**爬虫说明文档**

作者：徐康灵

**一、需求分析**

1)待爬取网站：广东省博物馆：<https://www.gdmuseum.com/col49/list>

2)数据格式要求

所有爬取的数据以utf-8编码保存为csv格式。建议使用csv，pandas等包进行读写，防止出现格式错误等问题。各组应自行检验导出csv文件可正常使用。

3)数据内容要求

应至少包含以下数据，文物基础数据（网站提供），如标题、时代、介绍等；文物对应的详情页面的URL；文物图片（原图，爬取请注意检查，部分网站可能需要手动解析原图地址）；原图下载链接。

需要提交一个说明文档，包括英文字段对应的中文说明，如time：文物时代。

**二、csv文档格式说明**

csv文档共包含5列。

第一列为名称，代表文物名称。

第二列为年代，代表文物年代。

第三列为类型，代表文物的类型。

第四列为图片，代表文物的原图链接。

第五列为详情，代表图片详情页面的链接。

其中，第一、五列数据必定存在。

第二列数据可能为空（如果文物年代不详，在csv中填入的为空字符串“”）。

第三列数据可能为空（如果文物的类型无法确定，在csv中填入空字符串“”）

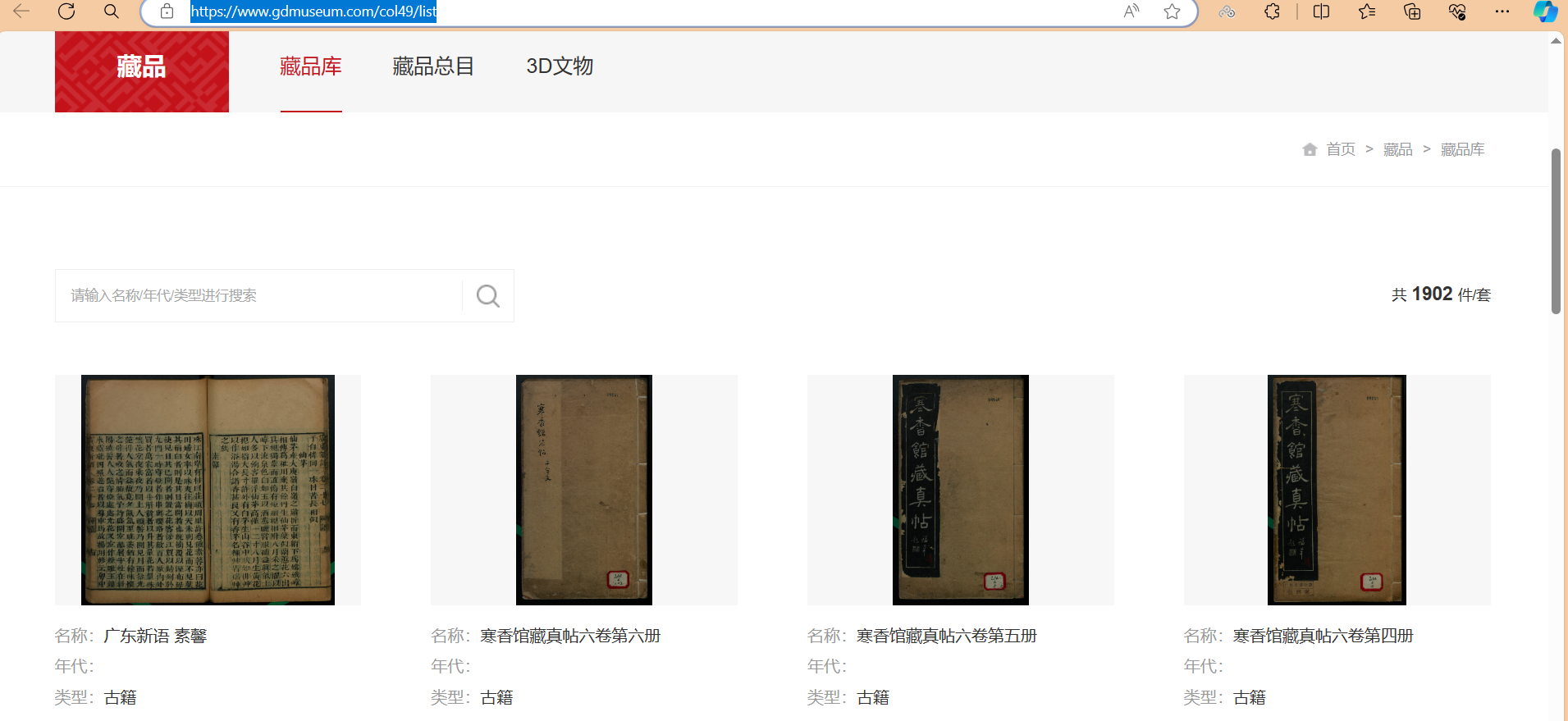
第四列数据可能为空（如果无图片，在csv中填入的为空字符串“”）。

**三、代码技术**

代码使用了requests库进行网站（html格式）的爬取，使用xpath对html源码进行解析。

**四、代码注解**

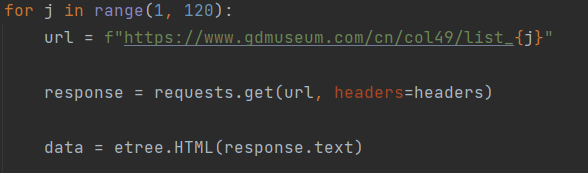
1.网页结构分析



通过开发者模式的网络界面查看，可以发现所有的页面均为GET请求。同时，可以发现代码均为HTML格式，而不是json格式，从而确定用xpath作为解析方式。

2.代码实现

每页16个藏品，共119页，因此可以设置循环范围（1,120），从而获得每页的内容。



再通过拷贝对应内容的xpath，轻松爬取数据，最后写入.csv文件中。

对于有些文物无年代、无类型、无图片的情况，我选择以空字符串代替写入。

