使用requests和re模块爬取i春秋论坛的精品贴（小爬虫）

原创

huanghelouzi 于 2018-11-06 20:54:33 发布 668 收藏

分类专栏： # 爬虫 # python 文章标签： 爬虫 python

本文链接：https://blog.csdn.net/huanghelouzi/article/details/83794132

爬虫 同时被 2 个专栏收录

2 篇文章 0 订阅

订阅专栏

python

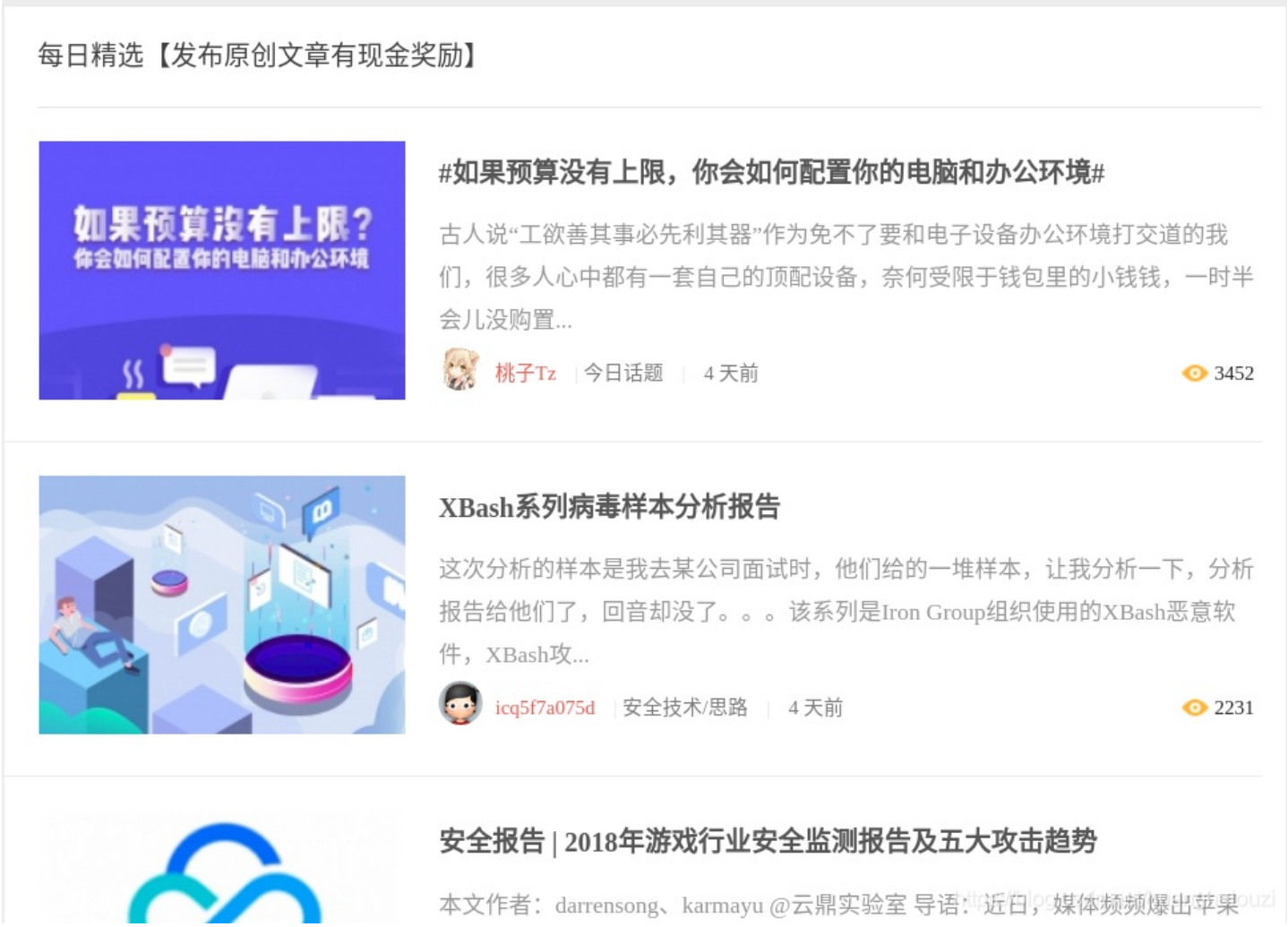
12 篇文章 2 订阅

订阅专栏

前言

下一篇是使用requests和re模块爬取某个学习站点的所有用户头像.

