Due: 2023/3/24 12:00

Homework 5

Student Number:
Name:

Problem 1. (25 points) Sketch the frequency-ordered postings for the data in Table 1.

	Doc1	Doc1	Doc3
car	27	4	24
auto	3	33	0
insurance	0	33	29
best	14	0	17

Table 1: tf values for documents.

Problem 2. (25 points) Let the static quality scores for Doc1, Doc2 and Doc3 in Table 2 be respectively 0.25, 0.5 and 1. Sketch the postings for impact ordering when each postings list is ordered by the sum of the static quality score and the Euclidean normalized tf values in Table 2.

	Doc1	Doc2	Doc3
car	0.88	0.09	0.58
auto	0.10	0.71	0
insurance	0	0.71	0.70
best	0.46	0	0.41

Table 2: Euclidean normalized tf values for documents.

Problem 3. (25 points) Explain how the common global ordering by $g(d)$ values in all high and low lists helps make the score computation efficient.

Problem 4. (25 points) When discussing champion lists, we simply used the r documents with the largest tf values to create the champion list for t. But when considering global champion lists, we used idf as well, identifying documents with the largest values of $g(d) + tf - idf_{t,d}$. Why do we differentiate between these two cases?