V1V1V1V1V1V1V1V1V1V1V1V1V1V1V1V1V1V1V1V1V1V1V1V1V1

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

Dated: February 13, 2023 Marks: 20
Time: 20 min.
Std-ID: _____Sol_____

Question No. 1

What do we mean by Extended Boolean Retrieval Model? Which drawbacks of Boolean IR model it removes? **[5]**

The goal of the Extended Boolean Model (EBM) is to overcome the drawbacks of the Boolean model that has been used in information retrieval. The two issues that EBM really solved are as below:
1. The Boolean model doesn't consider term weights in queries, and the result set of a Boolean query is often either too small or too big. It tries to incorporate the frequency of the term as a weight.
2. The Boolean model gives a flat results- all qualifying documents at the same level (rank or related). It gives higher rank to documents based on the frequency of query terms present in it.

Question No.2

Define the process of token normalization. Suggest what normal form should be used for these words(Tokens) – including the word itself as a possibility? **[5]**

Token normalization is the process of canonicalizing tokens so that matches occur despite superficial differences in the character sequences of the tokens. Normalization is helpful in reducing the number of unique tokens present in the text, removing the variations in a text. and also cleaning the text by removing redundant information.

Token	Normalized Feature
- *O'Rourke*	orourke
- *Anti-discriminatory*	antidiscriminatory
- *U. S. A.*	USA
- *Co-Education*	coeducation
- *cont'd*	contd

Question No.3

Suppose a query has a total of 4 relevant documents in the collection. System A and System B have each retrieved 10 documents, and the relevance status of the ranked lists is shown below:

System A: [- + - - - - - - - -]

System B: [+ + - - - - - - - -]

where the leftmost entry corresponds to the highest ranked document, and the rightmost entry corresponds to the lowest ranked document. A "+" indicates a relevant document and a "-" corresponds to a non-relevant one. For example, the top ranked document retrieved by System A is non-relevant, whereas the top ranked document retrieved by B is relevant. **[5]**

Answer the following queries:

Precision of System A	1/10
Precision of System B	2/10
Recall of System A	1/4
Recall of System B	2/4

Question No.4

Explain the following type of queries from an Information Retrieval prospective with an example. Also suggest suitable data structures that get the required answer without false positive. [5]

a. A bi-word query

Consider the query "labor policy" it is an example of bi-word query. The intent of the user is to get the documents that contains both the words "labor" and "policy" adjacent as given in the query should be in the documents. Positional Index can be used to answer this type of query without false positives in the result-set hence no post processing required.

b. A proximity query

Consider the query "labor policy /k" it is proximity query. The intent of the user is to get the documents that contains both the words "labor" and "policy" within k words apart in the documents. Positional Index can be used to answer this type of query without false positives in the result-set hence no post processing required.