V4V4V4V4V4V4V4V4V4V4V4V4V4V4V4V4V4V4V4V4V4V4V4V4

National University of Computer & Emerging Sciences
FAST-Karachi Campus
CS4051- Information Retrieval
Quiz#1

Dated: February 15, 2023 Marks: 20
Time: 20 min.
Std-ID: _______Sol_________

Question No. 1

Enumerate a list of problems that you can come across when processing document collection for information retrieval. **[5]**

Document processing is a major step in all information retrieval systems. This activity encompasses many challenging problems like:

- Identifying the document format, language and encoding.
- Identifying the tokenization's strategies
- Identifying the index able features from the documents.
- Normalization and Performing language specific treatment for the tokens

Question No.2

Consider a Boolean query given below:

Q = T1 AND T2 AND T3 AND T4

If we know the following facts:

Frequency of T1 > Frequency of T3; and Frequency of T3 > Frequency of T4;

What will be an efficient query order in the form of query work? Justify your answer.

As per the given information we have T4<T3<T1 but no information about T2, from the set of possible execution the place of T2 is not confirm for the optimal order. The four possible ordering would be (i) T2 AND T4 AND T3 AND T1 (ii) T4 AND T2 AND T3 AND T1 (iii) T4 AND T3 AND T2 AND T1 and (iv) T4 AND T3 AND T1 AND T2

Question No.3

In an IR System there were 60 relevant documents for a given query “q”. The system returned 108 documents in response to the same query. If 50% documents in the result-set are relevant, compute the Precision and Recall of the system? **[5]**

we know,
precision = (relevant-retrieved) / (total-retrieved)
=> precision = (relevant-retrieved) / (result-set) ---------- eq(A)
recall = (relevant-retrieved)/ (total-relevant) ---------- eq(B)
we need to find total relevant documents in result-set,
(result-set) = 108 documents
(relevant documents in result-set) = 108 X .50 = 54= (relevant-retrieved)
From eq(A) …. **precision = 54/108 = 0.5**
From eq(B) …. **recall = 54/60 = 0.9**

Question No.4

Illustrate the differences between following pair of terms.

Dictionary	**Thesaurus**
A dictionary contains an alphabetical list of words that includes the meaning, etymology and pronunciation. Organization of words in dictionary in lexicographic order. Dictionary is used to see the meaning, type and pronunciation of word. Dictionary may show use of the word in a sentence.	A thesaurus is a book that contains relationships between words like: synonyms and antonyms. Organization of words in thesaurus in generally in thematic order (conceptual order). Thesaurus is used to see the similarity and differences between pair of words or groups. Thesaurus may show the right usage or different context or sense of words.
Inverted Index	**Positional Index**
- Inverted index is a data structure for processing queries in IR systems. - It has two main components dictionary and posting lists. It only keeps terms and document IDs in which these terms occurred. - It can be used to process general Boolean query and failed on specialized queries like: proximity and phrase queries.	- It is also a data structures to support queries in IR systems. - It keeps term, document IDs and the position information of each term/token. - It can be used to answer very special queries like proximity, and general phrase queries.