最近在刷i春秋论坛的帖子，发现论坛首页每天都会推送一些精品文章，但是有时候好几天也没有更新首页的推送，总不能每天都去刷新吧。所以有了这个脚本或称之为小爬虫（如果它能被称为爬虫的话），去爬取精品文章的标题，链接以及简介。



正文

需要看懂这个脚本大概需要学会简单的正则表达式，requests模块和re模块的基本使用，如果不会请自行学习。大佬绕行。

第一步是爬取整个首页，分析源代码，这一步需要需要到`requests`模块，当然其他的模块可能也是可以的。

```
import requests

url = "https://bbs.ichunqiu.com/portal.php"
headers = {
            'Host': 'bbs.ichunqiu.com',
            'Connection': 'close',
            'Cache-Control': 'max-age=0',
            'Upgrade-Insecure-Requests': '1',
            'User-Agent': 'Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Ubuntu Chromiu
m/67.0.3396.99 Chrome/67.0.3396.99 Safari/537.36',
            'Accept': 'text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8',
            'Accept-Language': 'zh-CN,zh;q=0.8',
            }
def get_html(url, headers):
    response = requests.get(url=url, headers=headers, timeout=5)
    html = response.content
    return html.decode("utf-8")

if __name__ == '__main__':
    html = get_html(url, headers)
    print(html)
```

在返回的源代码中发现每一篇精品文章的的大概在这几个标签之间

```
<h3 class="clr" style="width: 100%; float: left;"><a href="https://bbs.ichunqiu.com/thread-47215-1-1.html" targe
t="blank" class="ui_colorG" style="color: #555555;">一键安装藏隐患，phpStudy 批量入侵的分析与溯源</a></h3><p class="
cdg ovh" style="width: 100%; float: left; color: #999999;margin-bottom: 28px;">一、前言近日，腾讯安全云鼎实验室监测
到大量主机被入侵并添加了一个名为“vusr_dx$”的隐藏帐号；同时，云鼎实验室还监测到此类帐号被大量创建的同时存在对应...</p>
```

以这篇文章为例，文章的名称在这个a标签之间

```
<a href="https://bbs.ichunqiu.com/thread-47215-1-1.html" target="blank" class="ui_colorG" style="color: #555555;
">一键安装藏隐患，phpStudy 批量入侵的分析与溯源</a>
```

但是每篇文章的url是改变的，所以为了方便，我选取的范围是

```
target="blank" class="ui_colorG" style="color: #555555;">这个文章的标题</a>
```

同样的道理，每一篇精品文章的url在这个范围中间

```
<h3 class="clr" style="width: 100%; float: left;"><a href="这个文章的url" target="blank" class="ui_colorG" style=
"color: #555555;">
```

简介在这个范围中间

```
<p class="cdg ovh" style="width: 100%; float: left; color: #999999;margin-bottom: 28px;">这个是文章的简介</p>
```

到现在位置我们已经确定了我们需要爬取的三个的具体位置，接着就可以使用正则表达式去匹配。下面是完整的匹配的代码

```
url = "https://bbs.ichunqiu.com/portal.php"
headers = {
            'Host': 'bbs.ichunqiu.com',
            'Connection': 'close',
            'Cache-Control': 'max-age=0',
            'Upgrade-Insecure-Requests': '1',
            'User-Agent': 'Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Ubuntu Chromium/67.0.3396.99 Chrome/67.0.3396.99 Safari/537.36',
            'Accept': 'text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8',
            'Accept-Language': 'zh-CN,zh;q=0.8',
            }
re_title = 'target="blank" class="ui_colorG" style="color: #555555;">(.+?)</a></h3>'
re_url = '<h3 class="clr" style="width: 100%; float: left;"><a href="(.+?)" target="blank" class="ui_colorG" style="color: #555555;">'
re_overview = '<p class="cdg ovh" style="width: 100%; float: left; color: #999999;margin-bottom: 28px;">(.+?)</p>'

def get_html(url, headers, re_title, re_url, re_overview):
    '''
        返回i春秋bbs的精品文章对应的一个元祖(标题，链接，简介)
    '''
    try:
        response = requests.get(url=url, headers=headers, timeout=5)
        html = response.content
        titles = re.findall(re_title, html.decode("utf-8"))
        urls = re.findall(re_url, html.decode("utf-8"))
        overviews = re.findall(re_overview, html.decode("utf-8"))
    except Exception as e:
        print(e)
    return titles, urls, overviews

if __name__ == '__main__':
    html = get_html(url, headers, re_title, re_url, re_overview)
    print(html)
```

