Python 爬虫之高考作文

文/刘子元

摘要

在互联网日益发展的今天，计算机应用成为生活中不可或缺的一部分。本文所介绍的网络爬虫程序，是从一个庞大的网站中，将符合预设条件的对象“捕获”并保存的一种程序。如果将庞大的互联网比作一张蜘蛛网，爬虫程序就像网上游弋的蜘蛛，将网上一个个“猎物”摘取下来。

【关键词】高考作文 Python 爬虫 互联网

1 概述

平时所说的爬虫，就是网络爬虫，大家可以理解为在互联网网页上爬行的一只蜘蛛。可以把互联网比作一张大网，爬虫就是在这张网上爬来爬去的蜘蛛，如果遇到需要的有价值的资源，它就会爬取下来。想抓取什么，就可以用代码控制抓取你想要的内容，不需要的那部分可以忽略不计。

当然网络爬虫并不是所有的网页都可以抓取，因为有些网页安全性极高，一般无法获取到他们的页面信息，抓取的资源（数据）也不是想怎么处理就怎么处理的，而是要在合法的范围内对数据进行一定的处理，如果对这些数据随意的散播或者用于不正当交易，是要负法律责任的。当然，一些保密的数据也会做好安全措施，不能轻易的被爬取。

网络爬虫 , 最近几年被广泛用于互联网搜索引擎和其他类似的网站，通过代码可以让程序自动采集所有能够访问到的页面内容，最终获取或更新这些网站的内容和检索方式。从功能上来讲，爬虫一般分为三部分，包含数据的采集、数据的处理和存储。前几年的网络爬虫是从一个或者多个初始网页的URL地址开始，得到最原始网页上的 URL，并且在抓取网页的过程中，不间断的从当前网页上获取新的 URL 放入队列 , 达到系统要求的停止条件才会终止。

2 实验目的

对于高三的学生来说时间就是金钱，在紧张的高三生活中，如何快速获取有参考价值的信息就显得比较重要，历年的真题基本上都是高三学生必练的。而语文科目在高考中占据很大比例，语文中的作文又是重中之重，因为语文中的作文成绩在整个语文成绩中起着举足轻重的作用，大家都知道想写好作文就要多阅读，但是那么多文章，怎么在短时间内去选取有用的作文呢？近几年全国各省份的高考满分作文是大家都要参考和阅读的，这些作文以及题目收集起来比较麻烦。所以，本次试验是运用网络爬虫技术，从某网站帮助学子们快速便捷地抓取最近两年全国各地的高考满分作文，以供日后的参考学习。

3 实验过程

3.1 实验环境

Windows10 64 位操作系统，Python2+requests+BS4, 所用工具 PyCharm。

3.2 技术分析

3.2.1 如何浏览网页

网页是由什么组成的？一般现在我们看到的网页都是 Html+CSS 组成的，Html 是一种超文本标记语言，CSS 简单的说就是样式，只有 Html，写出的网页基本上是黑白的，没有好看的样式，所以我们常常看到的五颜六色的颜色就是 CSS 起的作用，既然是语言，那它就会具备语言独有的特点，正是因为它有自己的特点，我们才能分析特点然后进行页面的抓取。我们访问网页的过程，首先会向服务器发送一个请求，然后经过服务器的解析，返回给客户端，用户就可以看到文字图片以及一些动态的效果。

因此，用户通过浏览器看到的网页实质是由 HTML 代码写出来的，爬虫爬来的便是这些内容，通过对获取到的 HTML 源码的分析和过滤，实现对图片、文字等资源信息的获取。本系统是通过抓取某网站的 HTML 代码，并进行筛选，最终获取自己需要的信息。

3.2.2 URL 的含义

URL 即统一资源定位符，也就是我们经常在地址栏输入的网址，这个网址就是对可以从互联网上得到的资源的位置和访问方法的简洁表示方法，是互联网上标准资源的统一地址。互联网上的每个文件都有唯一的 URL 地址，从它包含的信息可以看出文件的位置以及浏览器应该怎么处理它。

Language Rank	Types	Spectrum Ranking
1. Python		100.0
2. C		99.7
3. Java		99.5
4. C++		97.1
5. C#		87.7
6. R		87.7
7. JavaScript		85.6
8. PHP		81.2
9. Go		75.1
10. Swift		73.7

