文本分类模块使用流程

1.创建项目文件夹

进入data文件夹, 执行 `sh create_new_project.sh PROJECT_NAME` 命令创建项目文件夹，该同时会生成默认配置文件params_dict.py和规则文件rule.txt；

2.导入训练测试数据

训练和测试数据需按照固定格式放置在data/PROJECT_NAME文件夹下，格式参考《class_documentation》中的说明要求；

3.修改配置文件

打开配置文件params_dict.py，根据实际分类需求进行修改：

feature_method表示采用的特征提取方法，支持一个或多个；
feature_args表示采用特征提取方法对应的key、value参数字典，其数量必须与提取方法数一致，可以为空字典；
classifier表示采用的分类器或组合；
classifier_args表示采用分类器所对应的key、value参数字典，其含义请参考sklearn官方文档；

4.训练模型

进入src目录，运行以下命令进行训练：

```
python run_classify.py –t TRAIN_FILE_NAME PROJECT_NAME
```

如：

```
python run_classify.py –t train_mini.csv demo
```

其中train_mini.csv为data/demo/目录下的训练数据文件，完成训练后，data/demo/目录

下会生成”.model”后缀的模型文件；

5.测试和预测

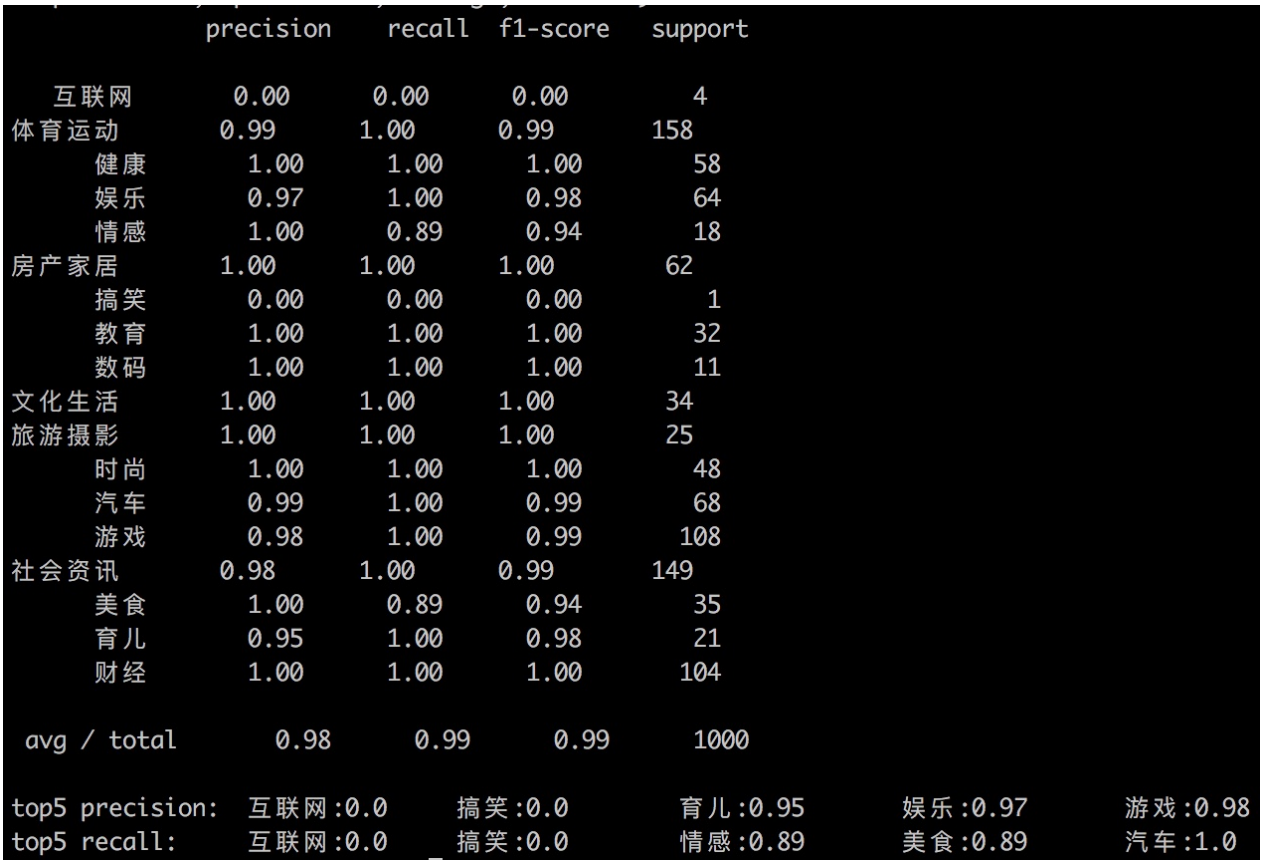
进入src目录，运行以下命令进行测试：

```
python run_classify.py -e -p TEST_FILE_NAME PROJECT_NAME
```

