Information Retrieval

Exercises week 6

Exercise 1: Relevance feedback

- Suppose we have 1000 abstracts(DF Leiden 20, DF University 50)
- In a first pass we retrieve 150 document titles for the query 'Leiden University' and provide relevance feedback for the top 5 documents

Relevance	Title
1	Leiden University
1	Leiden city of science 2022
1	Leiden Universiteit
0	3 October, Leiden liberation
0	Rembrandt of Leiden

a) What are p_i and r_i for the terms 'Leiden' and 'University' 1st pass search

b) Estimate p_i for the terms 'Leiden' and 'University' (cf. slide 26), for the second pass?

c) How and why can we smooth these estimates?

Solution Exercise 1

a) For the first pass search, we do not have relevance information. So we take p_i=0.5 for both Leiden and University. For r_i we will use the normalized DF, so r_i = 20/1000 and 50/1000 respectively (we assume all documents are not relevant)

b) Since we have relevance information now, we can refine the estimates . P_i = 1 and 1/3 respectively . R_i hardly changes, but could be refined as (17/997) and (49/997) respectively

c) Estimates should be smoothed to avoid dividing by zero and taking a log of zero. Smoothed estimates for 2nd pass: P_i (3+1/2)/(3+1) and (1+1/2)/(3+1) respectively, R_i (17.5/998) and (49.5/998) respectively

Exercise 2

- Consider the BIM RSV definition

$$RSV = \sum_{x_i = q_i = 1} c_i;$$

a) Explain why this ranking scheme can be implemented efficiently.

b) What is the key assumption that allows for this simple ranking model?

Solution Exercise 2

a) The ranking formula is a presence only scheme. Only the posting lists of query terms should be evaluated.

b) The key assumption is the linked dependence assumption (cf. slide 15) , this is in fact a weaker version of the conditional independence assumption which is commonly mentioned in deriving Naïve Bayes classifier. There is a more thorough explanation available by Victor Lavrenko at https://www.youtube.com/watch?v=ZPuWZ1bRsWA

Exercise 3

a) The k_1 parameter in the BM25 is a constant

Do you think that a term specific k_1 would be better?

Please motivate.

b) BM25 assumes binary query vectors, what would happen with the query: “wild wild world”, how to mitigate?

Solution Exercise 3

a) In fact the tf saturation function is an approximation of the theoretically motivated 2-poisson model, which contained three term specific parameters. So yes, term saturation should probably handled differently for e.g. high and low frequency terms. The reason a term independent approach was chosen was to create a robust term weighting function, without requiring a complex training procedure and lots of data. ChengXiang Zhai published some experimental results of estimating a term dependent k_1

b) Just ignoring the fact that this is probably a phrase query, BM25 will handle each query term additive, so the result list may become biased. Mitigation could be done by reweighting query terms just as documents (e.g. offer weights ‘https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-356.pdf and https://www.researchgate.net/publication/235277769_On_term_selection_for_query_expansion)