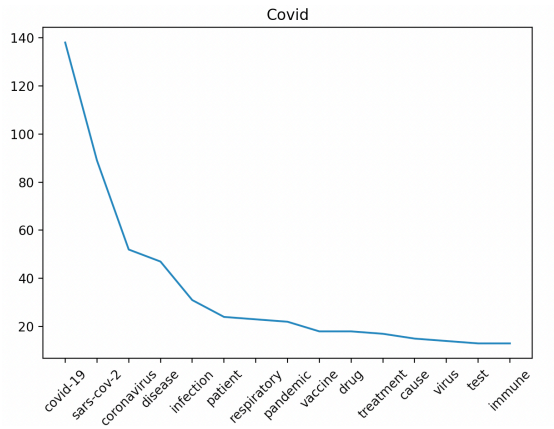
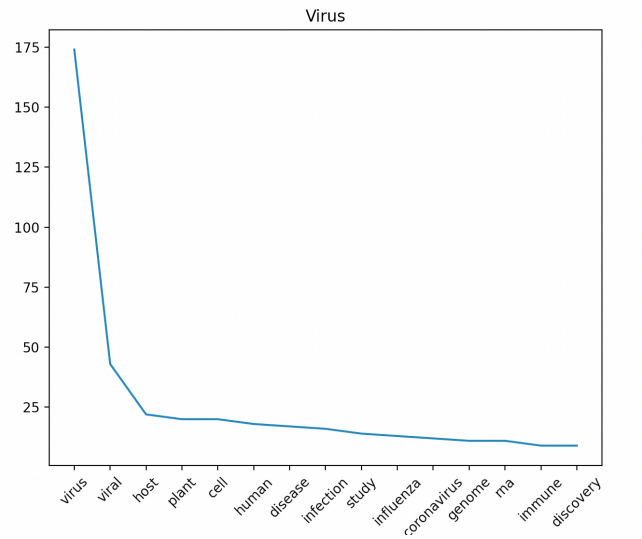
Project4 for the Biomedical Information Retrieval Course

Q56091079 李昱玟

- ## Environment
  - macOS
  - Python3
  - Flask
  - nltk
- ## Github
  - https://github.com/yyyyuwen/sentence_rank
- ## 功能
  - TF-IDF

TF-IDF是一種用於資訊檢索與文字探勘的加權技術，我們將兩篇文章裡頭每一篇小文章的words取出來分別做TF-IDF，再將算出來的分數加回去已經切好的每一行sentences裡頭算加權平均，分數較高者即為較重要的句子。圖為每一篇文章裡頭前五重要的單字的加總(取前15)





  - Data pre-Processing

1. 讀檔

讀取關鍵字為covid跟virus的文章各200篇，取Title、Label、AbstractText

2. 將文章分段轉成Sentences，取Stop word以及做Lemmatizer

首先先將文章切sentences，加上stop word，每個單字做詞性標註，最後再將字還原回去。

4. 建立字典

將各個單字建立字典，形成 {文章 : {sentence : split word}} 的格式。

5. 建立交集的字彙表並計算tf idf

將兩篇文章的共同單字取出來分別再做tf_idf尋找關聯。

6. 計算sentences加權

為了算sentences rank，我將重要單字的tf idf加權起來並做平均，最後取前三重要的句子來做比較。