CSE 6240
HW 1
Ke Wang
903058889

1. Positional Indexing

a. "fools rush in"

The phrase appears in document 2 (position 1 to 3), 4 (8 to 10), 7 (3 to 5).

b. "where angels fear"

The phrase appears in document 4 (11 to 13, 431 to 433), 7 (16 to 18).

c. "fools rush in" AND "angels fear to tread"

The phrase appears in document 4.

2. TF-IDF

a. We only consider the terms that appear in at least one document (if the term doesn't appear in any document, then TF==0, and we don't need to consider IDF), so *df* is at least one, and it is obviously an integer. The total number of documents is fixed (N), so $\frac{N}{df_i}$ is at most N, so idf is at most log(N). So it is always finite.

b. $idfOccurInEveryDocument = \log\left(\frac{N}{df}\right) = \log\left(\frac{N}{N}\right) = 0$. So the term that occurs in every document has the tf-idf value of 0. This is consistent with the intuition because the term occurring in every document is very likely to be a very common work like "the". Thus, by using the TF-IDF, we can filter out the common terms. This is like the use of stop word lists, but it is much better. First, we do not need to figure out the stop word lists, which can be time-consuming and buggy, in advance. Besides, we can dynamically determine the common terms.

c.

Tf-idf	Doc1	Doc2	Doc3
Car	44.48	6.59	39.54
Moto	6.24	68.61	0
Insurance	0	53.54	47.05
rent	21.07	0	25.58

d. Yes. Because we can construct a situation where TF>=1 and IDF>1. Say there exists 100 documents, and the term 'corgi' appears in 1 document, and its frequency is 100. Then tf-idf = tf*idf = 100 * log(100/1) = 200 > 1.

3. Evaluation

precision_of_S1_on_Q1 = 2/5
recall_of_S1_on_Q1 = 1/2
precision_of_S1_on_Q2 = 2/5
recall_of_S1_on_Q2 = 2/3

precision_of_S2_on_Q1 = 3/5
recall_of_S2_on_Q1 = 3/4
precision_of_S2_on_Q2 = 3/5
recall_of_S2_on_Q2 = 1

F_S1_Q1 = 4/9
F_S1_Q2 = 1/2
F_S2_Q1 = 2/3
F_S2_Q2 = 3/4

AP_S1_Q1 = 1/4(1+2/3) = 5/12
AP_S1_Q2 = 1/3(1+2/4) = 1/2
AP_S2_Q1 = 1/4(1+1+3/5) = 13/20
AP_S2_Q2 = 1/3(1+2/3+3/4) = 29/36

MAP_S1 = 1/2(5/12+1/2) = 11/24
MAP_S2 = 1/2(13/20+29/36) = 131/180

4. Evaluation

a. We would like to determine the mean value by paying more attention to the bad performance of one of 'precision' and 'recall'. The average mean is closer to the larger value of the two, so it's not a very good idea. Whereas, Harmonic mean is closer to the smaller value of the two, which is preferable. In the extreme case, when we just retrieve all the documents, 'recall' is equals to 1 and the average mean is larger than 0.5. The value is high but we are not expecting that.

b. It can be any value between 0 and 1 according to the definition.

c. Claim: a break-even point must exist.
Proof: Precision = tp/(tp+fp), Recall = tp/(tp+fn).
if tp==0 (this can happen when the first document retrieved is not relevant), then Precision==Recall, the break-even point exists.
If tp!=0, then we need to prove that fp==fn at some point. fp equals to 0 in the beginning, and fn>0 in the beginning (otherwise tp will equal to 0). So fn>fp at the beginning. fp monotonously increases as more documents are retrieved and fn monotonously decreases as more documents are retrieved. At the end, when all documents are retrieved, fn==0 and fp>=0, so fp>=fn. Thus, we can come to a conclusion that there always exists a break-even point.