NLP HW1

NCKU F64091130 楊采語

|  |  |  |
| --- | --- | --- |
| Environment types | Running environment | Python version |
| Colab | Colab |

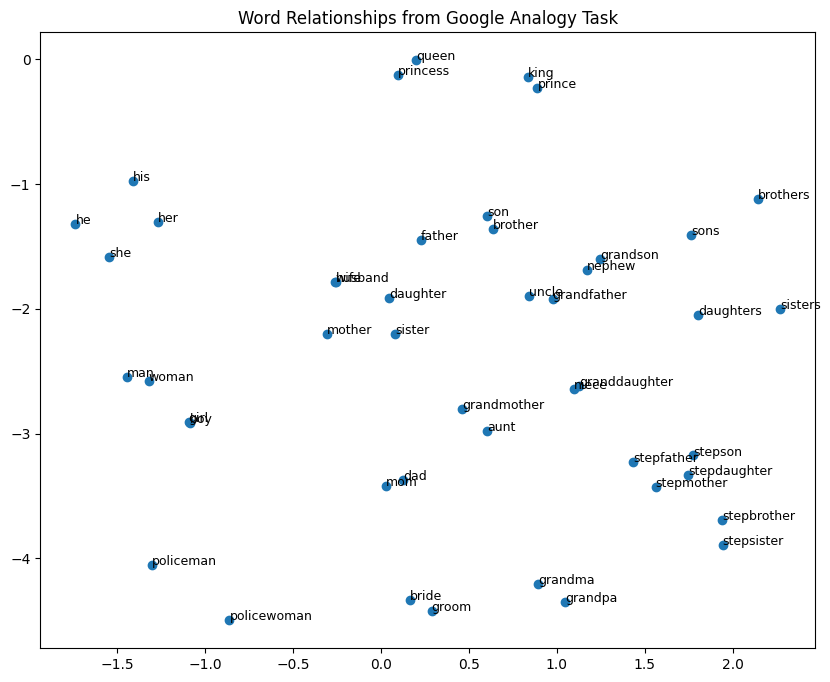
Part I & Part II

|  |  |
| --- | --- |
| 預處理過程 | 1. 分割資料每一行 2. 以” :” 判斷SubCategory 3. 計算” :”出現幾次，給予Category |
| Embedding model | glove-wiki-gigaword-100 |
| 備註 | #TODO2、#TODO3有問ChatGPT如何撰寫，並以他提供之範本進行修改 |

|  |  |
| --- | --- |
| Category | Accuracy |
| semantic | 2.2663208929980834% |
| syntactic | 14.341920374707259% |

|  |  |
| --- | --- |
| Sub-Category | Accuracy |
| capital-common-countries | 0.0% |
| capital-world | 0.0% |
| currency | 0.0% |
| city-in-state | 0.0% |
| family | 39.723320158102766% |
| gram1-adjective-to-adverb | 6.048387096774194% |
| gram2-opposite | 6.280788177339902% |
| gram3-comparative | 30.705705705705704% |
| gram4-superlative | 25.49019607843137% |
| gram5-present-participle | 13.446969696969695% |
| gram6-nationality-adjective | 0.0% |
| gram7-past-tense | 11.346153846153847% |
| gram8-plural | 15.54054054054054% |
| gram9-plural-verbs | 22.873563218390807% |

從Category來看，syntactic的準確度比semantic好上不少，因為semantic裡的子類別大多數準確率皆為0，只有family的準確度來到了40%，推測是因為語料庫中要出現首都、國家、貨幣的資料不如家庭詞彙多，且這些詞彙之間具有明確的語義關聯，可以使模型能夠較好地識別；而在syntactic內因為都是與語法的配對，可能會重複出現相同的語法、句型等，讓模型容易區分，但英文中有非常複雜的語法，還有可能因為口語、厘語等的特殊語法，使得模型容易混淆。



而從family的t-SNE可以發現圖中的詞語大多是與「家族」或「性別」相關的詞。相近位置的詞語在詞向量空間中具有相似的語義，而在man和woman、boy和girl以及policeman和policewoman等具有明顯的性別對應關係，並且這些詞語之間的距離較近，可以表示這些性別詞之間的語義關聯性，在家族關聯詞如father、mother、son、daughter會比較接近，也可以展現出他們語義間的關聯性。

