CSIT 5930 Search Engines and Applications

Fall 2021 Homework 1

Name: PENG, Zifan
Student ID: 20784377

1. Question 1

a) Precision is an important measure of retrieval effectiveness, it is meaningless unless compared to the level of Recall desired by the user. So, we need to ensure enough Recall Rate, in this case, the query need to be revised substantially, and in the meanwhile, the query may be vague so that the Recall rate is low.

So, in this paper, the lawyers who were to use the system for litigation support stipulated that they must be able to retrieve at least 75 percent of all the documents relevant to a given request for information

b) Question b

i. TRUE.

Although there is some difference between the results for each lawyer, the variance is not statistically significant at the .05 level. Although this was a very limited test, we can conclude that at least for this experiment the results were independent of the particular user involved.

ii. FALSE.

Unfortunately, the values for Recall and Precision for the substantially revised queries (23.9 percent and 62.1 percent, respectively) did not indicate a statistically significant difference.

iii. FALSE.

Although there is a marked improvement in the lawyer's Recall for requests 3, 4, and 5, and in the average Recall for all five information requests, the improvement is not statistically significant at the .05 level (z = -0.81). Hence, we cannot reject the hypothesis that both the lawyer and the paralegal get the same results for Recall.

c) Question c

Generate subset randomly, which contains relevant and irrelevant documents. And then by estimating the relevant rate in the subsets can estimate the whole database relevant documents number. And then we can compute the recall rate by these two numbers.

d) Question d

Full-text retrieval is difficult to use to retrieve documents by subject because we assume that users to foresee the exact words and phrases that will be used in the documents they will find useful. Because the effectiveness of full-text retrieval has not been substantiated by reliable Recall measures on realistically large databases. Some information requests placed almost impossible demands on the ingenuity of the individual constructing the query. And the most important terms and phrases w ere not used at all in relevant documents.

Example : “Company A’s comments concerning “steel quantity”.

2. Question 2

a) $idf_j = \log_2 (N/ df_j)$, and we have total of 10 documents, so the “N” is 10, we can easily compute this form as follows:

TF	DF	IDF
2	1	$\log_2 10$
0	2	$\log_2 5$
1	3	$\log_2(10/3)$
5	2	$\log_2 5$
2	10	$\log_2 1$

w_{ij} = tf/tf_{max} * idf = tf/tfmax * $\log_2 (N/ df_j)$, tf_{max} = 5, so the weights is as follows:

Weight = D = < 0.4*$\log_2 10$, 0, 0.2*$\log_2(10/3)$, $\log_2 5$, 0 >
= < 1.33, 0, 0.35, 2.32, 0 >

b) Query Vector, Q = < 1, 0, 0, 1, 0 >, and D = < 1.33, 0, 0.35, 2.32, 0 >This is inner product of D & Q $\mathrm{sim}(D_i, Q) = \sum_{k=1}^{t} d_{ik}\, q_k$, and the cosine similarity is $\frac{\mathbf{D} \cdot \mathbf{Q}}{|\mathbf{D}|\,|\mathbf{Q}|}$

So, D · Q = inner product = 0.4*$\log_2 10$ + $\log_2 5$ = 3.65,
| D | × | Q | = 2.69771 * 1.41 = 3.815
CosSim(D, Q) = 18.253497 ÷ 19.0739876698 = 0.956898 = 0.96

Inner product between Q and D is 3.65
Magnitudes of Q is 1.414
Magnitudes of D is 2.698
Cosine similarity values between Q and D is 0.96