Programming Assignment 1

姓名：黃奕瑋 學號：R14725055

1. 執行環境

- Jupyter Notebook (.ipynb 檔案)

2. 程式語言

- Python == 3.10.7
- nltk == 3.9.1

3. 執行方式

1. 所需套件：pip install nltk == 3.9.1
2. 用支援 Jupyter 格式的 ide 開啟 pa1.ipynb (我是使用 vscode)
3. 點擊 Run All 執行所有 cell
4. 程式會產出 result.txt，該檔案為最終答案。

4. 處理邏輯說明

1. 先使用 open("./1.txt", "r") 讀取 txt 檔案
2. 去除標點符號，使用 replace() 將可能出現的標點符號替換成空值。
3. 將字轉成小寫，使用 lower() 將文本字元轉成小寫。
4. 斷詞，因為文章為英文，所以用空格來區分每個單詞，以 split() 實作，並將分割後的單詞存進 list 中。
5. 字尾處理，如果出現 's 等字尾出現，我會將逗號以及逗號後的字全部移除，只保留逗號前的詞，例如: she's 只保留 she。
6. 連接詞處理，連接詞我會選擇將有無連結詞的版本都儲存，避免誤判，例如: two-hour 會同時保留 two-hour 以及 twohour。
7. Stopword removal，從 nltk.corpus 下載英文的停用字，接著將 list 中的 stopword 移除。
8. Stemming，使用 Porter's algorithm 做 stemming，從 nltk.stem import PorterStemmer，接著在依序將 list 中的詞做 stemming。
9. 輸出檔案，使用 join() 將 list 中的每一個單詞以空格分隔開，並組成 string，最後用 f = open('result.txt', 'w') 寫入 txt 中。