**弹幕消息情感分析**

目录

[1 程序模块说明 1](#_Toc491177691)

[1.1 底层功能模块 1](#_Toc491177692)

[1.1.1 基于机器学习的功能模块 1](#_Toc491177693)

[1.1.2 基于词典的功能模块 2](#_Toc491177694)

[1.2 图形界面模块 3](#_Toc491177695)

[2 文件组织结构 3](#_Toc491177696)

[2.1 数据标记过程 4](#_Toc491177697)

[2.2 词典 4](#_Toc491177698)

[2.3 分类器创建 4](#_Toc491177699)

[2.4 产生结果 4](#_Toc491177700)

[3 环境配置 5](#_Toc491177701)

[4 参考资料 5](#_Toc491177702)

# 1 程序模块说明

## 底层功能模块

### 1.1.1 基于机器学习的功能模块

1. textProcessing（文本预处理）

完成数据读取、中文分词、词性标注、句子切割、停用词过滤等功能。

**注意下里面需要用到用户词典、停用词词典。**

用户词典（用于中文分词）保存在D:/ReviewHelpfulnessPrediction/PreprocessingModule目录下，名称为userdict.txt ,可在里面增加一些新的词汇，每行的格式为：词语 词频。如果用户词典存储位置发生更改（不建议更改），则需要更改程序jieba.load\_userdict(userDictPath)。

停用词默认保存在D:/ReviewHelpfulnessPrediction/PreprocessingModule目录下，名称为stopword.txt，用于停用词过滤，如seg\_fil\_excel、seg\_fil\_txt等函数中。

1. unlabelDataProcessToLabel

完成客观语句过滤、重复评论删除、格式转换、已标记数据的检查与错误处理、标记数据合并等功能。

**注意下里面需要用到情感词典。**

情感词典默认保存在D:/ReviewHelpfulnessPrediction/SentimentDict目录下，不建议更改。

**经标记数据合并这一功能后产生的标记数据保存在D:/ReviewHelpfulnessPrediction/LabelReviewData目录下。**

1. selectBestClassifier

完成最佳分类器的选择、最佳特征维度选择的功能。

使用分类算法有：朴素贝叶斯、逻辑回归、支持向量机、k-最近邻、神经网络，使用评价指标是精度（正类、负类的加权平均值），分类效果在朴素贝叶斯上表现最好，可达到分类精度大致在89%左右。

为了优化效果考虑，最佳特征词语里面额外添加了人工标记的一些关键词，所以切不可在标记过程中随便添加关键词，否则可能会影响分类精度（也可修改程序去掉额外添加的人工标记关键词）。

注意下需要使用到已标记的训练数据来训练分类器，需要使用情感停用词来去除语句中的停用词。

标记的训练数据保存在D:/ReviewHelpfulnessPrediction/LabelReviewData目录下，不建议修改，情感停用词保存在D:/ReviewHelpfulnessPrediction/PreprocessingModule/sentiment\_stopword.txt文件路径下。

构建的效果最佳的分类器保存在D: /ReviewHelpfulnessPrediction/BuildedClassifier目录下，用于预测未知数据；最佳分类器名称以及最佳特征维度保存在D: /ReviewHelpfulnessPrediction/BuildedClassifier/bestClassifierDimenAcc文件路径下，用于读取最佳分类器。

1. predictDataPosNegProbility

完成预测未知数据类别、计算窗口范围内语句整体情感得分、寻找出含有不良内容语句等功能。

注意下需要使用到已标记的训练数据来提取特征，已标记的训练数据保存在D:/ReviewHelpfulnessPrediction/LabelReviewData目录下。

预测出的未知数据类别信息保存在D:/ReviewHelpfulnessPrediction/PredictClassRes目录下，绘制出的情感波动曲线图、类别成分占比图保存在D:/ReviewHelpfulnessPrediction/SentimentLineFig目录下，含有不良信息内容的语句保存在D:/ReviewHelpfulnessPrediction/StrangeWords目录下。

### 1.1.2 基于词典的功能模块

1. textProcessing（文本预处理）

同上

1. createDict

完成不同词典内容去重、合并功能（图形界面还缺乏这一内容）

1. sentimentAnalyzeBasedDict

完成预测未知数据类别、计算窗口范围内语句整体情感得分、寻找出含有不良内容语句等功能。

注意下需要使用已标记的训练数据来测试基于词典的情感分析的分类精度，已标记的训练数据保存在D:/ReviewHelpfulnessPrediction/LabelReviewData目录下。（看看计算精度算法）

预测出的未知数据类别信息保存在D:/ReviewHelpfulnessPrediction/PredictClassRes目录下，绘制出的情感波动曲线图、类别成分占比图保存在D:/ReviewHelpfulnessPrediction/SentimentLineFig目录下，含有不良信息内容的语句保存在D:/ReviewHelpfulnessPrediction/StrangeWords目录下。

