

**基于豆瓣影评数据的可视化**

**文本分析**

**需 求 规 格 说 明 书**

V 1.0

**版 本 历 史**

|  |  |  |  |
| --- | --- | --- | --- |
| 版本/状态 | 作者 | 日期 | 备注 |
| 1.0.0 | dyy | 2021-08-25 | 完成 |

目 录

[第一部分 引言 3](#_Toc16291)

[一、说明 3](#_Toc320)

[二、定义 3](#_Toc24668)

[①Ajax 4](#_Toc27421)

[②Cookie和Session 5](#_Toc22224)

[①请求头验证 6](#_Toc9028)

[②Cookie限制 6](#_Toc13552)

[③IP访问频率限制 7](#_Toc27079)

[④验证码限制 7](#_Toc24799)

[第二部分 综述 8](#_Toc300)

[一、项目背景 8](#_Toc31803)

[二、 建设目标 8](#_Toc28955)

[三、 建设原则 9](#_Toc12928)

[四、用户业务需求说明 10](#_Toc22454)

[2.1数据维护 10](#_Toc3987)

[2.2数据初始化 10](#_Toc1643)

[2.3影片管理 10](#_Toc18375)

[2.4短评管理 11](#_Toc20086)

[2.5系统维护 11](#_Toc25809)

[第三部分 需求分析 12](#_Toc17017)

[一、用例分析 12](#_Toc10115)

[二、界面风格 16](#_Toc7636)

[第四部分 验收标准 22](#_Toc11339)

[一、功能范围定义 22](#_Toc18328)

[二、性能指标定义 22](#_Toc4299)

[第五部分 环境和部署要求 23](#_Toc22100)

[一、网络部署图 23](#_Toc2818)

[二、应用部署 23](#_Toc25357)

第一部分 引言

### 一、说明

编写本说明书的目的是为了准确阐述项目具体业务需求和需求边界，本说明书的作者是【豆瓣数据分析】项目组，本说明书的确认者是【dyy】负责人，本说明书的读者是项目所有直接干系人。

本说明书是指导项目实施的重要指导性文件，也是用户最后进行验收（终验）的依据，说明书中内容一旦确认双方将以此为基础开展工作。如果需要变更说明书内容，必须走变更流程，变更必须得到甲乙双方书面确认，最后变更内容将作为本文的一部分，在项目实施过程中得以体现。

### 二、定义

#### 1. 数据分析

数据分析是指用适当的统计分析方法对收集来的大量数据进行分析，将它们加以汇总和理解并消化，以求最大化地开发数据的功能，发挥数据的作用。数据分析是为了提取有用信息和形成结论而对数据加以详细研究和概括总结的过程。

#### 2. 数据可视化

数据可视化主要旨在借助于图形化手段，清晰有效地传达与沟通信息。但是，这并不就意味着数据可视化就一定因为要实现其功能用途而令人感到枯燥乏味，或者是为了看上去绚丽多彩而显得极端复杂。为了有效地传达思想概念，美学形式与功能需要齐头并进，通过直观地传达关键的方面与特征，从而实现对于相当稀疏而又复杂的数据集的深入洞察。

#### 3. 数据挖掘

数据挖掘是指从数据库的大量数据中揭示出隐含的、先前未知的并有潜在价值的信息的非平凡过程。数据挖掘是一种决策支持过程，它主要基于人工智能、机器学习、模式识别、统计学、数据库、可视化技术等，高度自动化地分析企业的数据，作出归纳性的推理，从中挖掘出潜在的模式，帮助决策者调整市场策略，减少风险，作出正确的决策。

#### 4. 数据传输加密技术

数据传输加密技术的目的是对传输中的数据流加密，通常有线路加密与端—端加密两种。线路加密侧重在线路上而不考虑信源与信宿，是对保密信息通过各线路采用不同的加密密钥提供安全保护。端—端加密指信息由发送端自动加密，并且由TCP/IP进行数据包封装，然后作为不可阅读和不可识别的数据穿过互联网，当这些信息到达目的地，将被自动重组、解密，而成为可读的数据。

