# Project2 for the Biomedical Information Retrieval Course

Q56091079 李昱玟

- **Environment**
  - macOS
  - Python3
  - Flask
  - Nltk
  - Bootstrap5
- **Github**
  - https://github.com/yyyyuwen/project2-for-BIR-Course
- **主要功能**
  - 顯示XML(download from pubmed)以及Json(download from Twitter)內容
  - 顯示Picture

分別展示出4000、2000、1000筆資料做完Zipf distribution、Porter's algorithm以及經過Stop Word的Zipf distribution、Porter's algorithm

  - 相似字搜尋

打出關鍵字會搜尋出相似度高的詞排成列表，點選列表中的單字可以找出位於文章哪裡並標示出來

- **心得與想法**

此次專案中我們能利用Porter's algorithm 與 Stop words將文本的關鍵字更有效的搜尋，且使用Zipf Distribution看出其中的差異性。Porter's algorithm能夠將**動詞類**的單字字尾去掉去做分析。像是在我下載的文檔中，test, testing, tests, tested就佔了出現頻率前50中的四個，經過Porter's algorithm後能夠將這些單字合併再一起成為一個單字(test)，其頻率就會比之前單獨的還要高。而Stop words主要能表示一些**定冠詞**、**介系詞代名詞**類的單字，因為這些單字在文章中沒有什麼特別的資訊，且出現的頻率遠大於其他的單字。若是將這些單字留下很有可能會干擾其他重要單字的分析，因此我們可以使用Stop words找出那些字詞並把這些單字去掉，這樣一來剩下來的都是較有資訊的單字，方便讓我們在後續做分析。另外我將對不同數量的資料去做Porter's algorithm 與 Stop words並且經過Zipf distribution後去比較其差異性，不過我發現其差異並不大，分佈的位置也極為相近。

關鍵字搜尋的部分，我是將有相關的字都一一列出來，出現頻率由大到小列出來，當我們點選列表單字時，也能夠展示出對應的文本位置並Highlight，再將其傳至前端做展現(使用的是Flask以及Jinja2樣板)。

對於字的搜尋還有許多地方需要改進，例如如何切的更漂亮，可能下次還要再做normalization的處理等等，網站的部分因為是第一次寫也有許多該改進的地方。