<1111111111111111111111111111111111111111111111111111111111111111111111>

CS317- Information Retrieval
Quiz # 1

Dated: February 26, 2021   **Student ID: _____Sol_____**
Marks: 20   **Time: 15 min.**

Q1: What are the basic assumptions for Boolean Retrieval Model? [5]

There are three important assumptions in developing BM for IR, these are:
1. All the documents are machine readable.
2. The users know what features are presents in the documents.
3. Users know how to formulate queries using observed features.

Q2: Indicate True/False with a brief justification. [5]

    a. Lemmatization increases vocabulary size.

       False- Lemmatization mapped different morphological words to a root word. Hence it in fact decreases the vocabulary.

    b. Skip pointers in posting lists can reduce intersection cost of the queries.

       True. Skip pointers facilitate the fast navigation of intersect algorithm using skip-pointers, hence the time is reduced.

    c. Phrase queries can be easily supported by a positional index?

       True. The general phrase queries can easily be answered using position index. It contains position information about a word-feature in a document.

Q3: There are 16 relevant documents in a collection for a given query "q". The precision of the query is 0.40 and recall of the query is 0.25. How many documents are in the results-set (number of documents returned by the system against the query)? [10]

we know,

    precision = (relevant-retrieved) / (total-retrieved)

    => 0.4 = (relevant-retrieved) / (result-set)

    => (result-set) = (relevant-retrieved) /0.4 ---------- eq(A) similarly,

    recall = (relevant-retrieved)/ (total-relevant)

    => 0.25 = (relevant-retrieved)/ 16

    => relevant-retrieved= 0.25 * 16 = 4 hence

    eq(A) => result-set = 4 /0. 4 = 10

    Result-set has 10 documents.

<2222222222222222222222222222222222222222222222222222222222222222222222222>

CS317- Information Retrieval
Quiz # 1

Dated: February 26, 2021   **Student ID: ____Sol____**
Marks: 20   **Time: 15 min.**

Q1: What are the limitations of Boolean Retrieval Model? [5]

- Exact matching is used hence no support for partial matching.
- All the results from the retrieval are at the same level, no ranking of the results.
- All features are equally weighted – that is all terms have same importance.
- Users has to formulate a Boolean query to get its required documents.

Q2: Indicate True/False with a brief justification. [5]

a. Lemmatization produces human readable terms.

True. Lemmatization transform each morphological word to its root (lexeme) hence a human readable word is used from the dictionary.

b. Skip pointers are not useful for queries of the form x OR y.

In queries of the form (x OR y), it is essential to visit every entry of posting lists of two terms x and y. Hence skip pointers are not any helpful in speeding up the process.

c. Proximity queries can be easily supported by a positional index?

Proximity queries are of the form x y /k, we need to find documents contains both x and y, k words apart. Positional index is very helpful in determining the k for these two terms.

Q3: We have a two-word query. For one term the postings list consists of the following 16 entries: [4,6,10,12,14,16,18,20,22,32,47,81,120,122,157,180] and for the other it is the one entry postings list: [47]. Work out how many comparisons would be done to intersect the two postings lists with the following two strategies. Briefly justify your answers:

a. Using standard postings lists
In standard merge algorithm, the processing stopped as soon either list finish. In our cases there will be 11 comparisons.

b. Using postings lists stored with skip pointers, with a skip length of $\sqrt{P}$, as suggested in the textbook.
The skip pointers are available in larger posting list with a skip value 4, hence there will be 6 comparisons in deciding the results set. These are (4,47), (14,47), (22,47), (120,47), (81,47), and (47,47)