#### 5. 动态网页爬取

有时候在使用requests爬取网页的时候，会发现抓取的内容和浏览器显示的不一样，在浏览器里能看到要爬取的内容，但是爬取完之后的结果却没有，这就与网页是静态的还是动态的有关了。

静态网页是相对于动态网页而言，是指没有后台数据库、不含程序和不可交互的网页。而动态网页则是基本的html语法规范与Java、VB、VC等高级程序设计语言、数据库编程等多种技术的融合，以期实现对网站内容和风格的高效、动态和交互式的管理的网页。

二者的区别在于：

1.静态网页随着html代码的生成，页面的内容和显示效果基本上不会发生变化了——除非修改页面代码。

2.动态网页的页面代码虽然没有变，但是显示的内容却可以随着时间、环境或者数据库操作的结果而发生改变。

①Ajax

Ajax不是一门编程语言，而是利用JavaScript在不重新加载整个页面的情况下可以与服务器交换数据并更新部分网页内容的技术。

因为使用Ajax技术的网页里的信息是通过JavaScript脚本语言动态生成的，所以使用requests-BeautifulSoup的静态页面的爬取方法是抓不到数据的，我们可以通过以下两种方法爬取Ajax数据：

通过查找浏览器开发者工具里的Network-XHR/JS栏找到储存着数据的动态页面，使用requests请求网页，再将得到的数据使用json()方法进行转化就可以提取了。

使用selenium指挥浏览器，直接对数据进行抓取（或者通过selenium获取到渲染完整的网页源代码，再使用BeautifulSoup进行解析和提取数据）。

②Cookie和Session

Cookie是网站为了辨别用户身份进行Session跟踪而储存在用户本地终端上的数据（通常经过加密），由用户客户端计算机暂时或永久保存。

Session称为“会话”，存储着特定用户会话所需的属性及配置信息。

很多情况下，页面的更多信息需要登陆才能查看，所以在面对这类网页时，需要先模拟登陆后才可以对网页作进一步爬取。当我们模拟登陆时，客户端会生成Cookie并发送给服务器。因为Cookie保存着SessionID信息，所以服务器就可以根据Cookie判断出对应的SessionID，进而找到会话。如果当前的会话是有效的，那么服务器就会判断用户已经登陆并返回请求的页面信息，这样就可以对网页作进一步爬取了。

#### 6. APP爬取

除了web端，Python还可以爬取APP数据，不过这需要用到抓包工具，比如Fiddler。

相比web端而言，APP的数据爬取其实更容易，反爬虫也没那么强，返回的数据类型大多数为json。

#### 7. 多协程

当我们在做爬虫项目的时候，如果需要爬取的数据非常多，因为程序是一行行依次执行的，所以爬取的速度就会非常慢。而多协程就能解决这一个问题。

使用多协程，我们就可以同时执行多个任务。实际上，使用多协程时，如果一个任务在执行的过程中遇到等待，那它就会先去执行其它的任务，当等待结束，再回来继续执行之前的那个任务。因为这个过程切换得非常迅速，所以看上去像多个任务被同时执行一样，用计算机里的概念来解释的话，这其实就是异步。

我们可以使用gevent库来实现多协程，使用Queue()创建队列，spawn()创建任务，最后joinall()执行任务。

#### 8. 爬虫框架

在遇到比较大型的需求时，为了方便管理及拓展，我们可以使用爬虫框架实现数据爬取。

有了爬虫框架之后，我们就不必一个个地去组建爬虫的整个流程，只需要关心爬虫的核心逻辑部分即可，这样就大大提高了开发效率，节省了很多时间。爬虫框架有很多个，如Scrapy、PySpider等。

#### 9. 分布式爬虫

使用爬虫框架已经大大提高了开发效率，不过这些框架都是在同一台主机上运行的，如果能多台主机协同爬取，那么爬取效率就会更近一步提高。将多台主机组合起来，共同完成一个爬取任务，这就是分布式爬虫。

