Project #2 for the Biomedical Information Retrieval Course

Due: Oct 26, 2021

General Guideline

This homework is basically an individual homework. Each student has to do it all by himself (or herself). The final score will be evaluated from the system performance and individual demonstration.

Homework Overview

First of all, implement the Zipf Distribution computation (or frequency spectrum for terms) for both a set of text documents from PubMed (and Tweeter as control group) with *same subject*. The size of text document sets could range from 10 to 50000, depends on your intention. You have to preprocess the text index (token) set from document collection. Secondly, implement the *Porter's algorithm* as a functional module on your own software platform for the same set of text documents. Then, compare the difference.

In this project, basically your system or software platform will be able to compute text distribution from a set of documents. Match (or partial matching) process can be done using the dynamic programming-based *Edit distance* computation. Your retrieval results can be displayed in a format of indicating the location(s) and/or partial matching of the query keywords in each document, etc. Computer languages are not limited.

System Description

1. I suggest that you can test your system in advanced using certain gene or disease name, e.g. "covid-19", which contains approximately 50000 documents. Each individual student builds its own system components using basic matching scheme from the IR course. *Spelling correction* is necessary for this project.
2. In the final evaluation, each individual should present system description, running results, and system demo to verify system performance.