图 1

URL 由以下几部分组成：

（1）第一部分是协议，比如 http、www 协议。

（2）第二部分是服务器的地址，指出所在服务器的域名。

（3）第三部分是端口，有的 URL 并不是，访问一些资源的时候需要对应的端口号。

（4）第四部分是路径，指明服务器上某资源的位置（其格式与 DOS 系统中的格式一样，通常由目录 / 子目录 / 文件名这样的结构组成）。与端口一样，路径并非是必须的。

爬虫在爬取数据时需要有一个目标的 URL 才可以进一步获取数据，因此，它是操作爬虫获取数据的基本依据，准确深入的理解它的含义对爬虫的学习有很大帮助。

3.2.3 使用的库 Requests、BS4

负责连接网站，处理 http 协议。它允许发送 HTTP 请求，无需手工劳动。不需要手动为 URL 添加查询字串，也不需要对 POST 数据进行表单编码。Keep-alive 和 HTTP 连接池的功能是完全自动化的，一切动力都来自于根植在它内部的 urllib3。

Bs4 负责将网页数据结构化，从而更方便的获取内部数据，通过一定的规则针对文件内容进行操作，比如获取数据，修改，删除等。Bs4 提供一些简单的方法、属性。它是一个工具箱，通过解析文档为用户提供需要抓取的数据，因为简单，所以不需要多少代码就可以写出一个完整的应用程序。

Bs4 自动将输入文档转换为 Unicode 编码，自动格式化为 utf-8 编码。用户不需要考虑编码方式，如果文档没有指定一个编码方式，Bs4 会自动猜测文件内容的编码，当然我们推荐自己指定编码格式，并且也是相当简单的。

Bs4 已成为和 lxml 一样出色的 python 三方包，为用户灵活地提供不同的解析策略或强劲的速度。

3.2.4 Python 语言

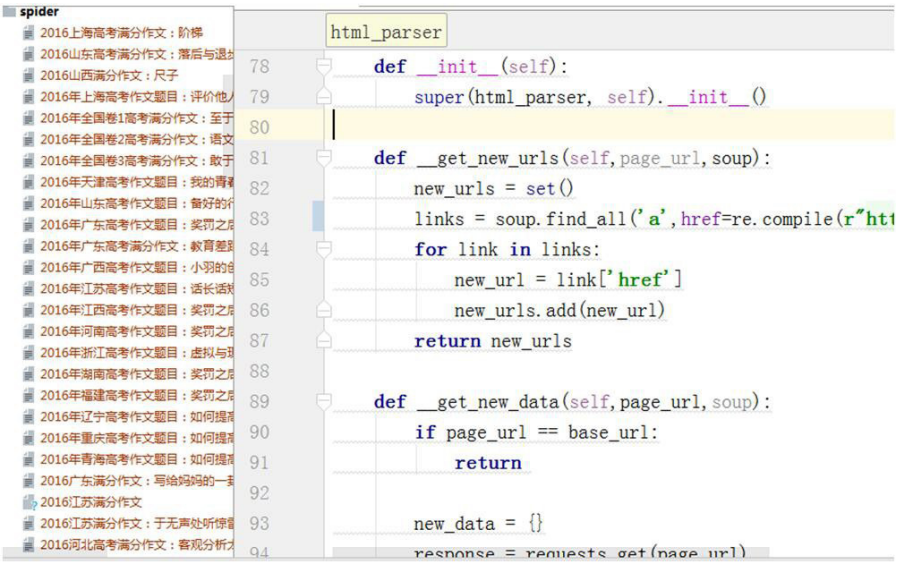

图 2

在形形色色的计算机语言中，C 语言、java 和 Python 一直是使用率最高的几种。根据国外某网站调查，2017 年，Python 成为当前全球使用率最高的计算机语言（如图 1 所示 IEEE 发布 2017 年编程语言排行榜 ）。所以利用该语言进行程序编写是符合当前形式的。本系统用 python 脚本语言开发，该脚本语言与其它编程语言相比的优势在于它的语法简单、系统库强大、实现功能容易、高效率的高层数据结构、简单的面向对象编程、代码结构清晰易懂。如今该语言被广泛的应用于系统后台处理和网页编程。由于此脚本语言有着这么多的优势，所以笔者通过该脚本语言实现了一个爬虫、敏感文件扫描和日志分析程序。爬虫通过任务队列、线程池实现多线程并发爬取网页，在爬取网页之后对网页进行解码分析，获取目录结构，对已知的目录结构进行敏感文件扫描。同时也通过脚本程序利用攻击规则库对用户请求的 web 日志进行安全分析，提取出日志中的 sql 注入攻击。

