Tutorial Information Retrieval

Universität Innsbruck - Department of Computer Science

E. Zangerle

2019-03-27

Exercise Sheet 2

Exercise 1 (Inverted Index) [2 Points]

The goal of this exercise is to build an inverted index.

a) Elaborate on the requirements in regards to the data structure underlying the inverted index.

b) Implement the inverted index and fill it the documents from the Cranfield Collection we provide via OLAT (make use of the tokenization and cleaning facilities implemented for Exercise Sheet 1). The SPIMI algorithm[1] (Single Pass In-Memory Indexing) might be helpful for this task.

c) Add a feature for persisting the created index and the according statistics on disk such that the statistics and the index only have to be computed once. After having created and persisted the index, only the index file has to be read to perform new searches on the index.

Exercise 2 (Boolean Retrieval Model) [4 Points]

Based on the inverted index created in the previous exercise, the next step is to employ the boolean retrieval model[2] on it. Therefore, we have to be able to parse queries and match these with the index. In order to do so, please follow the steps listed below.

a) Implement a query parser (utilizing the same methods as in Exercise 1).

b) Add boolean retrieval to your prototype. Please note that we are going to extend the prototype with other retrieval models as well, so please make sure to reflect this in your design choices.

c) Example queries for the Cranfield Collection are given in the queries.csv file. The file contains the query and the relevant documents (given as ids) as well as a relevance score. The relevance score can be ignored in this exercise sheet.

d) A very simple similarity measure $sim(q, d)$ between the query q and a document d that can be used on top of a boolean retrieval model is to use the size of the intersection of the query q and the document d:

$$sim(q, d) = \frac{|q \cap d|}{|q|} \tag{1}$$

Use this measure as a first attempt to ranking the retrieved set of documents.

e) Elaborate on the evaluation measures recall, precision and F_1-score. Compute these measures for the given queries in the test set. Please provide the overall precision, recall and F_1-score by aggregating the individual scores using the arithmetic mean.

[1] http://nlp.stanford.edu/IR-book/html/htmledition/single-pass-in-memory-indexing-1.html
[2] https://nlp.stanford.edu/IR-book/html/htmledition/boolean-retrieval-1.html

Exercise 3 (Text Processing) [4 Points]

The next step is to extend the current fulltext index to also perform text preprocessing in regards to stemming, lemmatization and stopword-removal:

a) Remove stopwords from both your index and your queries.

b) Apply a stemming algorithm to the tokens within your indexing pipeline.

c) Please also implement lemmatization.

How does the query accuracy evolve in comparison to the previous version of the index in terms of recall, precision and F_1-score? Please compare the index from Exercise 2 with a version only removing stopwords, a version removing stopwords and using stemming as well as a version removing stopwords and using lemmatization. Please make sure to present the results of your comparison accordingly (tables, plots). Elaborate on the chosen visualization (i.e. a boxplot) for showing differences.

Important: Submit your solution to OLAT and mark your solved exercises with the provided checkboxes. The deadline ends at 23:59 on the day before the discussion.