设计报告

网站数据采集子系统

一、引言

1.1系统设计目的

- 爬取网站数据，为整个项目提供基础数据，保证博物馆信息的完备。
- 设置定时爬取功能，并数据存入数据库中，保证信息的及时性。
- 对信息进行分析，提取用于搜索的关键信息。

1.2系统设计概述

- 通过使用爬虫技术，爬取目前204家国家一级博物馆的名称、简介、开闭馆时间、展览、展品、到达方式、地理位置等基本信息
- 采用定时爬取技术，将获取到的信息传送到数据库当中并且实现定时爬取及上传。
- 通过使用`NLP`技术，化繁为简，提取大段文字当中的关键信息和主题信息。
- 通过使用可视化技术，将爬取到的信息绘制成为图表，提供更好的用户体验。

1.3系统功能

- [x] 数据爬取
  - [x] 爬取全国一级博物馆的网站信息
    - [x] 博物馆基本信息（博物馆名称、博物馆简介、博物馆票价、如何前往）
    - [x] 参观信息
    - [x] 展览信息
    - [x] 经典藏品信息
    - [x] 博物馆的内景图片
    - [x] 博物馆的封面图片
    - [x] 博物馆的展览图片
    - [x] 博物馆的藏品图片
- [x] 数据加工
  - [x] 对博物馆信息进行过滤和加工
    - [x] 提取出简介中的关键词，提高模糊搜索的精度
  - [x] 对展览信息进行过滤和加工
    - [x] 抽取展览主题
    - [x] 抽取展览时间
    - [x] 抽取展览地点
    - [x] 抽取展览介绍信息等
- [x] 数据导入
  - [x] 将爬取到的数据存入excel表中
  - [x] 将爬取到的数据直接存入数据库

- [x] 数据更新
  - [x] 每24小时爬取一次新的数据
  - [x] 用新的数据更新原有
  - [x] 部署到服务器

1.4系统实现方法

- python 3.8
- `beautifulsoup`
- request
- selenium
- `lxrd`

二、接口设计

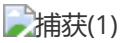
2.1需求规定

本系统需要提供博物馆的各种信息，具体包括博物馆名称、博物馆简介、博物馆的地址、博物馆的门票价格、博物馆封面图等内容。同时本系统对爬取到的数据进行加工处理，将无需使用以及存在异常数据的数据进行删除或者修订，确保信息的准确性和完整性，本系统同时为保证信息具备时效性，具备定时爬取的功能， 通过定时爬取官网信息，将信息及时上传到数据库当中，实现数据信息及时更新。本系统额外添加了数据可视化功能，将数据信息通过图的形式展示出来。

2.2运行环境

操作系统：`windows7`及其以上版本的操作系统或手机端`App`。

2.3基本设计概念和处理流程


捕获(1)

2.4用户与管理员的功能分配

	一般用户	管理员
查看基本信息	允许	允许
修改基本信息		允许
查看展览信息	允许	允许
修改展览信息		允许
查看藏品信息	允许	允许
修改藏品信息		允许

2.5人工处理过程

数据均在本地备份并且允许进行上传，防止发生数据缺失。

2.6尚未解决的问题

暂无尚未解决的问题。

三、接口设计

3.1用户接口

本系统的数据均只能又管理员进行操作，不对用户开放操作权限，用户只能进行查看。

3.2外部接口

本系统只能通过后台数据库进行操作，因此本系统的外部接口用于与数据库进行连接，实现数据交互和定时爬取。

3.3内部接口

本系统的各个程序存在内部接口，通过函数调用、参数传递和返回值的方式进行信息的传递。

四、运行设计

4.1运行模块组合

输入时启动接受数据模块，通过各模块之间的调用，读入并对输入数据进行格化。在接受数据模块得到充分的数据时，将数据通过传输代码上传至数据库内，并等待数据传输完成并确认上传是否成功。

数据库的连接必须在周期性处于联通状态，方便后台数据进行定时实时上传后台数据。

4.2运行控制

运行控制将严格按照各模块间函数调用关系来实现。在个事务中心模块中，需对运行控制进行正确的判断，选择正确的运行控制路径。

博物馆数据在发送上传数据库后，将等待数据库连接成功，然后上传数据，上传完成后确认数据库中已上传的数据是否与上传的数据一致。

4.3运行时间

A. 一般用户模块会经常运行。

B. 操作员模块使用次之。

C. 管理员模块使用出的最少。

五、系统数据结构设计

5.1逻辑结构设计要点

1.博物馆信息（编号、名称、类型、地址、门票、开放时间、建议游玩时间、博物馆等级、博物馆对应页面编号、简介、封面、如何前往、经纬度、周边景点）

2.内景图（编号、博物馆id、地址）

3.博物馆主题提取（编号、名称、简介）

4.展览（编号、博物馆id、藏品所属展览名字、展览简介、藏品名称、藏品图片地址、藏品简介）

5.2物理结构设计要点

1.博物馆信息

字段中文（注释）	字段英文	字段类型	主键	允许空	说明
博物馆编号	id	int	是	否	自增，不能为空
博物馆名称	name	varchar	否	是	
博物馆的类型	type	varchar	否	是	
地址	address	varchar	否	是	
门票	TicketPrice	int	否	是	
开放时间	OpeningHours	varchar	否	是	
建议游玩时间	Suggestedtraveltime	varchar	否	是	
博物馆等级	Museumlevel	varchar	否	是	
博物馆对应页面的编号	number	varchar	否	是	
博物馆简介	introduction	varchar	否	是	
如何前往	Howtogo	varchar	否	是	
周围的景点	Scenicspotsaround	varchar	否	是	
封面（图片）	cover	varchar	否	是	
经度	longitude	float	否	是	
维度	latitude	float	否	是	

2.内景图

字段中文（注释）	字段英文	字段类型	主键	允许空	说明
编号	id	int	是	否	自增，不能为空
博物馆id	Museumid	int	否	是	不能为空
地址	address	varchar	否	是	不能为空

3.博物馆主题提取

字段中文（注释）	字段英文	字段类型	主键	允许空	说明
博物馆编号	id	int	是	否	自增，不能为空
博物馆名称	name	varchar	否	是	
博物馆简介	introduction	varchar	否	是	

4.展览

字段中文（注释）	字段英文	字段类型	主键	允许空	说明
编号	id	int	是	否	自增，不能为空
博物馆id	Museumid	int	否	是	
藏品所属展览名字	Exhibitname	varchar	否	是	
展览简介	Exhibitsummary	varchar	否	是	
藏品名称	Collectionname	varchar	否	是	不能为空
藏品图片地址	Collectionimageurl	varchar	否	是	
藏品简介	Collectionsummary	varchar	否	是	

六、系统风险控制设计

6.1潜在风险

错误类型	错误详情
网址访问错误	无法访问指定的网址导致无法爬取相关信息
数据库访问错误	无法访问到数据库导致无法进行数据的上传
服务器部署错误	无法在服务器上部署定时任务导致无法进行定时任务的部署
数据格式错误	数据格式与数据库不一致导致无法上传数据

6.2规避风险方案

1.用多台电脑进行多次访问网址，确保网址是正确的网址，且能够爬取到正确的信息。

2.及时与数据库管理人员进行数据库的数据交互，确保数据库能够正常且及时访问。

3.定期测试服务器的定时服务系统是否正常。

4.数据格式严格按照数据库指定的类型进行上传。

6.3系统维护设计

定期进行本地备份，确保数据库一旦发生故障，依然能够有数据来源。