此语言是所有编程语言中最容易入门的语言，并且其应用面非常广泛，是一种非常有应用前景和研究前景的语言。其应用领域含括了后台开发、图像处理、数据挖掘、数据分析、机器学习、网络攻击、SDN、神经网络、自动化运维、计算机视觉、自然语言处理等。该语言作为一种“胶水语言”可谓无所不能，甚至能够开发安卓应用。

3.2.5 程序代码

以下为本系统部分程序代码：

coding:utf-8 改行代码是为了防止程序中的中文出现乱码；requests 和 bs4 是导入的抓取爬虫的库，引入 random 库是为了取到随机数，re 模块是为了使用正则表达式；re.compile(r"http://www.******.com/e/201[76]\d+/\S+.shtml") 上面这行代码用到了正则取到链接里面的 2016 和 2017 年的数据，当然如果取其他年份的可以写成 [7654]； for link in links: 然后利用循环实现对每个年份链接的读取，这样就能分别读取到定义好的年份链接，从而取到对应的数据。

count += 1if count == 300 print 'craw failed %d %s' % (count, str(e)) 以上代码可实现循环取到所要获取的作文的数量，如果获取到了 300 篇就中断，否则就输入，输出这里做了一个格式化的输入，会在每一个作文的前面加上序号，这样就能直观的看出来输出作文的数量，当然篇数 300 这个值是可以修改的，这里之所以写 300 是因为从该网站判断出每年的全国满分作文基本没有超过 100 篇的，这里获取 2 年的作文，如果写的是 100，那就只能获取到 100 篇。最终写个 main 函数，在该函数里面就是要调用写好的方法，程序最终运行的入口就在该函数里面。

3.3 实验分析

用户通过爬虫的入口向程序提供需要爬取的目标，爬取的深度和使用多少个线程爬取，如果没有定义线程数，程序会初始化为 9 个线程爬取。程序将爬取得到的网页内容进行解码分析，提取出里面的 URL，并将这些 URL 做一些处理后加入队列进行下一步爬取。目标爬取完毕之后会将结果保存下来，然后再调用敏感文件扫描模块对这些目录进行敏感文件扫描，对于存在的敏感文件程序会自动将结果保存下来。

3.3.1 目标网页爬取

在目标爬取的测试过程中，程序在获得一个目标站点后开始进行爬取。首先将这个目标 url 加入 urlQueue 队列中，在 start 函数中从 urlQueue 队列中获取第一个 url，随后调用线程中 addJob 函数将 url 和工作 work 函数同时加入线程池任务队列中。此时线程从线程池任务队列中获取任务，也就是获取到的 url 和 work 函数，随后线程开始执行 work 函数，work 函数即对 url 进行爬取，将 url 加入已爬取的任务队列 readUrls 中。爬取方法是调用 requests 模块中的 get 函数对目标进行网页抓取：html=requests.get（url），此方法返回一个 html 对象。该对象中的 content 属性为网页内容：htmldata=html.content。

接下来是对返回的网页内容进行解析分析，本程序采用的是 python 的第三方模块 bs4 对其解析。该模块是用 python 写的一个 HTML/XML 解析器，它可以很好的处理不规范标记并生成剖析树。通常用来分析爬虫抓取到的 web 文档。对于不规则的 html 文档，也有很多补全功能，节省了开发者的时间和精力。通过 BS4 对 htmldata 解析。接着遍历 allurl 列表，如果列表中的链接没有在 readUrls 中，就将其加入 urlQueue 队列中。如此循环操作，直到最后达到爬取的深度停止任务，完成网页爬取。

4 结论

图 2 为最终的实验结果，爬取到了所需的数据。

通过 Python 语言编程系统，可以迅速地将目标网站各网页中符合条件的信息抓取并储存到指定位置。通过本系统，高三学子能够快速准确地收集 2016 及 2017 年的高考满分作文，为以后学习成绩的进步打下坚实的基础。

参考文献

[1]https://www.liaoxuefeng.com.
[2]《用 python 写网络爬虫》作者：（澳）Richard Lawson.
[3]http://cn.python-requests.org/zh_CN/latest.
[4]百度百科.
[5]http://www.chinaz.com/news/2017/0724/792870.shtml.
[6]《Python 网路数据采集》作者：（美）Ryan Mitchell.

作者单位

山东淄博实验中学　山东省淄博市　255000