在上面的脚本中，需要正则表达式匹配的一共有三个参数，以文章的标题为例子

```
'target="blank" class="ui_colorG" style="color: #555555;">(.+?)</a></h3>'
```

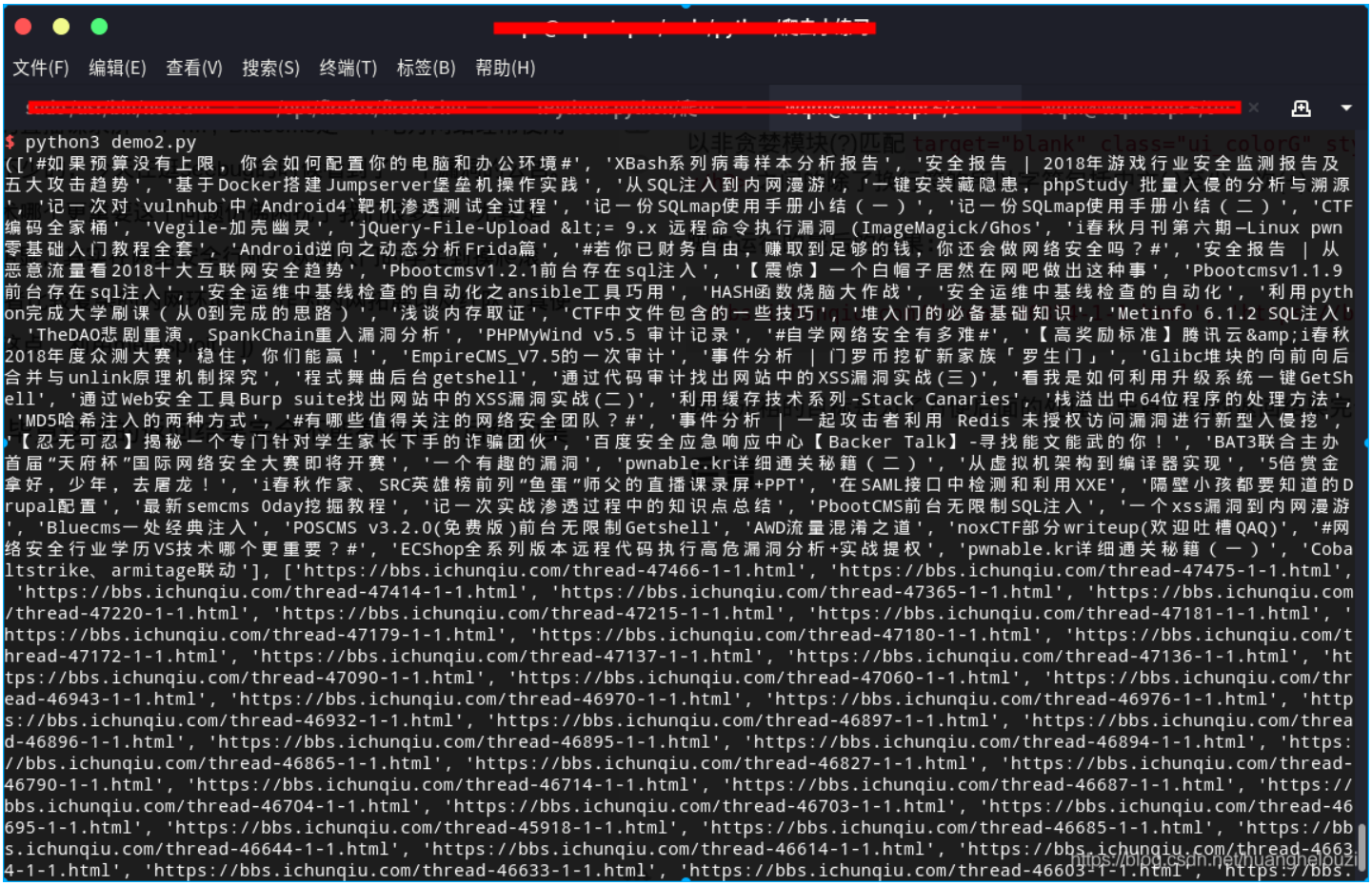
以非贪婪模块(?)匹配 `target="blank" class="ui_colorG" style="color: #555555;">` 和 `</a></h3>` 之间的除了换行符的所以字符包括中文(.)至少一次(+)。

