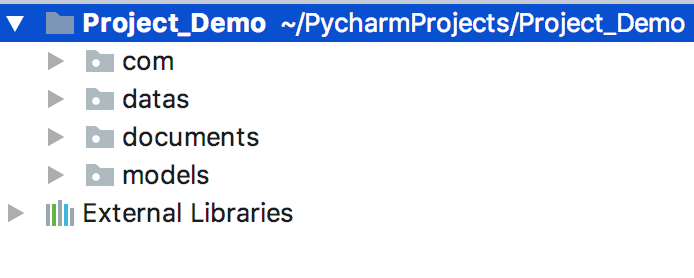
# 图片评论舆情判断模型代码试用说明

## 一、代码运行目标

该阶段项目根据对论坛图片的评论进行分析，通过对评论进行中性，差评，好评（分别标记为0，1，2）三种标记，从评论中的好中差成都来判断图片的（1，2，3，4，5）的等级。在本程序中，将评论的处理流程作为主要的工作，通过对评论的好中差程度的分析从而预测图片等级。本程序主要完成了对中文评论的处理，然后通过统计学习方法进行评论的预测。

## 二、代码主题结构

·com：存放了所有的py文件

·datas：存放了所有的数据文件

·model：存放了所有训练后的模型

·documents：存放着所有说明文件

## 三、功能模块介绍

（1）data\_manage.py

将原始的excel数据作为输入，通过程序的分词、向量化、样本平衡等操作输出训练用的数据集和用于测试用的数据集。

|  |  |  |  |
| --- | --- | --- | --- |
| **函数名称** | **功能描述** | **入参说明** | **出参描述** |
| copy\_list | 拷贝样本，按照较大的list将较小的list拷贝多次，使得新的较小的list与较大的list的大小相等 | large\_list：长度较长的list  small\_list：长度较短的list | small\_list\_copy：按照较长list拷贝后的新的较小的list |
| data\_processing | 涵盖了整个数据处理的流程 |  | train\_feature\_list：训练集特征  train\_tag\_list：训练集标签 test\_feature\_list：测试集特征  test\_tag\_list：测试集标签 |
| disperse\_samples | 将数据样本按照测试机与训练集1:9的比例，随机生成测试集和训练集 | List：整个list数据 | (train\_feature\_list)：训练集特征  (train\_tag\_list)：训练集标签(test\_feature\_list),：测试集特征  (test\_tag\_list)：测试集标签 |
| fenci | 使用jieba分词将中文评论进行分词 | Sentence：中文评论 | cutted\_sentence：切分之后的list，返回该句话的list |
| generate\_tag\_feature\_list | 从数据源中将数据读取出来，并且将特征向量与类别区分出来 | data\_source：数据源 | generate\_list：数据list |
| sample\_balance | 样本平衡，按照大的样本，将其他小的样本复制多份，非整份的则随机抽取 | matrixs\_list：所有的特征和样本值 | list\_2：样本平衡后的list |
| sentence\_2\_vec | 计算每一句的向量，用平均向量来表示这句话的向量 | cutted\_sentence：被切分后的评论 | cutted\_sentence\_list\_mean：该评论的向量值 |
| sperate\_tag\_feature | 把特征list和标签list拆分出来 | List：带有所有样本的数据list | feature\_list：特征值  tag\_list：标签值 |
| stopwords | 读取停顿词 | stop\_words\_path：停顿词的路径 | Stopwords：停顿次的list |
| vip\_word\_dic | 建立此向量的字典，便于后面的词向量的获取（腾讯词向量没有下载下来，所以用的是临时的） | vip\_words\_path：词向量的路径 | vip\_dict：词向量的字典 |

（2）models.py

根据处理好的数据的特征向量来进行模型的训练，并保存模型，进行预测。

|  |  |  |  |
| --- | --- | --- | --- |
| 函数名称 | 功能描述 | 入参说明 | 出参描述 |
| confusion\_matrix | 生成模型预测结果的混淆矩阵 | expect：原标签值  predict,：预测的标签值  str1：模型文件 |  |
| generate\_models | 将训练好的模型进行保存 |  |  |
| predict\_with\_models | 通过训练好的模型，进行保存 |  |  |
| write\_2\_txt | 把结果写入到txt中 | filename：文件名  str2：需要写入的字符串 |  |

## 四、具体功能实现流程

（1）将新闻向量化

a)数据处理模块data\_manage.py，主要完成了对图片的中文评论使用结巴分词进行分词，然后将分词出来的结果进行词向量化，将每句评论的向量取平均值，作为这句评论的特征向量。

b)由于标注的数据主要分为0，1，2三类（三类数据的样本数量不同，需要将较少的数据平衡到与较多数据一样的标准，在抽样过程中，才能保证随机性），标注为2的数据样本较大，通过复制0和1的数据样本达到数据平衡。

（2）评论舆情的正负判断

评论舆情判断模块models.py，主要完成了对评论数据的读取、处理和预测，通过训练过的模型，得到图片评论正负舆情判断的结果，写入到result文件中。

## 五、改进（下一步工作）

1. 评论向量化：在处理中文评论中，可以用tf-idf来获取重要词汇作为词向量字典，在进行中文向量化时使用该字典中的含有的词；同时，在向量化的过程中，本版本程序是通过将每个词向量化后取平均值，在改进中，可以只取权重比例前N的词，进行向量化，长度超过N则截取，不足N则用0填补。
2. 特征值：目前只有中文评论这一个维度，可以再找一些维度作为特征。
3. 模型训练：本文使用了基本的机器学习算法来进行模型训练与预测，接下来可以使用light\_BGM、Bert、fast-test、TextCNN等神经网络和深度学习算法。
4. 进行图片判断，目前设想是将评论加权平均，此时则应该将评论好、中、差标记为1，0，-1三种标记，便于评论正负的计算；除此之后也可以引入算子来对评论进行加权计算，算子的增加需要更进一步思考，可参考有一篇“电影推荐算法”文章的算子引入。