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：第四周：自动爬取某up主所有视频并通过弹幕分析特征 | | | |
| 实验学时：5 | | 实验日期：2020年11月28日星期六 | |
| 硬件环境：  PC | | | |
| 软件环境：  Python | | | |
| 实验内容：  首先要实现如何通过输入up主的名称来进行搜索    搜索李子柒，可以发现搜索中可以限定视频、番剧、用户等，所以选择用户，此时网址变为[https://search.bilibili.com/upuser?keyword=李子柒](https://search.bilibili.com/upuser?keyword=李子柒可以找到15)  可以找到15个用户，要选出真正的知名up主，我们可以设定排序和用户特征。选择粉丝数由高到低排序，网址变为<https://search.bilibili.com/upuser?keyword=>李子柒&page=1&order=fans&order\_sort=0  再选择“认证用户”，可以看到usertype=3，全部用户的usertype=0，普通用户usertype=1，up主的usertype=2。      现在选出来的就是我们真正想要的了。  查看网页源代码：    可以看到李子柒的用户id在链接中有  uid = self.browser.find\_element\_by\_xpath('//\*[@id="user-list"]/div[1]/ul/li/div[2]/div[1]/a[1]').get\_attribute('href')  uid = uid.split('/')[-1].split('?')[0]  于是可以进入up主的投稿主页面，网址为<https://space.bilibili.com/userid/video>  每一页有30个视频，视频多的up主有好几页视频，可以看到可以获取每个视频的基础信息。        要获得up主所有视频，还要知道他所有的页数并进行跳转页。    最后将结果保存到json中，共跑了270个视频的基本信息。    并得到视频的bv号：  再进行弹幕获取（超出文件限制了）：    词云结果如图： | | | |
| 结论分析与体会：  1. Beautiful Soup 提供一些简单的、python 式的函数用来处理导航、搜索、修改分析树等功能。它是一个工具箱，通过解析文档为用户提供需要抓取的数据，因为简单，所以不需要多少代码就可以写出一个完整的应用程序。  Beautiful Soup自动将输入文档转换为 Unicode 编码，输出文档转换为utf-8 编码。你不需要考虑编码方式，除非文档没有指定一个编码方式，这时，Beautiful Soup 就不能自动识别编码方式了。  2. 更深入的学习了网页解析。 | | | |
| 下周目标：  基于该up主的所有视频弹幕进行情感分析。 | | | |