Homework1

1. Data preprocess

I. 清洗資料:

A. "http\S+|www\S+"換成 "<url>"

B. "@\S+" 換成 "<user>"

C. "#\S+" 換成 "<hashtag>"

II. 將各句子以 (Word, Tag) 存為陣列

III. 定義 indexer: 轉換 Word 及 Tag 成編號

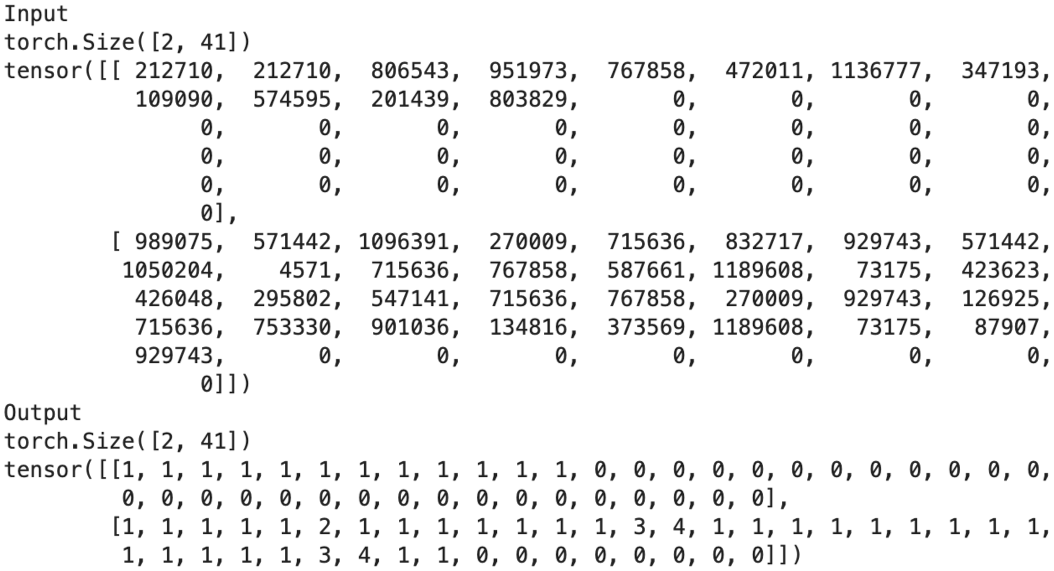
A. word_dict 加入 {"<PAD>": 0, "<UNK>": 1}

B. tag_dict 加入 {"<PAD LABEL>":0}

IV. pad the sequence to same length, "<PAD>"的 label 為"<PAD LABEL>"

2. Model architectures

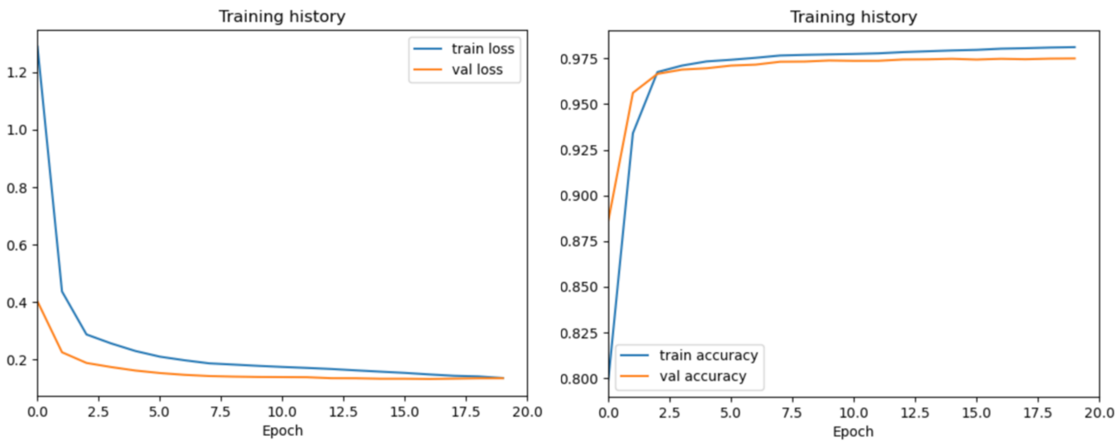
- embedding dimension = 200
- hidden dimension = 200
- word embeddings: 使用 glove.twitter.27B.200d.txt
- dropout: nn.Dropout(p=0.2)
- LSTM: nn.LSTM(200, 200, num_layers=3, bidirectional = True) LSTM model
- hidden2tag = nn.Linear(400, 22), 輸出句子中每個字的 label
- 模擬輸入及輸出：



