博物馆新闻采集分析子系统用户手册

**目录**

1. 导言...................................................................................................................3

1.1 版本修订历史.........................................................................................3

1.2目的.............................................................................................................3

1.3范围.............................................................................................................3

1. 概述....................................................................................................................3
2. 运行环境...........................................................................................................4
3. 安装与配置......................................................................................................4
4. 操作说明...........................................................................................................4
5. 技术支持信息..................................................................................................5
6. **导言**
   1. **版本修订历史**

|  |  |  |
| --- | --- | --- |
| **版本号** | **修订内容** | **修订者** |
| 1.0 | 用户使用手册 | 张紫旭 |
| 2.0 | 用户使用手册 | 张紫旭 |
| 3.0 | 用户使用手册 | 张艺朦 |

* 1. **目的**

该文档的目的是描述新闻采集分析子系统项目的用户使用说明，其主要内容包括：

（1） 运行环境

（2） 安装与配置

（3） 操作说明

本文档的预期读者是：

1. 其他子系统的开发人员
   1. **范围**

该文档定义了系统提交的产品的使用说明，主要描述了产品的操作流程，以及配置说明。

1. **概述**

博物馆新闻采集分析子系统是整个课程设计的后端支持系统之一。使用了Python语言，运行在本地服务器上，数据库系统是Mysql。该系统的开发目的是为前端的页面展示提供博物馆新闻数据。系统的处理过程主要包括数据获取、加工、情感分析服务。基于以上功能，设计的数据项包括（新闻编号，博物馆名称，新闻标题，新闻作者，新闻时间，新闻正文，新闻连接，新闻情感分析分数）。

该系统为前端提供数据源，明晰开发结构，提高开发效率，在开发过程中有注意到数据更新的影响并加以修改。

1. **运行环境**
2. 系统运行硬件环境

客户机：普通PC机

服务器：

CPU：2.0GHz

内存：2GB以上

硬盘：100G以上

网卡：千兆

1. 系统运行软件环境

操作系统：Unix/Linux/Win10

数据库：MySQL

浏览器：Microsoft Edge/Firefox/Google Chrome

1. **安装与配置**
2. 配置Python环境在服务器上
3. 下载远程数据库navicat
4. 本机配置虚拟机
5. 下载一些安装包，如scrapy、snowlp、fake\_useragent
6. **操作说明**
7. 获取指定博物馆新闻流程如下：

法一：

使用Python或其他语言操作数据库->连接数据库->借助于表newsall查找指定信息(单表查找，通过博物馆名称找到对应的所有全部新闻)->获取类似于json或者DataFrame格式的数据并处理显示。

法二：

使用命令行运行firstscrapy文件

scrapy crawl first -a museum=博物馆名称 -a startTime=开始时间 -a endTime=结束时间

例如：scrapy crawl first -a museum=故宫博物馆 -a startTime=2021-01-01 -a endTime=2021-05-15

这个爬虫的几种运行方式说明：

* scrapy crawl first //默认输入爬取的是 **博物馆** 新闻 在所有时间范围内
* scrapy crawl first -a startTime=2020-05-17 -a endTime=2020-05-18//爬取的是 **博物馆** 在指定时间内的新闻
* scrapy crawl first -a museum=故宫博物馆//爬取的是 **故宫博物馆** 在所有时间范围的新闻
* scrapy crawl first -a museum=故宫博物馆 -a startTime=2020-05-17 -a endTime=2020-05-18://爬取的是 **故宫博物馆** 在指定时间范围内的新闻

运行时参数为 museum startTime endTime不要写错，博物馆名称不用加引号，时间格式为：%Y-%m-%d

（2）系统部署在服务器上之后，会在每天的特定时间自动更新数据库内容

（3）后端服务器地址

组内数据库：

# host='localhost',

port=3306,

user='root',

password='mysql',

db='group2-zjf-news',

charset='utf8'

团队数据库：

host='151.106.117.0',

port=3306,

user='u987603792\_news',

password='Newsnews2',

db='news',

charset='utf8'

1. **目录说明**

**6.1 firstscrapy0520用于爬取数据入库**

│ scrapy.cfg

│ main.py //爬虫运行

└─firstscrapy//爬取第一层页面

│ items.py // 提取要爬取的字段

│ middlewares.py // 下载中间件和爬虫中间件，可以设置下载时所用的代理

│ pipelines.py // 管道文件，用于对数据进行处理或保存

│ settings.py // 设置文件，例如User-Agent，headers，下载等

│ \_\_init\_\_.py // python 文件包需要

├─spiders // 该目录下存放爬虫文件

│ first.py // 新闻爬取

└─secondscrapy//爬取第二层页面

**6.2 firstscrapy用于展示数据定制**

│ hellow.txt // 存储爬取的数据

│ scrapy.cfg

│ main.py //爬虫运行

└─firstscrapy

│ items.py // 提取要爬取的字段

│ middlewares.py // 下载中间件和爬虫中间件，可以设置下载时所用的代理

│ pipelines.py // 管道文件，用于对数据进行处理或保存

│ settings.py // 设置文件，例如User-Agent，headers，下载等

│ \_\_init\_\_.py // python 文件包需要

├─spiders // 该目录下存放爬虫文件

│ first.py // 新闻爬取

**7.技术支持信息**

本系统开发小组提供维护和升级，使用中如果有任何问题可以及时反馈给系统负责人。