Boolean Retrieval

NPFL124 – Natural Language Processing

Assignment

- Implement the inverted index with a hash used for the dictionary part of the index.
- Implement algorithms for postings intersection and union.
- Index the provided document collection.
- Write a query parser for AND, OR, and NOT.
- Process the provided set of boolean queries and submit the results.

Programming Language

For solving this assignment I used Python 3.10.3 and its standard library.

Report

For the program to finish its execution it takes approximately 2 minutes 30 seconds. Out of this, building the inverted index takes the most time, while the query processing is almost instantaneous. This is probably due to the fact that the main data structure used is a dictionary of strings and a set, which uses hash tables for its implementation.

The program is written in the following way:

The inverted index was built in a way that the program would simply go through the whole .xml file, and for every line that needed to be indexed, it would tokenize it (by removing the delimiters and spaces in between the words and converting the line into an array) and then add word by word to the dictionary.

After that, another function is called which loops through the file that contains queries and splits it into individual words and operators. Later on, every query will go through a simple recursive function which further calls a function responsible for evaluating a simple query with only 1 operator.

To implement the boolean operators OR, AND, and AND NOT, I used the basic union, intersection, and difference algorithms for 2 sets.

Adilović Una, April 2022