

## 前言：

收藏了多年的csdn免费文章，忽然收费或者突然被作者删除了怎么办？

## 文章目录

前言：

### 1. 工具

1.1需要使用到的模块

1.2 需要安装的工具

### 2. 获得文章内容的html（去除无关内容）

2.1 打开浏览器右键检查进行分析

2.2 开始数据解析

2.3 组装html

### 3. 将获得的html转成pdf

3.1 检查工具是否成功安装

3.2 使用pdfkit模块

3.2.1 先进行简单的配置

3.2.2 开始转换

== 大功完成了！ ==

小福利

## 1. 工具

### 1.1需要使用到的模块

pdfkit, requests, parsel

### 1.2 需要安装的工具

wkhtmltox-0.12.5-1.msvc2015-win64.exe 工具

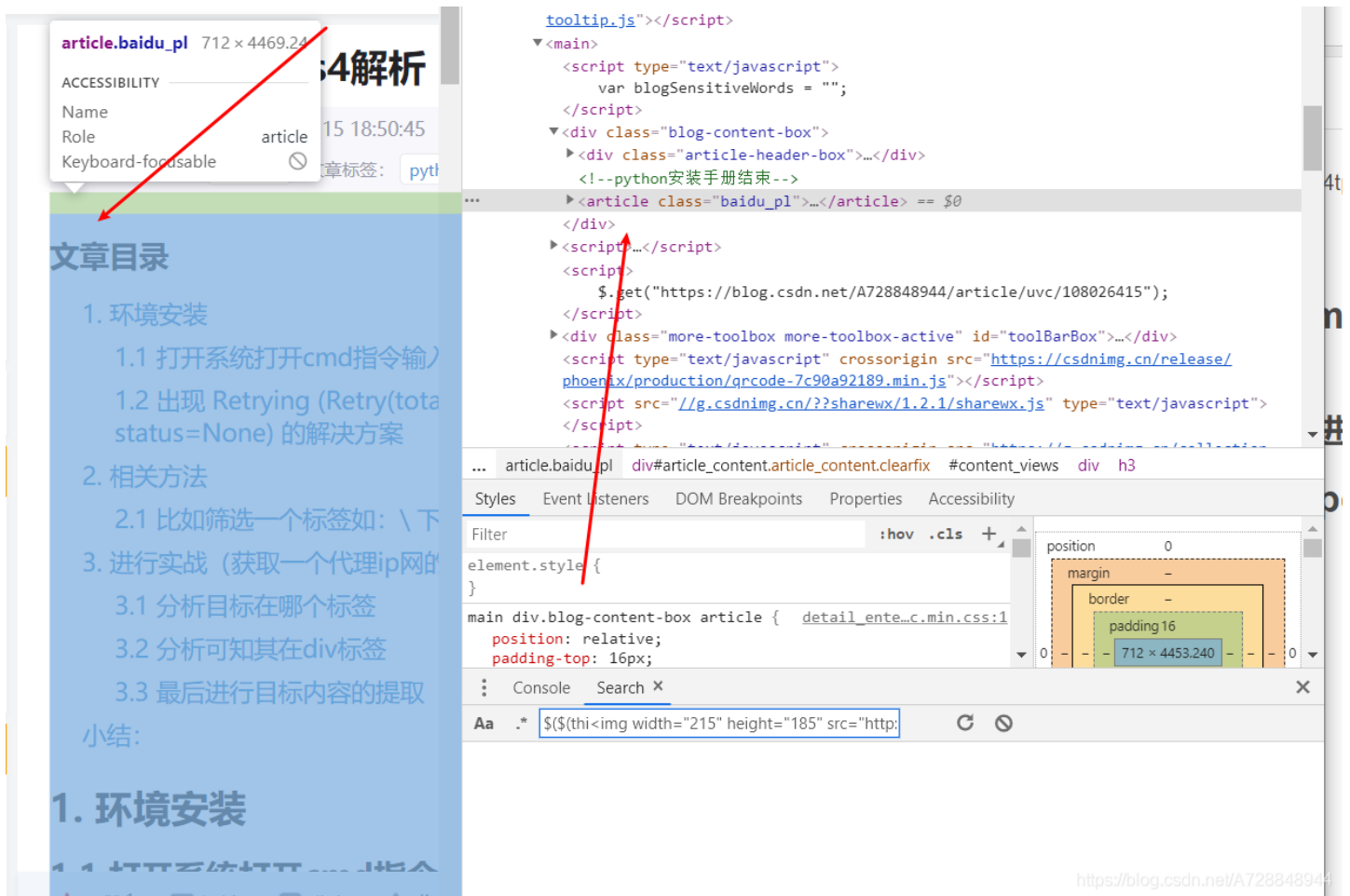
链接：[https://pan.baidu.com/s/1e\\_0\\_4tpyxlU8IHqJF56BhA](https://pan.baidu.com/s/1e_0_4tpyxlU8IHqJF56BhA)

