骗了钱想去启明学院对面吃螺狮粉队 种子杯复赛报告

参赛队伍详细信息

队伍名称 骗了钱想去启明学院对面吃螺狮粉

队长 李一宸 计算机学院17级信安1701班 弱鸡

队员 皮皓文 计算机学院17级信安1702班 弱鸡

队伍设备：

- I7 7700HQ
- GTX 1070
- 16GB RAM

使用语言以及运行环境

python 3.6 （anaconda）

附加包

- pandas
- numpy
- liblinear-muticore svc
- sklearn

linux内核版本 4.15

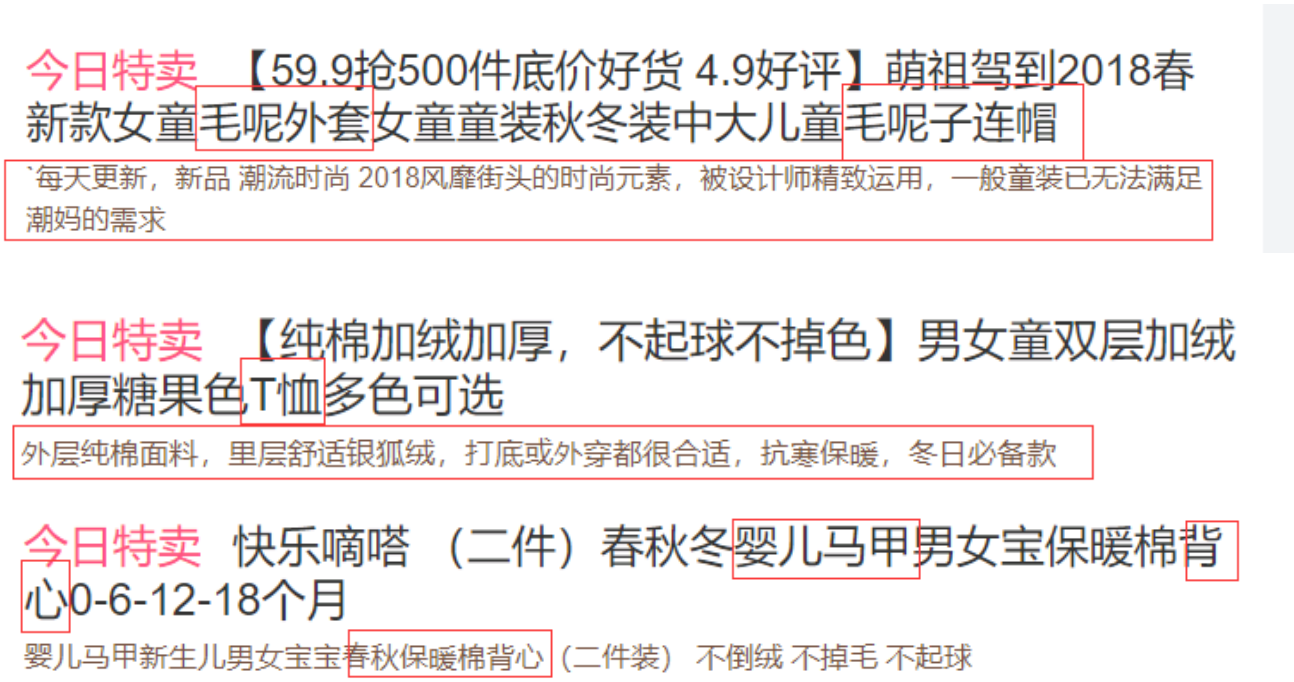
变量含义以及代码接口

变量含义均有详细注释，可读性很强

代码接口

```
python train.py valid 生成一个用训练集训练，预测验证集来获取模型评价指标
python valid.py 输出得分
python train.py train 训练模型，生成结果
python output.py 输出结果txt文件
```

特征提取思路



> 数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限而已

通过观察贝贝网原始数据，发现仅用description很难判断出商品种类（见上图）

而title中也几乎没有语序，大多是名词、形容词的反复堆叠，由于给定的数据是脱敏数据集，很难从网上获取现成的embedding结果。所以判断根据给定数据集使用传统的embedding的方法很难提取到特征，事后用一些word2vec的方法加上相对简单的神经网络也验证了这一猜想。所以根据数据特征，选取了跟语序无关的，更适用于网购商品title特征的 TF-IDF编码方式来对title进行编码，对出现率过高的停用词进行drop处理，将ngram_range设定在一个相对小的范围内（1-2），对于词频较低的单词（如1次）均进行了保留处理，对于编码进行了归一化，避免TF-IDF过度倾向于长文本。

详见下图sklearn中TF-IDF编码的参数表

```
min_df = 1
max_df = 0.6
use_idf = 1
smooth_idf = 1
sublinear_tf = 1
```

预测模型的选取

我们先后尝试了naive-bayes、bayes网络，多层全连接神经网络、xgboost、lightbgm,和embedding后的fasttext，均没有达到令人满意的效果。

受限于只能使用要兼顾日常学习生活的单台PC而没有服务器（当时不知道谷歌有羊毛薅），PC内存小计算能力差，没有办法全天候计算，我们没有进一步尝试更复杂更难收敛的深层神经网络（唉）

考虑到本次任务是分类任务，而分类任务有以下几个特点

- 高维特征空间
- 稀疏文档空间
- 语义级别上的稀疏
- 词典中词项频率的分布满足Zipf定理

文本文档集的统计学特征和统计分类模式定义了其在统计上是可分的，在几何上存在超平面。用统计的方式得到的超平面用SVM也可以得到，即使文本文档集的特征空间维度很高。

结合以上考虑，我们最终选择了linearsvc作为我们的分类器，并对sklearn库中的linearsvc进行了进一步的多线程优化，大大加快了其训练速度，使得我们有更多的调参选择。由于观测classification report发现其他分类器与其具有强关联性，不论使用何种模型融合方式都没有获得更好的结果，所以最终选择了这个模型。

对于模型参数的选择与优化思路

由于linearsvc参数量较小，加上我们优化之后的训练时间及其短，我们使用grid search很容易就找到了相对好的模型参数，其中最重要的即惩罚参数C 置为0.6

对sklearn中linearsvc的多线程优化

我们知道linearsvc的作者发布了支持多线程加速的liblinear的库，我们修改了相关接口，实现了多线程训练，对比传统sklearn中的linearsvc的速度提升如下表

线程数	训练时间	训练设备
单核	1911.7秒	I7 7700HQ
多核（8线程）	352.44秒	I7 7700HQ

表 数据集B的linearsvc训练时间