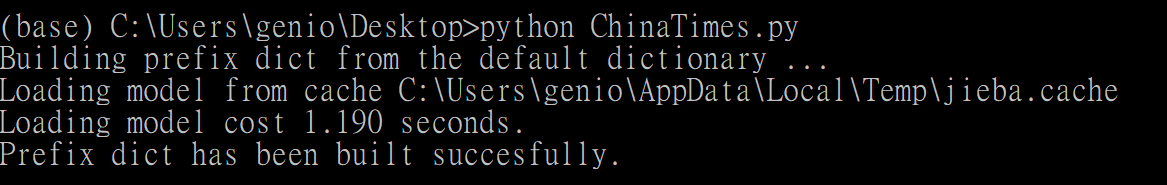
**快速操作**

1. **準備好ChinaTimes.py檔(請記得要修改py檔裡的辭庫(UTF-8)和文章(ANSI)路徑**

****

**\*文章可以多篇放入一個text檔，節省改路徑次數**

1. **在Anaconda Prompt進入放置py檔正確的目錄後執行**



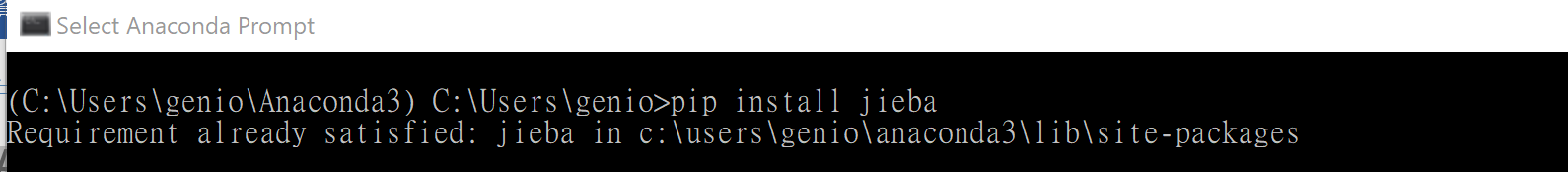
1. **最後呈現檔案如下**

****

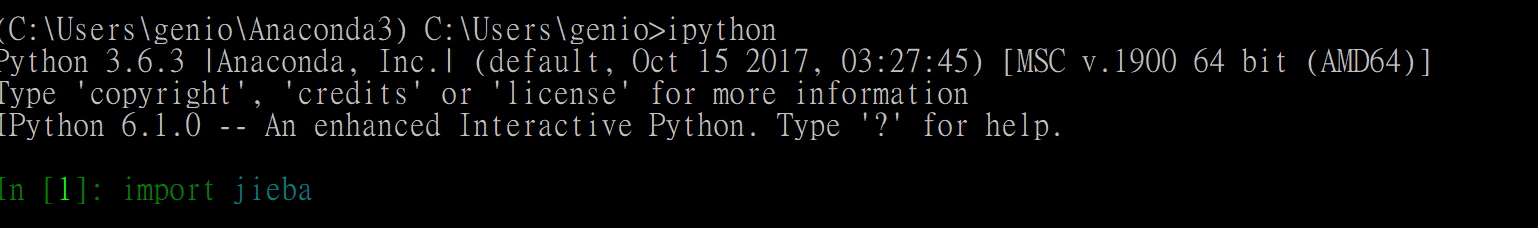
**注意:文章內容，有時文章裡會有不相關的文字連結，或就算頻率高亦不是造成股價漲幅變動大的主因，請各位利用自己的專業判斷去找關鍵字~**

