一、实验内容及功能说明

应用（4）：基于“新浪新闻”的多功能搜索引擎

实验内容：

1. 基于urllib.request、 html.parser和urllib.parse的网页链接爬虫
2. 基于numpy的pagerank算法
3. 基于BeautifulSoup和requests的网页内容爬虫
4. 基于tkinter的GUI
5. 基于PIL和io的图片显示器
6. 基于re和jieba的词语识别（中英文）
7. 基于wordcloud的词云显示

功能：

1. 基于热度（pagerank）排序生成新闻标题、时间、部分正文
2. 基于时间（网页自带）排序生成新闻标题、时间、部分正文
3. 基于相关度（正文词频、关键词、标题加权得出）排序生成新闻标题、时间、部分正文
4. 生成热搜词云图以及热搜词汇排行榜
5. 优化界面（欢迎图片、词云图片、单选器、分区显示、用户友好、傻瓜模式、滚动条、后台监测）
6. 爬取网页为：URL中包含“/2019-”或“/2018-”且非特殊网页（视频等）的所有各种格式网页

二、设计方案与设计思路

我首先从应用（1）做起，接着做应用（2）和应用（3），最后集成并进行优化。在查阅大量资料后完成作品。

应用（4）的设计主要注重两点：多功能和用户友好。最初的集成成品，只支持<https://news.sina.com.cn/>开头的新闻网页爬虫，排序只支持pagerank算法排序。为了满足更多的需求和我自己的开发欲望，我加入了更多的东西。

第一种按热度排序，当输入为网址时（http开头），直接按pagerank排序输出，当输入为词语时，将正文、关键词、标题加权后分词排序，若查询词不在前10％的词语中时剔除，之后输出结果。

第二种按时间排序，对于不同种类的网站，分别用不同种的方式爬取时间，并根据公式计算后排序，当输入为网址时（http开头），直接按此排序输出，当输入为词语时，将正文、关键词、标题加权后分词排序，若查询词不在前10％的词语中时剔除，之后输出结果。

第三种按相关度排序，将正文、关键词、标题加权后分词排序，将每个网站的该查询词的比例排序，最终得到结果。

词云图产生前，我用网上下载的停用词表和我自己总结的停用词表进行剔除。词云生成基本时默认属性，背景白色，字体msyh.ttc词语数量为200个。

为了使得用户友好，支持输入带空格/制表符、点击关键词查询、报错提示等很多小功能，还加入了后台进度显示，使等待的时间可视化。

三、程序运行效果

给出程序各种情况下运行结果的截图，并**配以文字说明**。

1. 初始界面：



1. 报错界面



1. 热度（PageRank）排序界面（以中国为例）



1. 时间排序界面（以中国为例）



1. 相关度排序界面（以中国为例）



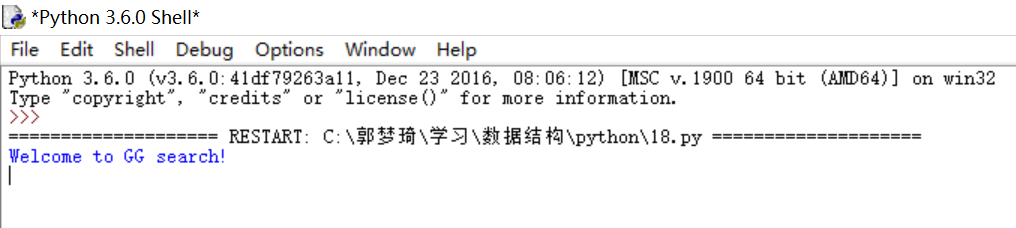
1. 热度（PageRank）排序界面（以https://news.sina.com.cn/为例）

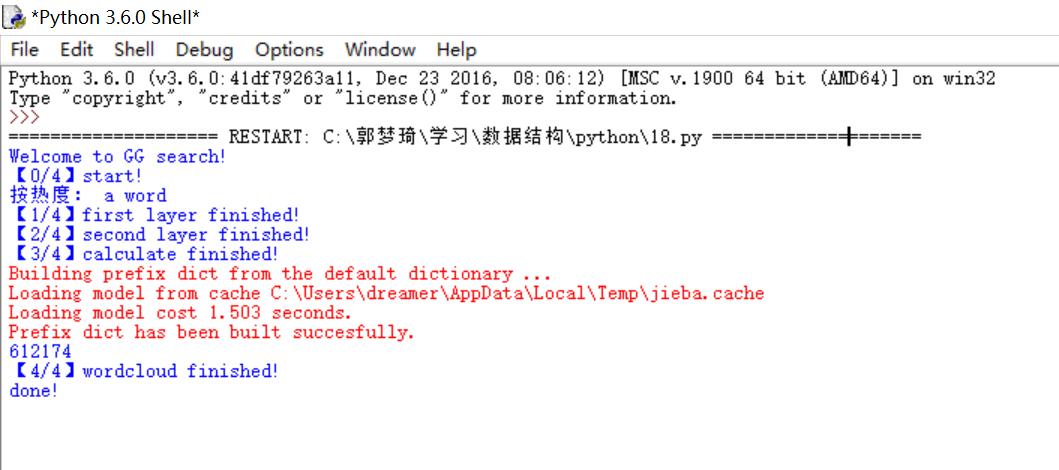


1. 时间排序界面（以https://news.sina.com.cn/为例）



1. 后台界面





四、设计亮点（自选应用必填）

说明为什么设计相应的功能；介绍设计的亮点。

1. 为了增加搜索引擎的覆盖面，支持多种（约6种）格式的新闻网页爬虫（爬取网页为：URL中包含“/2019-”或“/2018-”且非特殊网页（视频等）的所有各种格式网页）
2. 为了满足不同需求，支持三种不同的排序方式（格式化生成新闻标题、时间、部分正文）
3. 生成热搜词云图以及热搜词汇排行榜
4. 为了优化界面，采用滚动条并加入图片，且实现了分区显示
5. 实现傻瓜模式，可以点击热搜词查询，输入栏自动剔除空格或制表符
6. 为了提高程序的维护，提供后台监测进度

五、实验总结

（1）在实验中遇到了哪些问题？是如何解决的？

1. 网页种类太多，单一爬取方式无效

观察不同网页格式，改用多种爬取方式同时用try-except语句。

1. 停用词表不足

观察词云图和热搜词，添加必要的停用词表

1. wordcloud生成的图片过大，无法导入

四处搜索，发现自动修改图片大小的代码，偷过来

1. TK不支持给Frame加滚动条

四处搜索，发现TK支持Frame-Canvas(滚动条)-Frame，修改GUI代码

1. 发现词频产生的相关度排序，以及热词精准度不够

爬取关键词，加入标题，玄学调权重。

1. 。。。。。。

瞎找-乱写-成功了？？？

（2）收（jiao）获（xun）和体（tu）会（cao）

1. 成就感爆棚，一开始觉得自己做不出来，做个应用（1）或应用（2）就可以了，但是做着做着就爱不释手了，连肝四天，做出了一个令现在的我十分满意的作品^\_^（当然留了不少bug）
2. 之前学到python网课没白学，刚好wordcloud用上了，部分代码直接搬运（划掉）
3. 上网搜停用词表，上网搜停用词表，上网搜停用词表
4. 不要用TK，不要用TK，不要用TK
5. 为啥爬虫要那么多种，看了好久才知道怎么下手
6. 为啥网页要那么多种，格式统一不好吗
7. 爬虫好慢