Homework #4 for the Information Retrieval Course

Deadline: Nov 25, 2025

General Guideline

This homework is basically an individual-oriented work. Each student has to do it by yourself. The final score will be evaluated from analysis and demonstration.

Homework Overview

In this project, you are asking to implement and analyze "term weighting" technology for text documents in the vector space *before* executing the Porter's algorithm. At least 2~3 types of TF-IDF and/or *modified TF-IDF methods*[1], such as sentences or paragraphs are considered in this project. Then, you need to rank the most important key sentences in the paragraphs of documents based on term weighting and similarity measure, in which you have to choose one reasonable ranking/similarity computation method.

System Description

1. This homework uses the medical document data for representation purpose, for example using the "African Swine Fever Virus " as search term for document collection.
2. Each work deals with several small size categories (e.g. about 100~200 documents per category) as test examples.
3. Feel free to adopt any methods from the preprocessing stage of the IR system paper.

Note:

1. *modified TF-IDF methods could be: TF in sentence, TF in paragraph, IDF in sentence, IDF in paragraph, etc.*
   [Ref] Condensing biomedical journal texts through paragraph ranking, JH Chiang, HH Liu, YT Huang, Bioinformatics 27 (8), 1143-1149, 2011