Tutorial Information Retrieval

Universität Innsbruck - Department of Computer Science

E. Zangerle

DBIS
DATABASES AND INFORMATION SYSTEMS

2019-03-20

Tutorial 1

Exercise 1 (Text Corpus Statistics) [7 Points]

During the next weeks, we are going to construct a fulltext index step by step. In a first step towards this index, we will set up a set of methods for reading in text data, tokenization and the computation of statistics about a given text corpus. This corpus is constructed based on a set of text files containing the texts to be indexed.
Please remove all punctuation during parsing and tokenizing the input file and lower-case all input tokens (make use of e.g., SpaCy[1] or NLTK[2]). Implement a script that computes the following statistics for a given corpus:

a) Total number of terms

b) Total number of unique terms

c) Frequency of each term

d) List of the top-50 terms with the according frequency and rank over the whole text corpus.

e) Number of stopwords contained[3]

Please do not rely on any database for this task. Make sure to choose a modular architecture as this system is going to be extended during the next weeks. As for input texts, you can make use of any arbitrary text freely available at Project Gutenberg[4]. However, please make sure to add a few novels such that the index is filled with a sufficient number of texts for future benchmarking and comparisons among different versions of this inverted index.

Exercise 2 (Reading) [3 Points]

Please read the following paper:

Dirk Bahle, Hugh E. Williams, and Justin Zobel. Efficient phrase querying with an auxiliary index. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 215–221, New York, NY, USA, 2002. ACM. ISBN 1-58113-561-0. doi: 10.1145/564376.564415. URL http://doi.acm.org/10.1145/564376.564415

[1] https://spacy.io/
[2] https://www.nltk.org/
[3] For a list of stopwords, you can rely on e.g., NLTK's or SpaCy's stopword lists.
[4] http://www.gutenberg.org/

Subsequently, answer the following questions:

- What is the goal of this paper?
- How do the authors reach their goal?
- How do they evaluate their approach?

Important: Submit your solution (.txt, .java or .pdf) to OLAT and mark your solved exercises with the provided checkboxes. The deadline ends at 23:59 on the day before the discussion.