TF-IDF Term Weighting

Lecture 4, Aug 27, 2019

Task

In this exercise we'll have a look at how the TF-IDF term weighting works.
There are 5 different documents in the collection:

D1 "If it walks like a duck and quacks like a duck, it must be a duck."

D2 "Beijing Duck is mostly prized for the thin, crispy duck skin with authentic versions of the dish serving mostly the skin."

D3 "Bugs' ascension to stardom also prompted the Warner animators to recast Daffy Duck as the rabbit's rival, intensely jealous and determined to steal back the spotlight while Bugs remained indifferent to the duck's jealousy, or used it to his advantage. This turned out to be the recipe for the success of the duo."

D4 "6:25 PM 1/7/2007 blog entry: I found this great recipe for Rabbit Braised in Wine on cookingforengineers.com."

D5 "Last week Li has shown you how to make the Sechuan duck. Today we'll be making Chinese dumplings (Jiaozi), a popular dish that I had a chance to try last summer in Beijing. This is the best recipe for Jiaozi."

Steps

First, create a document-term matrix and record the count of occurrences of each term in the documents ($c_{t,d}$). For simplicity, we work with a restricted vocabulary of 5 terms ("duck," "Beijing," "dish," "wine," "recipe"). When counting, consider the 's form also the same word (e.g., "rabbit's" should be taken to be the same as "rabbit").

doc/term	t_1	t_2	t_3	t_4	t_4
	duck	Beijing	dish	wine	recipe
D1					
D2					
D3					
D4					
D5					

Table 1: Document-term matrix.

Compute TF term weights using L1 normalization, i.e.:

$$tf_{t,d} = \frac{c_{t,d}}{|d|}, \tag{1}$$

$c_{t,d}$ is the number of occurrences of term t in document d and $|d|$ is the sum of all term counts in the document; see the values in Table 1.

We also compute the IDF values for each term using this formula:

$$idf_t = \log \frac{N}{n_t}, \tag{2}$$

where N is the total number of document and n_t is the number of documents that contain term t. The base of the logarithm does not matter as long as the same one is used for all the IDF calculations.

doc/term	TF					IDF
	t_1	t_2	t_3	t_4	t_4	
	duck	Beijing	dish	wine	recipe	
D1						
D2						
D3						
D4						
D5						

Table 2: TF and IDF values.

Finally, we compute the TFIDF weights by multiplying each TF cell in Table 2 by the corresponding IDF value.

doc/term	TFIDF				
	t_1	t_2	t_3	t_4	t_4
	duck	Beijing	dish	wine	recipe
D1					
D2					
D3					
D4					
D5					

Table 3: Document TFIDF values.