AIMAS Homework 3

F94081076 郭立晨 資訊 113

程式過程:

1. 引入想要訓練的文本並對其做前處理，過程包含分離出其中的訓練文字、文字實體與種類的配對等等的資訊。
2. 將先前處理過的資料轉換成 CRF 需求格式的檔案(.data)並儲存作為本次的資料集。
3. 載入要使用的詞向量檔案，得到本次要訓練詞向量的 dictionary，我使用了課堂上提供的詞向量檔案以及從網路上下載的中文詞向量檔案
4. 分割先前建立的資料集為訓練集以及測試集。
5. 將訓練集、測試集中的詞彙轉換為詞向量再從中提取特徵。
6. 將訓練集、測試集的詞彙做出標籤。
7. 利用訓練集的特徵與標籤開始使用 sklearn-crfsuite.CRF 模型訓練，將訓練完成的模型利用測試集計算出 f1score。

採用特徵:

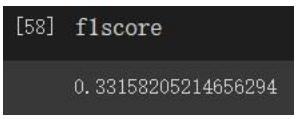
使用不同來源以及訓練規模的詞向量檔案提取特徵。

- 課程提供的詞向量檔案
- 網路上的詞向量檔案
  - ✓ https://fasttext.cc/docs/en/crawl-vectors.html
  - ✓ 使用 fastText 訓練詞向量(Common Crawl、Wikipedia 作為訓練資料)，再轉換成 word2vec 形式
  - ✓ using CBOW with position-weights, in dimension 300, with character n-grams of length 5, a window of size 5 and 10 negatives
  - ✓ vocabulary size: 2000000

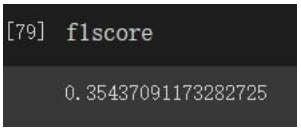
自行修改:

本次實驗我測試了兩種變因並得到四種結果，分別是，

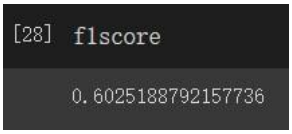
1. 去除角色名稱+原始詞向量檔案



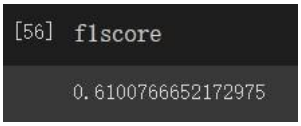
2. 保留角色名稱+原始詞向量檔案



3. 去除角色名稱+外來詞向量檔案



4. 保留角色名稱+外來詞向量檔案



實驗結果:

從實驗結果可以發現，角色名稱對詞彙分類有些微的影響，且與我的期待有些落差。我本來認為去除角色名稱這種類似雜訊的文字應該可以提升訓練模型的效率，讓準確率提升，但是在經過不同詞向量來源的實驗後，得知角色名稱去除不會增加模型的準確率反而有一些下降；而採用網路來源的詞向量模型經過實驗發現其確實大幅提升了模型的準確度到 60%，證明詞向量詞彙規模確實影響到模型的訓練成果。

實驗心得:

第一次學習處理語言的模型，比起直覺的影像處理辨識我覺得語言的分類辨識有點難以想像，而且資料的前處理也相對麻煩許多，需要對字串處理的函示很熟悉才能加快前處理的過程。再來就是，如果使用詞向量作為特徵提取的方式，如何建立詞向量的字典就變成一件很複雜的過程，訓練詞向量的模型又可以分成兩三種方式，以及對於何種語言進行訓練。若是目標文本具有一種以上的語言那詞向量的廣泛程度就會大大影響模型的準確度。

因此，這樣累積下來，訓練一個模型的過程中可以調整的參數就變得非常的多，使用的方法也可以互相排列組合。人們每天使用的詞彙以及句子等都會有變化，流行語破音字等更是難以辨識，連人類使用破音字都有可能用錯，對電腦來說更是困難的任務。怎麼樣的組合可以對於目標語言有較好的預測效率與準確度是非常值得研究的問題。