**概要设计报告**

**酒店数据获取分析**

**小组：** **第五组**

**姓名： 黄锦波 学号：2017324236**

**姓名： 郭奇东 学号：2017324217**

**姓名： 沈小玉 学号：2017324218**

**姓名： 谢金英 学号：2017324225**

**姓名： 谢兴悦 学号：2017324204**

**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 | 新建 | 2020-9-11 |
| 2 | V2.0 | 确定项目名称以及研究方向，分配工作 | 2020-9-13 |
| 3 | V3.0 | 初步分析项目需求，爬取部分热门城市的酒店数据 | 2020-9-16 |
| 4 | V4.0 | 完成项目需求分析，整理爬取的部分热门城市数据 | 2020-9-19 |
| 5 | V5.0 | 完成概要设计，分析所爬取的酒店数据的房型 | 2020-9-21 |
| 6 | V6.0 | 完成详细设计，将分析整理好的数据分类归放 | 2020-9-24 |
| 7 | V7.0 | 小组成员相互讨论爬取所出现的问题以及解决方案 | 2020-9-26 |
| 8 | V5.0 | 将所有需要补充的数据爬取下来 | 2020-9-27 |
| 9 | V5.0 | 制作PPT，完成文档。 | 2020-9-29 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**目 录**

1 引言 1

1.1 目的以及背景 1

1.2 背景 1

1.3 术语表 1

1.4 参考资料 1

2 概要设计决策 2

2.1 设计原则 2

2.2 设计决策 2

3 总体设计 2

3.1 软件体系结构 2

3.2 系统技术构架 2

3.3 系统运行环境 3

4 其他说明 3

5 附件 4

# 引言

## 目的以及背景

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。

## 背景

艺龙酒店通过特有的系统联接上国际互联网向全球多姿多彩、声情并茂地展示自己的风貌、特色，推销自己客房和各种服务，并可依此组成酒店连锁业，结成战略联盟，以劲灵活的营销手段向广大市场进军。它可以向众多的客户提供面对面的营销方式。它开拓了市场的广度和深度，这些都是平常方式下的人力物力所无法与之比拟的。它的数据可以为酒店开发客源市场带来了限的商机。

## 术语表

列出本报告中专门术语的定义、英文缩写词的原词组和意义、项目组内达成一致意见的专用词汇，同时要求继承全部的先前过程中定义过的词汇。

|  |  |  |
| --- | --- | --- |
| 术语 | 含义 | 备注 |
| 评价 | 酒店评分 |  |
| 房型 | 在酒店中某一类房型占比例多少 |  |

## 参考资料

1. 基于Python的网络爬虫技术 赵文杰 古荣龙
2. [基于Python的深度网络爬虫的设计与实现](https://kns.cnki.net/kcms/detail/detail.aspx?filename=CXJL201808088&dbcode=CJFQ&dbname=CJFDTEMP&v=99xL1%25mmd2F7kQ%25mmd2B6Gqkh0IXp7srfWRCuiGsohaCpms8Aihey2p6fRV3WICDNQmgJyIf9q) 彭智鑫.信息记录材料.2018(08)
3. [基于Python的数据信息爬虫技术](https://kns.cnki.net/kcms/detail/detail.aspx?filename=ELEW201811134&dbcode=CJFQ&dbname=CJFD2018&v=PJzwKefxODukK2HgTQhc5qfsiieO%25mmd2BRWiYaaUmxBlSFywaAqrf10ISS%25mmd2FmPm0me145) 魏程程 电子世界. 2018(11)
4. [基于Python的用户协同过滤推荐系统的综合研究](https://kns.cnki.net/kcms/detail/detail.aspx?filename=JJJJ201904004&dbcode=CJFD&dbname=CJFD2019&v=3ao4wdwYlIrvL%25mmd2Fura4O3VAAtstAT8DQ49PrMuoijH0A8SEDvOBaUeH3ulShaq7kb) 张阳 技术与教育 2019(04)
5. [主题网络爬虫研究综述](https://kns.cnki.net/kcms/detail/detail.aspx?filename=JSYJ200710006&dbcode=CJFQ&dbname=CJFD2007&v=vnbHOjb2yOtI1GLJ60JEAWu5rK%25mmd2F7Al4FQ2%25mmd2BT9pUeuRbADMFUwbdcFdj1nMr2r1x%25mmd2F) 刘金红,陆余良 计算机应用研究 2007(10)
6. ][多线程并发网络爬虫的设计与实现](https://kns.cnki.net/kcms/detail/detail.aspx?filename=XDJS201901022&dbcode=CJFQ&dbname=CJFD2019&v=H5n1wHhegrIcXBZYThWxe9UbvpKj%25mmd2BAWQTYXZe1b8QJj02bo11WgRvtYZF%25mmd2FbkpzUt) 邵晓文 现代计算机(专业版) 2019(01)

列出编写本报告时参考的文件、资料、技术标准以及他们的作者、标题、编号、出版日期和出版单位。

列出编写本报告时查阅的Internet上杂志、专业著作、技术标准以及其网址。

# 概要设计决策

## 设计原则

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、房型、评价。再通过筛选，分析在所有的酒店中，某一类房型占的比例是多少，以及客户评分在某个区间占得比例。

## 设计决策

本节记录系统的所有设计决策。

|  |  |  |  |
| --- | --- | --- | --- |
| **设计决策** |  | **标识号** |  |
| **问题描述** | 分析和挖掘酒店数据，充分掌握酒店数据，为酒店运营创造更大的价值 | | |
| **假设条件** | 根据地理位置，获取评分较高的酒店名字 | | |
| **可选方案** | 根据酒店名称获取该酒店的房型，地理位置，评价 | | |
| **决策** | 将评分划分为五个区域，整理出评分较高的酒店信息。 | | |
| **理由** | 这种做法可以筛选出各项条件比较好的酒店进行参考 | | |
| **影响** | 可根据获取的数据更改酒店详细设计 | | |
| **相关决策** | 将评分划分为五个区域，整理出评分较高的酒店信息。 | | |

# 总体设计

## 软件体系结构

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。

## 系统技术构架

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，明确我们爬取的是携程网站的数据，但除了携程网站，我们还爬取了其他类似携程网站的数据，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。

## 系统运行环境

本项目在运用Python语言进行开发，需要有 Python 解释器的环境支持，我们使用了 Python IDLE 3.7.3 来搭建基本环境，因仅使用 Python 解释器进行开发，效率极低，我们需要选择一款高效快捷的 IDE 工具，于是便使用了 PyCharm 集成开发环境，在爬虫开发过程中，若要查看请求网站时身份、响应数据内容等，此时需使用 Fiddler 工具，除此之外,还安装了 requests 模块、lxml 模块、selenium 模块 。

# 其他说明

我们团队基本实现了最初提出的目标，完成了对酒店数据的获取以及分析。在项目进行过程中，我们小组也遇到一些问题，但都可以通过请教别人、查找资料、小组内讨论解决。

# 附件