Word Segmentation

一、應用目標

將中文文本斷詞並過濾部分詞性詞，結果可使用於Text Mining。

二、環境要求:

Python 3.9 (pkl檔要求，若使用其他格式則3.x即可)

安裝CKIP Tagger

三、事前準備流程

* 文本爬蟲
  + 重大訊息
* CKIP Tagger 安裝
  + 跟著github的文件即可成功完成
  + [ckiplab/ckiptagger: CKIP Neural Chinese Word Segmentation, POS Tagging, and NER (github.com)](https://github.com/ckiplab/ckiptagger)

四、事前準備的檔案

* 文本
  + info\_data/material\_information\_2021.csv (2021的重大訊息)
  + <https://drive.google.com/drive/folders/1nGFJEjWlKoR4dZq2NtsNpk3V97vuX6-u?usp=sharing>

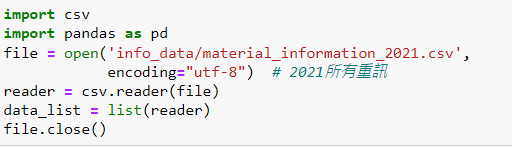
五、介紹

使用重大訊息作為分析文本。

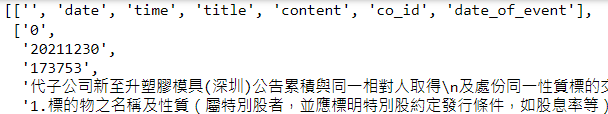
使用CKIP Tagger作為斷詞工具。

六、實作流程:

1.讀取文本資料



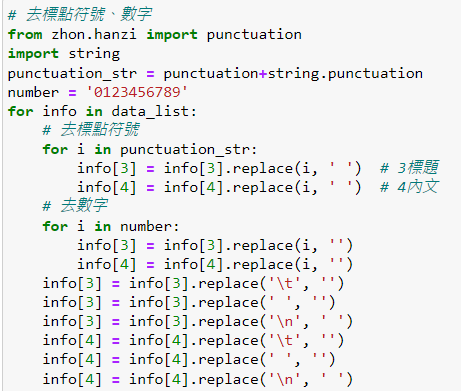
data\_list資料會是這個樣子



2.做文字的清理

因為後續的text mining是單純使用文字，並以詞為最小單位作為input進行分析、訓練，因此會拿掉數字與標點符號。

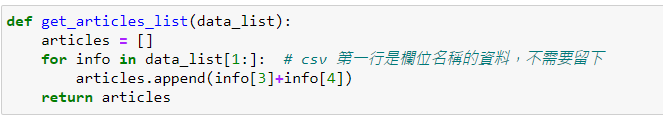
這裡要注意，根據不同的文本會有一些其他的符號需要去除，這部分需要依資料的狀況做微調。



3.整理一下資料

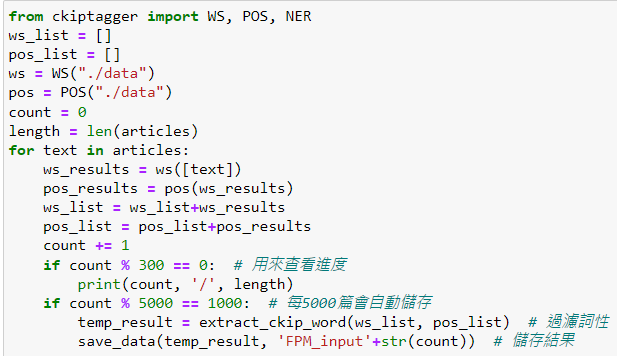
將要斷詞的文本單獨拿出來。





4.斷詞

斷詞，斷詞篇數每5000篇會自動過濾詞彙並儲存。



七、Source code

[stock/Word\_Segmentation.ipynb at main · samwang1228/stock (github.com)](https://github.com/samwang1228/stock/blob/main/%E6%B2%9B%E9%8C%A1/Word_Segmentation.ipynb)