Lab Assignment 3

Due date: 22 November 2021

Question: Design an inverted index for a particular directory in your system.

Use the "20Newsgroups" dataset (http://qwone.com/~jason/20Newsgroups/) (http://qwone.com/~jason/20Newsgroups/20news-19997.tar.gz).

The dataset has several directories. Choose any one directory to create the index of documents that are present in the directory.

Compare the sizes of inverted index generated following the two cases:

1. Without using any preprocessing techniques like stop words etc. for index construction.
2. Index constructed after using different preprocessing techniques (studied in class).

Implement two functions **f_or** and **f_and**, where **f_or** is able to compute **'or'** operation on postings and **f_and** can compute **'and'** operation on postings. Use the created index, to run different queries on postings of terms at runtime. For example, given any terms: term1, term2, term3 and term4, you may test it by running the following queries:

- (term1 and term2) or (term3 and term4)
- (term1 or term2) and (term3 or term4)

Compare the retrieval speeds on indexes built in points 1 and point 2.