**在Anaconda Prompt操作結巴演算法**

1. 安裝結巴套件在Anaconda Prompt裡



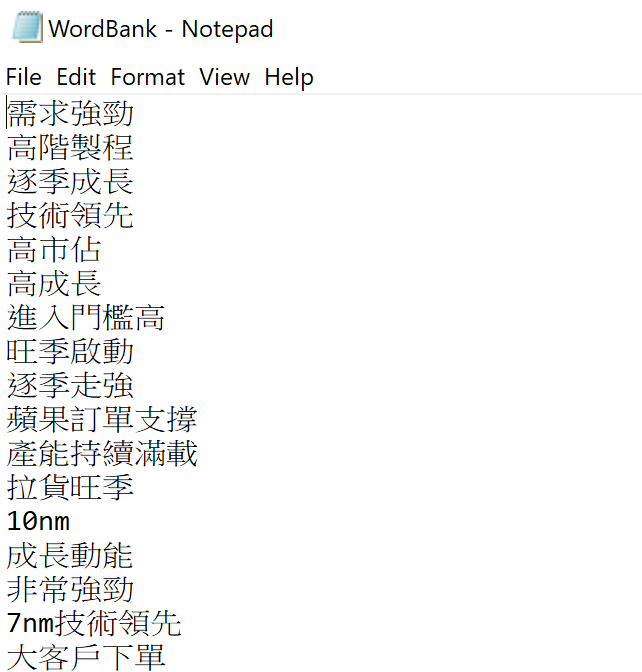
1. 打Ipython 進入python環境, 並import jieba



1. Import結巴後，將準備好的辭庫一起load進來

與之後要段詞斷句的新聞檔不同，此txt檔的encoding要設為UTF-8存檔





1. a)準備Notepad將複製的文件放入文本，放文本前可以先處理資料

b)可以留發布時間點以利之後對XQ15分線漲跌幅資料

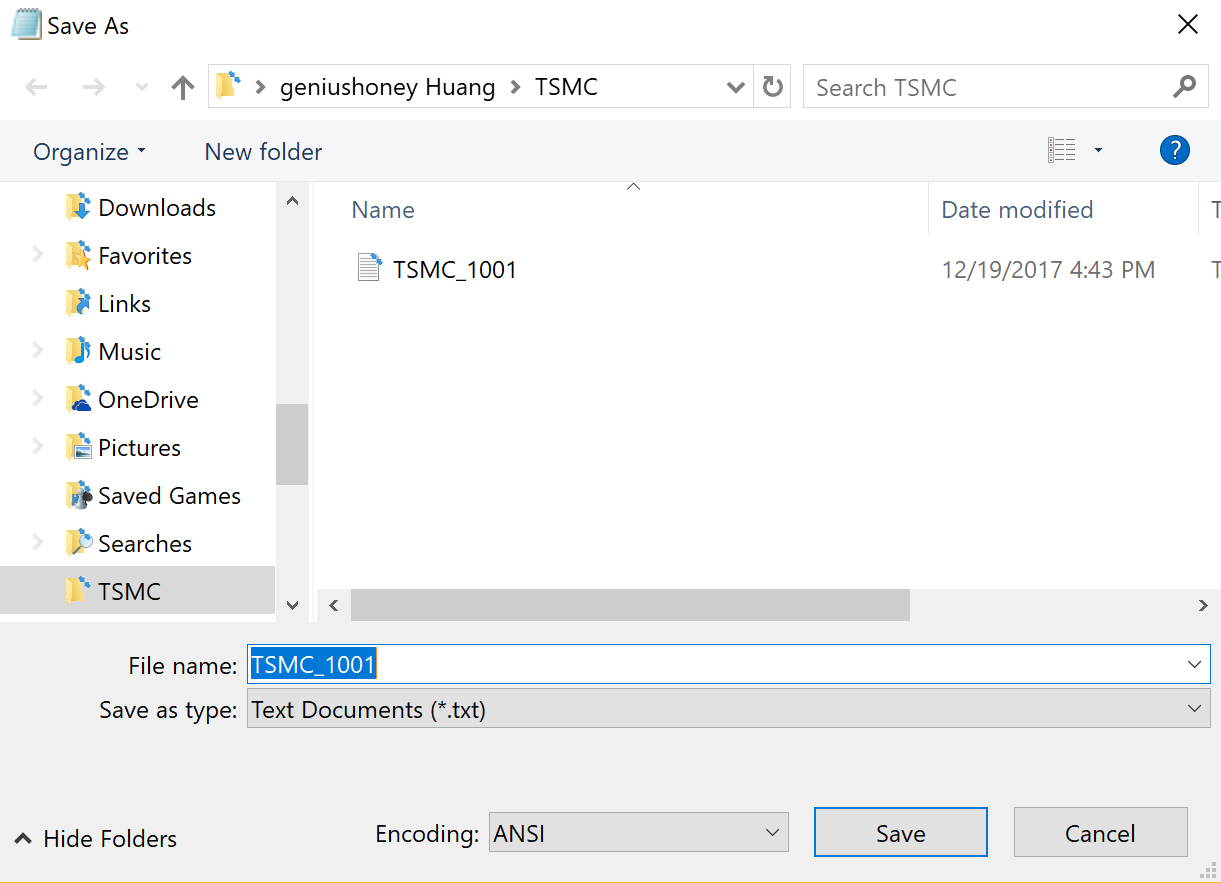


