**NLP學習筆記\_詞向量空間篇**

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)。

**自然語言的詞向量處理**

自然語言資料經過預處理作業程序後，我們得到一個梳理完畢的語料庫(有沒有很熟悉呢?)，接下來必須將語料庫中的每個字詞都與一個多維度向量空間(Vector Space)產生對應位置。

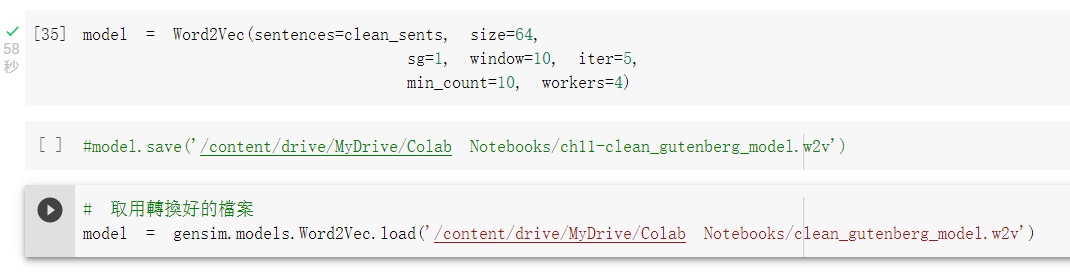
語料庫中的字詞在一開始僅能隨機定位，而深度學習模型會在分析語料庫中各字詞的關係(較常與某個特定字詞一起使用)後，便會移動字詞的位置，例如粉紅色與紫色在常見使用句上都因為是顏色的關係，所以會靠得較近。

**自然語言詞向量空間處理範例**

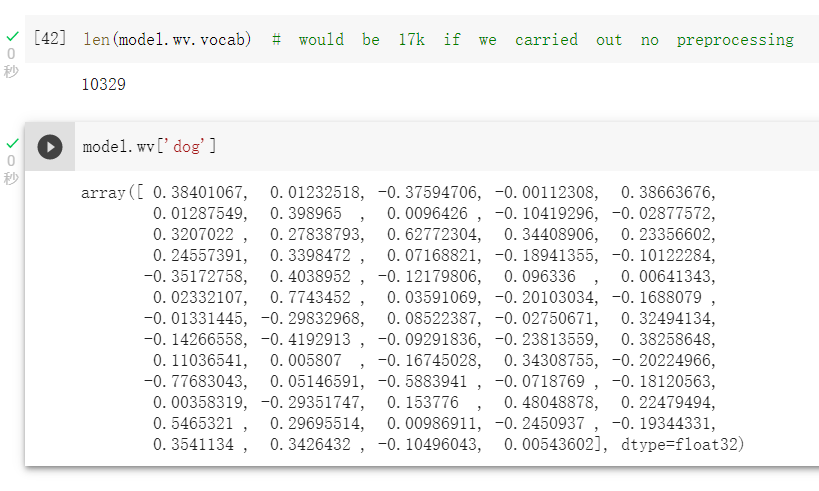
使用Cola開放平台(可參考文末說明)練習此次的課題。

實作流程如下:

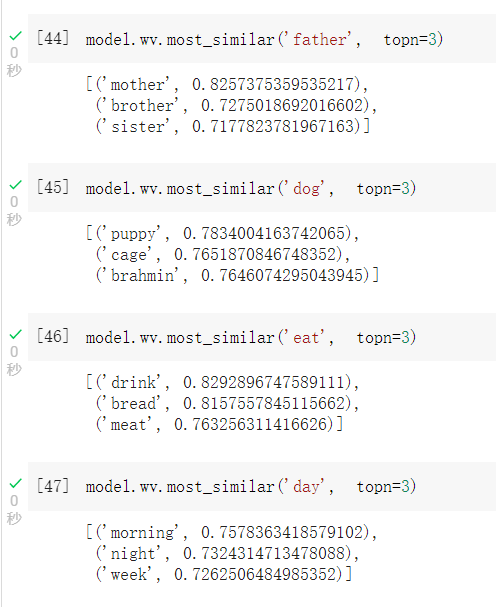
#### 步驟一， 載入相關模組，使用**gensim**自然語言函式庫的word2vec模組快速建立詞向量空間。



