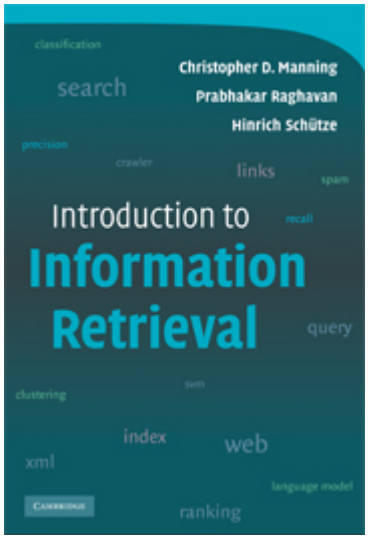
Reading Assignment: Text Preprocessing and Vectorization



Learning Goals

- Text preprocessing steps
- Document representations in vector space
- Document similarity

The reading assignment is based on the book Introduction to Information Retrieval by Christopher Manning, Prabhakar Raghavan and Hinrich Schütze from 2008. It is an older book, but still accurate - and available online. Although having information retrieval ("text search") as central topic, the concepts also apply for text classification.

The chapters relevant for the reading assignment are:

- Chapter 1: Boolean retrieval
- Chapter 2: The term vocabulary & postings list
- Chapter 6: Scoring, term weighting & the vector space model

Answering the questions below only requires the study of subsections **1.1, 2.1-2.2. and 6.2-6.3.2**. However, the left-out parts might be useful for context.

Carefully read through the sections stated above and answer the following questions (listed in reading order). Once finished, take the quiz in Moodle.

Questions

- Why is the matrix in Figure 1.1 called **binary** incident matrix?
- What is the difference between "word" and "term"?
- What concept does the term "corpus" denote?
- Which terms are used as synonym for "vocabulary"?
- What are the differences between "token", "type" and "term"?
- Find out what the standard encoding is in Python.
- Which preprocessing step described in section 2.2 do you think is mandatory for most applications?
- Find examples (that were not listed in the book) in your native language where tokenization is not straightforward, i.e. simply splitting at non-alphanumeric characters might not be enough.
- Which preprocessing steps decrease the size of the vocabulary?

- Is stemming or lemmatization more resource-intense?
- In your native language, find another example where normalization might be useful.
- What is the bag-of-words representation of a document?

- What is the problem with raw term frequency?
- Why is document frequency preferred to collection (or corpus) frequency?
- What is the inverted document frequency (idf) of a term that occurs in every document?
- Can the tf-idf weight of a term in a document exceed 1?
- Make sure to be able to follow the computations in Example 6.2 and Example 6.3.