文本分类实战

任务描述

Coronavirus tweets NLP - Text Classification是Tweets的情感多分类数据集，其中用户名称等信息已经脱敏化，尝试用多种方法以Original Tweet文本数据作为输入，进行情感Sentiment分类。在test dataset使用Accuracy作为评价指标对比模型效果，代码直接在Kaggle Notebook GPU 环境进行。

1. TF-IDF + 逻辑回归
    - `sklearn.fearture_extraction.text.TfidfTransformer`
    - `sklearn.linear_model.``LogisticRegression`
    - 参考代码：https://blog.csdn.net/u013230189/article/details/108371024
2. FastText https://flyai.com/article/681
3. word2vec + LSTM https://www.jianshu.com/p/edad714110fb
    - 不加载预训练词向量
    - 加载预训练词向量
    - 参考代码： https://www.kaggle.com/ziliwang/baseline-pytorch-bilstm
4. Bert
    - 参考文章：https://zhuanlan.zhihu.com/p/66057193
    - 参考代码：https://www.kaggle.com/sumitm004/simple-bert-model-for-text-classification

PS：

- 在kaggle中安装包，notebook下，执行!pip install ...
- 在kaggle中导入数据，打开data，上传相关数据文件，在notebook下添加你的data