IR: Homework 2

Name: 莊富成 Student ID: P77101016

Library: Regular expression, NLTK, tkinter(.ttk), string, glob, pandas, matplotlib

這次將不同步驟拆分成不同階段的檔案，減少無謂時間/IO 浪費

1. Get abstracts from CSV: 將 CSV 資料夾中，助教提供的 csv 檔整合成一個大檔案備著(all10k.csv)，裡面有兩欄位(Title, abstract)。實際上的全部內文利用 NLTK.word_tokenize 拆分成 tokens，並全部小寫化以及用 string.punctutation 剔除 tokenized 後的單獨英文標點符號，寫入 allwords.txt 中。
2. Freq: 計算頻率，製作出三檔案
    a. freq_normal.csv(原封不動的計算頻率)
    b. freq_nostop.csv(用 nltk.corpus 中的 stopwords，剔除掉英文的 stop words，計算頻率)
    c. freq_afterstem.csv(NLTK 有 PorterStemmer，直接用原始檔拿來做 stemming 後，統計頻率)
    d. 將有被 PorterStemmer 處理過的字另外統計起來，寫入 list_of_stemmwords.csv 中。
3. Porter analysis and Merge data: 重新跑一次 PorterStemmer，將會變的字根據 stemmed word 儲存成 list_to_stemwords.csv，後將 step2/3 兩個檔案合併起來成為 analysis.csv，裡面有 stemmed word, 轉變的次數，以及 original words 有哪些的分析表。
4. UI: 將頻率表及分布圖用 UI 表現，表只有前五千字，圖則用全表文字跑圖。最後 analysis 來表示被 Stemmer 處理的字，有幾筆，以及會變成甚麼字。
5. 分析結果：
    a. 在一般狀態下，常用的 stop words 果然是出現次數最多的前幾名，剔除掉 stop words 後，依序為 covid-19 以及 sars-cov-2 這兩個關鍵字，patients, coronavirus, vaccine 在後。
    b. Stemmed 後，”vaccin”大幅增加共 13330 筆，變化的原始文字從名詞(vaccine 單複數)、動詞三式、甚至形容詞以及衍生字都有。