脚本运行的最后的结果：

```
(['#如果预算没有上限，你会如何配置你的电脑和办公环境#', 'XBash系列病毒样本分析报告', '安全报告 | 2018年游戏行业安全监测报告及五大攻击趋势', '基于Docker搭建Jumpserver堡垒机操作实践', '从SQL注入到内网漫游', '一键安装藏隐患，phpStudy 批量入侵的分析与溯源', '记一次对`vulnhub`中`Android4`靶机渗透测试全过程', '记一份SQLmap使用手册小结（一）', '记一份SQLmap使用手册小结（二）', 'CTF编码全家桶', 'Vegile-加壳幽灵', 'jQuery-File-Upload &lt;= 9.x 远程命令执行漏洞 (ImageMagick/Ghos', 'i春秋月刊第六期—Linux pwn零基础入门教程全套', 'Android逆向之动态分析Frida篇', '#若你已财务自由，赚取到足够的钱，你还会做网络安全吗？#', '安全报告 | 从恶意流量看2018十大互联网安全趋势', 'Pbootcmsv1.2.1前台存在sql注入', '【震惊】一个白帽子居然在网吧做出这种事', 'Pbootcmsv1.1.9前台存在sql注入', '安全运维中基线检查的自动化之ansible工具巧用', 'HASH函数烧脑大作战', '安全运维中基线检查的自动化', '利用python完成大学刷课（从0到完成的思路）', '浅谈内存取证', 'CTF中文件包含的一些技巧', '堆入门的必备基础知识', 'Metinfo 6.1.2 SQL注入', 'TheDAO悲剧重演，SpankChain重入漏洞分析', 'PHPMyWind v5.5 审计记录', '#自学网络安全有多难#', '【高奖励标准】腾讯云&amp;i春秋2018年度众测大赛，稳住，你们能赢！', 'EmpireCMS_V7.5的一次审计', '事件分析 | 门罗币挖矿新家族「罗生门」', 'Glibc堆块的向前向后合并与unlink原理机制探究', '程式舞曲后台getshell', '通过代码审计找出网站中的XSS漏洞实战(三)', '看我是如何利用升级系统一键GetShell', '通过Web安全工具Burp suite找出网站中的XSS漏洞实战(二)', '利用缓存技术系列-Stack Canaries', '栈溢出中64位程序的处理方法', 'MD5哈希注入的两种方式', '#有哪些值得关注的网络安全团队？#', '事件分析 | 一起攻击者利用 Redis 未授权访问漏洞进行新型入侵挖', '【忍无可忍】揭秘一个专门针对学生家长下手的诈骗团伙', '百度安全应急响应中心【Backer Talk】-寻找能文能武的你！', 'BAT3联合主办 首届“天府杯”国际网络安全大赛即将开赛', '一个有趣的漏洞', 'pwna
```

