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：第二周：爬取b站弹幕 | | | |
| 实验学时：5 | | 实验日期：2020年11月14日星期六 | |
| 硬件环境：  PC | | | |
| 软件环境：  Python | | | |
| 实验内容：  一、爬取b站弹幕  第一周已经分析了爬取b站弹幕的方法，因此做起来比较顺手。  选取当日b站排行较高的视频作为示例，BV号为BV1f54y1r7HV  https://www.bilibili.com/video/BV1f54y1r7HV    属性有BV号（视频号），CV号（弹幕号，之后给出），headers    获取cid的方法，其中cid指的是b站所有视频对应的都有一个弹幕id，用来存储弹幕的位置，简单理解为弹幕号。我们在response中发现cid的位置：    因此用正则表达式取出cid：      但是用etree解析得到的文件为：    可以发现虽然条数都为max\_limit共3000条，但是内容不一样，再次刷新之后可以发现生成的文件又有刷新，所以应该时b站实时刷新，在这个文件中只保存最新的3000条。    查询资料，得到8个参数的意义  第一个参数代表弹幕出现的时间 以秒数为单位。这个时间就是视频播放的时间，也就是弹幕是在视频播放的第几秒发出的；  第二个参数代表弹幕的模式1..3 滚动弹幕 4底端弹幕 5顶端弹幕 6.逆向弹幕 7精准定位 8高级弹幕；  第三个参数代表字号12非常小,16特小,18小,25中,36大,45很大,64特别大；  第四个参数代表字体的颜色，以HTML颜色的十位数为准；  第五个参数代表Unix格式的时间戳。基准时间为 1970-1-1 08:00:00。也就是你发出弹幕的实时日期时间；  第六个参数代表弹幕池 0普通池 1字幕池 2特殊池 【目前特殊池为高级弹幕专用】；  第七个参数代表发送者的ID，用于“屏蔽此弹幕的发送者”功能；  第八个参数代表弹幕在弹幕数据库中rowID 用于“历史弹幕”功能。  二、jieba分词理论知识准备  1. 找到一个较全的词典库：“NLP民工的乐园: 几乎最全的中文NLP资源库”<https://github.com/fighting41love/funNLP>  包括：  B站可以发的弹幕设置如下，      停用词还应加上颜文字和emoji，不过颜文字可以拆分成一个个的通用符号，或者是建立颜文字表，且emoji有对应的Unicode编码，详细对应表见：<https://apps.timwhitlock.info/emoji/tables/unicode#block-6c-other-additional-symbols>  2. 深入认识jieba分词  1)结巴分词支持三种分词模式：  a)精确模式，试图将句子最精确地切开，适合文本分析；  b)全模式，把句子中所有的可以成词的词语都扫描出来, 速度非常快，但是不能解决歧义；  c)搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词  2)算法原理  基于Trie树(前缀树/字典树)结构实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图（DAG)  采用了动态规划查找最大概率路径, 找出基于词频的最大切分组合  对于未登录词，采用了基于汉字成词能力的HMM模型(隐马尔可夫模型)，使用了Viterbi算法  viterbi算法：相当于多步骤每步多选择模型的最优选择问题，其在每一步的所有选择都保存了前续所有步骤到当前步骤当前选择的最小总代价（或者最大价值）以及当前代价的情况下前继步骤的选择。依次计算完所有步骤后，通过回溯的方法找到最优选择路径  3)算法函数  a)jieba.cut()以及jieba.cut\_for\_search()返回的结构都是一个可迭代的generator，可以使用for循环来获得分词后得到的每一个词语(unicode)，也可以用list(jieba.cut(…))转化为list  b) 开发者可以指定自己自定义的词典，以便包含jieba词库里没有的词。虽然jieba有新词识别能力，但是自行添加新词可以保证更高的正确率  jieba.load\_userdict(file\_name) # file\_name为自定义词典的路径  c)补充停用词词库  d)词频统计  参考：https://taorui.blog.csdn.net/article/details/80128076?utm\_medium=distribute.pc\_relevant.none-task-blog-BlogCommendFromMachineLearnPai2-2.channel\_param&depth\_1-utm\_source=distribute.pc\_relevant.none-task-blog-BlogCommendFromMachineLearnPai2-2.channel\_param | | | |
| 结论分析与体会：  1. 考虑到b站是一个非常结合实时、十分有创意的网站，经常出现许多新梗，所以可能会出现传统词库中没有的词汇，应该自己进行增加。  2. 因为现在只爬取了当天的实时弹幕，且有数量限制，所以想先把基本步骤做完之后再做历史查询的功能增加数据量。  3. 这周了解了一些分词的原理，为下周正式操作做准备。 | | | |
| 下周目标：  实现分词，对弹幕内容进行统计分析，并制作好看的词云。 | | | |