NLP Assignment 1 - Named Entity Recognition

110524031葉軒宏

1. 檔案
2. source code資料夾: 程式碼 (包含原始資料集)
3. requirements.txt: Python 環境中已安裝庫的列表
4. test-submit.txt : 帶有預測實體標籤的 test-submit.txt
5. f1-score.txt : 用dev.txt 評估出的模型F1分數
6. 程式碼

資料夾:

1. dataset: 原始資料集 (train.txt、dev.txt、test-submit.txt)
2. pickle: 原始資料集生成的pkl檔
   1. dev\_eva.pkl: dev.txt不帶原始實體標籤 (用於評估F1分數)
   2. vocab.pkl: train.txt編碼單字
   3. vocab\_dec.pkl: vocab.pkl解碼單字
3. model: 儲存訓練出的模型
4. submit: 輸出資料
   1. dev\_eva.txt: 帶有原始與預測實體標籤的dev.txt
   2. test-submit: 帶有預測實體標籤的 test-submit.txt

程式檔(.py):

1. 110524031: 訓練程式，讀入train.pkl的資料，丟入模型中訓練，並計算及驗證損失值
2. data: dataloader，編寫讀取資料的方法
3. make\_dataset: 原始資料集依句子與(標籤)分欄寫入pkl檔

make\_dataset\_train、make\_dataset\_dev、make\_dataset\_dev\_eva、make\_dataset\_test\_submit

1. make\_vocab: 生成train.txt不重複單字並編碼

make\_vocab\_dec: 將vocab.pkl單字解碼

1. model: 定義了模型架構、損失函數以及驗證方法
2. utils: 定義了一些公用的變數和函式，以及訓練的基本參數
3. submit\_dev\_eva: 輸出dev\_eva.txt，帶有原始與預測實體標籤

submit\_test\_submit: 輸出test-submit.txt，帶有預測實體標籤

1. conlleval: 用於評估F1分數
2. 過程
3. 生成pkl資料: make\_dataset\_train.py、make\_dataset\_dev.py、make\_dataset\_dev\_eva.py、make\_dataset\_test\_submit.py、make\_vocab.py、make\_vocab\_dec.py

=> train.pkl、dev.pkl、dev\_eva.pkl、test-submit.pkl、vocab.pkl、vocab\_dec.pkl

1. 主程式訓練模型: 110524031.py
2. 資料代入模型生成文檔: submit\_dev\_eva(評估)、submit\_test\_submit(預測)

=> test-submit.txt (預測結果)、dev\_eva.txt (評估結果)

1. 生成F1分數 : python conlleval < ./submit/dev\_eva.txt > f1-score.txt

=> f1-score.txt

1. 資料處理
2. 字典: 將train.txt的單字編號，存成dict格式，打包到vocab.pkl
3. 文本: 以tab切分欄位依句子整理成dict格式

1. 排序: 將batch內資料進行大到小排序並做 padding，變成同樣長度的序列。
2. 模型
3. 模型分為三部分，Embedding(word2vec) 接LSTM再接Linear層
4. 損失函數使用cross entropy，優化器使用Adam
5. 模型參數: epoch = 64、batch size = 64、learning rate = 0.005、step = 5
6. 訓練
7. 每一個epoch會用資料對模型進行評估，算出該資料的遺失值，將訓

練次數定為64次，在訓練的過程可以看出training loss有逐漸下降的趨勢，但經過不到10次訓練後，雖然training loss持續緩慢下降，但validation loss卻不減反增，出現overfitting，最終經過篩選後將較好的模型帶入文本進行輸出。

1. 結果
2. Accuracy: 93.53% (All)、10.37% (non-O)
3. Precision: 26.45%
4. Recall: 11.04%
5. F1 score: 15.58