Part III

|  |  |
| --- | --- |
| 預處理過程 | 1. 先將字體轉成全小寫 2. 利用word\_tokenize先取出文本的token 3. 利用正則表達式剔除非英文的字 4. 比對nltk的stop\_word並移除 5. 移除文本的標點符號 6. 用lemmatizer將詞形還原 |
| Embedding model | Word2Vec |
| Hyperparameter  settings | vector\_size=100 window=5 min\_count=1 |
| 備註 | #TODO4有詢問claude有什麼做法可以不用一次獨入所有語料就能隨機sample  #TODO5 有詢問ChatGPT以甚麼特套件可以提供預處理功能，避免執行時間過久 |

|  |  |
| --- | --- |
| Category | Accuracy |
| semantic | 1.04859623407374% |
| syntactic | 3.4660421545667446% |

|  |  |
| --- | --- |
| Sub-Category | Accuracy |
| capital-common-countries | 0.0% |
| capital-world | 0.0% |
| currency | 0.0% |
| city-in-state | 0.0% |
| family | 18.379446640316203% |
| gram1-adjective-to-adverb | 1.411290322580645% |
| gram2-opposite | 0.6157635467980296% |
| gram3-comparative | 8.108108108108109% |
| gram4-superlative | 5.258467023172905% |
| gram5-present-participle | 2.367424242424242% |
| gram6-nationality-adjective | 0.0% |
| gram7-past-tense | 8.076923076923077% |
| gram8-plural | 0.0% |
| gram9-plural-verbs | 3.793103448275862% |

上面的數據因為colab會跳開，所以先使用5%的文本作為訓練用的資料，可以看見大部分字的準確度與使用glove-wiki-gigaword-100訓練的成果有相似的趨勢，syntactic的準確度比semantic好一點，也在family的類別中出現最高的準確率。

一張含有 文字, 螢幕擷取畫面, 圖表, 行 的圖片

自動產生的描述

可以發現自己train的model與glove-wiki-gigaword-100相比詞語較無分散開來，大部分是一大群一大群聚在一起，而且詞語間的對應關係較不明顯，在man和woman、boy和girl這些詞彙沒有出現如同glove-wiki-gigaword-100模型的明顯對應關係。

下面測試把將字體轉成全小寫、剔除非英文的字的預處理拿掉，看是否會較好判斷出國家名、地名等(因為開頭常為大寫)，並調整Hyperparameter settings加快訓練速度。

|  |  |
| --- | --- |
| 預處理過程 | 1. 利用word\_tokenize先取出文本的token 2. 比對nltk的stop\_word並移除 3. 移除文本的標點符號 4. 用lemmatizer將詞形還原 |
| Hyperparameter  settings | vector\_size=50 window=3 min\_count=5 |

|  |  |
| --- | --- |
| Category | Accuracy |
| semantic | 1.0147705491036194% |
| syntactic | 5.442622950819672% |

|  |  |
| --- | --- |
| Sub-Category | Accuracy |
| capital-common-countries | 0.0% |
| capital-world | 0.0% |
| currency | 0.0% |
| city-in-state | 0.0% |
| family | 17.786561264822133% |
| gram1-adjective-to-adverb | 2.9233870967741935% |
| gram2-opposite | 1.354679802955665% |
| gram3-comparative | 10.06006006006006% |
| gram4-superlative | 7.397504456327986% |
| gram5-present-participle | 3.882575757575758% |
| gram6-nationality-adjective | 0.0% |
| gram7-past-tense | 15.320512820512821% |
| gram8-plural | 0.0% |
| gram9-plural-verbs | 5.057471264367816% |

這次train完發現比經過字體轉成全小寫、剔除非英文的字的預處理完的成果還好，尤其在gram7-past-tense出現顯著的提升，甚至比使用glove-wiki-gigaword-100還好，這次甚至降低了vector size及window，並提高min\_count，理論上會損失部分語義信息，推測可能也是因為範圍變小讓各種不同時態的字詞比較不會被其他字影響。

Family：

一張含有 文字, 螢幕擷取畫面, 數字, 圖表 的圖片

自動產生的描述

gram7-past-tense：

一張含有 文字, 螢幕擷取畫面, 筆跡, 字型 的圖片

自動產生的描述

可以發現gram7-past-tense的參數距離都較遠(坐標軸差距接近10)，可能可以代表一群一群的分類的蠻清楚的，所以準確率相較較高。