隨堂練習6

第一堂課時，彥谷學長分享了精彩的經驗分享，但是 Peter 很可惜的第一堂無緣錯過。所以想請各位透過 GitHub 中大家的心得分享擷取彥谷學長分享的核心概念!

目標:

(1) 把GitHub留言利用 jieba 套件拆字詞

(2) 找出心得中頻率最高的五個字

輸入格式

```
from google.colab import auth
import gspread
from google.auth import default
import pandas as pd
import jieba
from snownlp import SnowNLP
import pandas as pd
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/10D1ji7egtNBK
dicts = gc.open_by_url('https://docs.google.com/spreadsheets/d/10D1ji7egtNBK4O
dicts = dicts.get_all_records()
dicts = pd.DataFrame(dicts)
```

```
# 讀取所有數據
rows = gsheets.get_all_records()
df = pd.DataFrame(rows)
====
請繳交此處之後的程式碼即可
```

輸出格式

	Word	Frequency
0	嘗試	35
1	努力	32
2	學習	31
3	興趣	28
4	挑戰	25

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/1d3lYzDGe5dgmYDVO1D91nwJtr4l6yhKp?usp=sharing

Answer&步驟

https://colab.research.google.com/drive/19SRBk3CCMH51EqaAK2zpABFiOVw5fN1D?usp=sharing

```
# 使用 Jieba 斷詞
df['Tokenized'] = df['留言'].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) >

# 排除停用詞（這裡提供一個簡單的停用詞列表，可根據需要擴展）
stop_words = set(['所以', '好', '因為', '大家', '的', '是', '了', '我', '也', '在', '和', '就', '
filtered_word_counts = {word: count for word, count in word_counts.items() if w

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