Information Retrieval

Exercises week 4

Exercise 1

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

Exercise 2

How does the base of the logarithm in the idf formula below affect the score calculation? How does the base of the logarithm affect the relative scores of two documents on a given query?

$$\text{idf}_t = \log \frac{N}{\text{df}_t} \qquad \text{Score}(q,d) = \sum_{t \in q} \text{tf-idf}_{t,d}$$

Exercise 3

If we were to stem jealous and jealousy to a common stem before setting up the vector space, detail how the definitions of tf and idf should be modified.

Exercise 4

Calculate the tf-idf weights for the terms car, auto, insurance, best, for each document, using the idf and tf values from the figures below. Using these term weights, rank the three documents by computed score for the query "car insurance", for each of the following cases of term weighting in the query:

i. The weight of a term is 1 if present in the query, 0 otherwise.

ii. Euclidean normalized idf.

term	df_t	idf_t
car	18,165	1.65
auto	6723	2.08
insurance	19,241	1.62
best	25,235	1.5

	Doc1	Doc2	Doc3
car	27	4	24
auto	3	33	0
insurance	0	33	29
best	14	0	17

Exercise 5

- Compute the vector space similarity between the query "digital cameras" and the document "digital cameras and video cameras" by filling out the empty columns in the table below. Assume N = 10,000,000, logarithmic term weighting (wf columns) for query and document, idf weighting for the query only and cosine normalization for the document only. Treat and as a stop word. Enter term counts in the tf columns. What is the final similarity score?

	query					document			
word	tf	wf	df	idf	q_i = wf-idf	tf	wf	d_i = normalized wf	$q_i \cdot d_i$
digital			10,000						
video			100,000						
cameras			50,000						