DATA20021 Information Retrieval (Spring 2022)
Exercise Set 3. Submit solutions by Friday 11 February.

Solutions are to be submitted via Moodle.

1. Describe the advantages and disadvantges of the Boolean retrieval model vs. the ranked retrieval model.

2. What is the idf of a term that occurs in every document? Compare this with the use of stop words.

3. Consider the table below of term frquencies for three documents from a collection of 806791 documents. First comput the idf values for each term and then compute the tf.idf weights for the terms for each document using the document frequency values (df_t column) in the table.

Table 1: Term frquencies and document frequencies.

term	Doc 1	Doc 2	Doc 3	df_t
car	27	4	24	18165
auto	3	33	0	6723
insurance	0	33	29	19241
best	14	0	17	25235

4. How does the base of the logarithm in the calculation of inverse document frequency affect the tfi.idf score? How does the base of the logarithm affect the relative scores of two documents on a given query?

5. Compute the Euclidean normalized document vectors for each of the three documents above, where each vector has four components, one for each of the four terms.

6. With the term weights as computed in the previous question, rank the three documents above by computed score for the query car insurance, for each of the following cases of term weighting in the query:

   (a) The wieght of a term is 1 if present in the query and 0 otherwise.

   (b) Euclidean normalized idf.