**文本内容计算课程实验报告**

实验名称： 文本内容获取

姓 名： 覃婧丽

学 号： 201911690341

院 系： 信息科学学院

专 业： 数字媒体技术

二〇二 二 年 四 月

|  |
| --- |
| 1．实验目的 |
| 使用网络爬虫完成基本数据爬取 |
| 2．实验内容 |
| 任务：网页内容获取  • 以当天微博热搜为第一内容，抽取出当天所有热搜的具体信息（排名，标题，热度，标签）  • 抽取当天所有热搜点击后的第一页的所有评论信息（用户id，评论内容，转发数，评论数，点赞数）为第二内容  •该任务以微博为例，但不仅限于微博平台。 |
| 3．实验步骤 |
| • 以当天微博热搜为第一内容，抽取出当天所有热搜的具体信息（排名，标题，热度，标签）   1. 确定热搜首页，获取基本信息      1. 确定排名、标题、热度、标签的获取     排名      标题、热度        标签       1. 循环遍历获取数据      1. 生成excel表格     • 抽取当天所有热搜点击后的第一页的所有评论信息（用户id，评论内容，转发数，评论数，点赞数）为第二内容   1. 确定热搜首页，获取热搜url列表        1. 解析页面数据，循环爬取用户id，评论内容，转发数，评论数，点赞数 |

|  |
| --- |
| 4．实验结果与分析 |
| 热搜爬虫结果    热搜评论信息结果 |
| 5．进一步工作说明 |
| 爬取热搜的时候以html的class形式进行筛选，爬取评论的时候以div[i]的属性进行筛选，发现class形式清晰易懂，div[i]的形式非常容易出错 |