CSCI 6370 IR and Web Search ASSIGNMENT 2 Due is 06/15/2020 23:59

Ulvi Bajarani
Student ID 20539914
E-mail: ulvi.bajarani01@utrgv.edu

Questions and Answers:

Problem 1. This assignment is designed for you to get familiar with various evaluation mechanisms in IR. Given the set of data in the table,

1. **Compute the recall/precision rate at each relevant document point.**
2. **Compute F-Measure at each relevant document point.**
3. **Compute the corresponding E-measure with $\beta = 0.8$ at each relevant document point.**
4. **Compute the R-Precision.**
5. **Computer the average precision, which is the average of the precision values at the points at which each relevant document is retrieved.**

The total relevant documents in a document set is 5. The top 10 retrieved documents are listed with the ones that are relevant marked.

Order	**Doc #**	**Relevant**
1	586	X
2	357	
3	358	X
4	108	X
5	345	
6	114	
7	555	X
8	888	
9	860	
10	167	X

Answer 1.

1. The answer is provided in the tables.
2. The answer is provided in the tables.
3. The answer is provided in the tables.
4. R-Precision $= P(5) = 0.6$.
5. MAP = 0.816706349206349

Order	**Doc #**	**Relevant**	Number of found documents	**Recall**	**Precision**
1	586	X	1	0.2	1
2	357		1		
3	358	X	2	0.4	0.666666666666667
4	108	X	3	0.6	0.75
5	345		3		
6	114		3		
7	555	X	4	0.8	0.571428571428571
8	888		4		
9	860		4		
10	167	X	5	1	0.5

Order	**Doc #**	**Relevant**	Number of found documents	**F-Measure**	**E-measure (b=0.8)**
1	586	X	1	0.333333333333333	0.609523809523809
2	357		1		
3	358	X	2	0.5	0.470967741935484
4	108	X	3	0.666666666666667	0.316666666666667
5	345		3		
6	114		3		
7	555	X	4	0.666666666666667	0.356862745098039
8	888		4		
9	860		4		
10	167	X	5	0.666666666666667	0.378787878787879

Problem 2. **Suppose we use the following the method to reformulate the query vector in response to relevance feedback:**

$$q_m = q_0 + \sum_{d \text{ is relevant}} d - \sum_{d' \text{ is irrelevant}} d'$$

Consider the initial query vector is $q_0 = \{1, 2, 0, 4, 0, 1\}$**. The relevant feedback gives two relevant vectors** d_1 **and** d_2 **and one irrelevant vector** d_3 **as follows:**

$$d_1 = \{0, 1, 1, 2, 0, 2\} \; d_2 = \{5, 0, 2, 0, 2, 0\} \; d_3 = \{4, 2, 1, 2, 1, 3\}$$

Calculate the reformulated query vector q_1

.

Answer 2.

$q_m = \{1+(0+5)-4; 2+(1+0)-2; 0+(1+2)-1, 4+(2+0)-2, 0+(0+2)-1, 1+(2+0)-3\}$

$q_m = \{2; 1; 2; 4; 1; 0\}$

Problem 3. **Phrasal query is to retrieve documents with a specific phrase (ordered list of contiguous words). For example, phrasal query for Q="computer learning theory" needs to retrieve all documents containing the phrase "computer learning theory". We usually ignore cases of the letters in the query. We may also allow intervening stop words and/or stemming. Describe an algorithm to do phrasal query with help of an inverted index.**

Answer 3.

```
Find set of documents $D$ in which all keywords ($k_1 ... k_m$) in phrase occur (using AND query processing).
Intitialize empty set, $R$, of retrieved documents.
For each document, $d$, in $D$:
  Get array, $P_i$ ,of positions of occurrences for each $k_i$ in $d$
  Find shortest array $P_s$ of the $P_i$'s
  For each position $p$ of keyword $k_s$ in $P_s$
    For each keyword $k_i$ except $k_s$
    Use binary search to find a position (p – s + i) in the array $P_i$
  If correct position for every keyword found, add $d$ to $R$
Return $R$
```