# Project #3 for the Biomedical Information Retrieval Course

Due: Nov 12, 2024

**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**

Implement the Word Embedding Technique(word2vec) for a set of text documents from PubMed with *same subject*. The size of text document sets could range from 1000 to 10000, depends on your original intention. You have to preprocess the text set from document collection. In this project, you can choose one of the 2 basic computational models:

1. Continuous Bag of Word (CBOW): use a window of word to predict the middle word
2. Skip-gram (SG): use a word to predict the surrounding ones in window.

Window size is not limited. 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 10000 documents available, or even you can try “enterovirus”.
2. In the final evaluation, each individual should present system description, running results, and system demo to verify system performance.