c) **將text file的encoding存成ANSI(預設)並記得自己儲存路徑**

檔案存在以下路徑 C:\Users\genio\TSMC

d)資料命名請以 TSMC\_XXXX命名 (XXXX代表幾月第幾篇以4位數)每個人會給文章編碼

ex. TSMC\_1001 (10月第1篇文章)

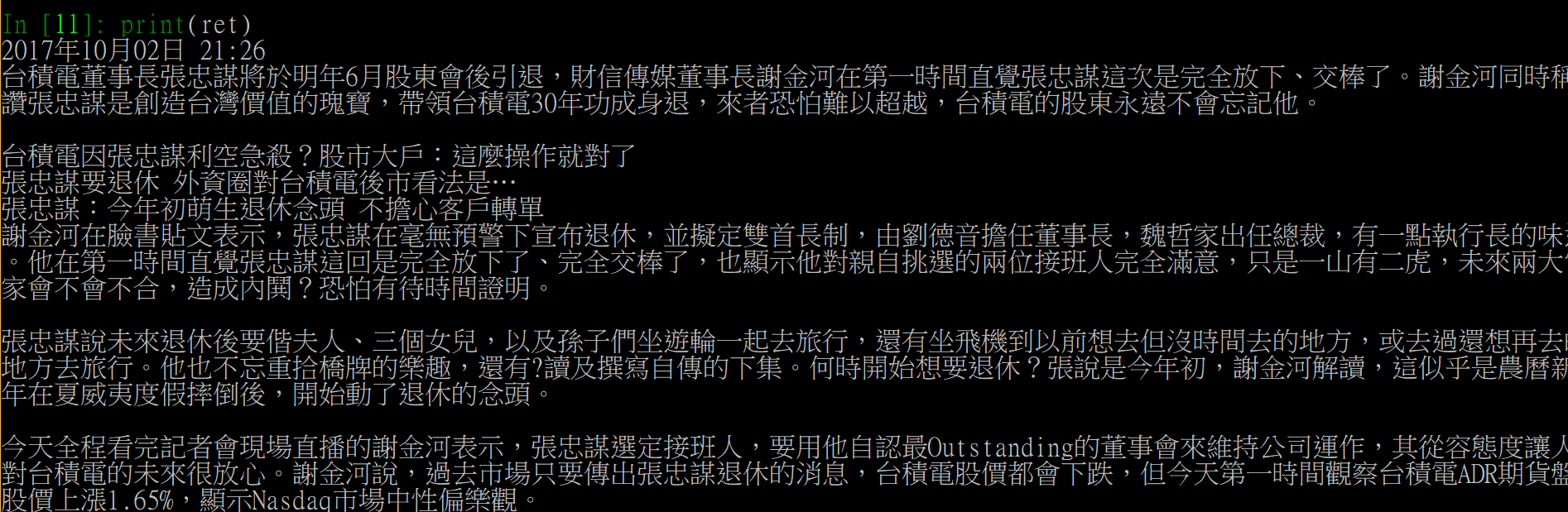


1. 輸入剛文檔所在路徑給python讀取文件

雙引號部分要自己打，word文本會自動變全形

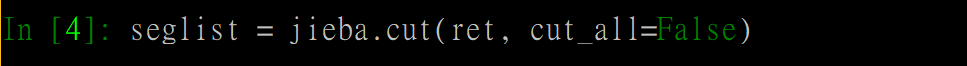
ret = open(“C:\\Users\\genio\\TSMC\\TSMC\_1001.txt”, “r”).read()



