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

Trier, 21.11.2022

Assignment 4 in the course Digital Libraries and Foundations of Information Retrieval

Winter Semester 2022

Deadline 12:15h on Monday, 28.11.2022

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: Boolean Retrieval 3+3+4+3+2 Points

We consider the following documents, given by their index terms:

- D_1: christmas, snow, skating, wine
- D_2: christmas, skiing, ropeway, wine, punch
- D_3: christmas, snow, cross-country, ice
- D_4: christmas, strudel, cinnamon, ice

(a) Construct the term-document incidence matrix for this document collection.

(b) Construct the inverted index for this document collection.

(c) Evaluate the query *christmas* ***AND*** *(ice* ***OR*** *punch)*. Show all intermediate results.

(d) Formulate a realistic information need that could be underlying this query. Determine all documents in the collection relevant for this information need. Determine precision and recall for the results computed in part c). If precision and recall are not perfect, give possible reasons for it.

(e) Explain the special role of the term *christmas* when answering queries, compared to the other terms in the vocabulary.

Task 2: Tokenisation 3+12 Points

(a) Explain the difference of tokens and terms.

(b) Consider the following text:

```
On June 28, 2016, Sinéad O'Connor attended the seventeenth award cere-
mony of the Peter-Paul-and-Mary award in the big ewent venue in Kürenz.
```

Use this example to demonstrate four typical problems when tokenizing a text.

Task 3: Document preprocessing 15 Points

Consider the following documents:

D_1 analyzing online schema extraction approaches for linked data knowledge bases

D_2 elements of computer science

D_3 artificial intelligence

Research the rules of the Porter stemmer, for example on `https://tartarus.org/martin/PorterStemmer/`. Apply these rules to reduce the tokens to their stem, after removing stop words. Document your solution by stating the rules applied for each token, in addition to the resulting stem.

These tasks will be discussed in the tutorial on December 5.

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.