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?

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?

Exercise 3

a) The k_1 parameter in the BM25 is a constant

   a) 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?