12/05 隨堂練習7

Liam 最近發現 Irene 總是看起來若有所思，但問 Irene 她也不說怎麼了，只會說沒事。今天，Irene 的日記被她不小心打開放在桌上就離開了，Liam 看到了其中一些內容，請幫Liam影響 Irene 心情的因素，幫助 Liam 分析 Irene 到底怎麼了？

請利用 Genmini 幫 Liam 分析 Irene是真的沒事嗎？還是另有隱情呢？

目標:

(1) 讀取 Irene的心情日記，利用 jieba 套件斷句

(2) 找出 Irene 最近在意的關鍵主題(他最常提到的話題)

輸入格式

```
##以下是嘗試區
from google.colab import auth
import gspread
from google.auth import default
import pandas as pd
import jieba
from snownlp import SnowNLP
from collections import Counter
import re

auth.authenticate_user()
creds, _ = default()
gc = gspread.authorize(creds)

# read data and put it in a dataframe
# 在 google 工作表載入 gsheets
gsheets = gc.open_by_url('https://docs.google.com/spreadsheets/d/1WWIx4CEWzp3h9jUER-3KzdI8jUfDqWBvFxqRjtk3lQA/edit?usp=sharing').sheet1
dicts = gc.open_by_url('https://docs.google.com/spreadsheets/d/1WWIx4CEWzp3h9jUER-3KzdI8jUfDqWBvFxqRjtk3lQA/edit?usp=sharing').get_worksheet(0)
dicts = dicts.get_all_records()
dicts = pd.DataFrame(dicts)

# 讀取所有數據
```

```
rows = gsheets.get_all_records()
df = pd.DataFrame(rows)

# 使用 Jieba 斷詞
df['Tokenized'] = df['內容'].apply(lambda x: list(jieba.cut(x, HMM=True)))

# 展平成所有詞語的列表
all_words = [word for tokens in df['Tokenized'] for word in tokens if len(word) > 1]

word_counts = Counter(all_words)

===
## 請修改停用詞、專有名詞列表與替換詞彙區
#定義停用詞列表(以下為舉例，可以做修改)
stop_words = set(['所以', '好', '因為', '大家'])

# 定義專有名詞列表（確保不被拆開）(以下為舉例，可以做修改)
proper_words = set(['專有名詞'])

# 定義需要替換的詞彙（如「天」替換為「天氣」）
replacement_dict = {'天': '天氣'}

===
# 合併專有詞彙為一個正則表達式
proper_words_pattern = '|'.join([re.escape(word) for word in proper_words])

# 將專有名詞標記為一個整體（不拆開）
all_words_str = ' '.join(all_words)
all_words_str = re.sub(r'(' + proper_words_pattern + r')', r'\1_', all_words_str)

# 進行斷詞（這裡簡單使用空格分開，實際情況中可以用適合的分詞工具）
split_words = all_words_str.split()

# 去掉專有名詞標記（即移除下劃線）
split_words = [word.replace('_', '') for word in split_words]

# 替換「意度」為「滿意度」
split_words = [replacement_dict.get(word, word) for word in split_words]

# 計數
word_counts = Counter(split_words)
```

```
# 排除停用詞
filtered_word_counts = {word: count for word, count in word_counts.items() if word not in stop_words}

# 將詞頻轉為 DataFrame 並排序
word_freq_df = pd.DataFrame(filtered_word_counts.items(), columns=['Word', 'Frequency'])
word_freq_df = word_freq_df.sort_values(by='Frequency', ascending=False).reset_index(drop=True)
word_freq_df.head(5)
```

輸出格式

請於google 表單中繳交

https://forms.gle/EhYTtXCJ9DyZb5hB6

Hint

請參考這個檔案進行修改(記得要複製回自己的雲端窩~)

https://colab.research.google.com/drive/1xdFOxPmTTeqrTdHrgdcB88eSMXy90WdK?usp=sharing