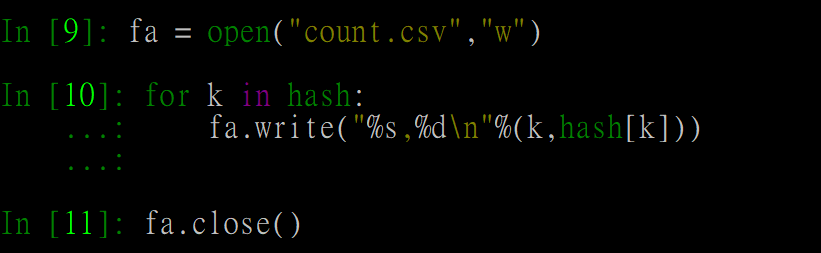
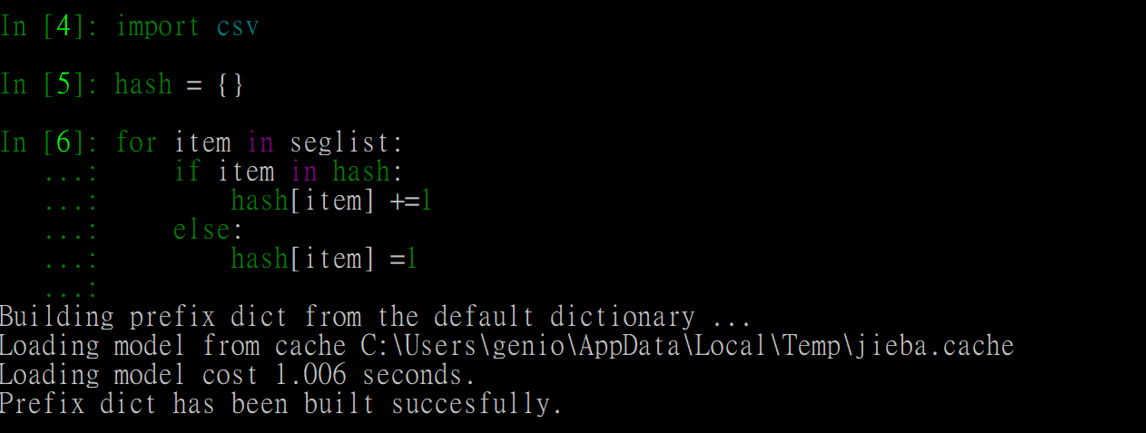


