**NLP學習筆記\_資料處理篇**

自然語言處理Natural Language Processing(NLP)為語言學，電腦科學及人工智慧等三大科技文明領域集結之應用。

常見的NLP應用包括:

1.文件分類/聚類:利用文件(如:郵件，網路評論等)中的文字意涵，進行分類及聚類。

2.電腦翻譯:就是我們常用的自動翻譯應用。

3.搜尋引擎:以文找文，以自然文句找文等非關鍵字搜尋。

4.語音辨識:Siri等Speech to Text(STT)應用，當STT走完流程後，還要進行ASR(語音識別），NLU(語意分析)，DST（對話狀態控制）， 與NLG(對話生成）。

5.聊天機械人(Chatbot):這是**自家卓越的產品**，大家一定都了解，就無需贅述了。

為了讓深度學習模型能夠處理自然語言，方法的是將語言轉換成數字形式，而常見的做法是one-hot編碼(one-hot encoding)與詞向量(Word vector)。

**自然語言的預處理**

自然語言資料預處理作業程序有下列幾種:

1.斷句或斷字(tokenization)

2.字母轉換小寫或全形轉半形

3.移除停用字(stop word)

4.刪除標點符號

5.字根提取(stemming): 字根提取能將意義差不多的單字歸類成同一token，例如house跟housing字根都是hous。

6.n-gram詞彙處理:例如詞New York含有兩個字，是2-gram，我們會獨立給一個token。

**自然語言預處理範例**

使用Cola開放平台(可參考文末說明)練習此次的課題。

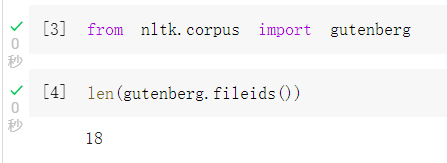
實作流程如下:

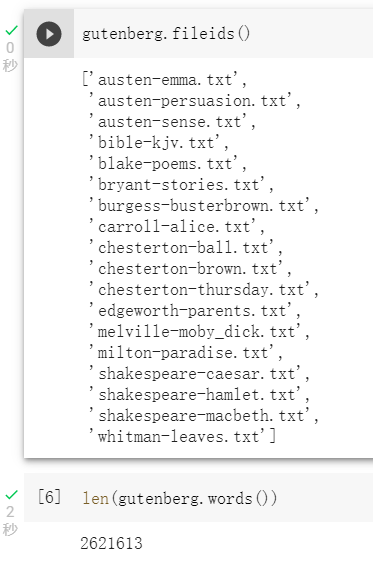
#### 步驟一， 載入相關模組，大多數的模組為**nltk**自然語言工具包與**gensim**自然語言函式庫。



