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

Trier, 05.12.2022

Assignment 6 in the course Digital Libraries and Foundations of Information Retrieval

Winter Semester 2022

Deadline 12:15h on Monday, 12.12.2022

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: Vector Space Model 2+3+3+2+1+4 Points

Consider the following term-document frequency matrix:

	term frequencies			
terms	tf_{d_1}	tf_{d_2}	tf_{d_3}	tf_{d_4}
biscuits (t_1)	0	3	0	4
stollen (t_2)	1	3	3	7
lebkuchen (t_3)	7	7	0	0
macaroons (t_4)	0	1	0	2
brownies (t_5)	0	1	0	0
cookies (t_6)	0	1	0	4
pastries (t_7)	0	0	4	0

(a) Compute the *inverse document frequency* idf_t for all terms t with the formula on slide 6-29.

(b) Compute the *document vectors* $\mathbf{d_i} = (w_{t_1,d_i}, \ldots, w_{t_7,d_i})^T$ for all documents d_i using the $tf \cdot idf$ weighting scheme.

(c) Normalize the document vectors to $\mathbf{d_i'}$ for all documents d_i.

(d) Compute the normalized query vector $\mathbf{q}$ for the query „lebkuchen cookies pastries“. Use term frequency 1 for the query terms.

(e) Make an educated guess what the best document will be for the query „lebkuchen cookies pastries“ (including a short reason for your guess).

(f) Compute a ranking of all documents for the query „lebkuchen cookies pastries“ using the cosine similarity. Compare the result of the query with your guess from task e); if there is a difference, give an explanation.

*Remark: Use the **base 2** for all logarithms computed while solving this task.*

Task 2: Vector space model 1+2+3+4+5 Points

Consider the following term-document frequency matrix:

	term frequencies			
terms	tf_{d_1}	tf_{d_2}	tf_{d_3}	tf_{d_4}
t_5	1	10	0	0
t_6	5	5	5	5
t_7	1	0	0	10
t_8	20	12	100	1000

In this task (with the exception of task a)) we use weights that are only based on tf, i.e., we ignore *idf* (or set $idf = 1$ for all terms).

(a) Explain the role of **idf** for the result of queries that consist of a single term.

(b) Compute for each document d the number of its term occurrences $|d|$ (i.e., its length as a document, not the length of its vector), i.e., the sum of the tf values of all terms for the document, and the length $||\mathbf{d}||$ of the corresponding document vector $\mathbf{d}$.

(c) Compute the result of the query that consists only of t_5, by computing the similarity of each document to the corresponding query vector.

(d) Compute the result of the query that consists only of t_6, by computing the similarity of each document to the corresponding query vector. Explain why not all documents have the same score. Is this a reasonable behavior of a search engine?

(e) Compute the result of the query that consists only of t_7, by computing the similarity of each document to the corresponding query vector. Compare the result with that of task (c), where also one document with $tf = 1$ and one document with $tf = 10$ is included in the result, and explain the difference.

These tasks will be discussed in the tutorial on December 19, 2022.

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.