CSIT 5930 Search Engines and Applications
Fall 2021 Homework 1

Due: See Canvas

Submission method: Submit your homework to Canvas. Submit a PDF file containing the written answer.

1. This question is about the case study conducted by Blair and Maron, which has been covered in the lecture. The paper can be downloaded from Canvas Assignment homepage.

   Read this paper to answer the questions.

   a. **[10 points]** In the experiment, the lawyers and paralegals are allowed to revise a query as many times as they want. What might be the reason for this experimental design instead of evaluating the performance for each submitted query?

   b. **[30 points]** State which of the following statements are true or false according to the paper and quote the text that support your answer.

      (i) The precision and recall obtained by the two lawyers are consistent with each other (i.e., the figures do not lead to contradictory conclusions).

      (ii) After the lawyers and paralegals have gone through rounds to formulate an information request into STAIRS' query, the precision and recall of the reformulated information request are significantly improved.

      (iii) If the lawyers are to write their own queries directly on STAIRS (not using the paralegals), the experiment found that the lawyers can get significantly better precision and recall than the paralegals.

   c. **[20 points]** Recall is time consuming to evaluate, because in principle the relevance of ALL of the documents needs to be judged against each query. Blair and Maron's experiment did not examine all the documents. Using your own words, describe how they chose subsets of the overall document collection for calculating recall?

   d. **[20 points]** In the paper, the authors gave a reason for the low recall of the experiment and some examples drawn from the document collection to illustrate the difficulty of using keywords to retrieve documents. Give a brief summary in, say, 2-3 sentences, to describe the reason and one example given by the authors. Do not copy the whole paragraph(s) from the paper.

2. **[20 points]** Suppose there are only 5 unique terms (numbered 1 to 5) in the collection, which contains a total of 10 documents. These five term's term frequencies in a document *D* and their document frequencies are given below:

TF	DF	IDF
$tf_{D,t1} = 2$	$df_{t1} = 1$	
$tf_{D,t2} = 0$	$df_{t2} = 2$	
$tf_{D,t3} = 1$	$df_{t3} = 3$	
$tf_{D,t4} = 5$	$df_{t4} = 2$	
$tf_{D,t5} = 2$	$df_{t5} = 10$	

(a) [5] Write down the document vector when tf/tf_{max} * idf weighting is used.

(b) [15] Given the query vector, *Q* = ⟨ 1, 0, 0, 1, 0 ⟩, compute the cosine similarity values between *Q* and *D* by first writing down the inner product and the magnitudes of Q and D.