多语言处理

具体流程

利用facebook预先训练的多语言的词向量，在训练集和测试集为不同的语言情况下，训练情感分析模型。

基本原理是，目前Facebook已经训练了在同一向量空间中下30多种语言的多语言向量而且目前有很多，

算法基于Convolutional Neural Networks for Sentence Classification pytorch实现版本。

这个算法主要是用卷积神经网络将每个词的词向量进行卷积

Facebook 训练的词向量是每个语言前200000个词汇和300维，同时是

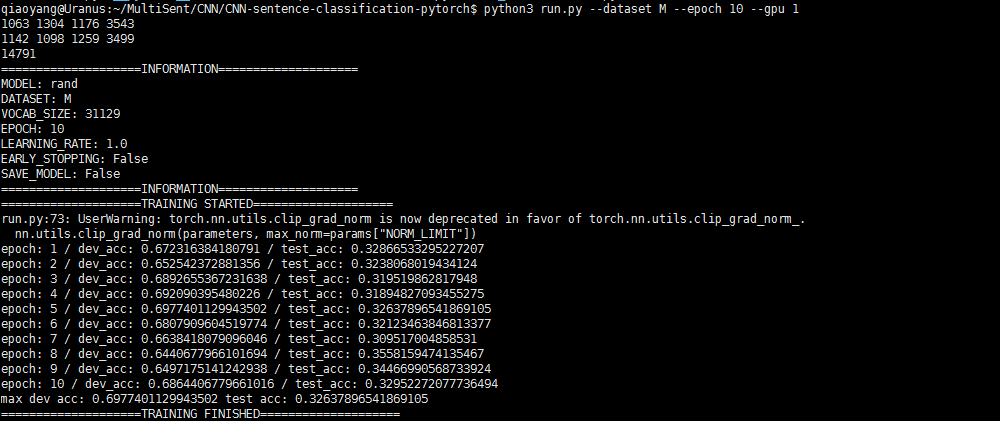
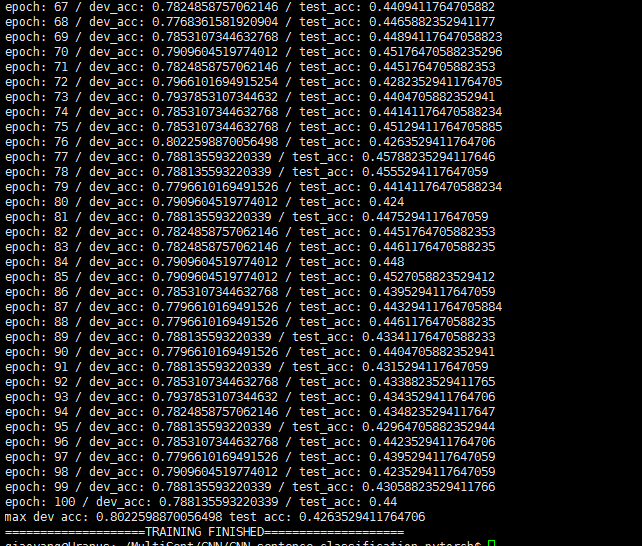
目前有好多个词向量