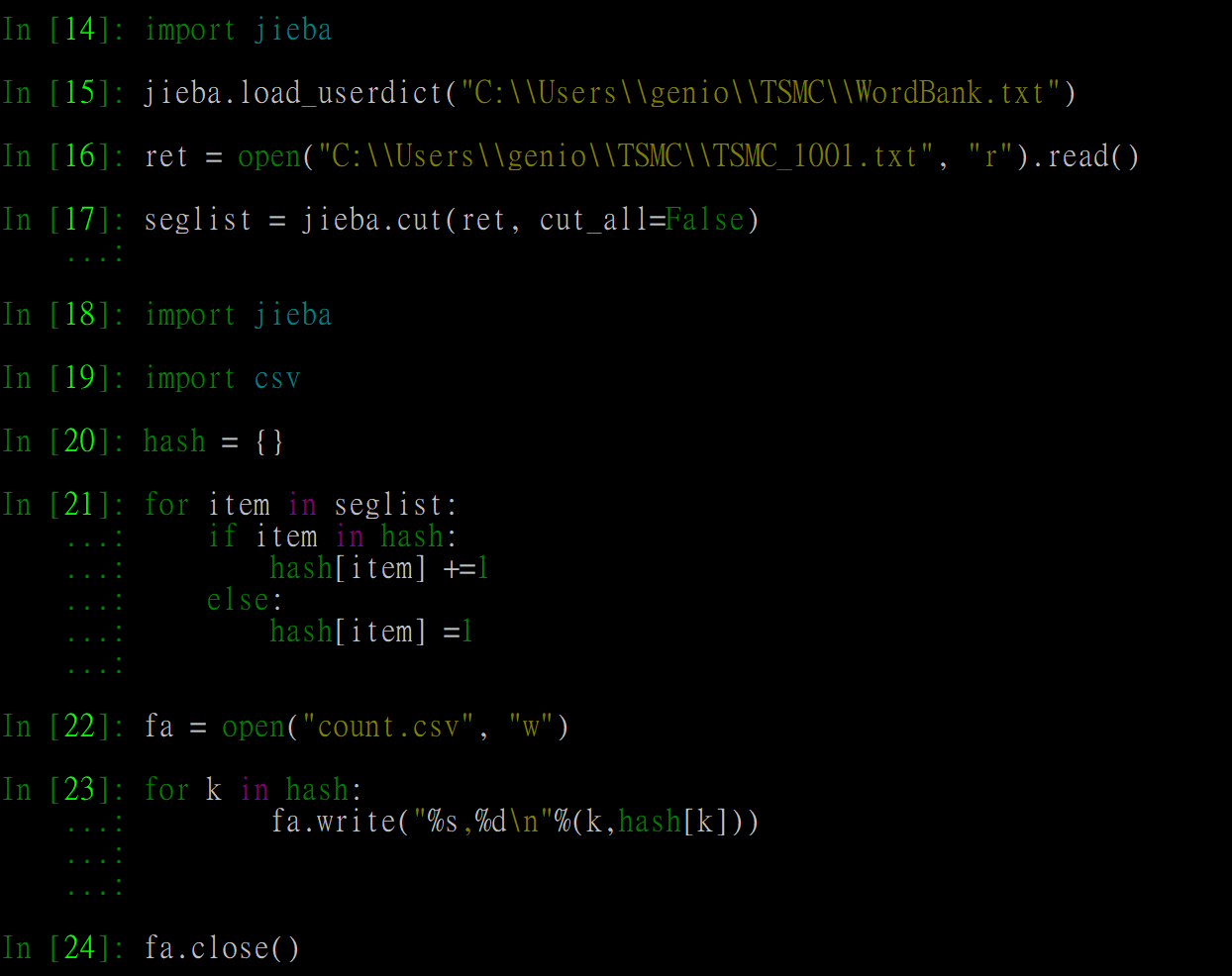
1. 套用結巴函數並做拆字的動作，並設成一個變數 seglist

seglist = jieba.cut(ret, cut\_all=False)

