IR: Homework 3

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
   1. freq\_normal.csv(原封不動的計算頻率)
   2. freq\_nostop.csv(用nltk.corpus中的stopwords，剔除掉英文的stop words，計算頻率)
   3. freq\_afterstem.csv(NLTK有PorterStemmer，直接用原始檔拿來做stemming後，統計頻率)
   4. 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：
   1. 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。
   2. 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。