以下为测试结果：

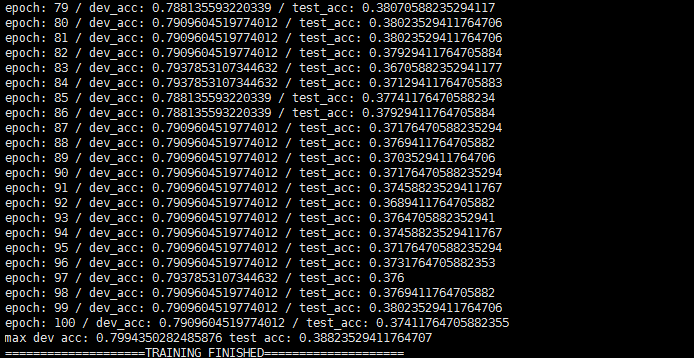
数据量：

法语3499条（positive1142,negative1098,netural 1259）

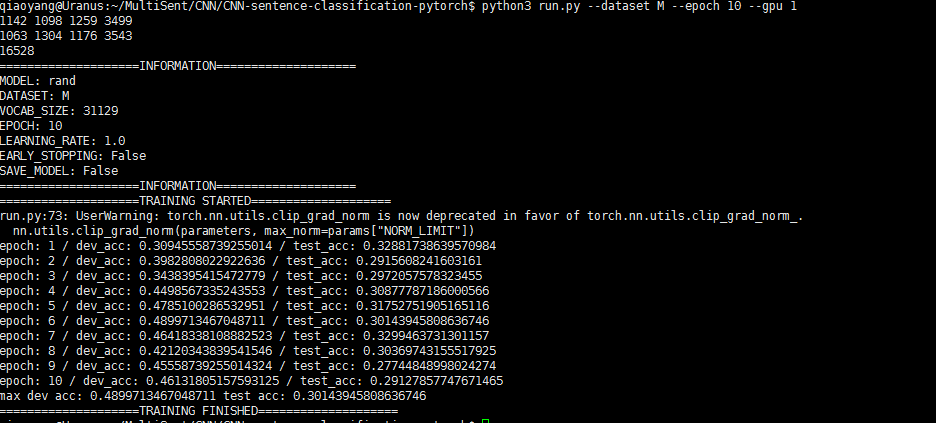
阿拉伯语 3543条(positive 1063,negative 1304,netural 1176)



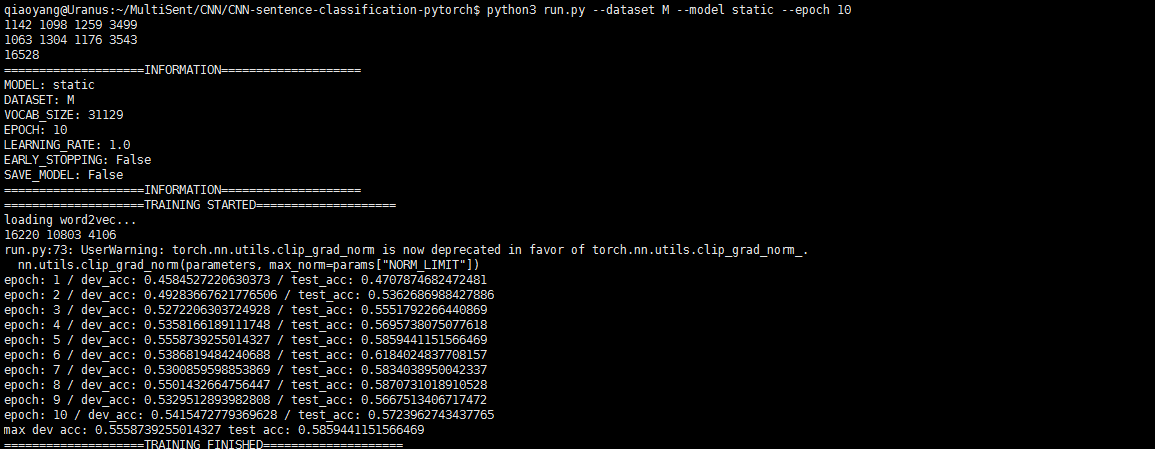
这是在



迭代次数为10次，法语预测阿拉伯语，无Facebook多语言向量，法语准确率：48%，阿拉伯语38%



迭代次数为10次，法语预测阿拉伯语，Facebook多语言向量，法语准确率：55%，阿拉伯语61%



由于

综上结果看出，同时在测试中发现有接近一半的词汇没有被预训练的词向量识别出来，估计是一些标点符号

而且文本中有

如果数据进行进一步处理，使得更多的词汇可以有更好的效果

开发思路：基于已有拥有的多语言的词向量