**系统开发说明文件**

1. **产品设计方案**
2. **产品目标**

用户通过搜索，借助生成的词云图，获取大致评价。既可以快速了解相关内容关键信息，也可作为是否值得欣赏的参考。

1. **产品内容总策划**

图 1 流程规划

1. 设计规范

首先需要一个图形页面，用户对查询模块进行选择。选择模块后，进入搜索页面，用户输入搜索内容，确认后后台开始自动进行对于内容的全部短评进行爬取。短评内容经过分词、去除无效信息后绘制词云图并储存。词云图通过图形页面弹出的方式向用户展示。

1. 测试规范

运行爬取短评及制作词云图代码，检验词云图生成效果（词云图尺寸、内容）；运行总程序代码，测试图形页面是否正常显示；最后对程序进行检测，测试程序能否正常运行。

1. 开发日程表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 编号 | 内容 | 日期 | 总负责人 | 备注 | 完成确认 |
| 1 | 讨论确定选题 | 6.3-6.4 | 郑绮雯 |  |  |
| 2 | 完成代码编写  （短评爬取、词云图绘制、图形页面） | 6.5-6.8 | 陈俊豪 |  |  |
| 3 | 短评爬取及词云图绘制代码测试并完善 | 6.8 | 郭可恩  郑绮雯 |  |  |
| 4 | 程序打包 | 6.7-6.8 | 方泽仪 |  |  |
| 5 | 程序测试并美化 | 6.9 | 范宜苑 |  |  |

表 1 开发日程表

1. 技术解决方案

|  |  |
| --- | --- |
| 待解决问题 | 解决方法 |
| 选择查询模块并进行搜索 | 利用easygui库制作图形页面 |
| 定位指定内容短评网址 | 利用正则表达式获取内容对应编号 |
| 爬取短评信息 | 利用Selenum库以及xpath |
| 词云图制作 | 利用jieba库进行分词，wordcloud库进行词云图绘制 |
| 词云图呈现 | 利用easygui库制作图形页面 |
| 代码打包为程序 | 利用pyinstaller库进行 |

表 2 技术解决方案

1. **产品实现方案**
2. **系统的主要功能**

* 自动采集海量评论的关键信息

传统搜索模式重复环节多，加之豆瓣成熟庞大的数据和评价体系，用户多次选择并浏览信息的模式极为低效。自动化大数据利用爬虫技术匹配、抓取，减少人工筛查导致信息错位的风险，并实现海量数据更广泛连接，形成更大的信息网络。该产品解决了用户人工操作及筛选的痛点，相比于人工对结果的寻找和分析，直接高效地智能抓取海量信息具有极大的优越性。

* 多层次多维度满足个性化需求

一方面，该产品拓宽搜索内容至视听读层面，综合豆瓣电影、音乐、读书三大网络平台，更大程度地满足用户的搜索需求，也避免了用户需求范围广时来回切换网站、进行搜索浏览的不便；另一方面，该产品智能抓取的评论信息基本涵盖了查找目标的主要内容、中心思想、观（听/读）者情感表达等多维度信息，给予用户多角度反馈。

* 实时更新计算评论生成信息

实时计算即评论处理和特征计算，该产品统计实时评论信息，数据量、信息量只增不减的情况下，也避免用户对评论信息产生固化印象，大大提高了该产品的使用性，增强其可持续性的参考价值。

* 自动统计热点关键字信息

罗列全部数据的意义不大，关键是对结果进行统计分析。该产品综合结巴分词及人工建立停用词库，自动过滤无关信息及无用信息，缩小展示范围，精准了解用户需求，提升用户使用搜索效率，同时也大大降低了个别用户的主观因素产生的干扰，更为贴近大众客观印象。

* 词云可视化反馈

对关键词信息进行总结、指引，精确缩小罗列范围，对词频信息实行排序策略，后反馈为词云图可视化效果，对相关评论信息自动进行强弱化处理，更为直观清晰，便于用户后期了解相关内容，并根据各个元素反映的宏观信息进行微观分析和判断。

