文本大作业

张配天-2018202180

2020 年 5 月 31 日

目录

1 文件结构及功能

```
projects/
├── TextClassification/ 保存短信处理时的各个文件
│   ├── segment.py 分词
│   ├── preprocess.py 预处理 (去除符号及单个字、英文全部转化为小写、连续的数字转为 x)
│   ├── training.ipynb 训练垃圾短信分类器，给 test 数据及打标签
│   ├── vectorize.ipynb 文本向量化，deprecated
│   └── naiveBates.py 使用 nltk 的方法进行朴素贝叶斯分类，deprecated
└── EsSearch/保存搜索引擎的各个文件
    ├── aggregate.ipynb 将训练集和测试集合并
    ├── buildSearch.py 构建搜索引擎
    ├── interface.py 用户查询接口
    └── searchEngine.ipynb 用于方便的测试 es 搜索引擎的输入
```

2 思路和实现方法

2.1 垃圾短信分类

I. 利用 **jieba** 对训练集、测试集短信进行分词，利用 **gensim 库的 preprocess 和 re 库**进行预处理(去除符号，英文转化为小写，去除停用词，连续的数字记为 x)，分别得到 test_preprocess.txt 和 train_preprocess.txt

II. 利用 **sklearn 的 CounterVectorizer 和 TfidfTransformer** 获取训练集中所有文档的 **tf-idf 矩阵**，以 7:3 的比例划分训练集和测试集进行交叉验证，并传入 **MultinomialNB** 模型中进行朴素贝叶斯计算，得到训练结果如 图 *1*所示

```
[[214760   1322]
 [  2019  21899]]
              precision    recall  f1-score   support

           0       0.99      0.99      0.99    216082
           1       0.94      0.92      0.93     23918

    accuracy                           0.99    240000
   macro avg       0.97      0.95      0.96    240000
weighted avg       0.99      0.99      0.99    240000
```

图 1: 交叉验证得到的分类器的准确率、召回率和 f1-score

III. 加入测试集的所有文档，重新构建语料库，计算 tf-idf，否则**由于单词总数不匹配**，使用训练集得到的 ConterVectorizer 对测试集的文档进行词频统计并转化为 tf-idf 后因维度不匹配无法传入模型。

IV. 使用**训练集的数据**训练分类器

V. 使用分类器预测测试集短信是否为垃圾短信，并将结果写入测试集

2.2 搜索引擎构建

I. 配置好环境后实例化 elasticsearch(下统称 *es*)，**创建 index，名为 message**

II. 由于要将原始文本数据传入搜索引擎，所以必须**修改 es 的默认 analyzer 和 search_analyzer，设置为中文的分词系统，否则无法查询**

III. 将打好标签的训练集和测试集合并，每一行提取出短信内容和标签，分别作为 **Field content 和 Field label** 保存入字典

IV. 用 **elasticsearch.helper.bulk** 向 es 实例中批量传入数据，不能用循环 index 传入，那样会卡住

2.3 用户接口

I. 用户输入搜索关键词 (用空格分隔)

II. 程序返回一共有 n 条匹配信息

III. 用户指定程序呈现前 k 条

IV. es 默认的字典保存前 10 条匹配结果，因此需要使用 scroll 游标保存大于 10 条的结果并返回

V. 程序返回结果，如 图 2 图 3所示

```
(base) PS C:\Pt_Python> & C:/apps/Anaconda/python.exe c:/Pt_Python/projects/esSearch/interface.py
type in your keys:
辅导 孩子 价格
there are 10000 results found in total , how many results you would like to attain:
10
****正常短信****女孩读高中18岁时有了孩子~孩子的辅导老师问
****正常短信****xx年义务辅导困难孩子“红蜡烛教育小组”
****垃圾短信****翰林辅导班开始报名中，小饭桌＋晚辅导，免费接送、价格优惠，欢迎前来报名…刘老师。
****垃圾短信****数学，由专业数学教师罗老师上课。辅导地址不变。辅导费每科xxx元，如果您孩子语数都需要辅导，一定给您再优惠xx元。作为您孩子
****正常短信****有些人怎么这么讨厌等你有空辅导辅导我们家孩子吧我给你钱这是钱的问题吗
****正常短信****法院上只要价格合适就轻易放弃孩子
****垃圾短信****翔宇教育明天下午辅导开始。下午作业辅导，一对一，周末小班，寒暑假辅导。祝孩子在新学期快乐成长学习进步！
****垃圾短信****对学生薄弱学科底进行强化辅导）确保孩子成绩有所提高！特设全托xxx元／月（早餐＋晚餐＋辅导＋住宿）；半托xxx元／月（晚餐＋
****正常短信****我们在新浦幼儿园辅导孩子们完成暑假作业
****正常短信****x算算小孩子的都在xx百的价格
(base) PS C:\Pt_Python>
```

图 2: 搜索结果

```
(base) PS C:\Pt_Python> & C:/apps/Anaconda/python.exe c:/Pt_Python/projects/EsSearch/interface.py
type in your keys:
促销 降价
there are 2665 results found in total , how many results you would like to attain:
20
****正常短信****目前一些加油站开始竞相降价促销
****垃圾短信****沃尔玛也将于x月xx日展开降价大促销活动
****正常短信****以350美元销售是为了应对索尼PS4的降价促销
****正常短信****今天的金盒特价是Gunnar视力保护眼镜60%降价促销
****正常短信****中介又打电话来游说降价降价降价
****正常短信****xxibmThinkpad机型大降价
****垃圾短信****正直春季，容易嗓子不舒服，苏梅爽含片降价促销，底价x.x特价销售，供货价x.x，预购从速
****垃圾短信****降价啦！艾尚雪答谢新老顾客，让利大酬宾大型促销活动开始了，时间截止到x月xx日，欢迎选购！
****正常短信****华为荣耀7降价2700
****正常短信****自营SafariBeige色大幅降价
****垃圾短信****美甲用品…冰点降价
****垃圾短信****您好！中东丰田汽车霸道xxxx xxxx 酷路泽xxxx xxxx xxxx年后降价促销，欢迎新老客户致电！ xxxxxxxxxxx 杨经理
****正常短信****blackberrypassportclassic均有小幅度降价
****正常短信****微软surfacepro平板电脑winx降价
****垃圾短信****魏县优彩饰家全体员工：祝新老客户羊年发财！     xxxx年所有产品（包括新款）一律降价，让利促销。   详询：xxxxxxxxxx
****正常短信****GSK中国宣布乙肝药降价三成或掀原研药降价|E药脸谱网|
****正常短信****“先涨价再降价”等猫腻再现
****正常短信****360自营上门手机维修降价咯
****正常短信****华为荣耀xplus啥时候降价啊
****正常短信****又要降价啦～明天来加油呀～～～～
****正常短信****高端手机市场首轮降价如期来临
```

图 3: 搜索结果

3 思考与改进

- elasticsearch 是个好东西，值得研究，大创里可能会用到
- sklearn 牛逼