Text Technologies for Data Science Assignment 1

Mark Nemec

1 Introduction

The purpose of this report is to outline decisions made in the implementations of search algorithms for information retrieval and discuss the performance on provided datasets and techniques used.

2 Overlap & tf-idf algorithms

Both of these algorithms have a fairly similar structure. Queries and documents are tokenized and turned into a dictionary where word is used as key and its count in query/document as value.

2.1 Overlap algorithm

This algorithm counts the number of words that were present in both the query and the document. The average precision of this search algorithm on `qrys.txt` and `docs.txt` was 0.1527.

2.2 tf-idf algorithm

This algorithm includes a pre-processing step where the total number of documents, tokens and average document length were computed. Moreover, we compute a dictionary of word to count mappings for the whole collection. These are later used together with the query and document dictionaries to compute tf-idf. The average precision of this search algorithm on `qrys.txt` and `docs.txt` was 0.3230.