#### 步驟二，斷句與斷字，標記token

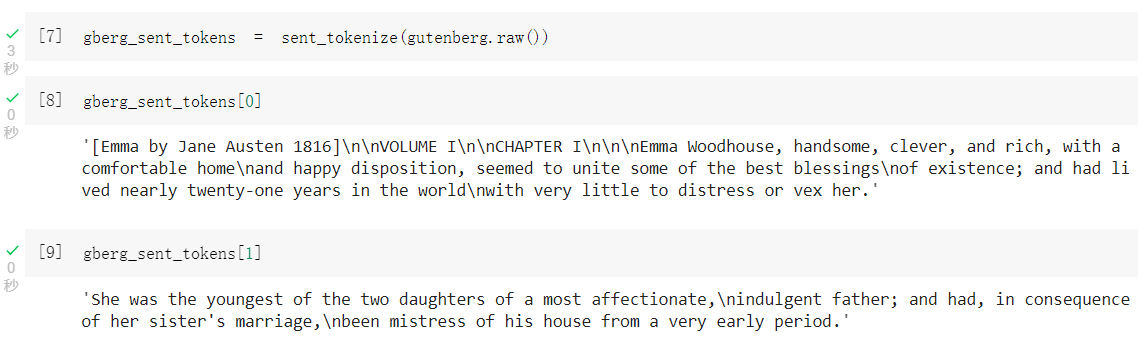
使用Project Gutenberg電子書的一部分作為語料庫。



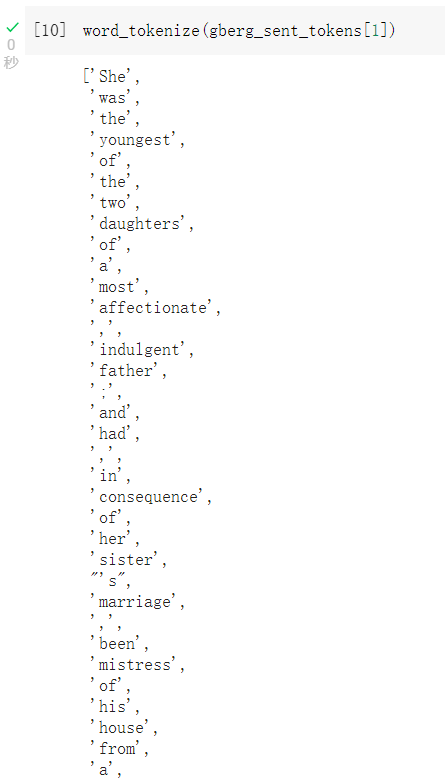


18部文學作品，合計2621613個單字。

#### 步驟三，斷句處理

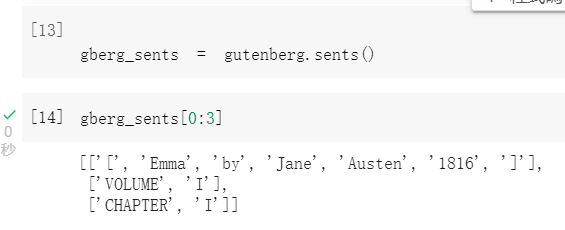
使用sent\_tokenize()函式將文字檔的每個句子分開。

#### 步驟四，斷字處理

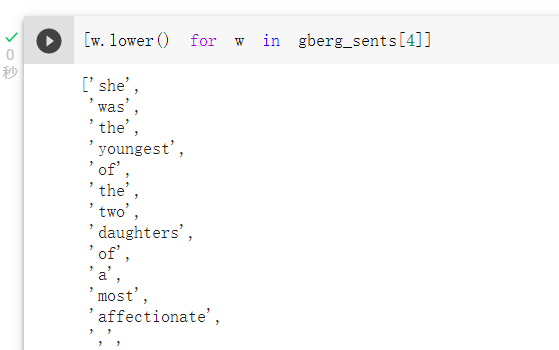


使用nltk的word\_tokenize()將句子拆分成一個個不含空格也沒換行符號的單字串列。

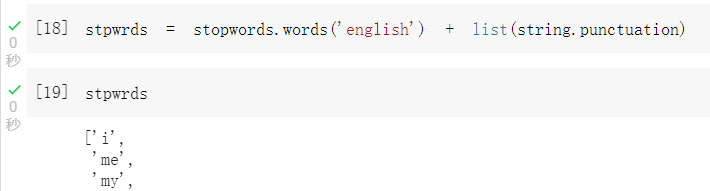
#### 步驟五，一次做完斷句斷字

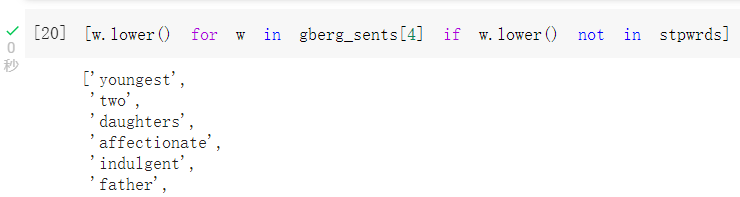


#### 步驟六，將大寫字母轉成小寫

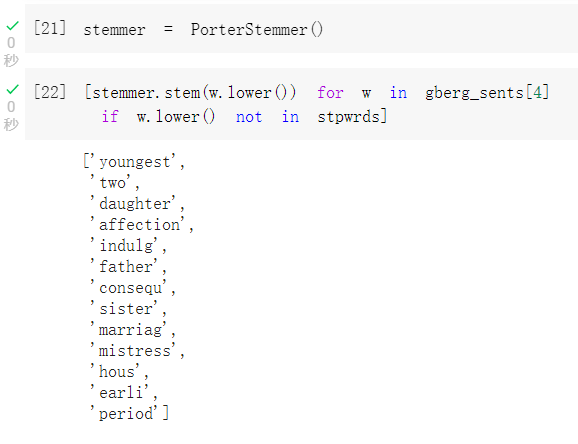


