虚假评论检测工作移交说明

1.爬虫算法：spiderfinal.py

2.主要参考算法见文件夹“Review-Helpfulness-Prediction-master”

（跟人进行修改后的程序已经加以注释，原程序可以在下文中给出的链接中下载）

**算法主要思想：**

* Python 文本挖掘：使用情感词典进行情感分析（算法及程序设计）

**<http://rzcoding.blog.163.com/blog/static/2222810172013101844033170/>** 及**后一篇博客**

情感词典法个人进行修改后的相关程序名称：

Textprocessing.py（预处理） pos\_neg\_senti\_dict\_feature\_ZJX.py

* Python 文本挖掘：使用机器学习方法进行情感分析

**<http://rzcoding.blog.163.com/blog/static/22228101720131019104850984/>** 及**后几篇博客**

机器学习法个人进行修改后的相关程序名称：

textprocessing.py store\_sentiment\_classifier\_ZJX.py pos neg ml feature.py

博主写的一篇相关的论文：面向评论效用评估的文本情感特征提取

（论文链接：<http://www.infotech.ac.cn/article/2015/1003-3513-31-7-113.html> ）

3.文件夹classify\_byRST：

**classify\_byRST.py**:根据从大众点评网上所爬取的内容，绝大多数评论中含有包含以下可以用于进行情感极性判断的数据rst1、rst2、rst3：【商品的rst1、rst2、rst3，如美食的是：口味、环境、服务；3者的评分取值范围为：1（一般）、2（好）、3（很好）、4（非常好）】，，只有极少数不含数据rst1、rst2、rst3；故而用以上的数据（rst1、rst2、rst3）通过加权平均来进行评论的情感极性判断，再用所得结果训练分类器；对极少数缺少以上信息(rst1、rst2、rst3)的评论，则通过分类器进行情感极性判断。

1. 文件夹review\_TEST\_22o：**review.xlsx**:利用爬虫算法获取的大众点评网上的部分数据（经过处理为xlsx形式）

**情况介绍**：用情感词典、利用博客中自带的训练集经过机器学习得到的分类器对从大众点评网上爬到的评论进行了分类，发现两者分类效果均不是很理想；故而采用**classify\_byRST.py**将有RST的评论分为积极和消极（RstReviewPos.xlsx、RstReviewNeg.xlsx,存在文件夹“classify\_byRST”中），再将（RstReviewPos.xlsx、RstReviewNeg.xlsx）作为训练集得到分类器sentiment\_classifier\_ZJX0719b.pkl，用此分类器对缺少RST的评论（noneRstReview.xlsx）进行分类，分类效果较好（结果直接进行显示，未保存）。