NYCU 313581016 丁驥

1. Which embedding model do you use? What are the pre-processing steps? What

are the hyperparameter settings?

1. 我使用的是 word2vec-google-news-300模型
2. 預處理步驟包括以下幾點：

* 將文本轉換為小寫
* 去除標點符號和數字
* 移除停用詞
* 進行lemmatization

1. hyperparameter settings：



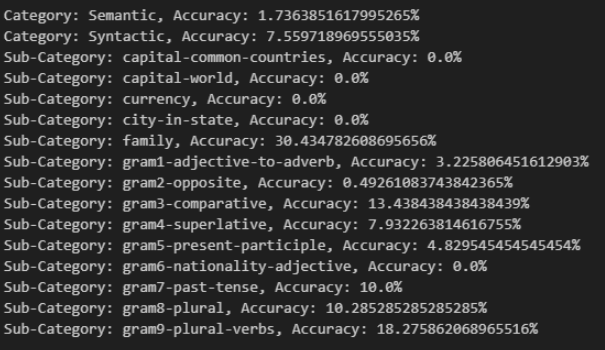
圖一：hyperparameter settings

* vector\_size：100
* window：5
* min\_count：5
* epochs：5
* 其餘依照Default

1. What is the performance for different categories or sub-categories?

我train 出來的model 其 accurancy 表現如下：

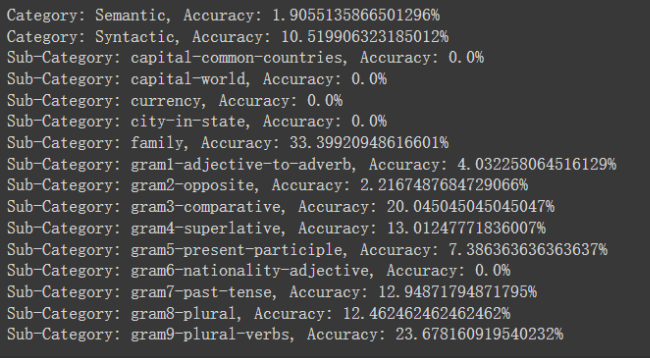
* 方法一



圖二：方法一的accurancy

說明：方法一想先確認整體架構，因此用較簡單的方式實作，在預處理方面僅處理stopwords 且train model 的時候，只有跑1個epoch， 因此除了family 的 acurrancy 有30 %外，其餘都相當不準確，而 0% 的原因是因為國家與地區大多用大寫來表示，我在最後針對此部分做調整。

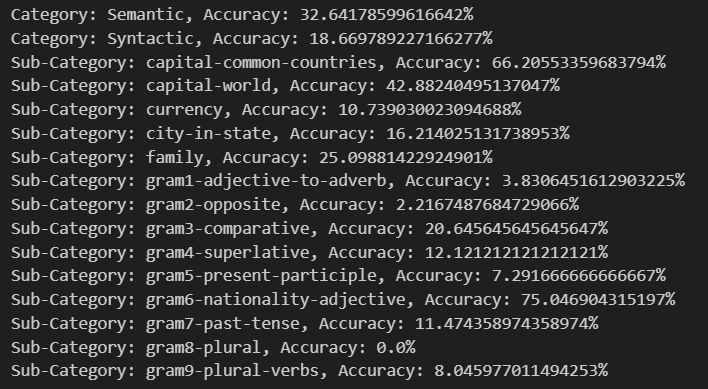
* 方法二



圖三：方法二的accurancy

說明：方法二在預處理方面與方法一相同，僅處理stopwords ，差別在於train model 的時候跑5個epoch， 可以看到Syntactic有明顯的增長，接著我們改變預處理的方式，並嘗試讓accuracy不要等於 0%。

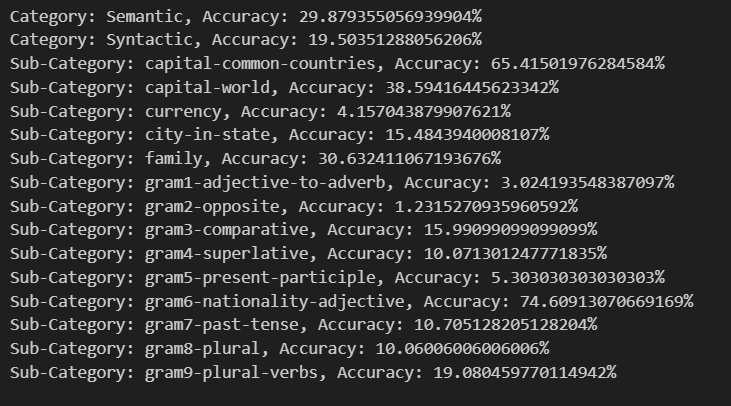
* 方法三



圖四：方法三的accurancy

說明：方法三在預處理方面做了將文本轉換為小寫、去除標點符號和數字、移除停用詞、進行lemmatization，且train model 的時候跑5個epoch，可以觀察到每一個種類都有明顯的增長，” capital-common-countries”與” gram6-nationality-adjective”兩項都有超過60%的accurancy，而” capital-world”也有42% 的accurancy，因此我們可以知道，或許地區相關的字具有固定的模式，模型可以捕捉到關係。比較可以注意的還有” gram8-plural” 為 0%，我們可以觀察到在方法一、二中還有數值，但在方法三卻是 0%，而我又在計算向量的過程中觀察到，很多常見的字都是miss(不存在模型)的，例如plays、says和swims等等，這個現象可能是lemmatization造成的，導致這些字被過濾掉，導致模型沒有這些字，無法計算。

