Information Retrieval

Language Modeling for IR

Assignment

Exercises

Book IIR: exercises 12.7, 12.8, 12.9

Exercise 12.7

Suppose we have a collection that consists of the 4 documents given in the below table.

docID	Document text
1	click go the shears boys click click click
2	click click
3	metal here
4	metal shears click here

Build a query likelihood language model for this document collection. Assume a mixture model between the documents and the collection, with both weighted at 0.5. Maximum likelihood estimation (mle) is used to estimate both as unigram models. Work out the model probabilities of the queries click, shears, and hence click shears for each document, and use those probabilities to rank the documents returned by each query. Fill in these probabilities in the below table:

Query	Doc 1	Doc 2	Doc 3	Doc 4
click				
shears				
click shears				

What is the final ranking of the documents for the query click shears?

How to compute parameters of a mixture model?

docID	Document text
1	click go the shears boys click click click
2	click click
3	metal here
4	metal shears click here

- Step1: Estimate parameters in unigram language models
- $\hat{P}(t|d) = \hat{P}_{mle}(t|M_d) = \boldsymbol{c}(t,d)/|d|$
  - P('click'|d1)= 4/8
  - P('click'|d2)=2/2
  - P('click'|d3)=0/2
- $\hat{P}(t|c) = \hat{P}_{mle}(t|M_c) = \boldsymbol{c}(t,c)/|c|$
  - P('click'|C) = 7/16
- Step2: Estimate the smoothed parameters
  - $\hat{P}(t|d) = 0.5\ \hat{P}_{mle}(t|M_d) + 0.5\ \ \hat{P}_{mle}(t|M_c)$
  - $\hat{P}(click|d3)$= 0.5*0+0.5*(7/16)

	RAW TERM FREQUENCIES				
	d1	d2	d3	d4	c
click	4	2		1	7
go	1				1
the	1				1
shears	1			1	2
boys	1				1
metal			1	1	2
here			1	1	2
	8	2	2	4	16

docID	Document text
1	click go the shears boys click click click
2	click click
3	metal here
4	metal shears click here

	SMOOTHED PARAMETERS P_sm(w|D)				
	d1	d2	d3	d4	
click	0,46875	0,71875	0,21875	0,34375	
go	0,09375	0,03125	0,03125	0,03125	
the	0,09375	0,03125	0,03125	0,03125	
shears	0,125	0,0625	0,0625	0,1875	
boys	0,09375	0,03125	0,03125	0,03125	
metal	0,0625	0,0625	0,3125	0,1875	
here	0,0625	0,0625	0,3125	0,1875	
click shears	0,058594	0,044922	0,013672	0,064453	

Exercise 12.8

Using the calculations in Exercise 12.7 as inspiration or as examples where appropriate, write one sentence each describing the treatment that the model in Equation (12.10) gives to each of the following quantities. Include whether it is present in the model or not and whether the effect is raw or scaled.

a. Term frequency in a document
b. Collection frequency of a term
c. Document frequency of a term
d. Length normalization of a term

$$\log P(Q|D) = \sum_{i=1}^{n} \log((1-\lambda)\frac{f_{q_i,D}}{|D|} + \lambda\frac{c_{q_i}}{|C|})$$

- Term frequency:
  - Present, scaled
- Collection frequency:
  - Present, scaled
- Document frequency
  - Not present (but…)
- Length normalization of a term
  - Not present

Exercise 12.9

In the mixture model approach to the query likelihood model (Equation (12.12)), the probability estimate of a term is based on the term frequency of a word in a document, and the collection frequency of the word. Doing this certainly guarantees that each term of a query (in the vocabulary) has a non-zero chance of being generated by each document. But it has a more subtle but important effect of implementing a form of term weighting, related to what we saw in Chapter 6. Explain how this works. In particular, include in your answer a concrete numeric example showing this term weighting at work.

Chapter 6: Vector space model

- *Tf. Idf weighting*
  - *Term frequency*
  - *Inverse document frequency*
- *Inverse document frequency, proportional to rareness of a term*
  - *1/DF*
- *Inverse document frequency behaves rather similar as the inverse collection frequency.*
  - *Why?*
  - *In which cases is there a difference?*
    - *Relation with macro / micro averaging (icf: micro, pooled average)*

Where is *tf.idf* Weight?

$$\log P(Q|D) = \sum_{i=1}^{n} \log((1-\lambda)\frac{f_{q_i,D}}{|D|} + \lambda\frac{c_{q_i}}{|C|})$$

$$= \sum_{i:f_{q_i,D}>0} \log((1-\lambda)\frac{f_{q_i,D}}{|D|} + \lambda\frac{c_{q_i}}{|C|}) + \sum_{i:f_{q_i,D}=0} \log(\lambda\frac{c_{q_i}}{|C|})$$

$$= \sum_{i:f_{q_i,D}>0} \log \frac{((1-\lambda)\frac{f_{q_i,D}}{|D|} + \lambda\frac{c_{q_i}}{|C|})}{\lambda\frac{c_{q_i}}{|C|}} + \sum_{i=1}^{n} \log(\lambda\frac{c_{q_i}}{|C|})$$

$$\stackrel{rank}{=} \sum_{i:f_{q_i,D}>0} \log\left(\frac{((1-\lambda)\frac{f_{q_i,D}}{|D|}}{\lambda\frac{c_{q_i}}{|C|}} + 1\right)$$

- - proportional to the term frequency, inversely proportional to the collection frequency

Numerical example

- Let's look at $\log\hat{P}(t|d)$ for a rare and common term from the example in 12.7
- Rank equivalant to Log (1 + (0.5 P(t|d)/0.5 P(t|c)))

- Rare term: 'metal'
  - log(P(metal|d4) =~ log (1 + 1/4 / 2/16) =~ log (1 + 2)

- Common term: 'click'
  - log(P(click|d1) =~ log (1 + 1/4 / 7/16) =~ log (1 + 4/7)

- It is more important to include 'metal' than 'click'…