Retrieval Models

Lecture 9, September 23, 2019

Exercise #2: Language Models

Given the term-document matrix of a small document collection in Table 1, answer the questions below. You may use a spreadsheet program for the computations.

term	D1	D2	D3	D4	D5
T1		1			1
T2		1			1
T3	3	2	2		1
T4			1	1	
T5			1	1	1
T6	2	1		2	

Table 1: Term-document matrix.

Use the following formula for computing the document language model, with the smoothing parameter $\lambda = 0.1$.

$$p(t|\theta_d) = (1 - \lambda)P(t|d) + \lambda P(t|C) \tag{1}$$

And this is the formula for scoring a given query:

$$P(q|d) = \prod_{t \in q} P(t|\theta_d)^{f_{t,q}} \tag{2}$$

Questions

Answers are given up to 3 digits. See the Excel sheet on GitHub for the computations.

1. What is the value of $P(t|d)$ for t=T2 and d=D2? 0.2
2. What is the value of $P(t|d)$ for t=T5 and d=D1? 0
3. What is the probability of the term T2 in the collection language model? 0.091
4. What is the probability of the term T6 in the collection language model? 0.227
5. What is the probability of T2 in the smoothed document model of D2 ($P(t|\theta_d)$)? 0.189
6. What is the probability of T5 in the smoothed document model of D1 ($P(t|\theta_d)$)? 0.014
7. What is probability of the query q="T3" given document D1? ($P(q|d)$)? 0.576
8. What is probability of the query q="T2 T1" given document D2? ($P(q|d)$)? 0.036
9. Which document has the highest probability for the query q="T6"? D4
10. Which document has the highest probability for the query q="T3 T1 T3 T2"? D2