如：

```
python run_classify.py -e -p test_mini.csv demo
```

其中test_mini.csv为data/demo/目录下的测试数据文件，如设置“-e”参数，则文件需包含标注信息（即“label”列），且测试过程中将打印各个类别及整体的准确率、召回率和F1值，如下图所示：



如果去掉“-e”参数，则表示只对样本进行预测而不做准确性评估；

完成测试后，data/demo/目录下将生成predict.csv文件和diff.csv文件（需“-e”）。predict.csv文件包含预测结果的“predict_label”列，diff.csv文件表示预测结果与标注不符的badcase，如下图所示：

```
text,item_info,label,predict_label,comment
你是个喜欢狡辩的人吗,,情感,社会资讯,喜欢 狡辩
儿子的午餐----妈妈的爱伴你健康快乐成长[楼],,美食,育儿,儿子 午餐 妈妈 爱 伴 健康 快
你是天生的劈腿高手吗,,情感,社会资讯,天生 劈腿 高手
阿里股票跌，马云蔡崇信大手笔回购阿里股份丨钛晨报,,互联网,娱乐,阿里 股票 跌 马云 蔡
天津大爆炸，看看哪些大型数据中心受损了？,,互联网,社会资讯,天津 大爆炸 看看 大型 数
享美食 丨 令人心痒难挡的炸物 你吃过几个？,,美食,体育运动,享 美食 令人 心痒 难 挡
扎克伯格：Google+只是小号Facebook,,互联网,游戏,扎克伯格 Google 小号 Facebook
相亲经典顺口溜 ， 笑尿了！,,搞笑,娱乐,相亲 经典 顺口溜 笑 尿
用新鲜上市的甜青豆做一碗西…,,美食,汽车,新鲜 上市 甜 青豆 做 一碗 西
"淘宝商城""光棍节""涉嫌欺诈",,互联网,游戏,淘宝商城 光棍节 涉嫌 欺诈
米饭这样炒，好吃的不得了！99%的看后都收了！,,美食,体育运动,米饭 炒 好吃 不得了 99
```

支持训练测试同时进行：

```
python run_classify.py -t train_mini.csv -e -p test_mini.csv demo
```

6.规则修正

若配置文件中设定classifier为lr_rule、svm_rule、multi_rule，表示分类器支持添加人工规则对分类进行修正。规则文件rule.txt位于data/PROJECT_NAME文件夹下。

参数关键字 “# gep_len: 5” 表示多关键词规则中关键词生效的最大间隔；”# rule_score: 0.3” 表示规则命中后的加分；

规则支持“一对一”、“一对多”、“多对一”和“多对多”的形式如：

```
word1 \t label1                  #一对一
word1 \t label1 label2           #一对多
word1 word2  \t label1           #多对一
word1 word2  \t label1 label2    #多对多
```

注：
1、关键词默认支持正则，请注意字符转义；
2、英文关键词请小写表示；
3、vim下注意tab键是否输出\t而不是四个空格符，否则用“ctrl+v、tab”代替。