Code Logs

4.10

Torchtext 中有相对官方的IMDB数据集，但torchtext太古老了，处理数据的方式不方便。最终在Hugging Face上下载。为方便使用，存储为csv格式。

* imdb\_train.csv
* imdb\_test.csv

4.11

对数据进行清洗。

* 考虑到imdb数据集特性会包含一些HTML标签，需要去除
* 考虑去除emoji，但统计发现含有emoji的数据量过小，可以忽略emoji的存在，故最终决定不去除
* 交叉验证的代码有点问题，预测结果不太对