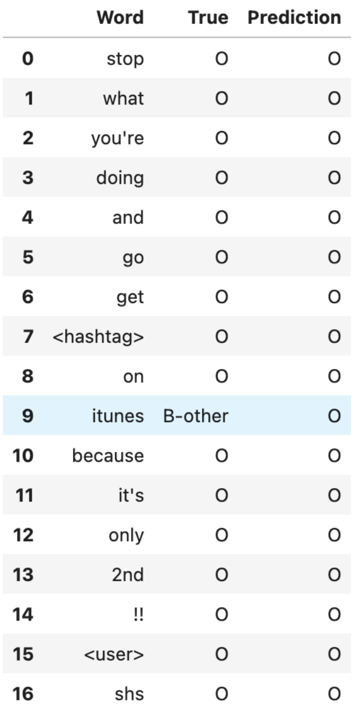
```
Input
torch.Size([2, 41])
tensor([[ 212710,  212710,  806543,  951973,  767858,  472011, 1136777,  347193,
          109090,  574595,  201439,  803829,       0,       0,       0,       0,
               0,       0,       0,       0,       0,       0,       0,       0,
               0,       0,       0,       0,       0,       0,       0,       0,
               0,       0,       0,       0,       0,       0,       0,       0,
               0],
        [ 989075,  571442, 1096391,  270009,  715636,  832717,  929743,  571442,
         1050204,    4571,  715636,  767858,  587661, 1189608,   73175,  423623,
          426048,  295802,  547141,  715636,  767858,  270009,  929743,  126925,
          715636,  753330,  901036,  134816,  373569, 1189608,   73175,   87907,
          929743,       0,       0,       0,       0,       0,       0,       0,
               0]])
Output
torch.Size([2, 41])
tensor([[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
         0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
        [1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 3, 4, 1, 1, 1, 1, 1, 1, 1, 1, 1,
         1, 1, 1, 1, 1, 3, 4, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0]])
```

3. Training process

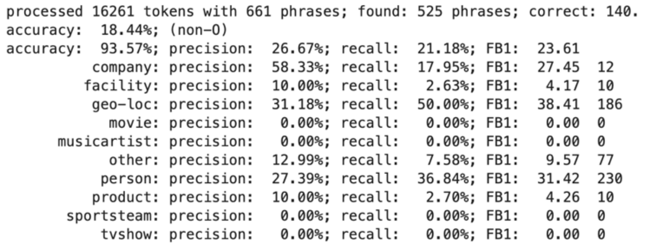
- parameters:
  - batch size: 64
  - epochs: 20
- optimizer: AdamW, learning rate =1e-3
- loss function: CrossEntropyLoss
- result



4. Evaluation scores



	Word	True	Prediction
0	stop	O	O
1	what	O	O
2	you're	O	O
3	doing	O	O
4	and	O	O
5	go	O	O
6	get	O	O
7	<hashtag>	O	O
8	on	O	O
9	itunes	B-other	O
10	because	O	O
11	it's	O	O
12	only	O	O
13	2nd	O	O
14	!!	O	O
15	<user>	O	O
16	shs	O	O



```
processed 16261 tokens with 661 phrases; found: 525 phrases; correct: 140.
accuracy:  18.44%; (non-O)
accuracy:  93.57%; precision:  26.67%; recall:  21.18%; FB1:  23.61
          company: precision:  58.33%; recall:  17.95%; FB1:  27.45  12
         facility: precision:  10.00%; recall:   2.63%; FB1:   4.17  10
          geo-loc: precision:  31.18%; recall:  50.00%; FB1:  38.41  186
            movie: precision:   0.00%; recall:   0.00%; FB1:   0.00  0
      musicartist: precision:   0.00%; recall:   0.00%; FB1:   0.00  0
            other: precision:  12.99%; recall:   7.58%; FB1:   9.57  77
           person: precision:  27.39%; recall:  36.84%; FB1:  31.42  230
          product: precision:  10.00%; recall:   2.70%; FB1:   4.26  10
       sportsteam: precision:   0.00%; recall:   0.00%; FB1:   0.00  0
           tvshow: precision:   0.00%; recall:   0.00%; FB1:   0.00  0
```