人工智慧於醫療應用與服務 Homework 3

NM6101072 張育嘉

1. 使用 NER model 進行名詞標註

   程式環境：Google Colab

   函式庫：

   numpy==1.21.6

   pandas==1.3.5

   tqdm==4.64.1

   sklearn-crfsuite==0.3.6

   流程：

   a. 將範例 sample_data.txt 處理成 BIO 格式。
   b. 將資料分成 Training 及 Testing。
   c. 將每個字轉成 word vector 並給予標註。
   d. 使用 CRF 預測並輸出 f1 score。
   e. 輸出預測結果至 output.csv。

   跟助教範例一樣使用 Google Colab 執行，需掛載檔案的雲端硬碟路徑。

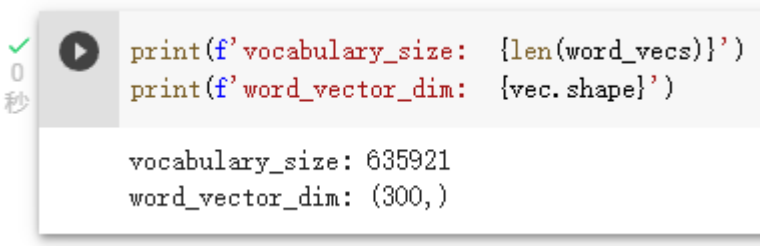
   改進 f1 score 可以有以下幾種方式：

   a. 抓取不同特徵，如 POS-tag，word_length，word_position，盡量描述資料分佈狀態使模型有更好的預測。
   b. 使用不同 NER model：神經網路模型去學習特徵，取代人工調整參數。
   c. 使用品質更好的 word2vec 特徵。

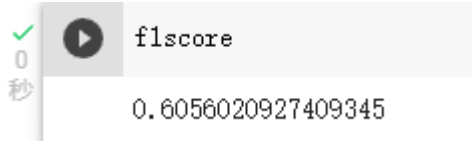
   修改的部分：

   本文不修改優化方式，使用 Chinese Word Vectors 中文詞向量[1]製作的 word2vec 當作輸入特徵依據；另外算出整個 data 的 f1 score。因此，原本使用的 cna.cbow.cwe_p.tar_g.512d.0.txt 改成

sgns.target.word-word.dynwin5.thr10.neg5.dim300.iter5_2 (1.69GB，下載連結如參考資料[2])，詞彙量及維度如下圖：



程式增加 Evaluation 區塊，使用 sample_data.txt 跑出來的測試結果如下圖：



2. 實驗

a. 比較簡體字及繁體字的預測結果

使用 Github 上的 Chinese Word Vectors 中文詞向量專案，該專案製作多個不同來源的中文詞向量，這邊使用百度百科所算出來的 word2vec 做實驗，分別做了簡體字及繁體字的預測實驗，簡體字版本使用 Word 將 sample_data.txt 由繁轉簡，存成 sample_data_S.txt，將程式執行了 5 次，每次分割出不同的 Training 及 Testing data，簡體字及繁體字版本的 f1 score 分數差不多，Test data 結果在 0.57 到 0.63 之間，Full data 結果在 0.68 到 0.73 之間。兩種版本皆取最後一次的執行結果，f1 score 如表 1。

表 1 百度百科 word2vec 在 sample_data 上的 f1 score

	Test data	Full data
簡體字版 sample_data.txt	0.6100	0.7138
sample_data.txt	0.5733	0.7194

b. 比較 lbfgs 及 l2sgd 的算法差異

實驗將算法改成 l2sgd，l2sgd 設置 calibration_eta 為 0.01，比較 lbfgs 及

l2sgd 的表現，結果如表 2。

表 2 lbfgs 及 l2sgd 在 sample_data 上的 f1 score

	Test data	Full data
lbfgs	0.5856	0.7298
l2sgd	0.5612	0.5395

3. 心得

實驗不同字體的結果，Test data 分數比 Full data 分數低，應該是在產生 Test data 的時候比較容易有 outliers。測試 l2sgd 算法時，發現 l2sgd 執行時間多了至少 1 倍，分數沒有比較高，因 lbfgs 屬於 L1Regularization 而 l2sgd 屬於 L2 Regularization，猜測 word2vec 應該是屬於較稀疏的特徵分佈，因此用 lbfgs 算法比較能描述資料分佈，比較不會產生 overfitting。

繁體字及簡體字的預測表現差不多，從輸出結果來看，簡體字輸出比較多詞彙，但簡體字的詞彙文法跟繁體字不一樣，簡體版判斷為相近詞彙，實際上在繁體版卻不是，猜測因此造成總體預測分數差不多。

article_id	start_position	end_position	entity_text	entity_type
3	281	284	9公分	med_exam
3	291	294	5公分	med_exam
3	298	301	9公分	med_exam
3	479	482	十公分	med_exam
3	575	586	2016年10月	time
3	619	622	三年半	time
3	1718	1721	两三年	time
3	2041	2046	超过9公分	med_exam
3	2300	2305	师：承太郎	name

article_id	start_position	end_position	entity_text	entity_type
3	281	284	9公分	med_exam
3	291	294	5公分	med_exam
3	298	301	9公分	med_exam
3	479	482	十公分	med_exam
3	575	586	2016年10月	time
3	619	622	三年半	time
3	978	983	今年的年初	time
3	1704	1707	近幾年	time
3	1716	1719	兩三年	time
3	2041	2044	9公分	med_exam
3	2286	2289	承醫師	name
3	2300	2305	承太郎醫師	name

article_id	start_position	end_position	entity_text	entity_type
24	1408	1410	明明	name
24	1857	1861	十二个月	time
24	1868	1872	十二个月	time
24	1986	1990	五个月后	time
25	467	469	两年	time
25	1396	1400	第四个月	time
25	1451	1456	5月30号	time
25	1499	1501	千寻	name
25	1580	1584	第四个月	time

article_id	start_position	end_position	entity_text	entity_type
24	1170	1172	路上	time
24	1196	1199	300	money
24	1246	1249	230	time
24	1986	1989	五個月	time
25	467	470	兩年多	time
25	1396	1400	第四個月	time
25	1451	1456	5月30號	time
25	1499	1501	千尋	name
25	1580	1584	第四個月	time

4. 參考資料

[1] Chinese Word Vectors 中文詞向量 https://github.com/Embedding/Chinese-Word-Vectors

[2] 百度百科 word2vec 下載連結

https://drive.google.com/file/d/1r7imzOnwNagXQ-IbniPcvvSJERYCltlB