提取码：2141

直接傻瓜式的默认安装即可

## 2. 获得文章内容的html（去除无关内容）

### 2.1 打开浏览器右键检查进行分析



通过简单的分析可以发现右边那箭头就是我们需要的内容，且都是与文章相关的，不相关的是没有的

## 2.2 开始数据解析

```
import requests
import parsel
url = 'https://blog.csdn.net/A728848944/article/details/108026415'
headers = {
    'user-agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/84.0.4147.125 Safari/537.36'
}
response = requests.get(url=url, headers=headers)
html = response.text
selector = parsel.Selector(html)
title = selector.css('.title-article::text').get()
article = selector.css('article').get() # 提取标签为article 的内容
print(article)
```

输出结果:

```
c_pdf x
C:\Users\安逸\AppData\Local\Programs\Python\Python38\python.exe C:\Users\安逸\IdeaProjects\python_workspace\csdn\c_pdf.py
<article class="baidu_pl">
  <div id="article_content" class="article_content clearfix">
    <link rel="stylesheet" href="https://csdnimg.cn/release/phoenix/template/css/ck-html5_views-963e387caa.css">
    <div id="content_views" class="markdown_views prism-atom-one-light">
      <!-- flowchart 箭头图标 勿删 -->
      <svg xmlns="http://www.w3.org/2000/svg" style="display: none;">
        <path stroke-linecap="round" d="M5,0 0,2.5 5,5z" id="raphael-marker-block" style="-webkit-tap-highlight-color: rgba(0, 0, 0, 0);"/>
      </svg>
    <p></p><div class="toc"><h3>文章目录</h3><ul><li><a href="#1__1" rel="nofollow">1. 环境安装</a></li></ul></div>
```

与上面的相互观察可指定数据提取是对的

## 2.3 组装html

由于提取出来的html是不全的，所以需要补充  
如下是一个标准的html结构

```
<!DOCTYPE html>
<html>
<head>
    <meta charset="UTF-8">
    <title>Document</title>

</head>
<body>
    相关内容
</body>
</html>
```

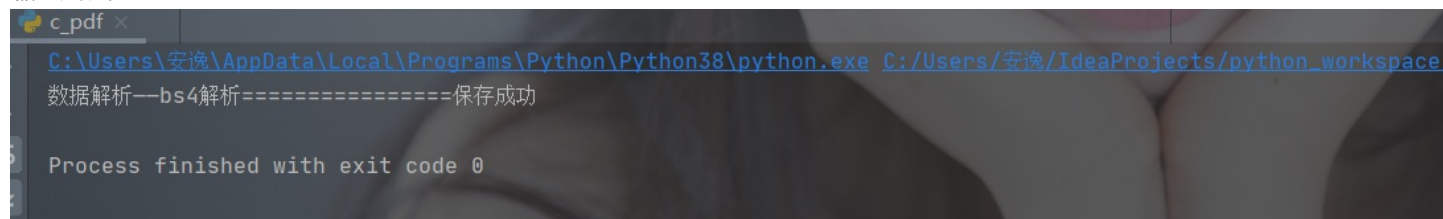
开始拼接，并保存下来

```
src_html = '''
<!DOCTYPE html>
<html>
<head>
    <meta charset="UTF-8">
    <title>Document</title>

</head>
<body>
    {content}
</body>
</html>
'''

with open(title+'.html', mode='w+', encoding='utf-8') as f:
    f.write(src_html.format(content=article))
    print(title+'=====保存成功')
```