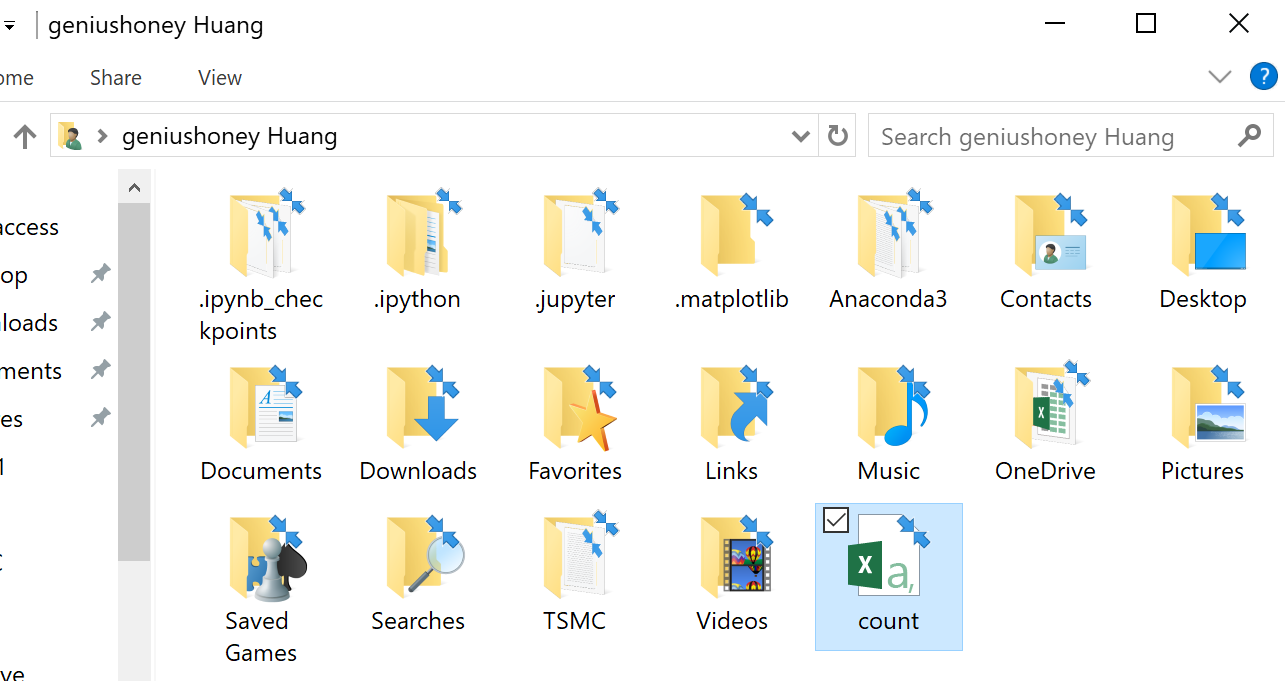


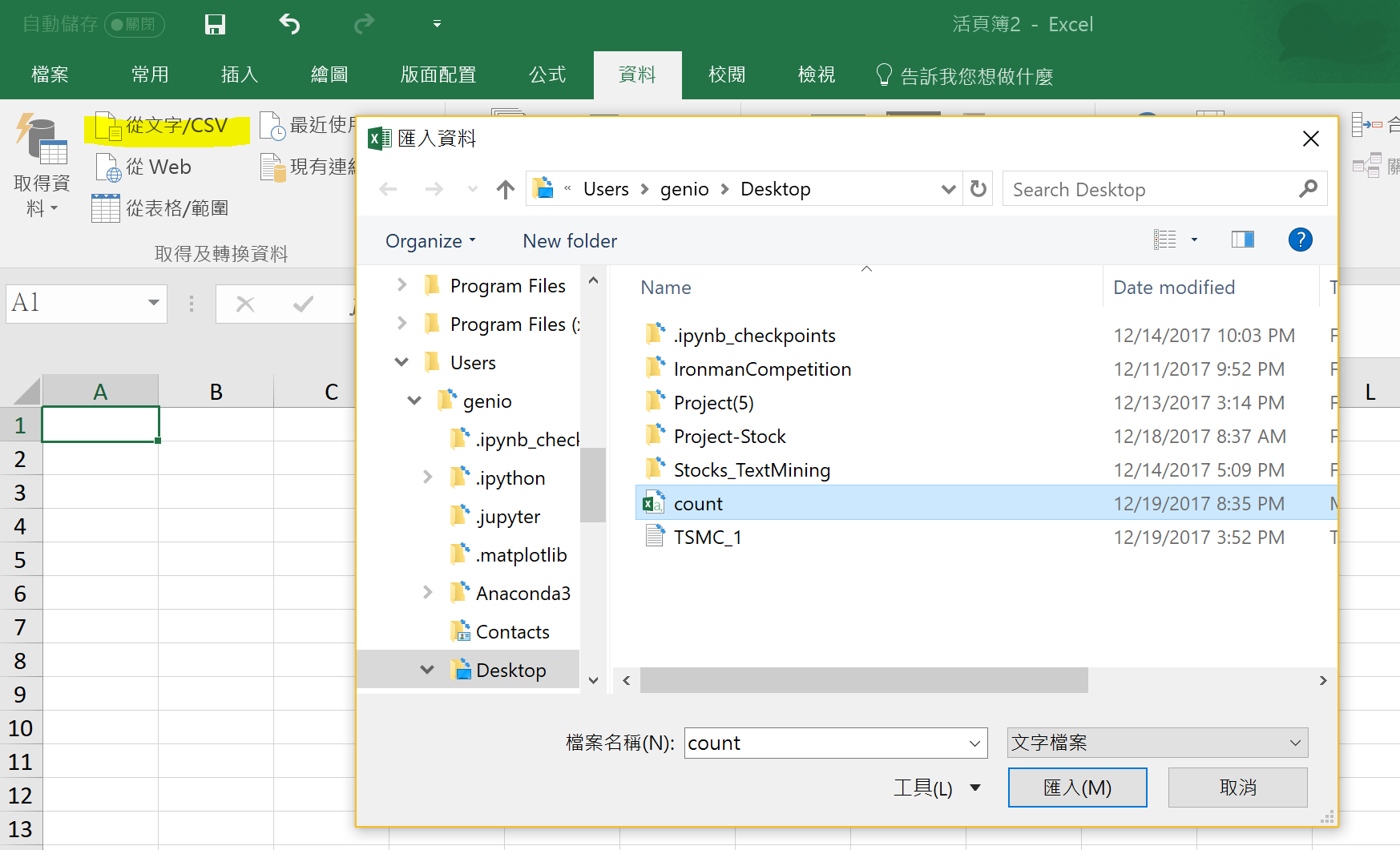
1. 寫迴圈計算該文章段詞字頻，並輸出成csv檔





1. 接下來搜尋count.csv檔案(先不要直接開啟它，因為有編碼問題(當時存的txt檔是ANSI，在excel開啟時要勾選