# Homework #4 for the Information Retrieval Course

Deadline: Nov 21, 2023

## 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*, such as sentences or paragraphs are considered in this project. Then, you need to rank the documents based on similarity measure, in which you have to choose one reasonable ranking and similarity computation method.

## System Description

1. This homework uses the medical document data for representation purpose.
2. Each work deals with at least 2 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, but you should be careful about it.