API列表

概述

Feature Extractor Module

FeatureExtractorBase

描述

特征抽取基类。

构造参数:
module_dir —— string类型, 必须要接受一个输出路径，输出模型到模块路径下，防止命名冲突，建议传入调用分类器的模块路径

方法

- **load_dict**()

  装载词典，有可能包括分词词典，关键词词典，词向量模型等

- **train**(*train_file_path*)

训练特征抽取模型，需要训练样本
，可能是统计tf、idf或者互信息量等信息。其中train_file_path是训练样本路径

- **gen_feature**(*text*, *item_info=None*)

  抽取特征

参数	类型	备注
text	string	文本
item_info	dict	item信息

  return: 返回sparse matrix类型或者dense matrix类型

- **gen_feature_batch**(*batch_item_lst*)

  批量抽取特征

参数	类型	备注
batch_item_lst	list	批量处理的item列表，item为(text,item_info)

  return: 返回sparse matrix类型或者dense matrix类型

- **gen_comment**(*text*, *item_info=None*)

  生成抽取特征备注信息

参数	类型	备注
text	string	文本
item_info	dict	item信息

  return: 返回string类型

WordsegExtractor

描述

继承于FeatureExtractorBase。
抽取的特征是基于分词，特征权重通过计算tfidf获得。

Classifier Module

所有分类器继承于ClassifierBase，其中，
**BasicClassifier表示一级分类器，其他表示二级分类器（业务用或者融合模型）。

ClassifierBase

描述

分类器基类。已经实现了预测、评估等方法，继承子类需要重写train、 load_model、classify、get_classify_comment这些方法。

构造函数参数包括：
train_file_path —— string类型，训练集路径，为空时采用默认值，默认值见配置conf.TRAIN_FILE_PATH
module_dir —— string类型, 模块输出路径，必须要接受一个输出路径，输出模型到模块路径下，防止文件名冲突，为空时默认采用分类器名称在data下建立新目录

方法

- **set_file_path**(*train_file_path=None*, *module_dir=None*)

  设置训练集和模块路径，ClassifierBase基类已实现

参数	类型	备注
train_file_path	string	训练集路径，csv格式， 表头包含“label”，“text”，“item_info”
module_dir	string	模块路径

- **train**()

  使用训练集样本训练模型

- **load_model**()

  装载训练好的模型

- **classify**(*text*, *item_info=None*)

  计算分类结果

参数	类型	备注
text	string	文本
item_info	dict	item信息

return: 分类结果的类型为dict类型，形式为{cate_name1:score1,cate_name2:score2}

- **gen_classify_comment**(*text*,*item_info=None*):

  生成分类备注信息，譬如可以输出 命中的规则，或者说抽取到了什么特征

参数	类型	备注
text	string	文本
item_info	dict	item信息

return: string类型,备注信息

- **predict**(*test_file_path*, *predict_file_path=None*)

  生成预测结果，ClassifierBase基类已实现

参数	类型	备注
test_file_path	string	测试集路径，csv格式， 表头包含“label”，“text”，“item_info”
predict_file_path	string	预测结果文件输出路径，csv格式，表头包含: “predict_label”，“text”，“item_info”,”comment” 为空时路径为“模块路径/ predict.csv”

- **evaluation**(*test_file_path*, *predict_file_path=None*, *diff_file_path=None*)

  评估结果并打印，ClassifierBase基类已实现

参数	类型	备注
test_file_path	string	测试集路径，csv格式， 表头包含“label”，“text”，“item_info”
predict_file_path	string	预测结果文件路径，csv格式，表头包含: “predict_label”，“text”，“item_info” 为空时路径为“模块路径/ predict.csv”

参数	类型	备注
diff_file_path	string	diff结果文件输出路径，csv格式，表头包含: “label”，“predict_label”，“text”， “item_info”，”comment” 为空时路径为“模块路径/ diff.csv”

diff结果示例：

```
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
```

评估结果示例：

```
              precision    recall  f1-score   support

   互联网       0.00      0.00      0.00         4
体育运动       0.99      1.00      0.99       158
      健康       1.00      1.00      1.00        58
      娱乐       0.97      1.00      0.98        64
      情感       1.00      0.89      0.94        18
房产家居       1.00      1.00      1.00        62
      搞笑       0.00      0.00      0.00         1
      教育       1.00      1.00      1.00        32
      数码       1.00      1.00      1.00        11
文化生活       1.00      1.00      1.00        34
旅游摄影       1.00      1.00      1.00        25
      时尚       1.00      1.00      1.00        48
      汽车       0.99      1.00      0.99        68
      游戏       0.98      1.00      0.99       108
社会资讯       0.98      1.00      0.99       149
      美食       1.00      0.89      0.94        35
      育儿       0.95      1.00      0.98        21
      财经       1.00      1.00      1.00       104

 avg / total       0.98      0.99      0.99      1000

top5 precision:  互联网:0.0     搞笑:0.0       育儿:0.95      娱乐:0.97      游戏:0.98
top5 recall:     互联网:0.0     搞笑:0.0       情感:0.89      美食:0.89      汽车:1.0
```

SvmBasicClassifier

描述

继承于ClassifierBase。svm模型构建的分类器

RuleBasicClassifier

描述

继承于ClassifierBase。规则构建的分类器

CommonClassifier

描述

继承于ClassifierBase。
内部实例化了SvmBasicClassifier和RuleBasicClassifier。
将子分类器的预测分数进行了合并统一。

Other

Preprocessor

方法

- text_decode(*text*)

  编码转换，将其他编码转换成utf-8