Trier University
Department IV – Computer Sciences
Prof. Dr.-Ing. Ralf Schenkel (H 539, schenkel@uni-trier.de)

Trier, 12.12.2022

Assignment 7 in the course Digital Libraries and Foundations of Information Retrieval

Winter Semester 2022

Deadline 12:15h on Monday, 19.12.2022

Note that the meetings on December 19, 2022 and January 2, 2023 will take place in the following Zoom meeting: https://uni-trier.zoom.us/j/83637205106?pwd=RnowWVR6SUNJeHQ3MWZyeUtOV3Z3dz09 (Meeting-ID: **836 3720 5106**, password: **fHqDg15P**). It may be necessary to login with your university account before accessing the meeting.

Task 1: Language models 4+4+3+4 Points

We consider the following documents (from which stop words were removed and for which a lemmatization was performed):

- d_1: Mosel Marx Porta Trier Dom
- d_2: Trier Porta Porta
- d_3: Trier Mosel Marx Trier Mosel Porta Dom Mosel

(a) Determine the *unigram language models* for the documents d_1,d_2 and d_3.

(b) Determine the ranking for the following queries with the *query likelihood model without smoothing*:

(1) Trier

(2) Trier Porta

(c) Determine the background language model (as unigram language model) for the collection consisting of d_1,d_2 and d_3.

(d) Determine the ranking for the following queries with the *query likelihood model with Jelinek-Mercer smoothing* (for $\lambda = 0.5$):

(1) Trier

(2) Trier Dom

Hint: For all logarithms that need to be computed to solve this task, use the ***base 2****.*

Task 2: VSM and BM25 3+4+3+5 Points

An analysis of classical fairy tales yields the following distribution of terms in $d_1 \dots d_7$:

term	d_1	d_2	d_3	d_4	d_5	d_6	d_7
father	0	5	0	0	0	1	0
mother	2	2	3	2	0	0	3
queen	0	0	0	0	8	1	0
dwarfs	0	0	0	0	4	0	0
princess	0	0	0	0	1	1	0
wolf	0	0	0	6	0	0	6
gold	2	0	1	0	1	0	0
house	2	5	1	3	4	1	1

(a) Determine $idf_t = \log \frac{N}{df_t}$ for the eight terms. Then determine the normalized weights of the term-document matrix with $tf \cdot idf$ (such that the length of each document vector is 1).

(b) Determine for each document its score for the query $\{mother, house\}$ in the vector space model with $tf \cdot idf$.

(c) Determine for each document its score for the same query with BM25. For the parameters, use $k = 2$ and $b = 0.75$.

(d) You particularly like d_3, so you want to retrieve similar documents. One can do this by using d_3 as a query. Determine the most similar document to d_3 in the vector space model with $tf \cdot idf$ and with $BM25$.

Hint: For all logarithms that need to be computed to solve this task, use the ***base 2****.*

Task 3: Probability Ranking Principle & Binary Independence Model 7+8 Points

We consider the following document vectors:

	document vectors						
terms	d_1	d_2	d_3	d_4	d_5	d_6	d_7
t_1	1	1	0	0	0	0	0
t_2	1	1	0	1	0	0	0
t_3	0	1	1	1	0	0	0
t_4	0	1	1	1	0	0	0

(a) Determine the ranking of the documents for the query $\{t_1, t_2, t_3, t_4\}$ using the similarity determined by the *Binary Independence Model* with smoothing. Assume that only documents with a positive similarity should be returned.

(b) A user now gives feedback on the result of task a). Documents d_1 and d_2 are *not relevant* for the query $\{t_1, t_2, t_3, t_4\}$, documents d_3 and d_4 are *relevant* for this query. Compute the ranking of these documents for the query $\{t_1, t_2, t_3, t_4\}$ using the similarity determined by the *Binary Independence Model* with smoothing again, now under consideration of this relevance information. Assume again that only documents with a positive similarity should be returned.

These tasks will be discussed in the tutorial on January 2, 2023.

General remarks:

- The tutorial group takes place on Mondays in the regular meeting in F55 at 12:15.
- The first meeting of the tutorial group will be on November 14, 2022.
- To be admitted to the final oral exam, you need to acquire at least 50% of the points in the assignments. In addition, you need to present at least one solution of a task in a convincing way during the tutorial.
- It is preferred to submit in groups of size up to two (but not larger); in that case, only one submission is sufficient for the whole group. Write the names of all group members on your solutions.
- Solutions must be handed in before the deadline
  - in Moodle (`https://moodle.uni-trier.de/`, course `DL-IR-2022`) as as a PDF or, if submitting multiple files, as an archive (.zip or comparable).

  Submissions that arrive after the deadline will not be considered. The name of at least one group member should occur in the file name of your submission. If you want to modify a previously uploaded solution, just re-upload your solution.
- Graded versions of your submissions will be returned in Moodle until the following tutorial.
- Announcements regarding the lecture **and** the tutorial group will be done in the StudIP course for the lecture.