Final for the Biomedical Information Retrieval Course

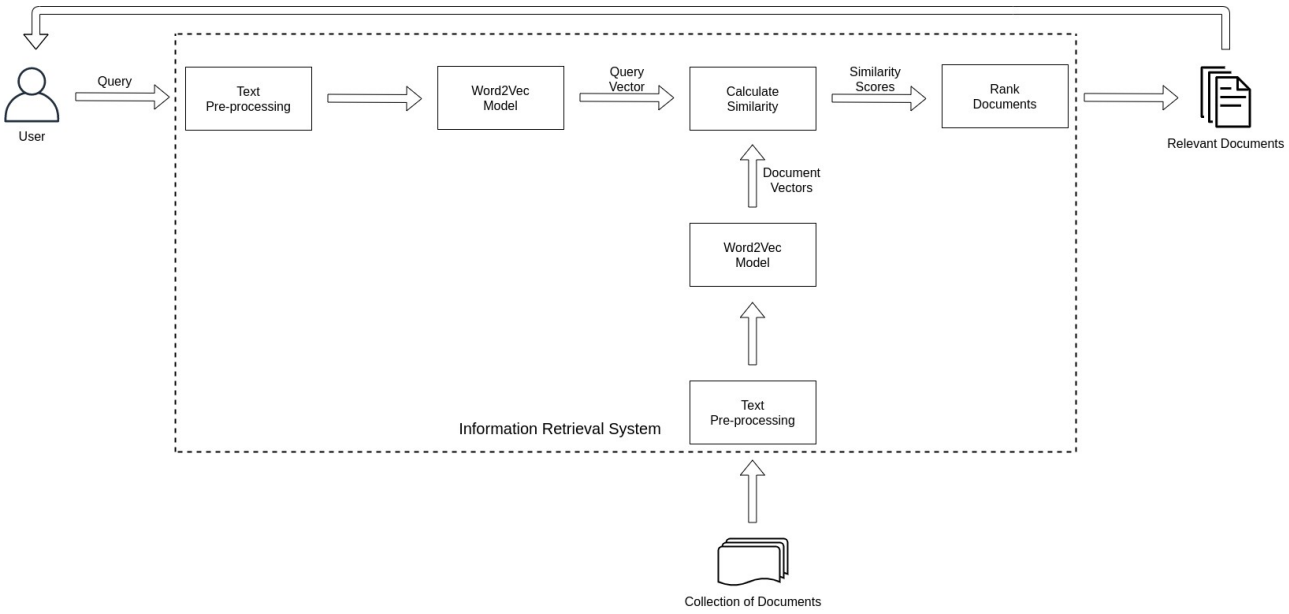
Q56091079 李昱玟 ｜P74094305 林渝育

- ## Environment
  - Python3
  - Flask
  - nltk
  - gensim
- ## Github

https://github.com/yyyyuwen/BIR-course-final

- ## 功能
  - Data pre-Processing
    1. 讀檔：

       讀取關鍵字為covid跟pneumonia跟heart disease的文章各1000 篇
    2. 將每一個word做前處理 : **Stop Word** -> **Lemmatizer**
  - Search & rank by cosine similarity



  - 文本聚類 (text clustering)

- tf-idf計算每篇文章的詞向量
- Kmeans進行文本聚類分析
- TSNE降維與視覺化

```
Cluster 1: heart, ventricular, echocardiography, coronary, myocardial
Cluster 2: cov, test, assay, sample, rt, detection
Cluster 3: covid19, sars, pcr, coronavirus, infection
Cluster 4: pneumoniae, mycoplasma, infection, pathogen, case
Cluster 5: diagnosis, lung, clinical, chest, pulmonary, bacterial
```

  - Word2Vec

將全部的文章集合起來做關鍵字的cosine similarity，為了分析，我們取了heart, coronavirus, pneumonia, fever, cough，經過TSNE降維可以看出來covid與pneumonia的距離比跟heart的還要近，因此可以推測出covid與pneumonia的關聯度較高。