2222222222222222222222222222222222222222222222222222222222222

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

Dated: February 21, 2019 Marks: 20
Time: 20 min.
Std-ID: _____Sol_________

Question No. 1

The following assumptions were made while developing Boolean Model for Information retrieval.

1. **The model assume that users know the features from the document.**
2. **The documents are available in machine readable format.**

Discuss how effective these assumptions are and what are their drawbacks?

First assumption that the users know the features of the document is not at all practical as user may not have any slight idea of what is in the documents. The second assumption that documents are available in machine readable format is far from simplifications some human languages poses many challenges in processing the digital documents for indexing the Boolean features.

Question No.2

What will be the best query processing order in the following cases: if the collection has cumulative frequency as below:

Term	**CF**	**Term**	**CF**
T1	**12389**	**T3**	**231**
T2	**231**	**T4**	**5166**

i. **T1 AND T4 AND T2**

The term frequencies for T1, T2 and T4 are 12389,231 and 5166 respectively. The most optimal order for Boolean query would be T2 AND T4 AND T1.

ii. **T2 AND T3 OR T2 AND T4**

The optimal order will be T2 AND (T3 OR T4).

Question No.3

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

1. **Lemmatization produce human readable features.**

   TRUE. In lemmatization an external lexicon/ thesaurus/dictionary is used and every token from the document is retrain for indexing if it is a dictionary word or lexeme or its variation. Hence the tokens are human readable and understandable, in fact these are lemmas from dictionary.

2. **A general phrase query can easily be answered with positional index.**

   TRUE. A general phrase query is of the form "w1 w2 …wn". A positional index can easily be used to fetch documents that contains these words from the query in the consecutive positions.

3. **Stemming increases, the size of the dictionary.**

   FALSE. Stemming maps derivational /inflectional morphemes to root word suing heuristic based algorithms hence does not increase size of the dictionary. It is in fact reduce the size of lexicon.

4. **Trailing wildcard queries can be answered with permuterm index.**

   TRUE. Trailing wildcard queries are of the form (abc*), a permuterm index maintains all possible rotations of a word with a secondary level access to the term in the dictionary. Hence we will be looking for *$abc in permuterm index to get all the documents that contains word start from abc.

5. **K-gram index cannot be used for proximity query.**

   TRUE. The proximity query is of the form (w1 w2 /k), in which the intention of the user is to get the documents in which w1 and w2 appears k words apart. The k-gram index contains all possible k-grams of the words (dictionary terms). It is effective for supporting wildcard queries but not good for proximity queries.

333333333333333333333333333333333333333333333333333333333

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

Dated: February 21, 2019 Marks: 20
Time: 20 min.

Question No. 1

What are some of the drawbacks of Boolean Retrieval Model?

There are several drawbacks of Boolean Model for IR.

From Users prospective: Users need training on query formulations, they need to understand Boolean queries. They need to have some clear idea about what features are there in the relevant documents. From System's prospective: The IR systems based on Boolean model considers all terms with same importance and independent of each other. It is based on exact matching and result-set is flat (that is all documents are equally ranked).

Question No.2

Compare and Contrast the following pair of terms

Stemming	Lemmatization
It is a heuristic- rule based approach, generally fast and use a single term. It generates unreadable tokens. Stemming algorithms err on the side of being too aggressive, sacrificing precision in order to increase recall.	It is a rigor process that uses a dictionary and uses context to determine the lemma, considered a slow approach. It generates readable lexeme from the dictionary. Lemmatization offers better precision than stemming, but at the expense of recall.
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.

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 reduce the size of lexicon(dictionary).**

   TRUE. The stemming process reduce different lexemes to a common representation based on stemming algorithm. Hence, stemming actually decreases the size of the vocabulary.

2. **Post processing of query required in some cases.**

   TRUE. In some cases, when the result-set of the retrieval contains false positive. The post processing step is unavoidable. The processing required for such cases is directly proportional to number of candidate documents in the result set.

3. **General wild card queries can be answered with bi-word index.**

   TRUE. A general wildcard query is of the form (abc*cd), in which the intention of the user is to get the documents in which starts with abc and ends with cd. The bi-word index maintains index information for each pair of bi-word (consecutive words that appears in the documents). This index is not helpful in any regards to general wildcard queries.

4. **Term independence is a feature of Boolean IR.**

   FALSE. In Boolean IR term independence assumption is not valid. In human language, the selection of words is very much dependent on the essence of subject for communication. It is in fact a limitation of the Boolean Model for IR.

5. **k-gram index cannot be used for proximity query.**

   TRUE. The proximity query is of the form (w1 w2 /k), in which the intention of the user is to get the documents in which w1 and w2 appears k words apart. The k-gram index contains all possible k-grams of the words. It is effective for supporting wildcard queries but not good for proximity queries.