ble.kr详细通关秘籍（二）', '从虚拟机架构到编译器实现', '5倍赏金拿好，少年，去屠龙！', 'i春秋作家、SRC英雄榜前列“鱼蛋”师父的直播课录屏+PPT', '在SAML接口中检测和利用XXE', '隔壁小孩都要知道的Drupal配置', '最新semcms 0day挖掘教程', '记一次实战渗透过程中的知识点总结', 'PbootCMS前台无限制SQL注入', '一个xss漏洞到内网漫游', 'Bluecms一处经典注入', 'POSCMS v3.2.0(免费版)前台无限制Getshell', 'AWD流量混淆之道', 'noxCTF部分writeup(欢迎吐槽QAQ)', '#网络安全行业学历VS技术哪个更重要？#', 'ECShop全系列版本远程代码执行高危漏洞分析+实战提权', 'pwnable.kr详细通关秘籍（一）', 'Cobaltstrike、armitage联动'], ['https://bbs.ichunqiu.com/thread-47466-1-1.html', 'https://bbs.ichunqiu.com/thread-47475-1-1.html', 'https://bbs.ichunqiu.com/thread-47414-1-1.html', 'https://bbs.ichunqiu.com/thread-47365-1-1.html', 'https://bbs.ichunqiu.com/thread-47220-1-1.html', 'https://bbs.ichunqiu.com/thread-47215-1-1.html', 'https://bbs.ichunqiu.com/thread-47181-1-1.html', 'https://bbs.ichunqiu.com/thread-47179-1-1.html', 'https://bbs.ichunqiu.com/thread-47180-1-1.html', 'https://bbs.ichunqiu.com/thread-47172-1-1.html', 'https://bbs.ichunqiu.com/thread-47137-1-1.html', 'https://bbs.ichunqiu.com/thread-47136-1-1.html', 'https://bbs.ichunqiu.com/thread-47090-1-1.html', 'https://bbs.ichunqiu.com/thread-47060-1-1.html', 'https://bbs.ichunqiu.com/thread-46943-1-1.html', 'https://bbs.ichunqiu.com/thread-46970-1-1.html', 'https://bbs.ichunqiu.com/thread-46976-1-1.html', 'https://bbs.ichunqiu.com/thread-46932-1-1.html', 'https://bbs.ichunqiu.com/thread-46897-1-1.html', 'https://bbs.ichunqiu.com/thread-46896-1-1.html', 'https://bbs.ichunqiu.com/thread-46895-1-1.html', 'https://bbs.ichunqiu.com/thread-46894-1-1.html', 'https://bbs.ichunqiu.com/thread-46865-1-1.html', 'https://bbs.ichunqiu.com/thread-46827-1-1.html', 'https://bbs.ichunqiu.com/thread-46790-1-1.html', 'https://bbs.ichunqiu.com/thread-46714-1-1.html', 'https://bbs.ichunqiu.com/thread-46687-1-1.html', 'https://bbs.ichunqiu.com/thread-46704-1-1.html', 'https://bbs.ichunqiu.com/thread-46703-1-1.html', 'https://bbs.ichunqiu.com/thread-46695-1-1.html', 'https://bbs.ichunqiu.com/thread-45918-1-1.html', 'https://bbs.ichunqiu.com/thread-46685-1-1.html', 'https://bbs.ichunqiu.com/thread-46644-1-1.html', 'https://bbs.ichunqiu.com/thread-46614-1-1.html', 'https://bbs.ichunqiu.com/thread-46634-1-1.html', 'https://bbs.ichunqiu.com/thread-46633-1-1.html', 'https://bbs.ichunqiu.com/thread-46603-1-1.html', 'https://bbs.ichunqiu.com/thread-46594-1-1.html', 'https://bbs.ichunqiu.com/thread-46409-1-1.html', 'https://bbs.ichunqiu.com/thread-46410-1-1.html', 'https://bbs.ichunqiu.com/thread-46567-1-1.html', 'https://bbs.ichunqiu.com/thread-46423-1-1.html', 'https://bbs.ichunqiu.com/thread-46354-1-1.html', 