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：第三周：统计分析弹幕数据，制作词云 | | | |
| 实验学时：5 | | 实验日期：2020年11月21日星期六 | |
| 硬件环境：  PC | | | |
| 软件环境：  Python | | | |
| 实验内容：  1. 进行jieba分词  1）将繁体转化为简体,需要下载langconv.py和zh\_wiki.py，见  <https://blog.csdn.net/wds2006sdo/article/details/53583367?utm_medium=distribute.pc_aggpage_search_result.none-task-blog-2~all~sobaiduend~default-1-53583367.nonecase&utm_term=langconv%20python%20%E5%AE%89%E8%A3%85&spm=1000.2123.3001.4430>  <https://github.com/skydark/nstools/blob/master/zhtools/langconv.py>  <https://github.com/skydark/nstools/blob/27c45fceebeef551368d3ca57dc995456dc2ff34/zhtools/zh_wiki.py>    2）将全角转换为半角    3）添加停用词表以及jieba无法识别的词语，例如：    停用词表这里选用哈工大停用词表：stop\_words.txt，里面包括符号等  <https://github.com/fighting41love/funNLP/blob/master/data/%E5%81%9C%E7%94%A8%E8%AF%8D/%E5%93%88%E5%B7%A5%E5%A4%A7%E5%81%9C%E7%94%A8%E8%AF%8D%E8%A1%A8.txt>  4)按词性分词  jieba.posseg.cut(text)  词性标注是在给定句子中判定每个词的语法范畴，确定它的词性并加以标注的过程。    5)制作词云    最开始出现全是框框，百度后才发现是中文的问题，需要增加字体  参考：<https://blog.csdn.net/liubinbin_521/article/details/79321378>  显示结果为：    想要做一个更加好看的，添加样式的词云：    最开始一直显示不出来背景，字体颜色根据背景图生成也是黑色的，查了很久资料才发现是忽略了UserWarning: mask image should be unsigned byte between 0 and 255警告，应该将图片转换为数组，而不是直接open。  同时最开始生成的图像很不清晰，可以用scale调节，例如scale=2, 按原图比列放大，数值越大 图像越清晰，在保存的时候才能看到。  应用的背景图为：结果如下：  PS:锦鲤池嘛这是? | | | |
| 结论分析与体会：  1.新学的函数  ·ord()：是chr()函数（对于8位的ASCII字符串）或unichr()函数（对于Unicode对象）的配对函数，它以一个字符作为参数，返回对应的 ASCII 数值，或者 Unicode 数值  ·全角半角转化  ·collection.Counter():collection在python官方文档中的解释是High-performance container datatypes，直接的中文翻译解释高性能容量数据类型。其中Counter中文意思是计数器，也就是我们常用于统计的一种数据类型.  UserWarning: mask image should be unsigned byte between 0 and 255. Got a float array warnings.warn("mask image should be unsigned byte between 0"  参考资料：  1. 中文jieba分词使用手册  <https://blog.csdn.net/alis_xt/article/details/53522435>  2. 中文分词词性对照表  <https://blog.csdn.net/accp_cn/article/details/71774109?utm_medium=distribute.pc_relevant.none-task-blog-searchFromBaidu-1.control&depth_1-utm_source=distribute.pc_relevant.none-task-blog-searchFromBaidu-1.control>  等  3.python把每个功能都包装得太好了… | | | |
| 下周目标：  自动爬取一个up主的前n个视频（例如10），根据所有弹幕热点内容进行分析该up主的风格。 | | | |