山东大学 软件 学院

信息检索 课程实验报告

|  |  |  |  |
| --- | --- | --- | --- |
| 学号：201800301299 | 姓名： | | 班级： 2018级AI班 |
| 实验题目：第一周：选定项目主题，查找相关资料 | | | |
| 实验学时：5 | | 实验日期：2020年11月7日星期六 | |
| 硬件环境：  PC | | | |
| 软件环境：  Python | | | |
| 实验内容：  1. 选题  灵感来源于在B站刷视频的时候看到弹幕吵架，于是有了想分析b站弹幕的想法。现在bilibili已经成为年轻人手机里必备的APP，bilibili网站（以下简称B站）除了动漫、鬼畜、生活等新颖有创意的内容以外，还有一大特色就是B站的弹幕。弹幕指的是悬浮于视频上方的实时评论，可以给观看视频的人“实时互动”的感觉，弹幕真正让B站从一个单向的视频播放平台，变成了双向的情感连接平台，进而形成了B站的社区文化，创造了许多好玩的梗。于是我想通过爬取B站的弹幕，来分析不同的分区不同的up主不同的视频中，有哪些“弹幕文化”，有哪些好玩的梗，同时分析大多数观众对该视频的感情色彩。这个主题较为新颖，同时也很贴近现实，可以通过大家弹幕中发表的言论，分析大家对该up主或该事件的态度，也可以通过大家发弹幕的时间位置判断视频的精彩瞬间，具有较大的现实价值。  2. 资料查找  首先需要做的是看一下B站弹幕列表。在网页版的B站中随意点击一个视频，进入如下界面：    只观察与弹幕有关的内容，主要有两部分，在左下方显示出该条视频的弹幕总数，右上方有具体的弹幕列表，展开弹幕列表如下：    点击(1)“时间”可以让弹幕列表按照出现时间进行排序，点击(2)“弹幕内容”可以让弹幕按照字典顺序进行排列，而在我查找资料的时间内，(3)“查看历史弹幕”按钮是禁用的，搜索后才发现需要登录。登录后即可按照日期选择当天的弹幕了。    查看网页的源代码，想找一下有没有弹幕的列表，但发现  弹幕列表上方、下方的up主信息、相关推荐都在源代码中显示出来，而弹幕列表却没有，B站果然不会让这么轻易地爬取下来。  于是右键检查元素，界面如下：    但是并没有找到类似于弹幕的文件，且没有通常会保存到的xml文件。  然后我才想到，弹幕应该是在视频播放的时候实时显示出来的，把视频打开不播放肯定没有相应的弹幕啊..果然，视频一播放就在网络—XHR下多了好多条记录，猜想弹幕一定在里面了。      可以看到有一串9位数字串出现了很多次，同时显示是该条视频的”cid”，而播放不同的视频可以看到这9位数字各不相同，可以认定是弹幕的id。  但是还是没有找到弹幕所在的位置，只好去网上查找方法。  <https://www.bilibili.com/read/cv304122/>通过看这篇文章，可以看到  <http://comment.bilibili.com/252900431.xml>就是这条视频弹幕所在的位置了。  更广泛地说，[http://comment.bilibili.com/<cid>.xml](http://comment.bilibili.com/%3ccid%3e.xml)是视频弹幕存放的位置，也就是只需要知道视频的cid就可以找到视频弹幕的位置了，解决了一个大问题。    可以看到显示出了弹幕列表，还有该文件的max-limit是1500，也就是文件中只包含1500条弹幕。 | | | |
| 结论分析与体会：  1.第一周选题过程比较曲折，最开始已经选定了以新闻爬虫和搜索引擎，但当想到了弹幕这个主题后，更想分析这个数据，所以又更换了主题。  2.想做的内容有：基于python的爬虫，数据分析，慈云和情感分析。  在爬出视频的弹幕列表后，对其进行时间点、数量、用户等的分析，并制作词云形象化显示不同的视频不同的关键词，更能彰显B站文化，最后对弹幕内容进行具体的情感分析。 | | | |
| 下周目标：  实现弹幕爬虫，并进行数据分析。  可能遇到的难点有：  1)如何获取视频的总弹幕数，而不限于max-limit的数量。  2)如果需要查看历史弹幕内容，需不需要实现自动登录。 | | | |