文章编号 :1004-6429 (2018) 02-0053-03 收稿日期 :2018-01-19

浅析基于 Python 爬虫技术的特性及应用

白雪丽

(山西传媒学院,山西太原,030619)

摘 要:在分析大数据网络时代爬虫技术重要性的基础上,介绍了基于 python 语言爬虫技术的基本情况及其特性,并列举了其在实际应用中的案例。

关键词 :大数据 ;爬虫技术 ;Python 语言 数据挖掘

中图分类号 :TP391.3 文献标识码 :A

1994 年,两位斯坦福大学的毕业生创建了 Google 公司。创业初期的 Google 公司仅拥有一个老服务器和一个 Python 网络爬虫,而现在的 Google 公司是世界上最强大的科技企业,在美国纳斯达克的股票市值达到 7 600 亿美元,其核心架构只有 2 点:一是隐藏在世界各个角落的全世界最庞大的数据仓库,二是一个网络爬虫。

随着大数据时代的发展,网络上有价值信息的读取和解析渐渐成为成功挖掘数据的基础性工作。面对互联网上各种格式的网页和网页开发者不同的设计思路,网络爬虫技术在近年来获得了人们越来越多的关注和研究,爬虫技术的发展驱动了现代的大数据网络时代。初创型企业为了节省成本,更需要利用网络上庞大的数据资源,爬虫技术正是一把利器,通过构造爬虫不仅可以连接数据和解析数据,还可以将这些数据进行分析并将分析结果利用图表进行展示。爬虫可以实现网页之间的连接,构建整个网站的地图映射。除此之外,爬虫不会遗漏扩展的链接,会继续对其进行跟踪。面对不同网站,其架构可能千差万别,这就意味着爬虫开发者需要具备足够的经验来应对这些问题。

1 基于 Python 语言爬虫技术的概况

很多计算机语言都有自己的“杀手”级应用。Ruby 被日本人松本行弘发明之后,多年都默默无闻,唯一值得称道的就是其动态语言的特性与元编程能力,随着其在 Web 端的“杀手”级应用 rails 的诞生,Ruby 语言一夜之间成了最流行的语言之一。相对而言,Python 语言则拥有更多“杀手”级应用。其中,在 Web 端的两款框架 django 和 flask 已经广为人知,同时在网页数据解析方面,也有 BeautifulSoup 和 Scrapy 这两款非常强大的“杀手”级应用。

BeautifulSoup 是一款快速抓取有效数据的 Python 爬虫库函数。BeautifulSoup 的原作者是 Leonard Richardson,其 3.0 之前的版本许可协议是基于 Python 软件基金会的许可证,4.0 版本是基于 MIT License 的许可证,目前的稳定版本是在 2016 年 8 月 2 日发布的 4.5.1 版本。对于广大开发者来说,BeautifulSoup 的 API 非常简单和友好,并且可以把很多难以阅读的标签处理得很好。但光有 BeautifulSoup 包是不够的,爬虫开发者还需要一些其他的库函数,比如 urlib2 或者 requests 包都是非常必要的补充。这些库函数的组合可以使开发者具备强大的爬虫处理功能,不仅可以下载网络上的数据,还可以解析其元素。同时 BeautifulSoup 有非常强大的社区,初学者可以通过社区来快速入门。BeautifulSoup 的学习曲线非常平缓,初学者很容易入门,在很多情况下,开发者可以下载所需要的 HTML 元素,相对于 urllib2 的特性来说,requests 的特性更加强大。

Scrapy 也是一款爬虫框架,其最大的特性是基于 Twisted 框架制作,这是一个异步并发的库,其爬虫的性能非常卓越。Scrapy 和 BeautifulSoup 一样可以基于 Python2 和 Python3 运行,兼容性均不是问题,其可以构造基于 HTML 数据包含 CSS 和 XPath 的解析。相对而

言，BeutifulSoup 更多应用于 HTML 的数据解析，Scrapy 则更多用于下载 HTML 并处理数据及存储。相对来说，Scrapy 的学习曲线很陡峭，除了需要阅读更多书籍和教程，还需要长时间的练习，才能成为 Scrapy 高手。

由此看来，如果开发者没有太多经验，需要“爬”的工作量不是很大，那么 BeautifulSoup 就是一个很好的选择 如果开发者想拥有一个强大的富开户端的爬虫，或者开发者已经在爬虫领域拥有足够的经验，则可以尝试 Scrapy。