'https://bbs.ichunqiu.com/thread-46370-1-1.html', 'https://bbs.ichunqiu.com/thread-46322-1-1.html', 'https://bbs.ichunqiu.com/thread-46321-1-1.html', 'https://bbs.ichunqiu.com/thread-46282-1-1.html', 'https://bbs.ichunqiu.com/thread-46250-1-1.html', 'https://bbs.ichunqiu.com/thread-46283-1-1.html', 'https://bbs.ichunqiu.com/thread-46260-1-1.html', 'https://bbs.ichunqiu.com/thread-46249-1-1.html', 'https://bbs.ichunqiu.com/thread-46131-1-1.html', 'https://bbs.ichunqiu.com/thread-46127-1-1.html', 'https://bbs.ichunqiu.com/thread-46149-1-1.html', 'https://bbs.ichunqiu.com/thread-46121-1-1.html', 'https://bbs.ichunqiu.com/thread-46172-1-1.html', 'https://bbs.ichunqiu.com/thread-46068-1-1.html', 'https://bbs.ichunqiu.com/thread-46100-1-1.html', 'https://bbs.ichunqiu.com/thread-46060-1-1.html', 'https://bbs.ichunqiu.com/thread-46072-1-1.html', 'https://bbs.ichunqiu.com/thread-46059-1-1.html', 'https://bbs.ichunqiu.com/forum.php?mod=viewthread&tid=46069&page=1&extra=#pid509804', 'https://bbs.ichunqiu.com/thread-46029-1-1.html', 'https://bbs.ichunqiu.com/thread-46026-1-1.html', 'https://bbs.ichunqiu.com/thread-44203-1-1.html'], ['古人说“工欲善其事必先利其器”作为免不了要和电子设备办公环境打交道的我们，很多人心中都有一套自己的顶配设备，奈何受限于钱包里的小钱钱，一时半会儿没购置...', '这次分析的样本是我去某公司面试时，他们给的一堆样本，让我分析一下，分析报告给他们了，回音却没了。。。该系列是Iron Group组织使用的XBash恶意软件，XBash攻...', 'CTF编码全家桶小程序提供Base64、Url、HTML实体、莫尔斯电码等编码转换工具，凯撒密码、栅栏密码、ROT13、MD5、SHA等加密工具，及IP地址查询、Whois信息查询等工...', 'jQuery-File-Upload 是 Github 上继 jQuery 之后最受关注的 jQuery 项目，该项目最近被披露出一个存在了长达三年之久的任意文件上传漏洞，该漏洞在随后发布的 v9...', '第六期月刊—Linux pwn零基础入门教程全套上线啦~！！根据上期读者给出的优化建议进行了全方位优化修改。漂亮的排版，便捷的书签，巩固知识的课后习题，完整的知...', '上期的Android逆向之动态分析so篇大家学习的如何啦？本期斗哥将带来Android逆向之动态分析Frida篇。主要内容有Frida环境搭建与Frida在Android环境下的运行与使用...', '前几周斗哥分享了基线检查获取数据的脚本，但是在面对上百台的服务器，每台服务器上都跑一遍脚本那工作量可想而知，而且都是重复性的操作，于是斗哥思考能不能找...', '本期讲解一下hash函数，由于之前在比赛中做到了一题hash有关的题目，引发了此次的深（烧）度（脑）研究，本来想讲讲原理，但是太难，看得很痛苦，所以此次通过结...', '安全运维工作中经常需要进行安全基线配置和检查，所谓的安全基线配置就是系统的最基础的安全配置，类比木桶原理的那块最短的木板，安全基线其实是系统最低安全要...', '简介：PHPOK企业站系统（以下简称系统或本系统），采用PHP+MYSQL语言开发，是一套成熟完善的企业站...', '讲百遍不如打一遍，网络安全的本质是攻防的对抗。11月16日至17日，首届“天府杯”国际网络安全大赛将在成都天府新区西博城举办。在为期两天的活动中，精彩激烈的...', '当一颗恒星病变老去，黑洞就会出现，时间便如吹熄火苗般湮灭，所有的世界被吸收、被碾压直至崩溃、不复存在。这是他们的狂欢时刻，暗星已经崛起，若你不战，银河...', 'i春秋作家、SRC英雄榜前列“鱼蛋”师父的直播课录屏+PPT...', 'Bluecms是一个地方网站经常使用的开源的cms，在很多地方性的网站上应用还是不少的，今天在逛seebug的时候看到了一个漏洞的公告。有公告但是这里还没有详情，很好...', '学历和技术哪个更重要这个问题仿佛困扰了我们很多年，尤其是当“鱼与熊掌不可得兼”的时候，这种纠结就更加明显，其实在网络安全行业，从刚入门的学生到摸爬滚打...', '在使用Cobaltstrike的时候发现他在大型或者比较复杂的内网环境中，作为内网拓展以及红队工具使用时拓展能力有些不足，恰恰armitage可以补充这点，利用metasploit...'])

返回元祖的目标是为了方便后面的处理，毕竟这样的返回结果完全不能看好吗？后继的美化处理略。



后言

共勉。