THEORY ASSIGNMENT, JAN-2024
Information Retrieval (CSE 4053)

Programme: B. Tech (CSE) **Semester: 7th**
Full Marks: 20 **Submission Dt.- 10. 01. 2024**

Subject/Course Learning Outcome	***Taxonomy Level**	**Ques. Nos.**	**Marks**
Outline the concepts and apply the basics of indexing and querying of an information retrieval system	L3	1	2
Understand the data corpus used in information retrieval systems.	L3	2	2
Illustrate various components and experiment with different compression techniques to compress the index of dictionary and its postings lists.	L3	3,4	4
Apply retrieval models to construct information retrieval systems.	L5	5, 6, 7	6
Understand the methods to enhance the retrieval system through the use of techniques like relevance feedback and query expansion.	L3	8	2
Apply text clustering and classification techniques for information retrieval.	L5	9, 10	4

*Bloom's taxonomy levels: Knowledge (L1), Comprehension (L2), Application (L3), Analysis (L4), Evaluation (L5), Creation (L6)

Answer all questions. Each question carries equal mark.

- **Assignment scores/markings depend on neatness, clarity and date of submission.**
- **Write your answers with enough detail about your approach and concepts used, so that the grader will be able to understand it easily.**
- **You are allowed to use only those concepts which are covered in the lecture class till date.**

1. For a 2 word query, the postings lists are as mentioned below: [2, 5, 7, 13, 16, 18, 23, 32, 39, 56, 59, 63, 68, 76, 80] and [12, 56]. Show how many comparisons (step by step representation) would be done to find out the intersection of the above two postings lists using skip pointers with a skip length of $\sqrt{P}$.
2. For a given document stored in the data warehouse, compress the words by applying following preprocessing technique separately.
   i. Normalization

ii. Stemming (Use Porter stemmer)

iii. Stop words removal

> Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on full text or other content based indexing. Automated information retrieval systems are used to reduce what has been called "information overload". Many universities and public libraries use IR systems to provide access to books, journals and other documents. Web search engines are the most visible IR applications.

3. Edit distance can be used for spelling correction in search queries.

   (i) Define edit distance.

   (ii) As an example of how to calculate edit distance efficiently, show how dynamic programming can be used to calculate the edit distance (using Levenshtein matrix) between “system” and “item”.

4. Suppose the vocabulary for your inverted index consists of the following 6 terms:

   elite, elope, ellipse, eloquent, eligible, elongate

   Assume that the dictionary data structure used for this index stores the actual terms using Dictionary-as-a-string storage with front coding and a block size of 3. Show the resulting storage of the above vocabulary of 6 terms.

5. Consider a collection made of the 4 following documents (one document per line) :

   D1: John gives a book to Mary

   D2: John who reads a book loves Mary

   D3: who does John think Mary love?

   D4: John thinks a book is a good gift

   - These documents are pre-processed using a stop-list and a stemmer. The resulting index is built to allow by applying vector-based queries. Give a (graphical or textual) representation of this index.
   - We now focus on 3 terms belonging to the dictionary, namely book, love and Mary. Compute the tf-idf based vector representation for the 4 documents in the collection.

- Consider the query “**love Mary”**. Give the results of a ranked retrieval for this query. What document is (are) considered to be the most relevant?

6. Compute the Jaccard matching score and the tf matching score for the following query-document pairs.

   q: [information on cars] d: “all you’ve ever wanted to know about cars”

   q: [information on cars] d: “information on trucks, information on planes, information on trains”

   q: [red cars and red trucks] d: “cops stop red cars more often”

7. The figure below shows the output of an information retrieval system on two queries. Crosses correspond to the relevant documents, dashes to non-relevant documents. Let the two documents contain 3 and 6 relevant documents, respectively, but only those shown in the figure are retrieved by the system, not the others.

Rank	**Q1**	**Q2**
1	X	-
2	-	X
3	-	-
4	X	X
5	X	-
6	-	X
7	-	X
8	-	-
9	-	X
10	-	X

   (a) Draw the precision-recall curve

   (b) Compute the Mean Average Precision

   (c) Compute the R-precision

8. Suppose that a user’s initial query is cheap CDs cheap DVDs extremely cheap CDs. The user examines two documents, d1 and d2. She judges d1, with the content **CDs cheap software cheap CDs** relevant and d2 with content **cheap thrills DVDs** nonrelevant. Assume that we are using direct term frequency (with no scaling and no document frequency). There is no need to length- normalize vectors. Using Rocchio relevance feedback (as specified in SMART Algorithm) what would the revised query vector be after relevance feedback? Assume $\alpha = 1$, $\beta = 0.75$, $\gamma = 0.25$.

9. Classify whether a given person is a male or a female based on the measured features. The features include height, weight, and foot size. Data set is shown below

sex	**height (feet)**	**weight (lbs)**	**foot size**
male	6	180	12
male	5.92 (5'11")	190	11
male	5.58 (5'7")	170	12
male	5.92 (5'11")	165	10
female	5	100	6
female	5.5 (5'6")	150	8
female	5.42 (5'5")	130	7
female	5.75 (5'9")	150	9

Below is a sample to be classified using Naive Bayes algorithm as a male or female.

sex	**height**	**weight**	**foot size**
sample	6	130	8

10. Based on the data below, estimate a multinomial Naive Bayes classifier and apply the classifier to the test document. Calculate the probability that the classifier assigns the test document to *c* = China or not.

	docID	words in document	In *c*= China?
training set	1	Taipei Taiwan	yes
	2	Macao Taiwan Shanghai	yes
	3	Japan Sapporo	no
	4	Sapporo Osaka Taiwan	no
	5	London	no
test set	6	Taiwan Taiwan Sapporo	?