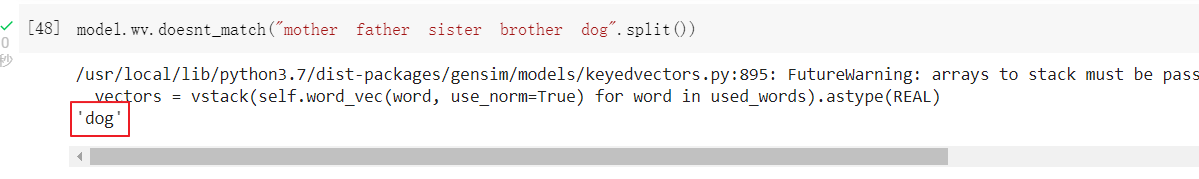
#### 步驟二，查看詞向量空間的內容



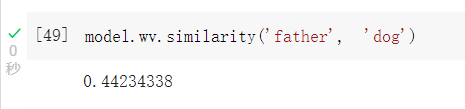
使用most\_similar()查看特定單字，看看附近的字是否意義相近，以評估生成的詞向量空間效果。



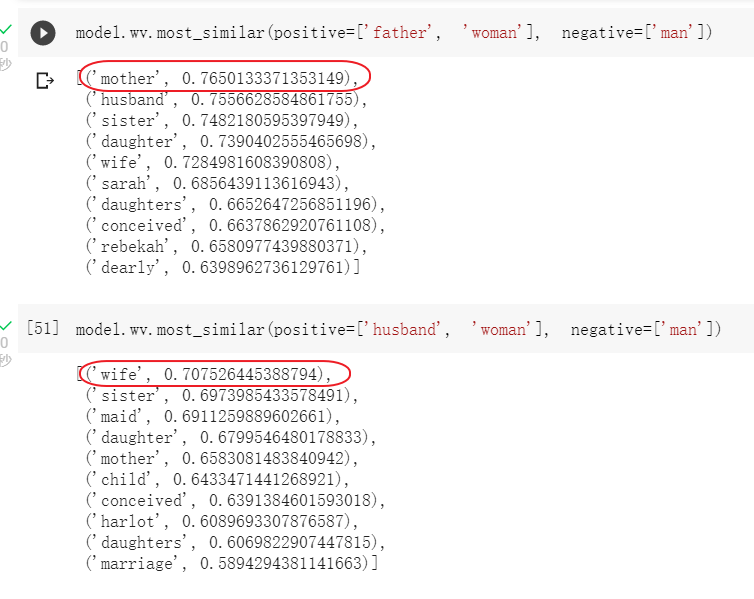
可以分析某個句子中，哪個字最與眾不同(此例為dog)