2　基于 BeautifulSoup 爬虫技术的特性

2.1　数据存储

爬虫获取的数据可以通过 csv 库将数据存储为 csv 格式。若 csv 文件的路径构造不存在，Python 会创建一个新的路径来存放 csv 文件 若 csv 路径存在，Python 则会重写这个 csv 文件。除此之外，还可以存储成 pdf 格式。pdf 格式的文件可以转换成文本格式，在纯文本的 pdf 文件转换过程中，基本不会出现差错，若是 pdf 含有图表或者特定格式，转换效果可能会受到影响。

2.2　处理脏数据

获取数据后，需要利用正则表达式或者一些判断语句来实现对脏数据的清洗和过滤，这样可以获得更加清晰的内容，比如说一些特殊字符格式，或者一些简单的词组，还有一些网站里的特殊标记符号等。

2.3　处理自然语言词语

通常情况下，人们会投入很多精力在感兴趣的词语和不感兴趣的词语 2 个方面。美国 Brigham Young 大学的语言学家曾构造了美国当代词语的词语库，其中包含了美国过去 10 年间各种流行出版物中的词语，词汇数超过 4.5 亿个，词语库中超高频率的 5 000 个词语是免费的。通过对获取文本的自然语言进行处理，将高频词语作为过滤容器，获得文本中出现频率最高的词语清单，从而在爬虫开发者并未阅读文本内容时，爬虫程序已经获取了此文本的特征结构。

NLTK 库是一款卓越的词频及语法多样性分析包，很多时候开发者甚至会怀疑这个分析包是不是威力过于巨大，很多情况下，一些唾手可得的分析代码就足以胜任。

2.4　正则表达式的权衡

正则表达式可以完成与爬虫相同的工作，很多开发者会疑惑为什么不使用正则表达式而使用爬虫。原因是使用爬虫代码的强健程度更高，正则表达式需要对页面很多细节做出相应的改变，但在运行速度上，正则表达式比爬虫代码有更大优势。这就是代码鲁棒性与性能的权衡，如果用正则表达式可以很容易完成，开发者就可以继续沿着这条路线前行，对于很复杂的页面来说，爬虫代码则更具优势。

3　爬虫在工程管理领域的应用

在工程管理领域，很多设计公司都有门户网站。通过爬虫将其多年的设计案例爬取下来，构建甲方的工程情报数据库是未来的一个发展趋势。以日本设计公司 OBAYASHI CORPORATION 为例，其设计方案非常繁多，遍布世界各个地区。通过在浏览器中对其门户网站的标签进行观察，可以找到该公司案例对应的各种标签，然后再通过爬虫库函数丰富的查询语句和解析功能，实现各个数据项的爬取，之后再使用电子表格实现数据结构存储，从而构建其相关设计案例的甲方情报数据库。

图 1 是 OBAYASHI CORPORATION 公司门户网站展示的一些设计案例，通过对其标签层级的观察，使用爬虫技术获取各个设计案例之后，开始构造哈希表数据结构，构造出的数据结构表包括工程项目名称、工程项目实施地国家、建设公司名称、具体实施地区、项目建设时间等 5 项。然后，在爬虫的迭代循环过程中将逐步完善的哈希表填充到最后的数组数据结构中。得到完整的数据库后，便可以对日本设计公司的数据进行挖掘和统计。比如要查询与此设计公司来往最密切的 5 家施工公司，便可以利用数据库的统计功能，统计出施工单位出现频次最高的 5 家；如果要单独查询日本福冈地区与设计公司来往最密切的 5 家施工公司，就可以利用数据库条件查询语句设定来实现。

