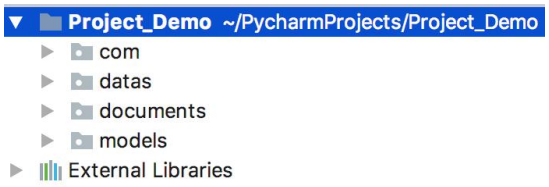
图片评论舆情判断模型代码试用说明

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。