#### 步驟七，移除停用字與標點符號

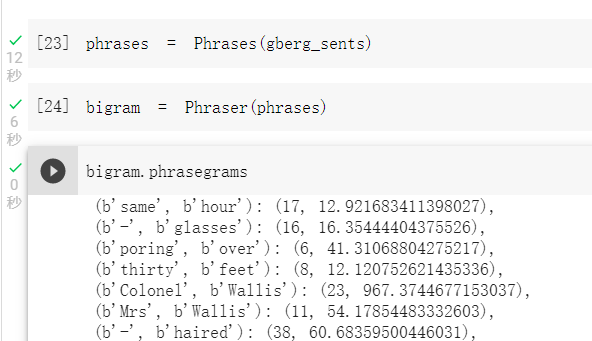




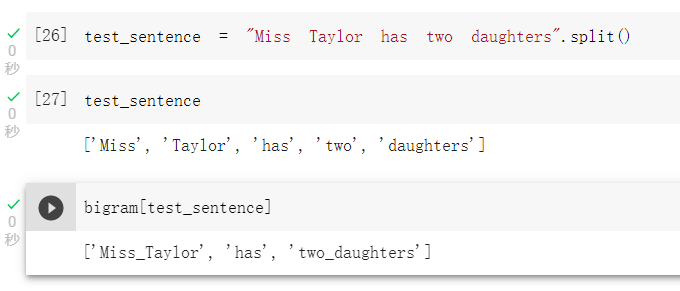
#### 步驟八，字根提取



#### 步驟九，分析n-gram 詞彙並串成單一詞彙



測試看看一小段

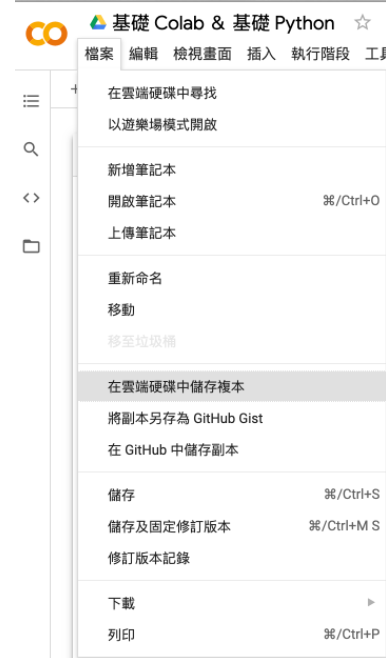


#### **附錄:前置作業Colab環境準備**

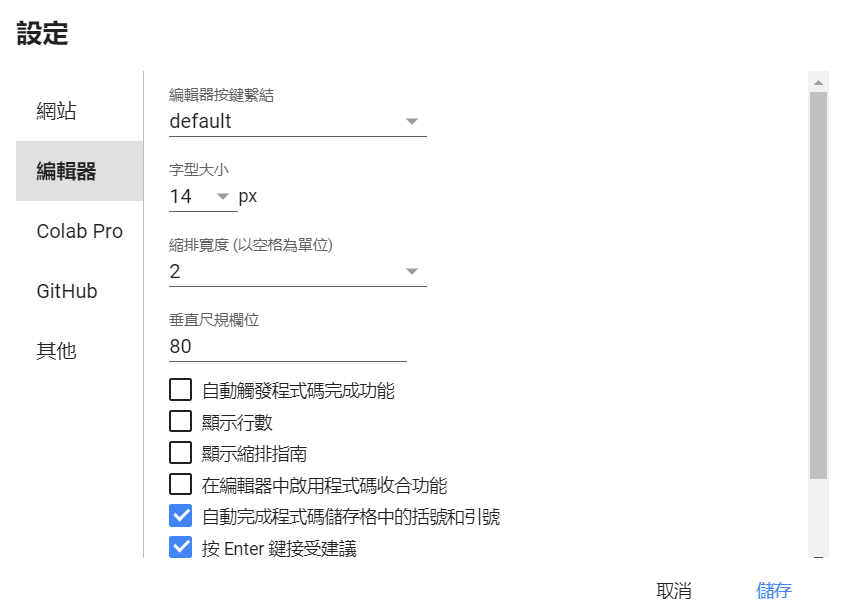
1.必須有自己的google帳號:使用自己的個人的 Google 帳號才可以執行自己 Colab上的程式。

2.先下載 Chrome瀏覽器，並登入 Google 帳號。

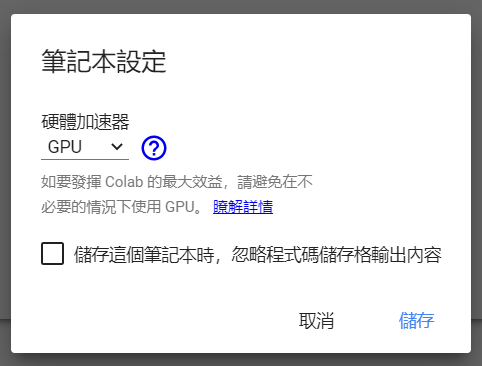
3.建立副本至個人目錄，否則不能存檔。



4.在⼯具 > 設定>編輯器中,反勾選「⾃動觸發程式碼完成功能」。



5.(必須做)設定執行階段類型。



6.療癒模式(非必須)，在⼯具 > 設定>其他, 勾選 「柯基⽝模式」和「貓咪 模式」，就有可愛的貓 貓狗狗來陪你寫程式了。



