Python版本3.6

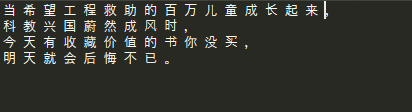
Tensorflow版本 1.8.0（估计1.4以上版本都可以）

先训练字典文件，把word换成id：

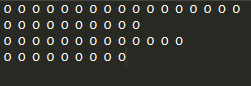
将测试文件和训练文件按照如下格式合并成一个文件，命名为corpus\_data：

训练、测试文件格式如下，utf-8编码

\_text文件：一句话为一行，字直接用空格隔开



\_tag文件：一句话一行，标签用空格隔开，对应着\_text文件里的字



运行word2id.py 文件：

python word2id.py --corpus\_py=\*\*\*\*\*(填corpus\_data的保存路径，默认为data\_path) --vocab\_path=\*\*\*\*\*（填希望生成的word.pkl文件所在的路径，默认为data\_path）--min\_count=\*\*(填最小阈值，若词频低于此阈值则不统计，默认为0)

字典文件为word2id.pkl，放在目录data\_path下。二进制存储的，用pickle.load读取。

训练文件为 train\_data\_text、train\_data\_tag，放在目录data\_path下

测试文件为 test\_data\_text、test\_data\_tag, 放在目录data\_path下

使用模型判别命名实体生成的文件为 demo\_data\_tag

训练好的模型为一串数字，放在目录data\_path\_save下

参数：

--train\_data：训练数据路径，默认为data\_path

--test\_data：测试数据路径，默认为data\_path

--demo\_data：生成判别命名实体的路径，默认为data\_path

--batch\_size：一个batch的大小，默认为64

--epoch：训练迭代轮数，默认为40

--hidden\_dim：每一个lstm cell中的维数，默认为300

--optimizer：可选Adam/Adadelta/Adagrad/RMSProp/Momentum/SGD，默认为Adam

--CRF：在rnn后面加crf，如果为false，则不加crf而是softmax，默认为true

--lr：learning rate 学习率，默认为0.001

--clip：解决梯度爆炸，默认为5.0。（clip作为一个阈值，如果梯度的l2范数大于这个阈值，则把梯度要除以他的l2范数）

--dropout：dropout的比例，默认为0.5

--update\_embedding：训练的时候是否更新embedding，默认为true

--pretrain\_embedding：使用随机的词向量还是预训练好的词向量，默认为random（表示没有预先训练好的词向量）

--embedding\_dim：词向量的维度，默认为300维

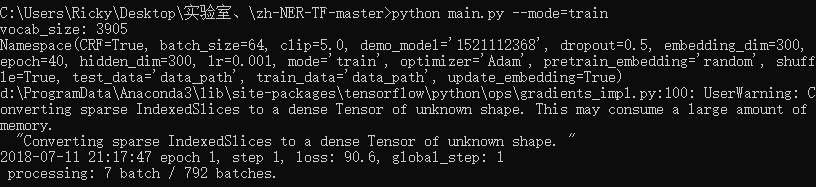
--shuffle：打乱训练数据顺序，默认为true

--mode：模型运行状态：train\test\demo，默认为demo

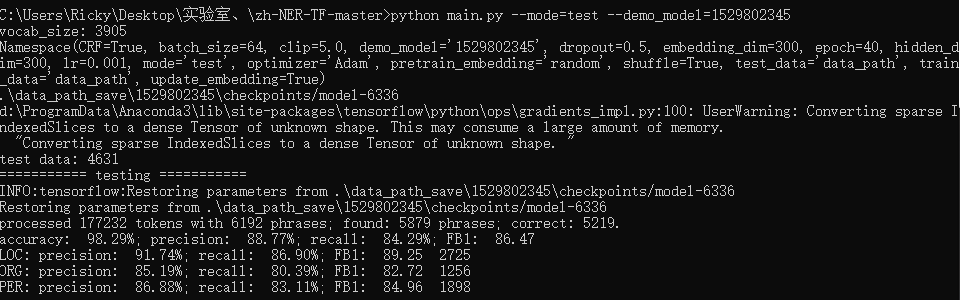
--demo\_model：在非train状态时，作为已经储存好的模型加载进来。默认是1521112368，是以训练结束时的时间戳来命名的文件。

程序运行：

Python main.py --mode=train



Python main.py --mode=test --demo\_model=\*\*\*(填训练结束后生成的模型，在data\_path\_save文件夹下)



Python main.py --mode=demo --demo\_model=\*\*\*(填训练结束后生成的模型，在data\_path\_save文件夹下)