#### 10. 反爬虫机制与应对方法

为了防止爬虫开发者过度爬取造成网站的负担或者进行恶意的数据抓取，很多网站会设置反爬虫机制。所以当我们在对网站进行数据抓取的时候，可以查看一下网站的robots.txt了解一下网站允许爬取什么，不允许爬取什么。

常见的爬虫机制有以下4种：

①请求头验证

请求头验证是最常见的反爬虫机制，很多网站会对Headers里的user-agent进行检测，一部分网站还会检测origin和referer。应对这类反爬虫机制可以在爬虫中添加请求头headers，将浏览器里对应的值以字典的形式添加进去即可。

②Cookie限制

有些网站会用Cookie跟踪你的访问过程，如果发现爬虫的异常行为就会中断该爬虫的访问。应对Cookie限制类的反爬虫，一般可以通过先取得该网站的Cookie，再把Cookie发送给服务器即可，可以手动添加或者使用Session机制。不过，针对部分网站需要用户浏览页面时才能产生Cookie的情况，比如点击按钮，可以使用selenium-PhantomJS来请求网页并获取Cookie。

③IP访问频率限制

一些网站会通过检测用户行为来判断同一IP是否在短时间内多次请求页面，如果这个频率超过了某个阈值，网站通常会对爬虫进行提示并要求输入验证码，或者直接封掉IP拒绝服务。针对这种情况，可以使用IP代理的方法来绕过反爬虫，如代理池维护、付费代理、ADSL拨号代理等。

④验证码限制

很多网站在登陆的时候需要输入验证码，常见的有：图形验证码、极验滑动验证码、点触验证码和宫格验证码。

·图形验证码：需要用到OCR技术，使用Python的第三方库tesserocr可以做到这一点。

·极验滑动验证码：需要使用selenium来完成模拟人的行为的方式来验证。

·点触验证码：需要在使用selenium的基础上借助第三方验证码服务平台来解决。

·宫格验证码：需要使用selenium和模板匹配的方法。

第二部分 综述

### 一、项目背景

在电影行业飞速发展的当今，电影已经成为了非常普遍的娱乐选择，中国电影最近几年也是突飞猛进，越来越多的人走进电影院。而信息媒体的发展也使得所有人都拥有了随时随地与世界各地的人讨论电影的机会，电影无疑已经进入了全民关注，全民讨论时代。

但是随着越来越多的电影上映，越来越大的阅片量，大家对电影质量和制作水平的要求也越来越高。电影出品方想要斩获高票房，就要把握好消费者的喜好，制作出符合市场期待的电影。

在国内，豆瓣、淘票票、猫眼三家网站都有自己所属平台的评分功能，俨然形成了国内的“电影评分三巨头”。电影评分网站上蕴含了大量的电影评论、排行、评分等信息，是消费者择片困难时的重要参考依据，也是电影出品方了解用户喜好的有效途径。

一份科学的电影消费者数据分析报告，是预测票房的重要依据之一，也能够帮助出品方规划未来电影拍摄计划，对出品方投资有着重要意义。所以，如何分析消费者喜好信息是电影出品方非常关注的事情。

### 建设目标

1、爬虫页面响应及内容加载在5秒内，内容展示响应速度时间小于5秒。

2、爬虫配置，系统维护功能简单易用，80%用户可以在经过简单培训后结合使用手册，完成爬虫的开发和平台的维护。

3、爬虫平均宕机运行时间≤10小时/月。平均故障恢复时间≤180分钟/次。

4、爬虫可以自动根据爬虫节点的负载进行负载均衡处理，同时当爬虫出现不可恢复的错误时，可以智能的选择其他节点进行重新运行，保证爬虫任务可以正确完成。

### 建设原则

#### 3.1平台高可用性

云爬虫平台的所有爬取节点和数据存储分析节点均匀的分布在多台物理节点上,单台机器的宕机不会引起整个爬取进程的中断，这种分布式架构提升了系统整体的健壮性。

#### 3.2抓取高效性

单机模式下的网络爬虫效率不高，不能满足大规模的抓取任务需求，云爬虫平台为爬虫租户分配多个爬取节点,通过读取共享任务池来共同执行抓取任务，每个爬取节点都可以看成是一个单机的网络爬虫，能大幅度的提高页面的抓取效率。

