1111111111111111111111111111111111111111111111111111111111

National University of Computer & Emerging Sciences
FAST-Karachi Campus
Information Retrieval (CS317)
Quiz#1

Dated: February 13, 2019 Marks: 20
Time: 20 min.
Std-ID: _____Sol_____

Question No. 1

What are some of the assumptions we made while developing Boolean Model for IR? Outline them. [5]

Boolean Model(BM) makes several assumptions regarding users, data and retrieval model.

User: BM assumes that user can visualize the documents and she can see features in it. Using these features she can come up with the Boolean queries.

Data: BM assumes that all the documents are available offline and they are static in nature. The crawler can (machine) read them and easily maintain index on the features present in the document.

Retrieval Model: BM can serve exact match queries.

Question No.2

In an IR System there were 60 relevant documents for a given query "q". The system returned 140 documents in response to the same query. If 40% 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) = 140 documents
(relevant documents in result-set) = 140 X .40 = 56= (relevant-retrieved)
From eq(A) …. **precision = 56/140 = 0.4**
From eq(B) …. **recall = 56/60 = 0.93**

Question No.3

Comments on the following statement as TRUE or FALSE with justification (1-2 line explanations). 5 X 2 marks each.

1. **Stemming is a slow process and increases the size of the lexicon(dictionary).**

   FALSE. Stemming is based on heuristic rules from linguistic experts and it is generally very fast to compute. Stemming maps derivational /inflectional morphemes to root word hence does not increase size of the dictionary.

2. **A general phrase query cannot be answered by bi-words index.**

   TRUE. The general phrase queries are if the form (w1 w2 w3 ..wn), the bi-words index cannot handle general form of phrase query. It may get false positive in the result set. Hence a post processing step is unavoidable.

3. **Extended bi-words index is not very useful.**

   FALSE. Extended bi-word index is specialized index in the sense to retrieval extended bi-words features for example: “coins are in pocket” is index as “coin pocket”, omitting the two middle terms “are in” This index is helpful in retrieval pair of terms that appears together not necessary as phrase/adjacent sequence of term.

4. **Positional Index is good for proximity query.**

   TRUE. Positional index maintains positional information of each term and it is very helpful in processing proximity based query. For example, (term1 term 2 /k), where intention is to retrieve documents that contains term1 and term 2 k words apart.

5. **Skip Pointers in posting list can improve the processing time for queries.**

   True. Posting list contains documents IDs of the documents that contains a term. For large dataset it is a long list of documents IDs. While processing queries (conjunctive queries) using posting list we need to process intersection or merging of two posting lists. Skip list as a posting list may guide the skipping of document IDs without missing a valid document. Hence it is faster than usual list.