big5)，開啟一個新的excel並把資料import至excel

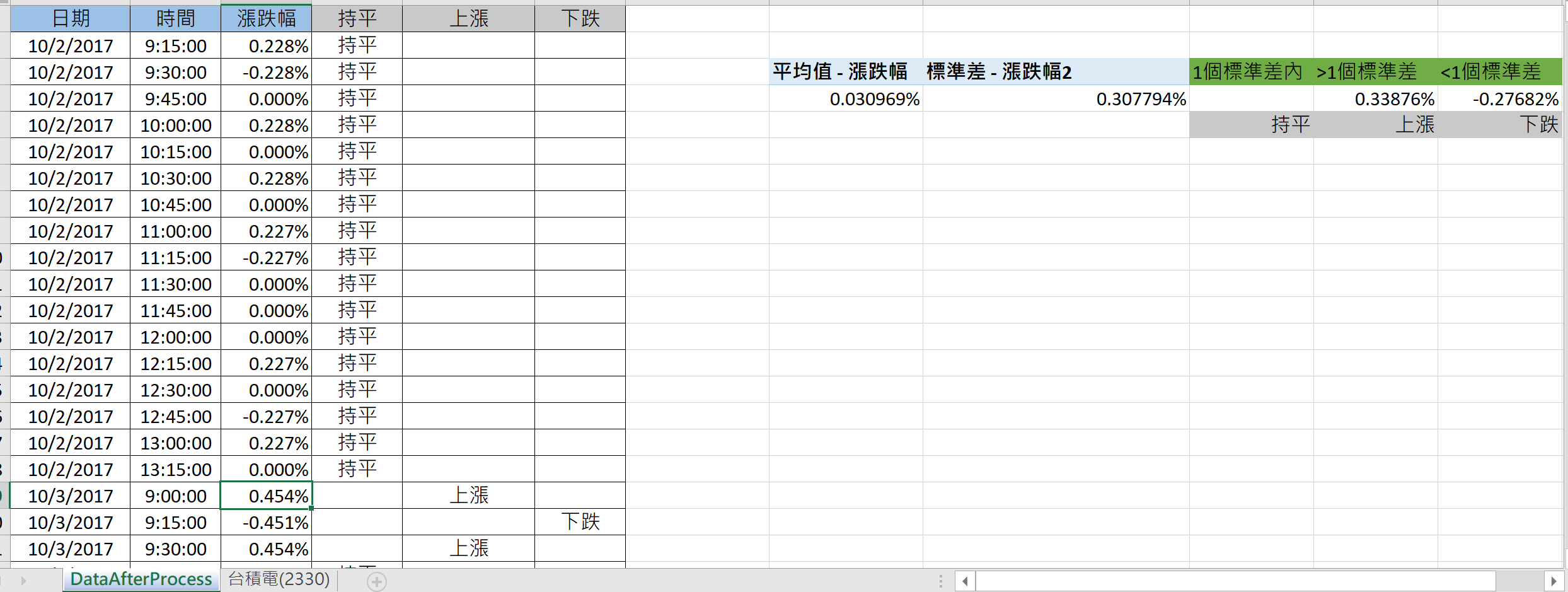




選擇950(Big5)



1. **顯示結果 -> 自行排序-> 篩選關鍵字 -> Target填空 (2330Updated1001-1031資料)**



若print出來會成下圖