#### 3.3高可扩展性

支持静态爬取和动态渲染的主流网站数据爬取，如天猫、京东、大众点评、豆瓣等，能够根据当前爬虫任务量动态地调节爬虫节点数量，比起传统爬虫方式灵活性更强。同时，租户在编写脚本时自定义程度高，允许租户根据不同的爬取需求自定义爬取范围。

### 四、用户业务需求说明

#### 1. 整体业务需求示意图

#### 爬虫

#### 需求详细说明

**2.1数据维护**

用户对数据库中的所有数据进行数据维护。

**2.2数据初始化**

用户点击数据初始化进入初始化数据，如果没有可初始化数据将不进行任何操作。

2.3影片管理

2.3.1添加影片信息

用户点击影片管理进入影片管理系统，输入需要添加的影片信息内容，添加到数据库中。

2.3.2删除影片信息

用户双击影片信息可以进行删除影片功能。

2.4短评管理

2.4.1选择影片

用户点击短评管理进入短评管理界面，从下拉列表中选择需要爬取的影片名词。

2.4.2爬取信息

用户点击爬取进行爬取选择的影片的内容。

**2.5系统维护**

**2.5.1退出**

用户点击系统维护选择退出即可退出整个系统。

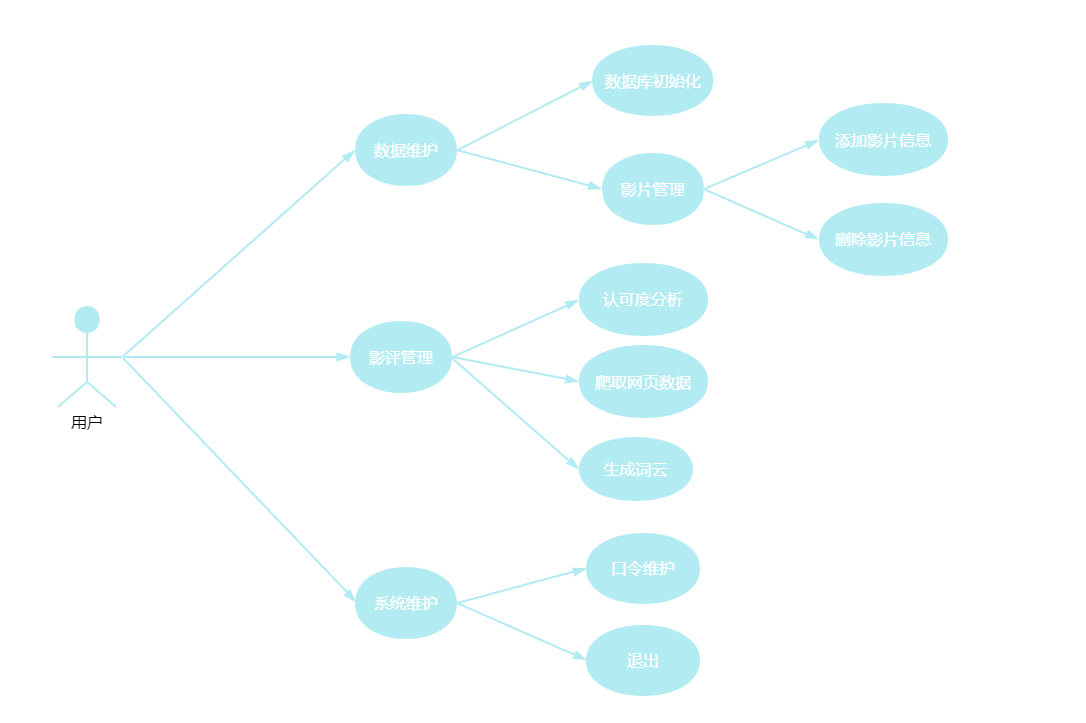
**2.5.2口令维护**

用户点击系统维护选择口令维护即可进行口令维护。

第三部分 需求分析

### 一、用例分析

#### 1、用户用例



**用户用例说明：**

##### 1.1、后台管理员数据维护用例描述

|  |  |
| --- | --- |
| ID | CH0001 |
| 用例名称 | 数据维护 |
| 父用例ID | - |
| 主要执行者 | 后台管理 |
| 前置条件 | - |
| 事件流 | 1. 输入建表语句 2. 创建表对应的字典 |
| 可选事件流 | 无 |
| 异常事件流 | 无 |
| 后置条件 | 无 |