可以計算兩個字間的相似性分數

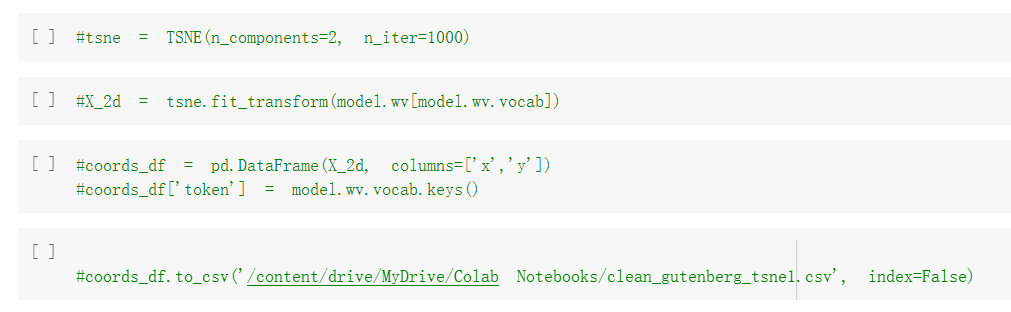


還可以對詞向量做猜題運算，例如父親詞向量-男人詞向量+女人詞向量，或是丈夫詞向量-男人詞向量+女人詞向量

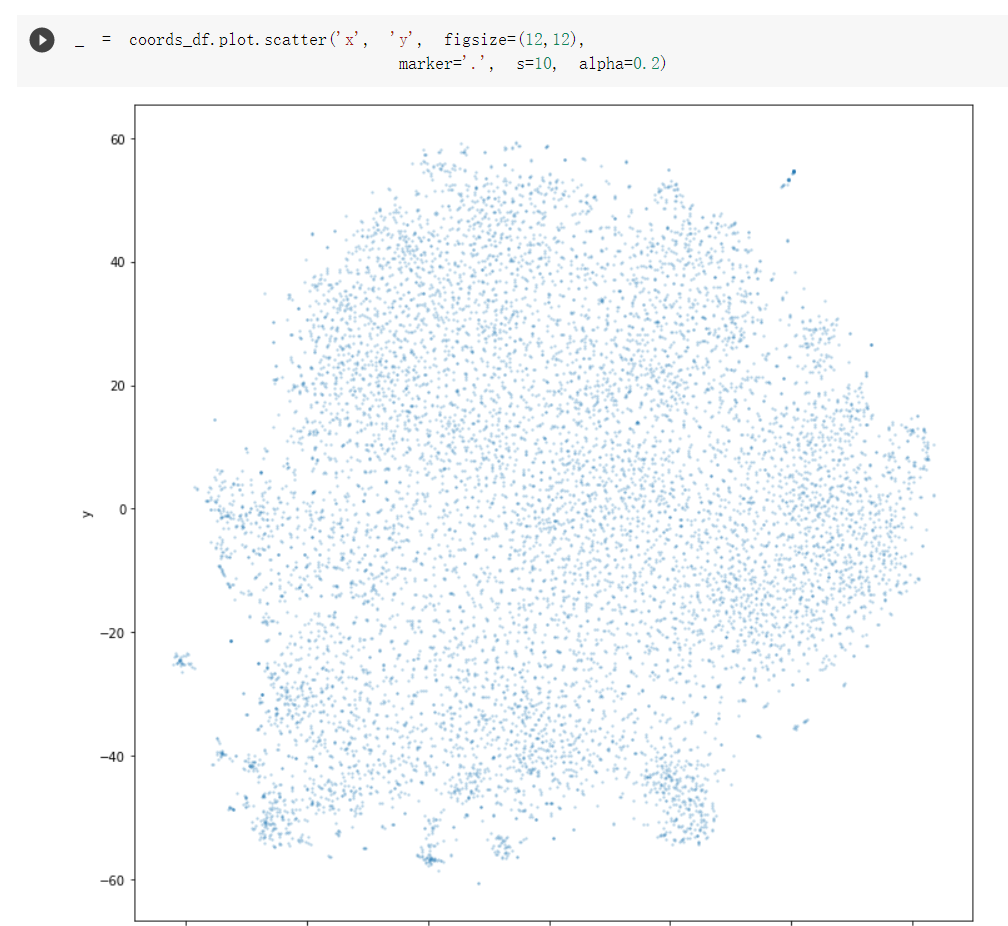


