作業（二）

最後編輯時間：2021/11/02

編輯者：侯則瑜

運行環境：

* ubuntu.
* python 3.7 (miniconda).

使用套件：

* selenium, pandas, tqdm, flask, numpy, plotly, levenshtein.
* localtunnel.

安裝套件方式：

* conda create --name hw2 python=3.6 -y
* source activate hw2
* pip install -r requirement.txt
* apt install nodejs
* npm install -g localtunnel

啟動：

* source activate hw2
* sh start.sh

描述：

* 資料夾 script 中為開發 / 測試腳本。
  + 腳本 download.py 根據指定關鍵字下載 pubmed 網站文章摘要。
  + 腳本 frequency.py 根據下載的文件進行文字處理，統計 word 的頻率。
    - 分別使用 1000 、 5000 、 10000 、 50000 篇文本資料來建立頻率表。
  + 腳本 plot.py 是針對指定的文章數量來建立 word frequency bar plot 。
  + 腳本 distance 是針對指定的 word 去計算 edit distance 。
* 檔案 service.py 是基於 flask 開發的 API server 。
  + 使用 sh start.sh 指令來啟動，預設是 8080 端口。
  + 以下為使用範例指令：
    - localhost:8080/plot?size=1000&top=20
      * 根據 1000 篇文章來計算字詞的頻率，圖表中顯示頻率最高的 20 個單字。
    - localhost:8080/search?word=covid&top=50
      * 分別從 1000 、 5000 、 10000 、 50000 建立的頻率表計算 covid 單字與其他單字的 edit distance ，顯示最相似的 50 個單字。