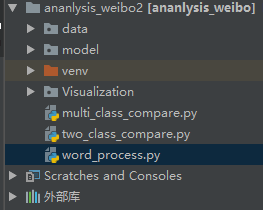
1. **文件结构**

文件结构主要分为数据、可视化、模型和具体处理



data文件夹中存放数据预处理、训练、测试所需的文件，stop.txt为数据预处理所需要的停留词文件，train.txt为模型训练文件，test.txt为模型测试文件。文件列表如下图：



model文件夹存放训练的模型，存放了二分类和多分类所训练的贝叶斯、SVM、Adaboost模型，模型文件列表如下图：



Visualization文件夹存放可视化界面的代码文件，测试模型文件生成的二分类、多分类结果对比图片与文档，文件列表如下图：



1. 功能实现
2. 数据预处理

对微博数据的预处理文件为word\_process.py，

|  |  |  |
| --- | --- | --- |
| 函数名称 | 相关说明 | |
| build\_key\_word() | 参数 | 文件路径 |
| 返回值 | 生成词向量所需关键词 |
| 功能 | 从文件中读取微博内容，对微博内容进行分词，并删除停留词等干扰项，统计分词结果的出现频次，取前20%生成词向量所需关键词 |
| loadDataSet() | 参数 | 文件路径或文本，加载模式 |
| 返回值 | 每条微博的分词与情感标签 |
| 功能 | 若加载模式为train或test，则从文件中读取微博内容，对微博内容进行分词，并删除停留词、空格等干扰项，返回每条微博的分词结果和情感标签；  若加载模式为text，则对文本进行分词，并删除停留词、空格等干扰项，返回文本的分词结果 |
| setOfWordsToVecTor() | 参数 | 关键词表、待处理的单条微博分词词序列 |
| 返回值 | 单条微博对应的一维词向量矩阵 |
| 功能 | 将文本向量化，根据关键词构建词向量，即进行词向量分量的标记工作，将词向量整合转化成词向量矩阵 |
| setOfWordsListToVecTor() | 参数 | 关键词表、待处理的多条微博分词结果 |
| 返回值 | 所有微博对应的二维词向量矩阵 |
| 功能 | 调用函数setOfWordsToVecTor()，将所有文本向量化，生成所有微博的词向量矩阵 |

1. 二分类训练

|  |  |  |
| --- | --- | --- |
| 函数名称 | 相关说明 | |
| cal\_average() | 参数 | 每次训练得到的错误率形成的数组 |
| 返回值 | 平均错误率 |
| 功能 | 求平均错误率 |
| Bayes\_error() | 参数 | 训练文本的词向量矩阵和情感标签，测试文本的词向量矩阵和情感标签 |
| 返回值 | 训练得到的贝叶斯模型，测试错误数量，函数运行时间 |
| 功能 | 计算贝叶斯的二分类错误个数 |
| SVM\_error() | 参数 | 训练文本的词向量矩阵和情感标签，测试文本的词向量矩阵和情感标签 |
| 返回值 | 测试错误数量，函数运行时间 |
| 功能 | 计算SVM的二分类错误个数 |
| Adaboost\_error() | 参数 | 训练好的贝叶斯模型，训练文本的词向量矩阵和情感标签，测试文本的词向量矩阵和情感标签 |
| 返回值 | 测试错误数量，函数运行时间 |
| 功能 | 计算Adaboost的二分类错误个数 |

1. 多分类训练

|  |  |  |
| --- | --- | --- |
| 函数名称 | 相关说明 | |
| cal\_average() | 参数 | 每次训练得到的错误率形成的数组 |
| 返回值 | 平均错误率 |
| 功能 | 求平均错误率 |
| Bayes\_error() | 参数 | 训练文本的词向量矩阵和情感标签，测试文本的词向量矩阵和情感标签 |
| 返回值 | 训练得到的贝叶斯模型，测试错误数量，函数运行时间 |
| 功能 | 计算贝叶斯的多分类错误个数 |
| SVM\_error() | 参数 | 训练文本的词向量矩阵和情感标签，测试文本的词向量矩阵和情感标签 |
| 返回值 | 测试错误数量，函数运行时间 |
| 功能 | 计算SVM的多分类错误个数 |
| Multinomial\_Adaboost() | 参数 | 训练好的贝叶斯模型，基分类器个数，模型学习率，方法，训练文本的词向量矩阵和情感标签，测试文本的词向量矩阵和情感标签 |
| 返回值 | 测试错误数量，函数运行时间 |
| 功能 | 计算Adaboost的多分类错误个数 |

