目 录

[1.概述 3](#_Toc487982368)

[2.测试题目 3](#_Toc487982369)

[2.1任务1—环评信息爬取 3](#_Toc487982370)

[2.2任务2—网页遍历 4](#_Toc487982371)

[3.交付物要求 6](#_Toc487982372)

[4.常用爬虫工具 6](#_Toc487982373)

# 1.概述

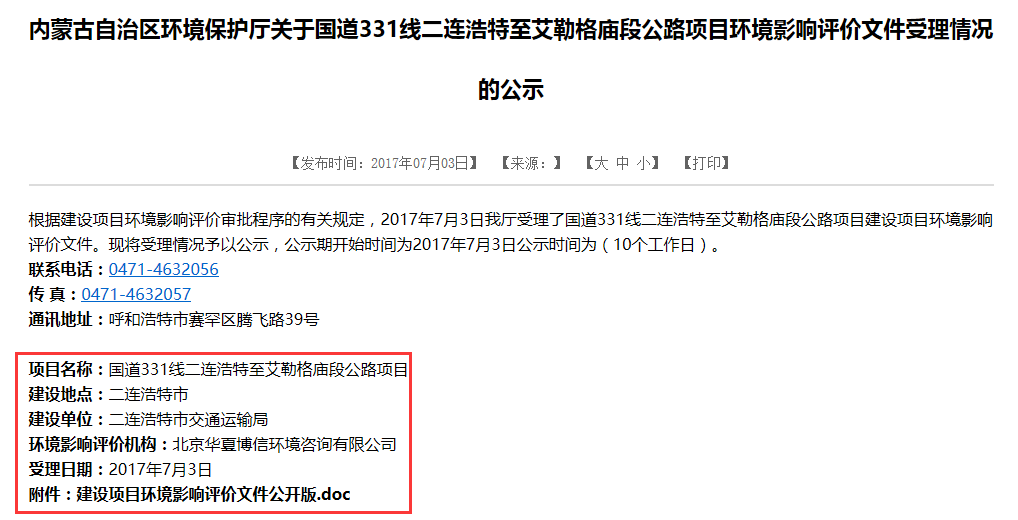
本文档包含了爬虫方向的测试题目，请各位同学使用Python或Python与JavaScript混合编程，按照以下要求完成测试。

# 2.测试题目

## 2.1任务1—环评信息爬取

**1）任务描述**

使用Scrapy框架爬取内蒙古自治区环境保护厅环评审批“项目受理情况”栏目中（入口：<http://www.nmgepb.gov.cn/ywgl/hjpj/xmslqk/index.html>），受理日期在2015/1/1 至今的施工项目的以下字段信息：“项目名称”、“建设地点”、“建设单位”、“环境影响评价机构”以及“受理日期”，其中一条项目记录如图2.1-1所示。

**图2.1-1 需爬取信息网页截图**

**2）输出要求**

将爬取结果以csv文件输出，文件名为“内蒙古自治区环评数据.csv”。文件中包括6个字段，分别是“项目名称”、“建设地点”、“建设单位”、“环境影响评价机构”、“受理日期”以及当前页面的url。输出结果参见附件“参考\_内蒙古自治区环评数据.csv”。

## 2.2任务2—网页遍历

**1）任务描述**

给定一个爬虫入口网址，实现在**无人工干预**的情况下自动遍历一个封闭（没有链接指向其他域名下的网页）的小型网站的所有页面，并记录各页面间的父子关系。

**2）示例站点**

“测试站点”文件夹中储存了一个由9个页面组成的示例站点，站点结构如图2.2-1所示。

P2L3

P1L2A

P1L2B

P1L3

P2L2A

P2L2B

P1L1

MAIN

P2L1

**图2.2-1 示例站点结构**

在“测试站点”文件夹下打开命令行（shift+右键弹出），输入指令“python server.py 端口号”即可在指定端口搭建本地服务器，在浏览器中访问“<http://127.0.0.1:端口号/>main.html”即可进入站点主页。

**3）通用性测试**

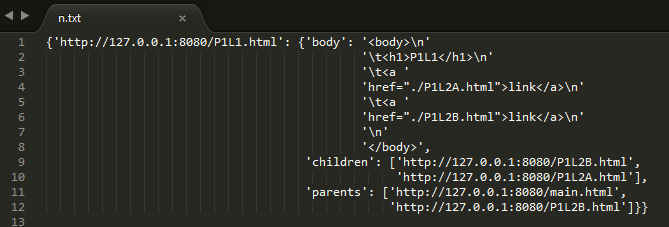
本任务的爬虫程序交付后**将在另一个随机生成的相似站点上进行通用性测试**，通用性测试结果将作为评判本任务完成质量的重要评判标准。关于目标站点已知的信息有：

* 站点入口网址与示例站点一致；
* 该站点封闭，规模大约为10~20个页面；
* 所有链接都位于a标签，所有a标签都是链接；
* 页面间跳转（链接）有可能依赖JavaScript脚本，类型包括但不限于示例站点P2L2A页面中使用的示例脚本。各页URL和页面内容各不相同。

**4）输出要求**

输出格式参考图2.2-2，要求每个页面对应一条记录，每条记录至少包括以下信息：

* 本页面的URL
* 页面源码
* 所有父页面（含有指向本页面的链接的页面）的URL
* 所有子页面（本页面上的链接指向的其它页面）的URL



**图2.2-2输出格式**

# 3.交付物要求

1）针对以上问题编写爬虫程序（需有必要的代码注释），需注意代码的规范性、可读性、模块化等；

2）仔细参考《0 交付文档（标准格式-爬虫方向）》，针对以上问题撰写一篇文档，文档中必须含有以下内容：

* 使用的软件/第三方库的版本信息、编写时使用的操作系统以及爬虫程序的配置及使用指南；
* 爬虫程序的主要实现思路与流程图。

# 4.常用爬虫工具

1）Python爬虫框架：Scrapy

2）Python HTTP库：Requests

3）Python驱动浏览器：Selenium

4）JavaScript驱动的无头浏览器：PhantomJS、SlimerJS、CasperJS

5）提取网页信息的Python库：lxml、BeautifulSoup