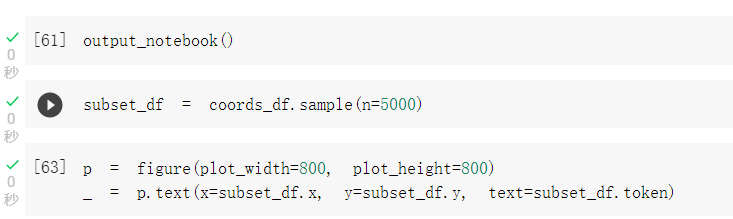
#### 步驟三，描繪詞向量空間

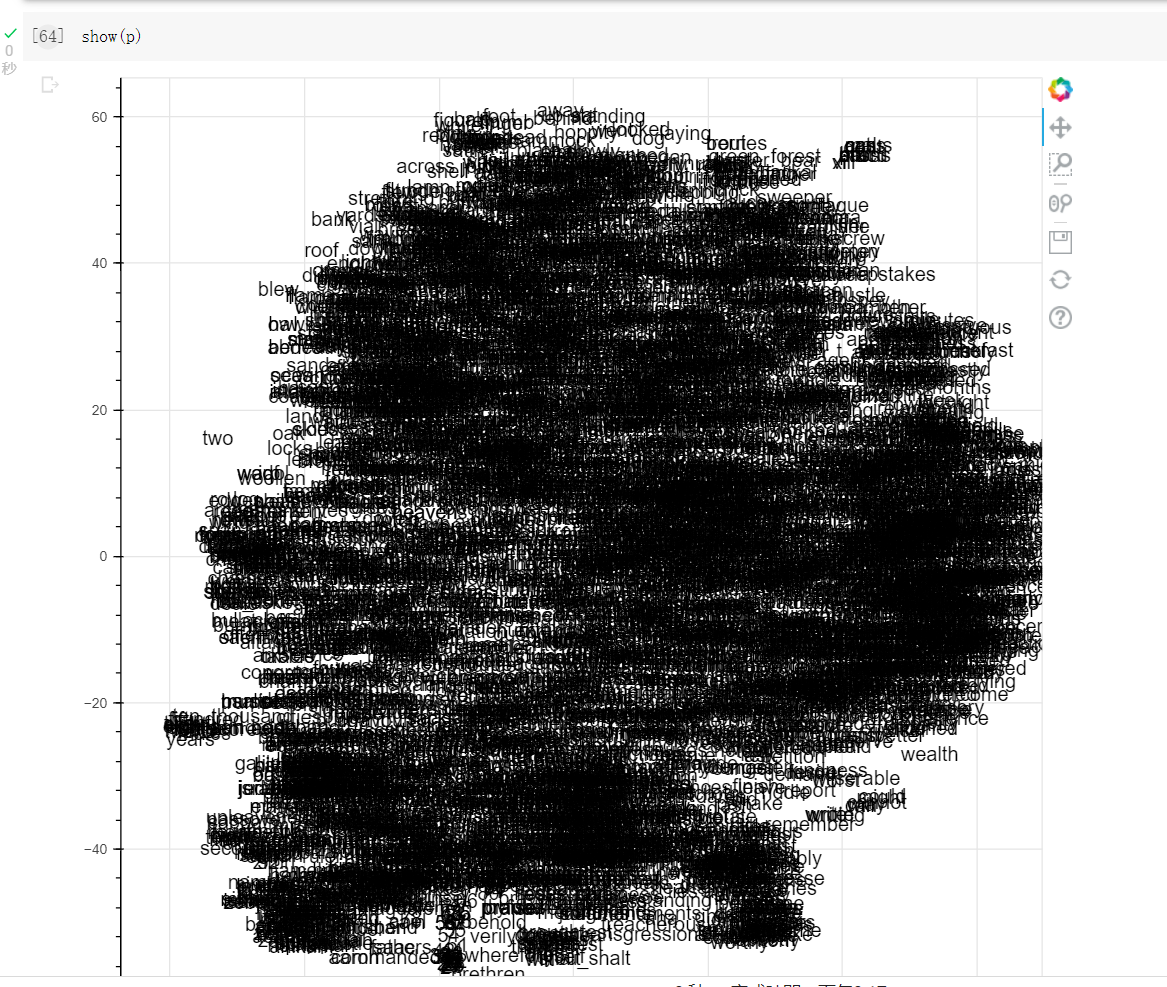
用t-SNE進行降維，將前面生成的64維詞向量空間降為2維(以利繪製成泡泡圖)



