Foundations of Information Retrieval

Exercise 3

Exercise session: 23.11.2017
Questions to: Markus Rokicki, rokicki@L3S.de

Levenshtein distance

a) Compute the Levenshtein distance between the terms „top“ and „stop“ using matrix. Use the matrix to determine the required transformations steps.

b) Compute bi-gram based similarity between the terms “top” and “stop”.

c) What are advantages and disadvantages in terms of efficiency, memory consumption, and precision using Levenshtein distance for spelling correction compared to an n-gram index?

Wildcard Query

Given is the wildcard query Sh*sp*re (Shakespeare).

a) For this wildcard query, create queries for a bigram index and a permuterm index.

b) For this query, which terms will be delivered by the permuterm index, but not by the bigram index?

   Which characteristics have the terms that will be delivered for this query by the bigram index but not by a permuterm index?

   If the terms exists, provide examples. Justify your answer!

Soundex

Compute the Soundex code of your last name.
Try to find another name with the same Soundex code.