##### 1.2、用户数据初始化用例描述

|  |  |
| --- | --- |
| ID | CH0002 |
| 用例名称 | 用户数据初始化 |
| 父用例ID | - |
| 主要执行者 | 用户 |
| 前置条件 | - |
| 事件流 | a.进入系统主界面  b.点击数据维护下的数据初始化 |
| 可选事件流 | 无 |
| 异常事件流 | 无 |
| 后置条件 | 无 |

##### 1.3、用户添加影片信息用例描述

|  |  |
| --- | --- |
| ID | CH0003 |
| 用例名称 | 添加影片信息 |
| 父用例 ID | - |
| 主要执行者 | 用户 |
| 前置条件 | - |
| 事件流 | a.进入主页面  b.点击数据维护下的影片管理  c.输入需要添加的影片信息 |
| 可选事件流 | 无 |
| 异常事件流 | 无 |
| 后置条件 | 无 |

##### 1.4、用户删除影片信息用例描述

|  |  |
| --- | --- |
| ID | CH0004 |
| 用例名称 | 删除影片信息 |
| 父用例 ID | - |
| 主要执行者 | 用户 |
| 前置条件 | - |
| 事件流 | a.进入主页面  b.点击数据维护下的影片管理  c.双击查询出来的影片信息 |
| 可选事件流 | 无 |
| 异常事件流 | 无 |
| 后置条件 | 无 |

##### 1.5、用户查询影片信息用例描述

|  |  |
| --- | --- |
| ID | CH0005 |
| 用例名称 | 查询影片信息 |
| 父用例 ID | - |
| 主要执行者 | 用户 |
| 前置条件 | 无 |
| 事件流 | a.进入主页面  b.点击短评管理下的短评管理  c.从下拉列表中选择需要查看的影片信息 |
| 可选事件流 | 无 |
| 异常事件流 | 无 |
| 后置条件 | 无 |

##### 1.6、爬取网页数据用例描述

|  |  |
| --- | --- |
| ID | CH0006 |
| 用例名称 | 爬取网页数据 |
| 父用例ID | - |
| 主要执行者 | 用户 |
| 前置条件 | - |
| 事件流 | a.进入主页面  b.点击短评管理下的短评管理  c.选择需要爬取的影片  d.点击爬取 |
| 可选事件流 | 无 |
| 异常事件流 | 无 |
| 后置条件 | 无 |

##### 1.7、口令维护用例描述

|  |  |
| --- | --- |
| ID | CH0007 |
| 用例名称 | 口令维护 |
| 父用例ID | - |
| 主要执行者 | 用户 |
| 前置条件 | - |
| 事件流 | a.进入主页面  b.点击系统维护下的口令维护 |
| 可选事件流 | 无 |
| 异常事件流 | 无 |
| 后置条件 | 无 |

##### 1.8、退出用例描述

|  |  |
| --- | --- |
| ID | CH0008 |
| 用例名称 | 退出 |
| 父用例ID | - |
| 主要执行者 | 用户 |
| 前置条件 | 用户已经登录 |
| 事件流 | a.进入主页面  b.点击系统维护下的退出 |
| 可选事件流 | 无 |
| 异常事件流 | 无 |
| 后置条件 | 无 |

