作業（一）

編輯時間：2021/10/12

侯則瑜

Environment: Ubuntu
Language: Python
Library: os, json, xml, pandas, re, nltk

給定一個資料夾，根據該資料夾進行文件逐一搜尋/處理，如果檔名有包含 xml ，就執行 xml 資料格式處理，如果有包含 json ，就執行 json 資料格式處理。

- xml
  - 每個文件中有多篇小文章，置於 <PubmedArticle> 中，計算有幾個，即代表有幾篇小文章。
  - 將所有的字串從 tag 中提出，將多餘的空白符號、換行符號刪除，計算所有字元的數量。
  - 去掉標點符號後，用空格符號對字串進行切割，計算字的數量。
  - 使用 porter stem 作法將 word 進行時態上的還原，計算總共出現幾個單字（unique）。
  - 非逗號、句號、問號、驚嘆號、分號刪除，將多餘空白刪除，根據空白進行切割，計算長度代表句子數量。
- json
  - 每個文件中有多個小篇幅，用一個 dictionary 包住，所以計算其中的有幾個 dictionary ，來計算有幾個小篇幅。
  - 提取所有的 key 中所包含的字串內容，整合在一起，去掉多餘的空白符號，計算文章中有幾個字元。
  - 去掉標點符號，用空白符號進行切割，計算有多少字。
  - 使用 porter stem 作法將 word 進行時態上的還原，計算總共出現幾個單字（unique）。
  - 非逗號、句號、問號、驚嘆號、分號刪除，將多餘空白刪除，根據空白進行切割，計算長度代表句子數量。

將上面算出的數據以及文章內容整合至 dataframe 當中，定義一個方法 search 來根據關鍵字尋找字詞，尋找方法去逐一去搜尋 dataframe 中的文章內容欄位，輸出內容包含這個關鍵字在每個搜尋的文章中是否有出現，以及出現的次數，格式為 dataframe 。