Project #1 for the Biomedical Information Retrieval Course

Due: Sept 24, 2024

General Guideline

This homework is basically an individual-oriented work. 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

Implement a full-text retrieval tool (i.e. search engine) for a set of text documents. Specifically, your system will be able to perform document retrieval according to specified keyword(s) and then display in an easily visualization way. The tool is able to calculate the document *statistics* (such as number of characters, number of words, number of sentences(EOS), etc.) and to determine how many sentences in documents using smart method (for example, rule-based approach). Computer languages are not restrictive.

System Description

1. This homework uses both the standard PubMed Biomedical Data which can be obtained at http://www.ncbi.nlm.nih.gov/ under XML format and twitter data at http://twitter.com/ under JSON format for information retrieval purpose.
2. Each individual student builds its own system components using basic matching scheme from the IR course.
3. Basic document preprocessing algorithms can be obtained from public domain (e.g. Porter's algorithm, stop-words, stemming algorithm, etc.). In the final evaluation, each individual presents system description, running results, and system demo to verify performance.