Language Modeling

Lecture 2, Aug 28, 2017

Task 1

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

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

Task 2

Given the following toy document collection in Table 2, first create the term-document matrix for each field separately, then answer questions using the Mixture of Language Models approach. You may use a spreadsheet program or write Python code for the computations.

term	(1) title	(2) body	(3) anchors
D1	T1	T1 T2 T3 T1 T3	T2 T2
D2	T4 T5	T1 T3 T4 T4 T4 T5	T5 T3
D3	T1 T3 T5	T1 T1 T5 T3 T5 T3 T3	T1 T1 T5

Table 2: Document collection with fielded documents.

Use the following formula for computing the field language models. The smoothing parameter is 0.1 for all fields ($\lambda_i = 0.1$).

$$P(t|\theta_{d_i}) = (1 - \lambda_i)P(t|d_i) + \lambda_i P(t|C_i) \tag{3}$$

The document language model is computed using Eq. 4. Use the following field weights: title: $\mu_1 = 0.1$, body: $\mu_2 = 0.7$, anchors: $\mu_3 = 0.2$.

$$P(t|\theta_d) = \sum_i \mu_i P(t|\theta_{d_i}) \tag{4}$$

Scoring the query goes by substituting Eq. 4 back into Eq. 2.

Questions

1. What is the probability of T1 in the field language model of D3 for field 2 ($P(t|\theta_{d_2})$)?
2. What is the probability of T2 in the field language model of D2 for field 1 ($P(t|\theta_{d_1})$)?
3. What is the probability of T4 in the collection language model of field 1 ($P(t|C_1)$)?
4. What is the probability of T4 in the collection language model of field 3 ($P(t|C_3)$)?
5. What is the probability of T2 in the (smoothed) document model of D2 ($P(t|\theta_d)$)?
6. What is the probability of T5 in the (smoothed) document model of D1 ($P(t|\theta_d)$)?
7. What is probability of the query q=“T3” given document D1? ($P(q|d)$)?
8. What is probability of the query q=“T2 T1” given document D2? ($P(q|d)$)?
9. Which document has the highest probability for the query q=“T4”?
10. Which document has the highest probability for the query q=“T1 T2 T4”?