### 二、界面风格

##### 1.主页面：



##### 2.数据维护页面

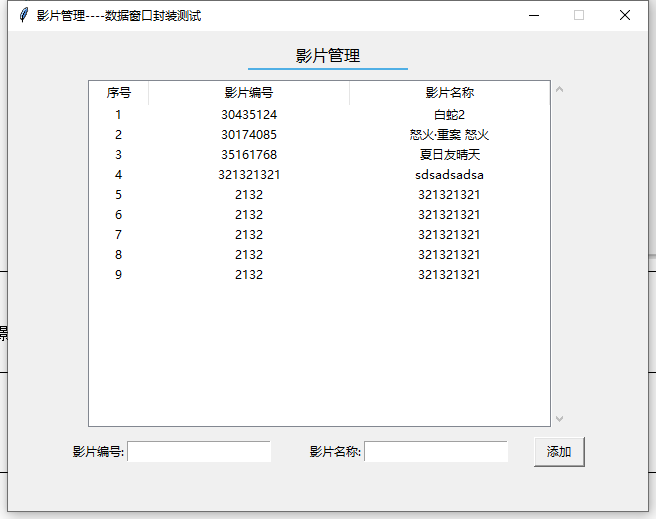


##### 3.系统维护页面



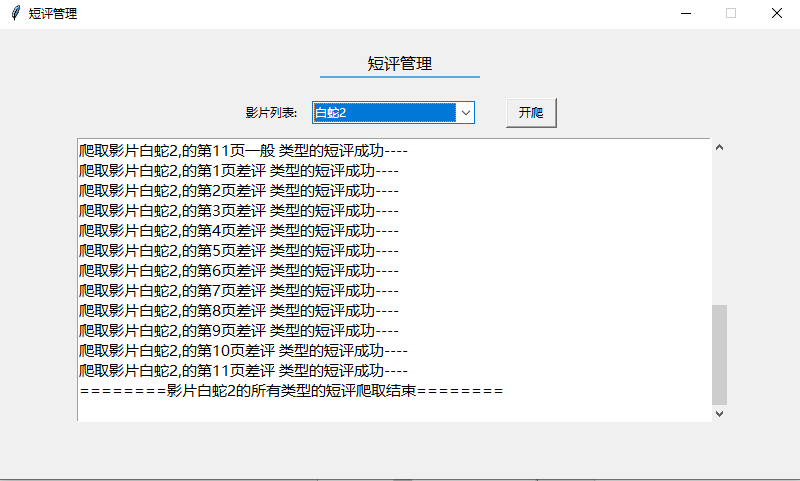
##### 4.影评管理功能页面：

##### 5.影片管理功能页面：

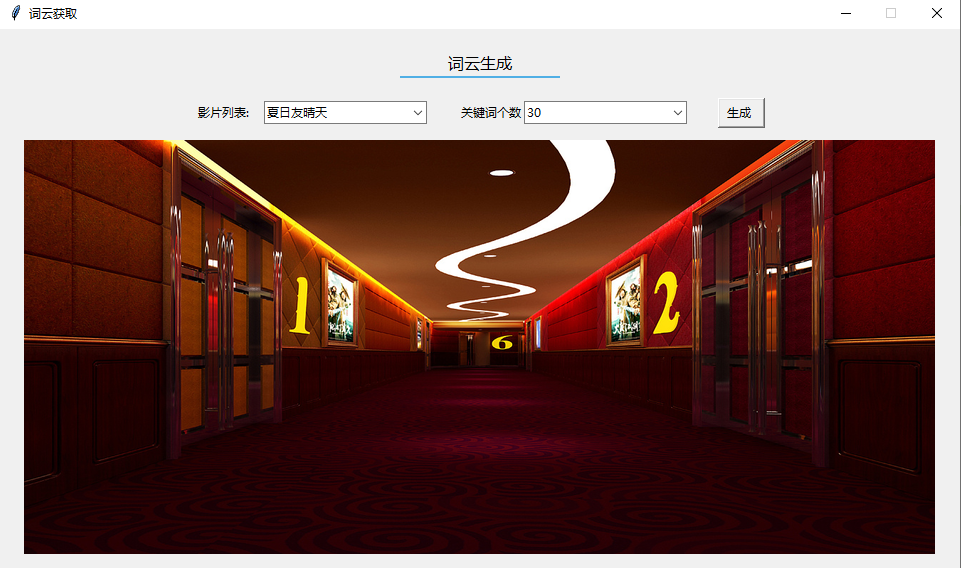


##### 6.短评管理功能页面：



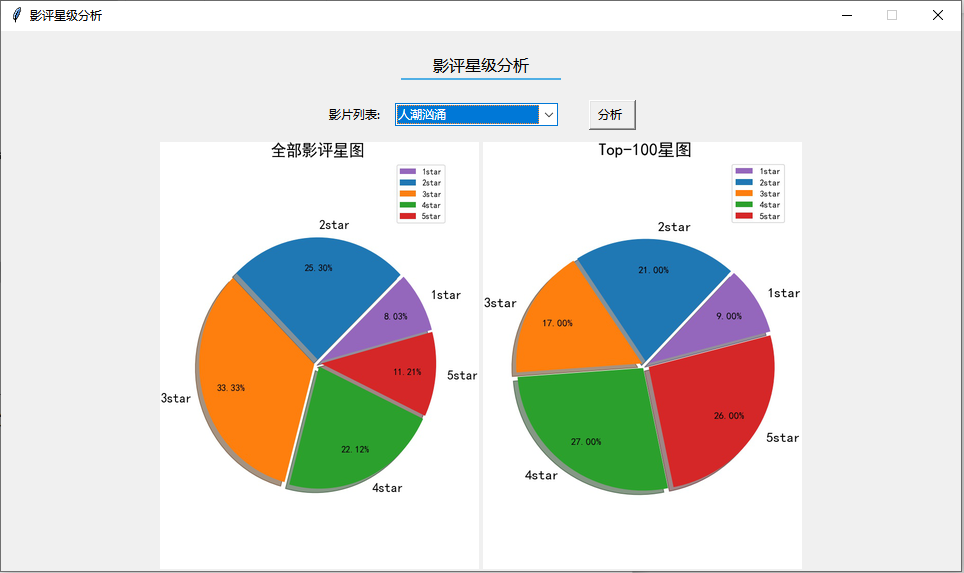


##### 7.词云获取功能页面：





##### 8.影评等级分析功能页面：



第四部分 验收标准

### 一、功能范围定义

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| # | 产品 | 模块 | 组件 | 规格/型号 | 角色 | 接入 |
| 1 | 豆瓣数据分析项目 | 系统维护单元 | 口令维护 | 维护数据库中的数据 | 用户 | Web |
| 2 | 退出 | 用户登录爬取数据后退出系统 | 用户 | Web |
| 3 | 数据维护功能单元 | 数据初始化 | 初始化数据库中的数据 | 用户 | Web |
| 4 | 影片管理 | 添加删除豆瓣中的影片 | 用户 | Web |
| 短评管理 | 爬取影片中的短评 | 用户 | Web |

