Information Retrieval Homework 2

Ondřej Měkota

January 4, 2020

Overview

- Using Whoosh search engine for Python
- Run-0 – only regular expression based tokenizer
- Run-1 – BM25, lemmatisation/stemming, stopwords, lowercasing
- For English, preprocessing was done only using the framework

Notes

- Pseudo relevance feedback did not work.
- Lemmatisation for English did not work either.
- Removing numbers did not show improvement.
- BM25 tuning improved MAP by about 0.03.

Tables with results

Run	Czech	English
Run-0	0.0366	0.0764
Run-1	0.3342	0.4002

Table: map

Run	Czech	English
Run-0	0.0560	0.1040
Run-1	0.3360	0.4600

Table: P_10