1. **关键技术与技术难点**
2. 关键技术

图 2 关键技术

1. 技术难点
2. 爬虫

* ChromeDriver与Chrome版本兼容问题

升级了Chrome，使得ChromeDriver与Chrome版本不对应，从而利用不了webdriver进行爬虫。对此将from webdriver\_manager.chrome import ChromeDriverManager 导入，使ChromeDriver与Chrome版本兼容。如果本就版本对应，可不导入。

* 使用正则表达式获取信息

正则表达式提取网址信息时出现反斜杠，为避免了多次转义造成的反斜杠困扰。在Python正则表达式前加上 r 表示原生字符串（rawstring），该字符串声明了引号中的内容表示该内容的原始含义。

* 对网页进行翻页时selector发生改变

通过观察，发现豆瓣评论网页在第一页里的“后一页”位置上的selector与后面24页的“后一页”位置上的selector不同，对此通过条件语句解决问题。

* 反反爬虫措施

通过selenium完全模拟浏览器操作。

为了防止请求过于频繁，网站限制IP访问。故而导入 time 库，利用time.sleep(5) 使每爬一页评论后就推迟执行5秒或2秒。

* 异常处理

利用traceback库跟踪异常返回信息，便于调试。

1. 词频统计

* 人工停用词选取

需要大量多次运行程序，在词云图上选取出无用的词语，工作量比较繁重。

1. 词云图生成

* 图片大小的调试

因为要结合第四部图形界面作词云图展示，为了合适与美观，多次对词云图大小进行调试，最终确认width=800, height=400 为合适的大小之一。

1. 图形界面

* 不同环节中图形界面的运用

考虑程序运行的合理性，适当调整并添加各环节提示信息，如在爬取信息时增加“词云生成中”信息提示，调用buttonbox模块中choices参数增加多方面选择，增强交互体验感。

1. 程序打包

* Pyinstaller安装失败

在cmd中依次键入了：  
（1）pip install pywin32  
（2）pip install wheel  
（3）pip install -U setuptools  
（4）pip install pyinstaller

或是其他方法，具体情况具体分析。

* 打包出错

权限问题（某些文件拒绝访问或无权限执行某些操作）：使用管理员权限运行cmd；

运行exe后只出现命令提示框，无运行结果：生成的exe文件和调用的依赖和数据在同一个文件夹下；  
“no such file or directory”：重新打包；

运行打包好的软件时，提示找不到模块（module）：指定 -p 参数，后跟自定义模块的路径。

* 程序图标的利用

发送快捷方式到桌面，属性窗口中修改设计好的图标。

1. **已完成的改进和存在的问题**

* 已完成的改进：

1. 实验初期，该产品的目标网站局限于豆瓣读书，即对豆瓣读书的书评进行爬取和分析，而后在豆瓣读书网站运行成功的基础上，团队将目标网站拓展至豆瓣读书、豆瓣电影、豆瓣音乐三大平台，极大地提高了丰富度、适用性和用户使用感。
2. 实验中期，结巴分词精确对长词进行切分，虽提高召回率，但仍存在不少无关信息如“这”“那”等评论用户使用频率高等但无实际意义的词，而后团队在不断尝试中逐步完善人工停用词表，目前词云展示效果具有明显的提升。
3. 实验中期，easygui图形用户界面虽能成功运行，但美观性不高，而后团队使用Photoshop软件制作窗口背景图及后期应用软件图标，并调整窗口显示大小比例，提升视觉美观度，增强了用户体验感。
4. 实验后期，为进一步提升用户使用感，团队将现有多个py文件打包为exe可执行文件，使其完全可以被分发到目标机器上直接执行，无须在目标机器上安装 Python 解释器环境。

* 存在的问题：

1. 程序运行时间慢
2. 为进一步完善程序，解决黑框窗口的弹出缺陷，我们将dos窗口（黑框）控制台删去，制作新版本的可执行文件。执行该文件发现发现dos窗口仍然存在（会在词云成功生成之前自动关闭，无需手动关闭黑框），且缺少了原版本中词频显示信息，效果弊大于利，故还是保留原始版本。目前原始版本dos窗口弹出问题仍待优化。
3. 目前产品可应用于豆瓣读书、影视、音乐三大版块，具有较强的适用性，可进一步优化多渠道、多网站、多模块的通路，打造用户获取多平台信息的新入口，为实现信息资源覆盖化提供更完备的技术支持。
4. **测试报告**
5. **测试目的**