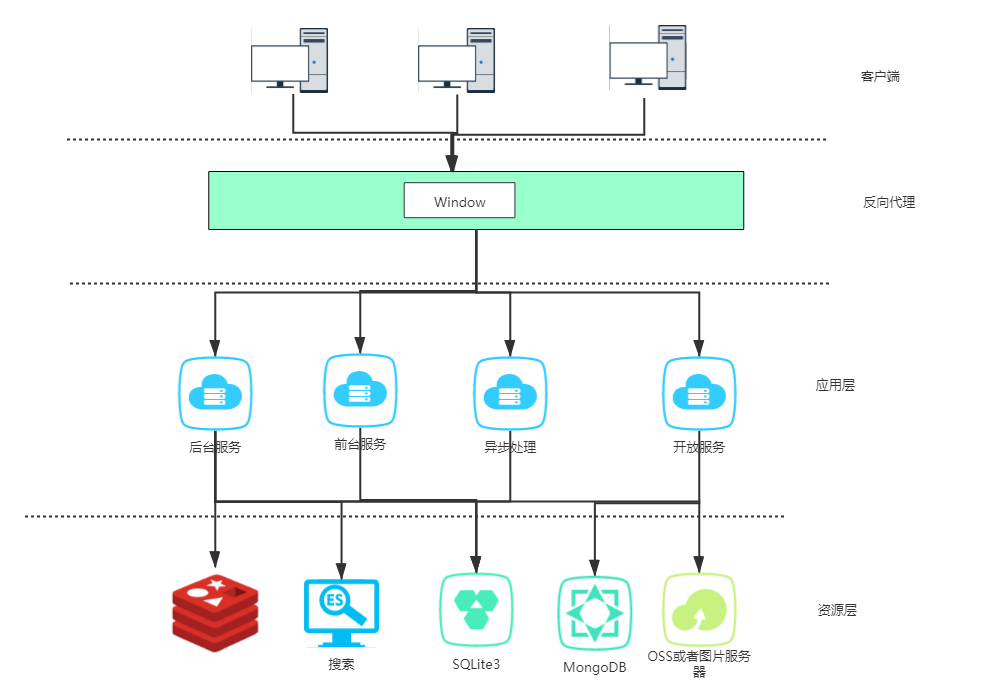
### 二、性能指标定义

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| # | 产品 | 模块 | 组件 | 规格/型号 | 性能级别 |
| 1 | 豆瓣数据分析项目 | 系统维护单元 | 口令维护 | 维护数据库中的数据 | A |
| 2 | 退出 | 用户登录爬取数据后退出系统 | A |
| 3 | 数据维护功能单元 | 数据初始化 | 初始化数据库中的数据 | A |
| 4 | 影片管理 | 添加删除豆瓣中的影片 | A |
| 5 | 短评管理 | 爬取影片中的短评 | A |

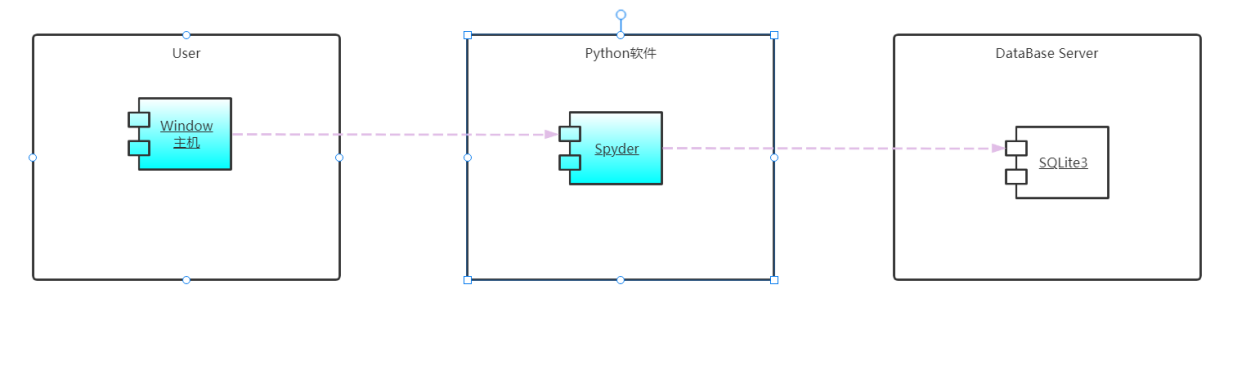
**说明：**级别（A:表示非常重要必须达到的技术性能要求,B:表示重要推荐达到的技术性能要求,C：表示非重要可以弱化的技术性能要求.）

第五部分 环境和部署要求

### 一、网络部署图



### 二、应用部署



运行环境说明

#### 服务器

1.1OS：Windows XP

1.24GB以上内存

1.31T以上的硬盘空间，不包括数据存储空间

1.4集成显卡

1.5硬盘转速7200RPM

1.6硬盘类型为SAS

#### 2. 客户机器环境

2.1 系统：Windows XP以上

2.2 内存：512M以上

2.3 硬盘：10G以上

2.4 浏览器最低要求：IE浏览器版本12.0、0360浏览器、谷歌浏览器