通过对图 1 中该公司设计案例的数据进行爬取，将得到的结果存储到 csv 格式的文件中，并输出在客户终端，显示效果如图 2 所示。

4　结语

大数据时代的兴盛，极大地推动了爬虫技术的发展。目前从事爬虫技术的开发者不计其数，爬虫技术已

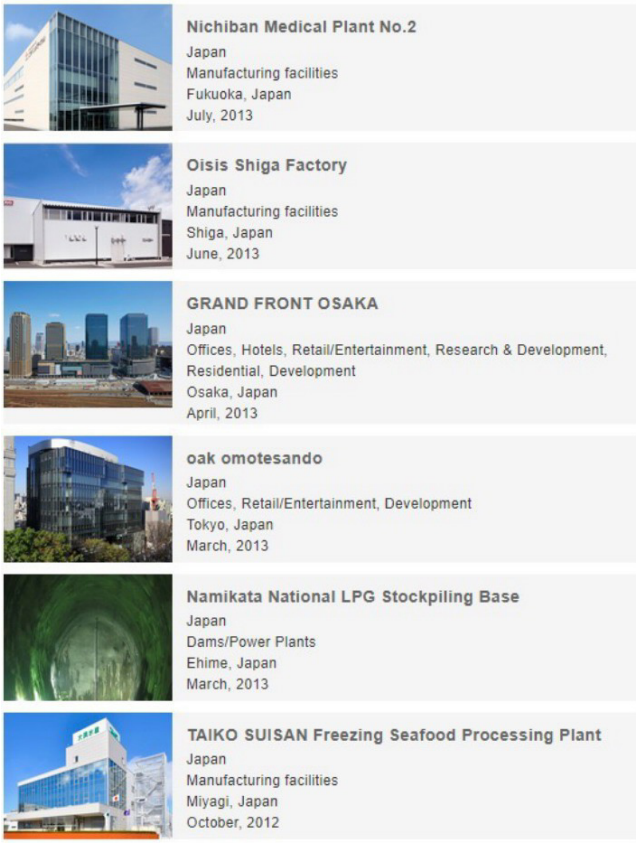

图 1　日本 OBAYASHI CORPORATION 设计公司网站部分设计案例

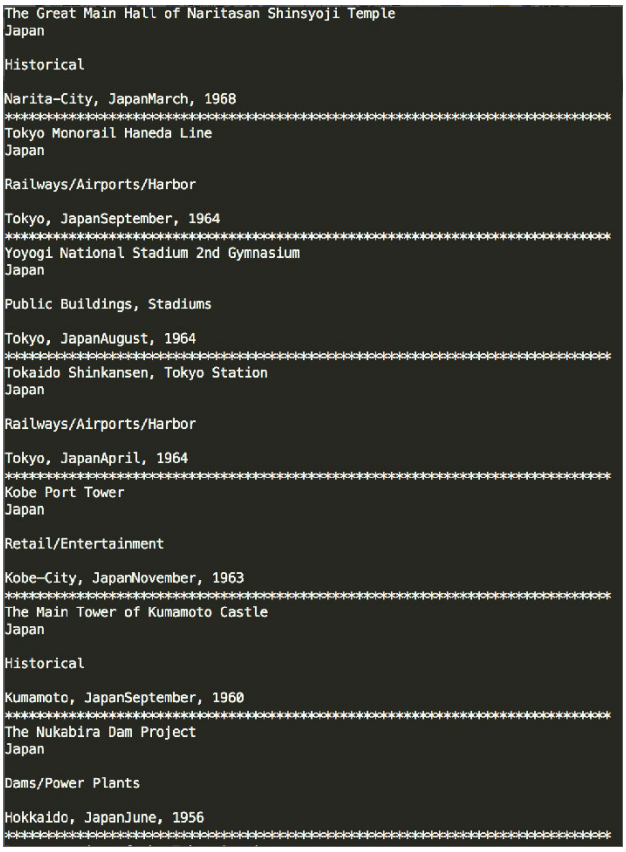

图 2　爬取到的信息储存为 csv 格式后在客户终端显示效果图

经成为数据挖掘企业的一把利器，各种爬虫库也应运而生，用来处理特定领域的各种数据。有了 csv 格式的文件，便可以将其存储至 SQL 数据库中，或者类似 MongoDB 的 Nosql 数据库中，建立企业自身的数据库。同时，也可以将 csv 格式的文件导入 Matlab 或 R 中，实现数据的特定算法研究和绘图展示，这都为爬虫技术的实际应用提供了便利。

参考文献

[1]　郭丽蓉.基于 Python 的网络爬虫程序设计 [J] .电子技术与软件工程，2017 (12)：23-24.

[2]　李璋.基于 Hadoop 的互联网数据营销系统的设计与实现 [D] .北京：中国科学院大学，2017：4-5.

[3]　王锦阳. 主题网络爬虫的并行化研究与设计 [D] .成都：西南石油大学，2017.

（责任编辑：王　欣）

作者简介：白雪丽，女，1988 年生，山西传媒学院助理工程师。

Analysis on the Characteristics and Application of Crawler Technology Based on Python

BAI Xueli

ABSTRACT: On the basis of the analysis of the importance of crawler technology in the era of large data network, this paper introduces the basic situation and characteristics of crawler technology based on Python, and lists the cases of its practical application.

KEY WORDS: large data; crawler technology; Python language; data mining