**详细设计说明书**

**酒店数据收集分析**

**小组：** 第五组

**姓名： 黄锦波 学号： 2017324236**

**姓名： 谢兴悦 学号： 2017324204**

**姓名： 郭奇东 学号： 2017324217**

**姓名： 沈小玉 学号： 2017324218**

**姓名： 谢金英 学号： 2017324225**

**版本：V1.0**

**发布时间：2020-09-10**

**岭南师范学院 信息工程学院**

**深圳信盈达科技有限公司 人才实训学院**

文件变更记录

|  |  |  |  |
| --- | --- | --- | --- |
| **修改编号** | **版本** | **修改条款及内容** | **修改日期** |
| 1 | v1.0 | 撰写详细设计说明书 | 2020-09-15 |
| 2 | v2.0 | 修改文件的格式 | 2020-09-17 |
| 3 | v3.0 | 增改系统详细设计说明部分 | 2020-09-20 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**目 录**

1 引言 1

1.1 目的 1

1.2 背景 1

1.3 词汇表 1

1.4 参考资料 1

2 系统结构 2

2.1 需求概述 2

2.2 总体设计 2

3 系统详细设计说明 2

3.1 包及类结构设计 2

3.1.1 city(城市名称) 2

3.1.2 hotel(酒店名称) 2

3.1.3 price(酒店价格) 2

3.1.4 grade(酒店评分) 2

3.1.5 address(酒店地址) 2

3.1.6 hotel\_url(酒店网址) 3

3.2 模块设计 3

4 其他设计要求 3

4.1 尚未解决问题 3

5 附件说明 3

# 引言

## 目的

编写本文档的目的是为了使开发人员在完成概要设计说明书的基础上，完成概要设计的各项功能规定；为程序员写出实际的程序代码提供依据。它是软件设计阶段所有任务和所有相关设计人员的参考资料。预期读者为系统的开发者，系统的使用者，以及指导老师。

## 背景

技术在革新，时代在变迁，在智能信息时代，人们的生活在“数据化”，而酒店的运营管理活动也在被“数据化”。分析和挖掘数据是酒店经营者必须要掌握的基本技能，更充分利用数据掌握酒店内部和外部信息，提高个人的影响力和控制力，从而为酒店创造更大的价值。“信息即权力”，在酒店掌握信息越多的人，就会拥有更多的话语权。而信息来自对数据的加工，如果不能从数据中萃取信息用于经营决策，那么采取的数据只会成为酒店的成本；如果能够充分发挥数据的力量，在数据中挖掘价值，那么数据就会成为酒店的资产。

整个系统围绕着获取所需的酒店数据(酒店名称、地理位置、价格、评分)进行开发设计，从而实现了整个酒店数据收集分析系统的设计。

## 词汇表

|  |  |  |
| --- | --- | --- |
| 词汇名称 | 词汇含义 | 备注 |
| city | 城市名称 | hotel.py |
| hotel | 酒店名称 | hotel.py |
| price | 酒店价格 | hotel.py |
| grade | 酒店评分 | hotel.py |
| address | 酒店地址 | hotel.py |
| hotel\_url | 酒店网址 | hotel.py |

## 参考资料

1．本小组编写的《需求分析报告》

2．本小组编写的《概要设计说明书》

# 系统结构

## 需求概述

1. 在广东省的区域内，根据城市的名字，获取该城市下的评分4.5以上的酒店的名称。
2. 根据酒店名称获取该酒店的房型，地理位置，评价（score）。
3. 通过获取的数据来分析酒店房间的价格。
4. 酒店评分：评价从0~5.0，我们划分为五个区域（0~1，1~2，2~3，3~4，4~5），客户评分在某个区间占的比例(1~5(取整))
5. 系统结构合理，模块化结构，界面清晰，操作方便，文档规范。

## 总体设计

在本项目中，首先需要得到艺龙酒店的网址，然后依据其爬取酒店数据，要获取的内容有：酒店名称、地理位置、价格、评价。再通过筛选，分析在所有的酒店中，某一酒店房间的价格是多少，以及客户评分在某个区间占得比例。

本项目运用了多次XPath语言，XPath (XML Path Language)是一门在 XML 文档中查找数据信息的语言，可用来在XM文档(包括 HTML)中对元素和属性进行遍历，XPath 通常是使用节点路径表达式来选取 XML或 HTML文档中的节点或节点集。

从大体上说，爬取酒店的数据需要完成的基本操作步骤有：①明确目标；②爬；③取；④存。首先，明确我们爬取的是艺龙酒店的数据，但除了艺龙酒店，我们还爬取了其他类似的数据，此时便有多个链接地址，将多个链接存入到 URL列表进行统一爬取；接着，就是爬网站的数据内容了，通常爬的数据内容一般是网页源代码；爬下需要的数据，就使用 Python 爬虫技术进行提取有价值的数据内容，用的是分析 HTML 源码；将数据提取完毕，就把这些数据内容存储起来，将数据内容保存在文件中。

# 系统详细设计说明

## 包及类结构设计

## 3.1.1 city(城市名称)

用于存储城市名称

## 3.1.2 hotel(酒店名称)

用于存储酒店名称

## 3.1.3 price(酒店价格)

用于存储酒店的价格

## 3.1.4 grade(酒店评分)

用于存储酒店的评分所在的范围

## 3.1.5 address(酒店地址)

用于表示酒店的地址以获取数据

## 3.1.6 hotel\_url(酒店网址)

用于找到相对应的酒店

## 模块设计

|  |  |  |
| --- | --- | --- |
| 界面 | py | 功能描述 |
| 酒店数据界面 | hotel.py | 获取酒店各数据内容 |
| 城市获取界面 | get\_city.py | 获取城市名称 |
| IP获取界面 | get\_ip\_agent.py | 获取ip代理地址 |
| 用户代理界面 | user\_agent.py | 获取用户代理 |
| 代理池界面 | useragenttool.py | 制作代理池 |

# 其他设计要求

## 尚未解决问题

a.待定阶段问题：无法爬取携程酒店的数据。解决方案是换一个网站，爬取艺龙酒店的数据。

# 附件说明