 1. What is stemming and lemmatization? Describe the differences. How do the techniques affect retrieval?

B. What are permuterm index? Generate permuterm for "great". **5**

C. Create inverted index for following documents: **5**
D1: Tropical Freshwater Aquarium Fish
D2: Tropical Fish, Aquarium Care, Tank Setup
D3: Keeping Tropical Fish and Goldfish in Aquariums, and Fish Bowls.
D4: The Tropical Tank Homepage-Tropical Fish and Aquariums.

D. Explain the soundex algorithm for phonetic corrections. Interpret the results of soundex code for "Hilbert" and "Heilbronn". **5**

2. (Attempt Any 3 Questions) **15**

A. What is snippet? Consider the document with 500 sentences. Interpret that the word "activity" with frequency 35 is significant or not. **5**

B. Evaluate content based recommendation system with its advantages and limitations. **5**

C. Elaborate on variable length encoding as posting compression technique **5**

D. Justify the need of distributed indexing. How it is created? **5**

3. (Attempt Any 3 Questions) **15**

A. Discuss Probabilistic relevance feedback. **5**

B. Describe weighted zone index? Consider the query "Ramkrishnan" in a collection in which each document has three zones: author, title and body. Three weights g1=0.2 , g2=0.31 and g3=0.49, respectively corresponding to the title, conclusion and body zones. Find document score for the query if it appears in author and body. **5**

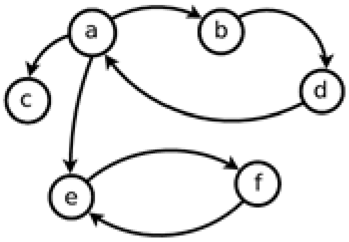
C. Given a document containing terms with the given frequencies A(3) ,B(2), C(1). Assume document collections 10,000 and document frequencies of these terms are A(50), B(1300), C(250). Compute TF-IDF. **5**

D. An IR system returns 1000 relevant documents, and 800 non-relevant documents. There are a total of 2000 relevant documents in the collection. Calculate precision, recall and F-measure. **5**

4. (Attempt Any 3 Questions) **15**

A. Differentiate between SEO and Paid placement **5**

B. Compute probability transition matrix for following graph assuming teleportation probability as 0.5 **5**



C. Evaluate the working of URL frontier for web crawler. **5**

D. Elaborate on CPC, CPM and click spam with respect to advertising on web. **5**

5. (Attempt Any 3 Questions) **15**

A. **Generate Levenshtein distance for 'Honda' and 'Hyundai'** **5**

B. Discuss the concept invisible web. **5**

C. What is parametric and zone index? Determine their need using suitable example. **5**

D. **Describe Hubs and Authorities.** **5**