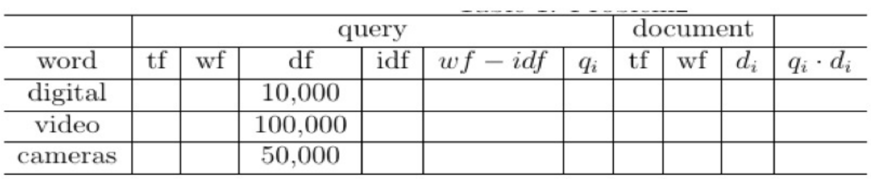
秦格华-5140219335
01/06/2017

Homework 6

Exercise 1 Compute the vector space similarity between the query “digital cameras”and the document “digital cameras video cameras”by filling out the empty columns in following table. Assume N = 10, 000, 000, logarithmic term weighting(wf columns) for query and document, idf weighting for the query only. Enter term counts in the tf columns, What is the final similarity score?



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

Solution.

	query						document			
word	tf	wf	df	idf	$wf-idf$	q_i	tf	wf	d_i	$q_i \cdot d_i$
digital	1	1	10,000	3	3	0.7936	1	1	0.5206	0.4131
video	0	0	100,000	2	0	0	1	1	0.5206	0
cameras	1	1	50,000	2.30	2.30	0.6084	2	1.3	0.6768	0.4118

Exercise 3 Compute the top scoring documents on the query best car insurance for each of the following weighing schemes:

- nnn.atc (nnn for documents, atc for query)
- ntc.atc (ntc for documents, atc for query)

Solution.
For term frequency, we use augumented tf as $tf_{t,d} = 0.5 + \frac{0.5tf_{t,d}}{\max_t(tf_{t,d})}$
According to the question’s setting, we have the nnn weights for documents:
So we have:

$$Score(q, doc1) = 0.56 * 27 + 0.353 * 3 + 0 + 0.51 * 14 = 23.32$$

	Query(atc weight)			
Term	tf	idf	$tf-idf$	atc weight
car	1	1.64	1.65	0.56
auto	0.5	2.08	1.04	0.353
insurance	1	1.62	1.62	0.55
best	1	1.5	1.50	0.51

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

$$Score(q, doc2) = 0.56 * 4 + 0.353 * 33 + 0.55 * 33 + 0 = 32.04$$

$$Score(q, doc3) = 0.56 * 24 + 0 + 0.353 * 29 + 0.51 * 17 = 38.06$$

Doc3>Doc2>Doc1.

For ntc.atc:

ntc weight for Doc1, Doc2 and Doc3:

	Doc1				Doc2				Doc3			
Term	tf	idf	$tf-idf$	normalized weight	tf	idf	$tf-idf$	normalized weight	tf	idf	$tf-idf$	normalized weight
car	27	1.65	44.55	0.897	4	1.65	6.6	0.075	24	1.65	39.6	0.595
auto	3	1.08	6.24	0.125	33	2.08	68.64	0.786	0	1.08	0	0
insurance	0	1.62	0	0	33	1.62	53.46	0.613	29	1.62	46.98	0.706
best	14	1.50	21	0.423	0	1.50	0	0	117	1.50	25.5	0.383

So we have:

$$Score(q, doc1) = 0.56 * 0.897 + 0.353 * 0.125 + 0 + 0.51 * 0.423 = 0.762$$

$$Score(q, doc2) = 0.56 * 0.075 + 0.353 * 0.786 + 0.55 * 0.613 + 0 = 0.657$$

$$Score(q, doc3) = 0.56 * 0.595 + 0 + 0.55 * 0.706 + 0.51 * 0.383 = 0.916$$

Doc3>Doc1>Doc2.