## 图形界面模块

1. basicParmSetDialog

完成基本参数设置功能，包括窗口大小、情感得分边界、积极分值边界、消极分值边界这四个参数，用于静态分类。

1. dynamicParmSetDialog

完成基本参数设置功能，包括窗口大小、情感得分边界、积极分值边界、消极分值边界、更新时间间隔这五个参数，用于实时抓取之当前静态图、实时抓取之全部静态图。

1. dyLineParmSetDialog

完成基本参数设置功能，包括窗口大小、情感得分边界、积极分值边界、消极分值边界、更新时间间隔、显示消息数、动态速度这七个参数，用于绘制动图。

1. BDSA

所有模块功能的集成，完成显示图形界面、处理各种事件的功能。

图形界面框架包括菜单栏、显示区、状态栏、进度条。

菜单栏包括设置、预处理、机器学习、字典这四个功能模块。

显示区分为文本显示区和图形显示区。

状态栏显示运行过程中的一些状态信息。

进度条用于显示实时抓取数据过程中的更新进度。

# 文件组织结构

程序运行过程中需要使用的文件以及所产生的文件均保存在D:/ReviewHelpfulnessPrediction目录下。

## 数据标记过程

1. D:\ReviewHelpfulnessPrediction\BulletData

用于保存经过客观语句过滤后的原始弹幕数据

1. D:\ReviewHelpfulnessPrediction\LabelingData

用于保存正在标记过程中的弹幕数据

1. D:\ReviewHelpfulnessPrediction\SpeNameLabeledData

用于保存已标记完的特定名字的弹幕数据

1. D:\ReviewHelpfulnessPrediction\LabelReviewData

用于保存已标记的全部弹幕数据

1. D:\ReviewHelpfulnessPrediction\KeyWords

用于保存人工标记的一些关键词语

## 词典

1. D:\ReviewHelpfulnessPrediction\SentimentDict

用于保存积极消极情感词典以及程度词词典

1. D:\ReviewHelpfulnessPrediction\PreprocessingModule

用于保存中分分词过程中使用到的用户词典以及停顿词词典

## 分类器创建

1. D:\ReviewHelpfulnessPrediction\BuildedClassifier

用于保存构建出的分类器以及最佳分类器所在的文件路径

## 产生结果

1. D:\ReviewHelpfulnessPrediction\PredictClassRes

用于保存原始弹幕数据的预测类别

1. D:\ReviewHelpfulnessPrediction\SentimentLineFig

用于保存情感波动曲线图以及类别成分占比图

1. D:\ReviewHelpfulnessPrediction\StrangeWords

用于保存含有不良信息内容的弹幕数据

1. D:\ReviewHelpfulnessPrediction\StrangeTimes

用于保存出现不良信息内容的时间段

# 环境配置

1. 编译器：Python2.7.12
2. 集成开发环境：Pycharm2017.1.5
3. 需要使用到的库

**nltk、jieba** 用于中文分词、词性标注、自然语言处理

**sklearn** 提供了大量的分类算法，例如朴素贝叶斯、支持向量机、神经网络等

numpy 用于数值计算、矩阵处理

**matplotlib** 用于图形绘制

**PyQt4** 用于图形界面设计

xlwt、xlrd用于excel文件的读写

time 用于时间的获取、运行时间的计算

# 4 参考资料

1. Review-Helpfulness-Prediction

<https://github.com/Azure-rong/Review-Helpfulness-Prediction>

使用文本挖掘技术 包括自然语言处理，情绪分析和机器学习技术。

语言：Python。（标准库：xlrd，jieba，gensim，nltk，scikit-learn）

该项目包括以下主要部分： 查看数据集、文本预处理模块 、查看有用特征提取模块、审查有用性预测模块。

1. 知网情感词典

<http://www.keenage.com/html/c_bulletin_2007.htm>

由知网发布的词典，包括中文情感词典和英文情感词典。

1. snownlp—中文文本内容处理算法

<https://github.com/isnowfy/snownlp>

SnowNLP是一个python写的类库，可以方便的处理中文文本内容，是受到了TextBlob的启发而写的，由于现在大部分的自然语言处理库基本都是针对英文的，于是写了一个方便处理中文的类库，并且和TextBlob不同的是，这里没有用NLTK，所有的算法都是自己实现的，并且自带了一些训练好的字典。注意本程序都是处理的unicode编码，所以使用时请自行decode成unicode。

1. 使用python+机器学习方法进行情感分析

<http://www.10tiao.com/html/284/201607/2652389939/1.html>

详细介绍了应用机器学习方法进行情感分析的大致过程

1. 徐琳宏, 林鸿飞, 赵晶. 情感语料库的构建和分析[J]. 中文信息学报, 2008, 22(1):116-122.

邓扬, 张晨曦, 李江峰. 基于弹幕情感分析的视频片段推荐模型[J]. 计算机应用, 2017, 37(4):1065-1070.

介绍了数据标记规范以及相应注意事项

1. 基于词典的中文情感倾向分析算法设计

<http://site.douban.com/146782/widget/notes/15462869/note/355625387/>

1. 应用PyQt4进行图形界面设计

http://zetcode.com/gui/pyqt4/

1. 文本分类技术

<http://www.blogjava.net/zhenandaci/category/31868.html>