通过测试验证程序已经达到设计规范。

**2.测试环境**

硬件环境：网络连接正常的电脑

软件环境：win10系统

**3.测试方法**

通过多次运行程序，记录测试项目数据并分析。

**4.测试项目**

a）程序安装与卸载

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 测试内容 | 测试方法 | 预期测试结果 | 测试效果 |
| 1 | 程序安装 | 使用解压软件解开压缩包，找到词云生成器.exe | 词云生成器.exe成功打开并运行 | 程序成功运行 |
| 2 | 程序卸载 | 删除词云生成器.exe或直接删除压缩包 | 程序或压缩包顺利删除 | 程序/文件  成功删除 |

b）程序功能测试

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 测试内容  （功能模块） | 测试方法 | 预期测试结果 | 测试效果 |
| 1 | 图形页面正常显示 | 打开词云生成器.exe,运行程序 | 程序成功打开，各页面正常显示，且页面按钮正常工作 | 各页面能正常弹出并  顺利运行 |
| 2 | 图书模块词云图  生成时间及效果 | 点击首页面  “图书”按钮，并根据指引输入书籍名 | 在一定时间内，词云图正常弹出，图上能较直观反映大众评论的关键信息 | 词云图正常呈现，能快速从词云图获取评论关键信息，若能优化爬取短评时间效果更优。 |
| 3 | 影视模块词云图绘制  生成时间及效果 | 点击首页面  “影视”按钮，并根据指引输入影视名 | 在一定时间内，词云图正常弹出，图上能较直观反映大众评论的关键信息 | 词云图正常呈现，能快速从词云图获取评论关键信息，若能优化爬取短评时间效果更优。 |
| 4 | 音乐模块词云图绘制  生成时间及效果 | 点击首页面  “音乐”按钮，并根据指引输入音乐名 | 在一定时间内，词云图正常弹出，图上能较直观反映大众评论的关键信息 | 词云图正常呈现，能快速从词云图获取评论关键信息，若能优化爬取短评时间效果更优。 |

c）操作失误运行效果测试

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序号 | 测试内容 | 测试方法 | 预期测试结果 | 测试效果 |
| 1 | 误选模块后可返回  并重新选择或退出 | 任意点击一模块，在输入页面按下“Cancel”按钮，待报错页面出现后单击页面，进入下一页面后点击“继续”或“退出”按钮 | 若选择“继续”，可再次返回首页面进行选择；若选择“退出”，弹出“欢迎再次使用”页面 | 误选后程序能正常运行，相应页面能顺利弹出 |
| 2 | 搜索内容无对应短评  正常报错 | 在输入页面输入错误信息 | 最终弹出重试页面，单击页面后可选择是否继续使用 | 成功弹出重试页面，后续内容与预期效果一致 |

1. **产品安装和使用说明**
2. **设计说明**

程序有三大模块，分别为书籍作品评价词云图生成模块、电影作品评价词云图生成模块、音乐作品评价词云图生成模块。各个模块各自有三个组成部分，分别为图形界面生成部分、网络爬虫部分和词云生成部分。

1. **安装说明**

使用解压软件解开压缩包，打开压缩包中的评论词云的文件夹，继续打开此文件夹中的dist文件夹中的主程序文件夹，找到词云生成器.exe即可直接使用。

若想安装到桌面，右键点击此可执行文件，在桌面创建快捷方式即可。然后右击属性，选择更改图标，按浏览，回到刚才的主程序文件夹，选择其中的图标.ico，选择作为图标即完整安装。最后一步更改图标的操作可以忽略，可以自主选择自己喜欢ico文件作为图标。

1. **功能介绍**

本产品拥有图形界面生成功能、搜索功能、打开chrome浏览器并爬取作品评论功能以及筛选词语制作词云功能，能向使用者快捷直观得展示所搜索的作品评价。

1. **使用说明和注意事项**

本产品适配chrome浏览器，请在使用前先安装chrome浏览器。打开程序后请按照图形界面的指示进行操作。

注意：程序运行中请勿手动关闭浏览器，否则将造成爬取评论失败。（待爬取评论后浏览器将自动关闭，无需手动关闭）