Exercise 6.10

Consider the table of term frequencies for 3 documents denoted Doc1, Doc2, Doc3 in Figure 6.9. Compute the tf-idf weights for the terms car, auto, insurance, best, for eachdocument, using the idf values from Figure 6.8.

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

▶ Figure 6.9 Table of tf values for Exercise 6.10.

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

▶ Figure 6.8 Example of idf values. Here we give the idf's of terms with various frequencies in the Reuters collection of 806,791 documents.

Solution

	Doc1	Doc2	Doc3
car	44.55	6.6	39.6
Auto	6.24	68.64	0
Insurance	0	53.46	46.98

Best	21	0	25.5

Exercise 6.12
How does the base of the logarithm in (6.7) affect the score calculation in (6.9)? How does the base of the logarithm affect the relative scores of two documents on a given query?

Solution

For any base b > 0, $\text{idf(b)} = \log_b(\frac{N}{df}) = \frac{\log_{10}(\frac{N}{df})}{\log_{10}(b)}$,

$\frac{\text{idf(b)}}{\text{idf(10)}} = \frac{1}{\log_{10}(b)}$is a constant.

So changing the base affects the score by a factor $\frac{1}{\log_{10}(b)}$, and the relative scores of two documents on a given query are not affected.

Exercise 6.15
Recall the tf-idf weights computed in Exercise 6.10. Compute the Euclidean normalized document vectors for each of the documents, where each vector has four components, one for each of the four terms.

Solution
doc1 = [0.8974, 0.1257, 0, 0.4230]
doc2 = [0.0756, 0.7867, 0.6127, 0]
doc3 = [0.5953, 0, 0.7062, 0.3833]

Exercise 6.17
With term weights as computed in Exercise 6.15, rank the three documents by computed score for the query car insurance, for each of the following cases of term weighting in the query:
1. The weight of a term is 1 if present in the query, 0 otherwise.
2. Euclidean normalized idf.

Solution
1. q = [1, 0, 1, 0]
score(q, doc1)= 0.8974, score(q, doc2) = 0.6883, score(q, doc3) = 1.3015
Ranking: doc3, doc1, doc2
2. q = [0.4778, 0.6024, 0.4692, 0.4344]
score(q, doc1) = 0.6883, score(q, doc2) = 0.7975, score(q, doc3) = 0.7823
Ranking: doc2, doc3, doc1

Exercise 6.19
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 Table 6.1. 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?

Solution

Word	Query					document			qi*di
	tf	wf	df	idf	qi=wf-idf	tf	wf	di=normalized wf	
digital	1	1	10,000	3	3	1	1	0.52	1.56
video	0	0	100,000	2	0	1	1	0.52	0
Cameras	1	1	50,000	2.3	2.3	2	1.3	0.68	1.56

Similarity score = 1.56+1.56 = 3.12

Exercise 6.20

Show that for the query **affection**, the relative ordering of the scores of the three documents in Figure 6.13 is the reverse of the ordering of the scores for the query **jealous gossip**.

Solution:

For the query **affection**, score(q, SaS) = 0.996, score(q, PaP) = 0.993, score(q, WH) = 0.847, so the order is SaS, PaP, WH.

For the query **jealous gossip**, score(q, SaS) = 0.104, score(q, PaP) = 0.12, score(q, WH) = 0.72, so the order is WH, PaP,SaS.

So the latter case is the reverse order of the former case.

Exercise 6.23

Refer to the tf and idf values for four terms and three documents in Exercise 6.10. Compute the two top scoring documents on the query **best car insurance**for each of the following weighing schemes: (i) `nnn.atc`; (ii) `ntc.atc`.

Solution

(i) nnn.atc

nnn weights for documents

Term	Doc1	Doc2	Doc3
Car	27	4	24
Auto	3	33	0
Insurance	0	33	29
Best	4	0	17

	Query				Product		
Term	tf(augmented)	idf	tf-idf	atc weight	Doc1	Doc2	Doc3

Car	1	1.65	1.65	0.56	15.12	2.24	13.44
Auto	0.5	2.08	1.04	0.353	1.06	11.65	0
Insurance	1	1.62	1.62	0.55	0	18.15	15.95
Best	1	1.5	1.5	0.51	7.14	0	8.67

Score(q, doc1) = 15.12 + 1.06 +0 + 7.14 = 23.32, score(q, doc2) = 2.24 + 11.65 + 18.15 + 0 = 32.04, score(q, doc3) = 13.44 + 0 + 15.95 + 8.67 = 38.06
Ranking: doc3, doc2, doc1

(ii) ntc.atc
ntc weight for doc1

Term	tf(augmented)	Idf	tf-idf	Normalized weights
Car	27	1.65	44.55	0.897
Auto	3	2.08	6.24	0.125
Insurance	0	1.62	0	0
Best	14	1.5	21	0.423

ntc weight for doc2

Term	tf(augmented)	Idf	tf-idf	Normalized weights
Car	4	1.65	6.6	0.075
Auto	33	2.08	68.64	0.786
Insurance	33	1.62	53.46	0.613
Best	0	1.5	0	0

ntc weight for doc3

Term	tf(augmented)	idf	tf-idf	Normalized weights
Car	24	1.65	39.6	0.595
Auto	0	2.08	0	0
Insurance	29	1.62	46.98	0.706
Best	117	1.5	25.5	0.383

	query				Product		
Term	tf(augmented)	idf	tf-idf	atc weight	Doc1	Doc2	Doc3
Car	1	1.65	1.65	0.56	0.502	0.042	0.33
Auto	0.5	2.08	1.04	0.353	0.044	0.277	0
Insurance	1	1.62	1.62	0.55	0	0.337	0.38
Best	1	1.5	1.5	0.51	0.216	0	0.19

Score(q, doc1) = 0.762, score(q, doc2) = 0.657, score(q, doc3) = 0.916
Ranking: doc3, doc1, doc2

Exercise 6.24
Suppose that the word **coyote** does not occur in the collection used in Exercises 6.10