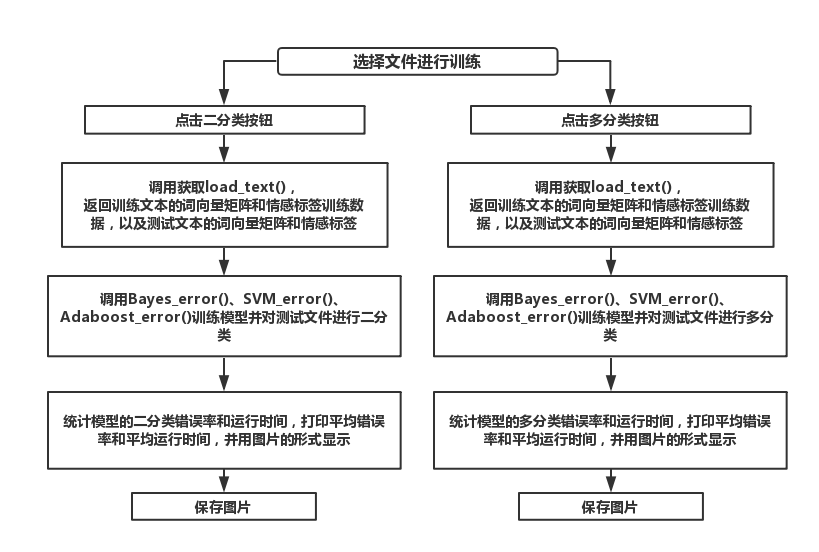
1. 可视化

由于可视化部分函数较多，在这里只介绍部分主要函数的功能。

|  |  |  |
| --- | --- | --- |
| 函数名称 | 相关说明 | |
| load\_text | 功能 | 根据训练文件生成关键词词库，并根据参数mode实现不同功能：   * 若mode=train，对训练文件生成词向量矩阵，从训练数据中随机切分出一部分数据作为测试数据，返回训练文本的词向量矩阵和情感标签训练数据以及测试文本的词向量矩阵和情感标签 * 若mode=test，对测试文件生成词向量矩阵，返回测试文本的词向量矩阵和情感标签 * 若mode=text，对输入文本进行分词，返回关键词表和文本分词结果 |
| file\_2\_class() | 功能 | 调用load\_text()获取训练文本的词向量矩阵和情感标签训练数据以及测试文本的词向量矩阵和情感标签，调用Bayes\_error()、SVM\_error()、Adaboost\_error()训练模型并对测试文件进行二分类，统计模型的二分类错误率和运行时间，打印平均错误率和平均运行时间，并用图片的形式显示 |
| file\_m\_class() | 功能 | 调用load\_text()获取训练文本的词向量矩阵和情感标签训练数据以及测试文本的词向量矩阵和情感标签，调用Bayes\_error()、SVM\_error()、Adaboost\_error()训练模型并对测试文件进行多分类，统计模型的多分类错误率和运行时间，打印平均错误率和平均运行时间，并用图片的形式显示 |
| test\_file\_2\_class() | 功能 | 调用训练好的贝叶斯、SVM、Adaboost模型对测试文件进行二分类，并打印每个模型的二分类平均错误率，并保存到相关文件中 |
| test\_file\_m\_class() | 功能 | 调用训练好的贝叶斯、SVM、Adaboost模型对测试文件进行并保存到相关文件中多分类，并打印每个模型的多分类平均错误率，并保存到相关文件中 |
| two\_Bayes() | 功能 | 对输入文件调用训练好的贝叶斯模型进行二分类，打印分类结果，并保存到相关文件中 |
| two\_SVM() | 功能 | 对输入文件调用训练好的SVM模型进行二分类，打印分类结果，并保存到相关文件中 |
| two\_Adaboost() | 功能 | 对输入文件调用训练好的Adaboost模型进行二分类，打印分类结果，并保存到相关文件中 |
| multi\_Bayes() | 功能 | 对输入文件调用训练好的贝叶斯模型进行多分类，打印分类结果，并保存到相关文件中 |
| multi\_SVM() | 功能 | 对输入文件调用训练好的SVM模型进行多分类，打印分类结果，并保存到相关文件中 |
| multi\_Adaboost() | 功能 | 对输入文件调用训练好的Adaboost模型进行多分类，打印分类结果，并保存到相关文件中 |
| two\_c\_result() | 功能 | 将输入文本和各模型二分类结果写入相关文件中 |
| multi\_c\_result() | 功能 | 将输入文本和各模型多分类结果写入相关文件中 |

可视化部分相关流程图如下：

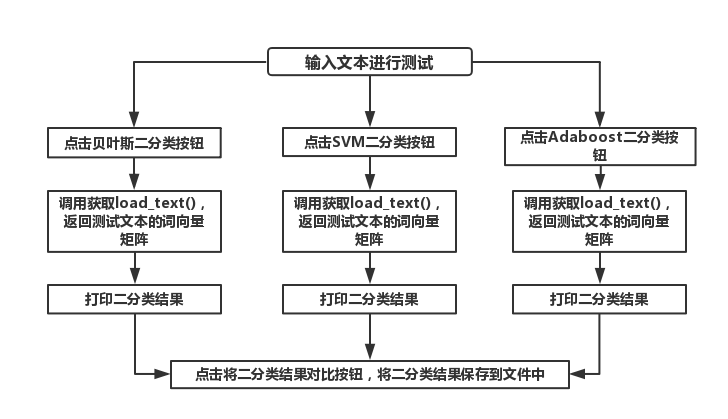
1. 选择文件训练模型，并进行二分类和多分类，打印和保存分类结果



1. 选择文件进行测试，进行二分类和多分类，并打印和保存分类结果



1. 输入文本进行测试，进行二分类，并打印和保存分类结果



1. 输入文本进行测试，进行多分类，并打印和保存分类结果