* 方法四



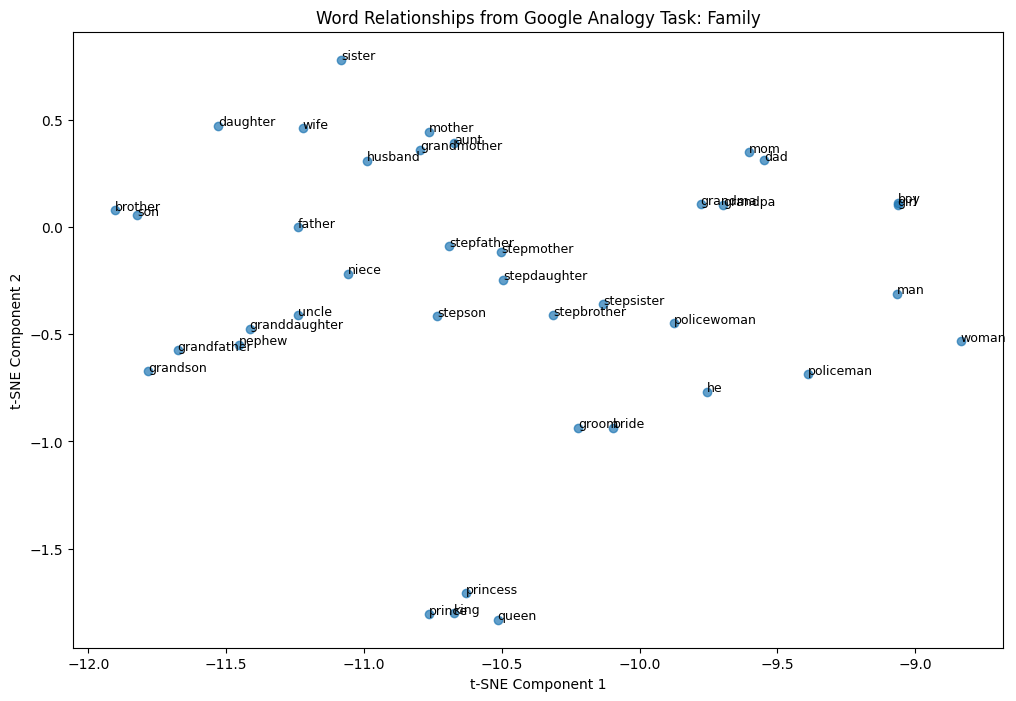
圖五：方法四的accurancy

說明：方法四是使用沒有lemmatization的dataset 進行模型訓練，並只有sample 10% 的資料(回答第五題使用)，可以觀察到所有類別都有accurancy 數值，代表之前的猜測是正確的。

1. What do you believe is the primary factor causing the accuracy differences for your approach?

我認為這題需要做部分的驗證來確認答案，不過礙於模型訓練需要不少時間，因此我以我有更改過的因素進行討論，我認為訓練的dataset與測試的dataset很重要，像是我前面有相當多accuracy的數值是 0 ，這是因為兩筆資料大小寫與lemmatization 的問題，若沒有處理好，不論怎模調整hyperparameter都不會有好的結果。

1. What’s your discovery from your t-SNE visualization plots?



圖五：T-SNE visualization plots

說明：可以看到類似的字詞都聚集在一起，且可以視覺化的看到向量相減的大致結果，例如我們可以看到左上角的daughter 與 其右邊的wife，我們接著也可以往下看到son 的右邊是 father，可以發現到右邊是雙親(father和 mother)，左邊是小孩(daughter和 son)，上面是女生，下面是男生，具有對應的關係。我覺得這相當有趣，我記得也有看過不同語言中，其意思一樣的字分布也會類似，若我之後有機會可以來實做看看。

1. What’s the difference in word representations if you increase the amount of training data?

因為增加資料再訓練會比較多時間，因此我以 sample 10% (第二題中的方法四)和 sample 20% (第二題中的方法三)進行比較，我們可以發現就每個分類來說，較多資料的方法四表現比較少，其中方法三有幾項數值比較高，其原因是因為沒有lemmatization 造成，因此不列入考慮。我認為直觀的來看，隨著訓練數據的增加，模型能夠接觸到更多的上下文信息，這意味著模型在學習詞的向量表示時，同時可以考慮到詞在更多語境中的用法，因此會更穩定。

若有一些詞彙比較不常出現，也可能出現較多次，不容易被min\_count 篩掉，不過我覺得現在的training data 夠多了，主要是要增加資料的多樣性，以及避免 imbalance 的問題，讓模型可以涵蓋更多範圍。

**環境**

System win11

CPU AMD Ryzen 7 4800H with Radeon Graphics

**Python version**

Python 3.10.11

參考資料：本作業使用ChatGPT進行架構說明與部分的程式修改(詳細使用方式寫在註解)，並參考以下網路資料。

<https://sfhsu29.medium.com/nlp-%E5%B0%88%E6%AC%84-1-2-%E5%A6%82%E4%BD%95%E8%A8%93%E7%B7%B4%E8%87%AA%E5%B7%B1%E7%9A%84-word2vec-5a0754c5cb09>

<https://www.kaggle.com/code/mehmetisik/text-preprocessing-visualization-wikipedia-nlp>