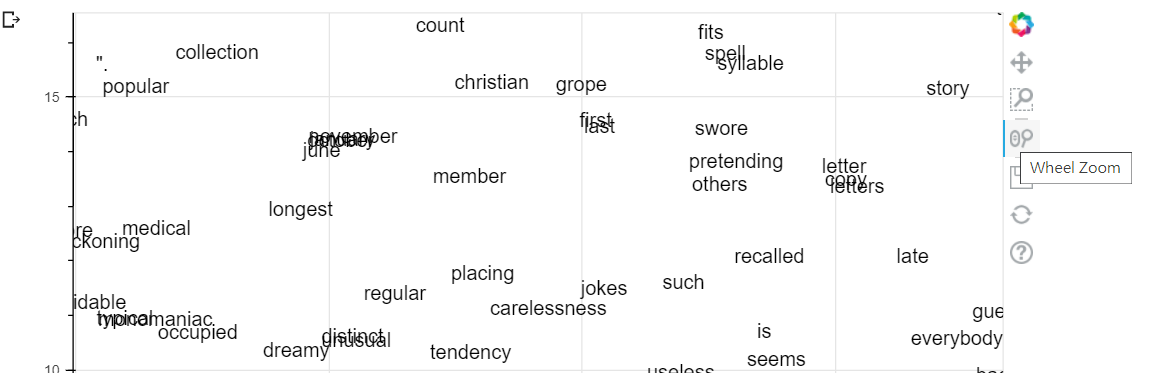
#### 步驟四，繪圖



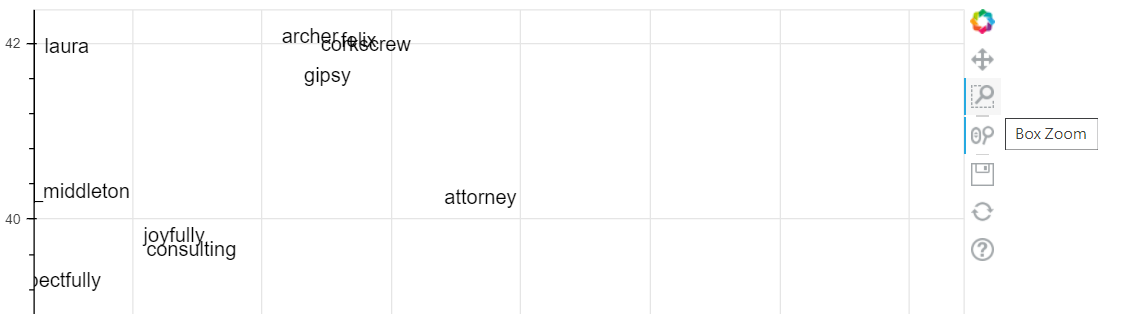




太密了，來縮放



或是局部觀察



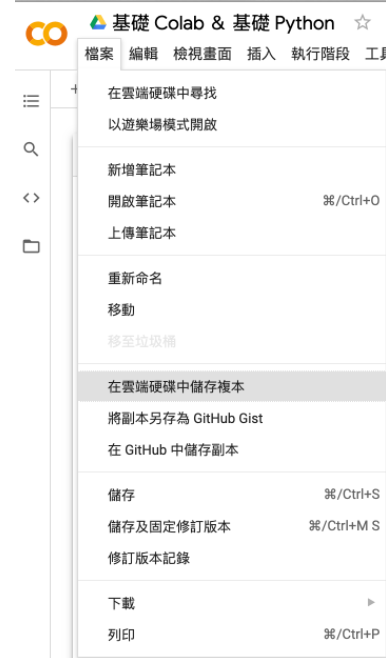
很有趣吧。

#### **附錄:前置作業Colab環境準備**

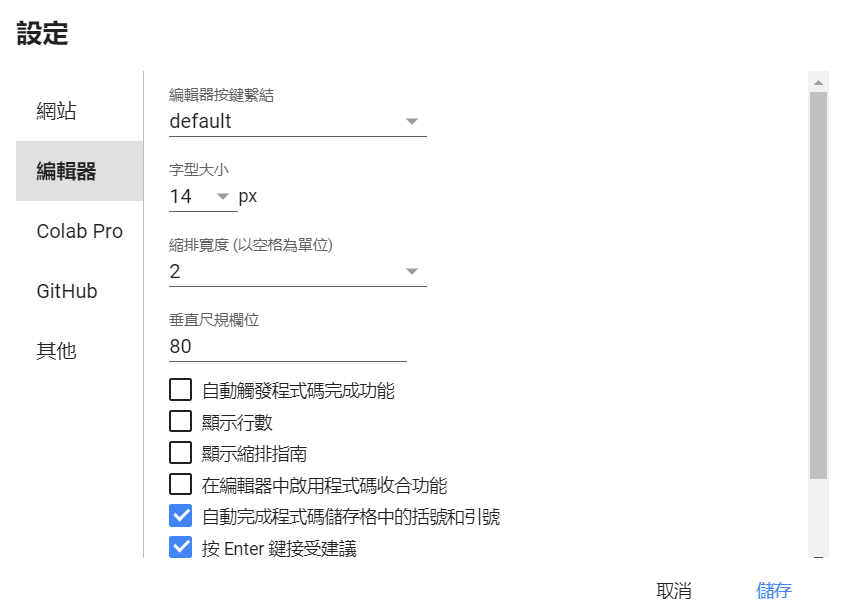
1.必須有自己的google帳號:使用自己的個人的 Google 帳號才可以執行自己 Colab上的程式。

2.先下載 Chrome瀏覽器，並登入 Google 帳號。

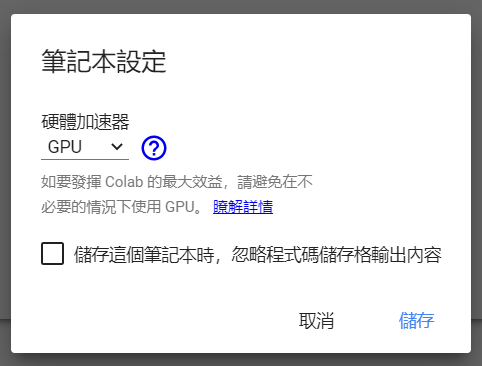
3.建立副本至個人目錄，否則不能存檔。



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。