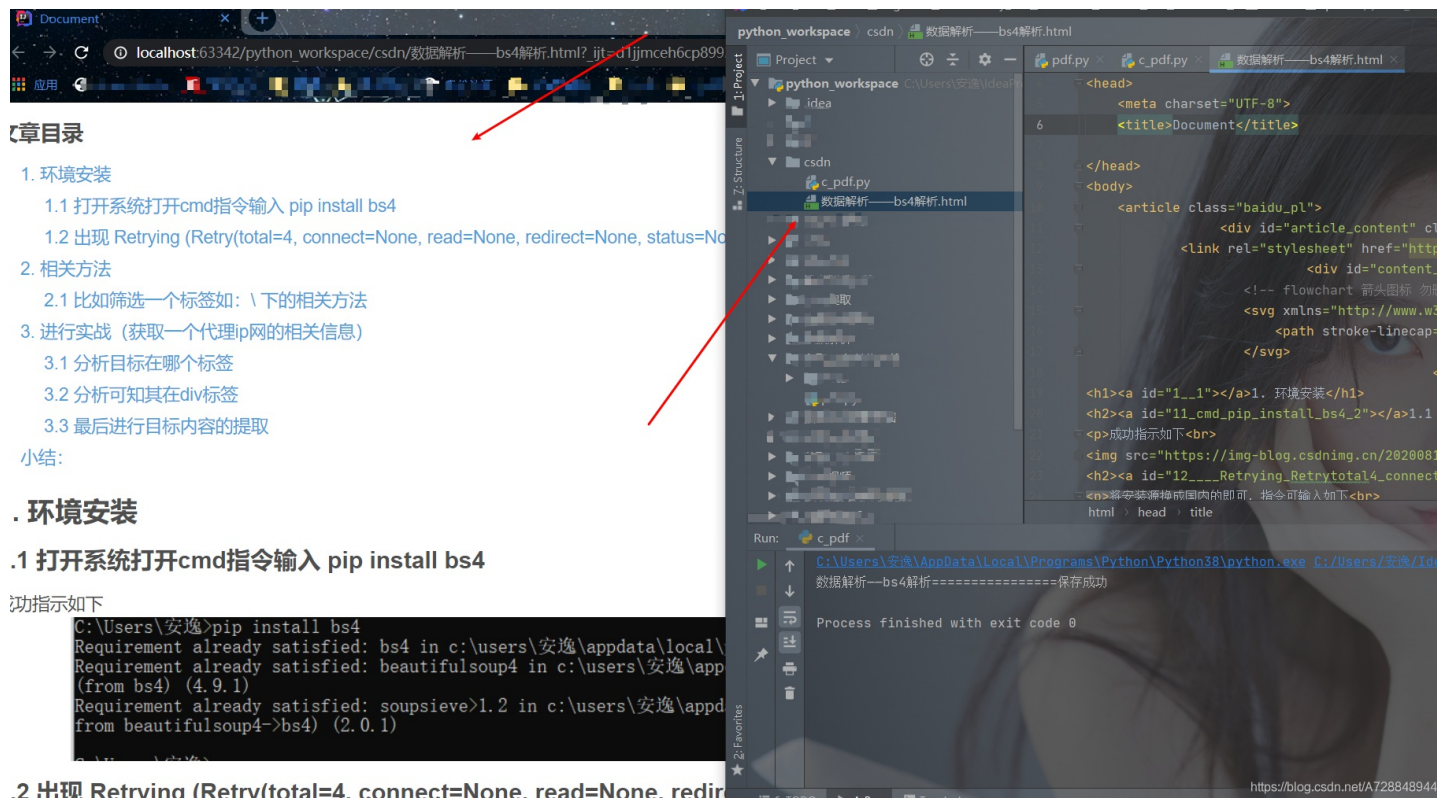
输出结果:



```
c_pdf x
C:\Users\安逸\AppData\Local\Programs\Python\Python38\python.exe C:/Users/安逸/IdeaProjects/python_workspace
数据解析--bs4解析=====保存成功

Process finished with exit code 0
```

看保存出来的html

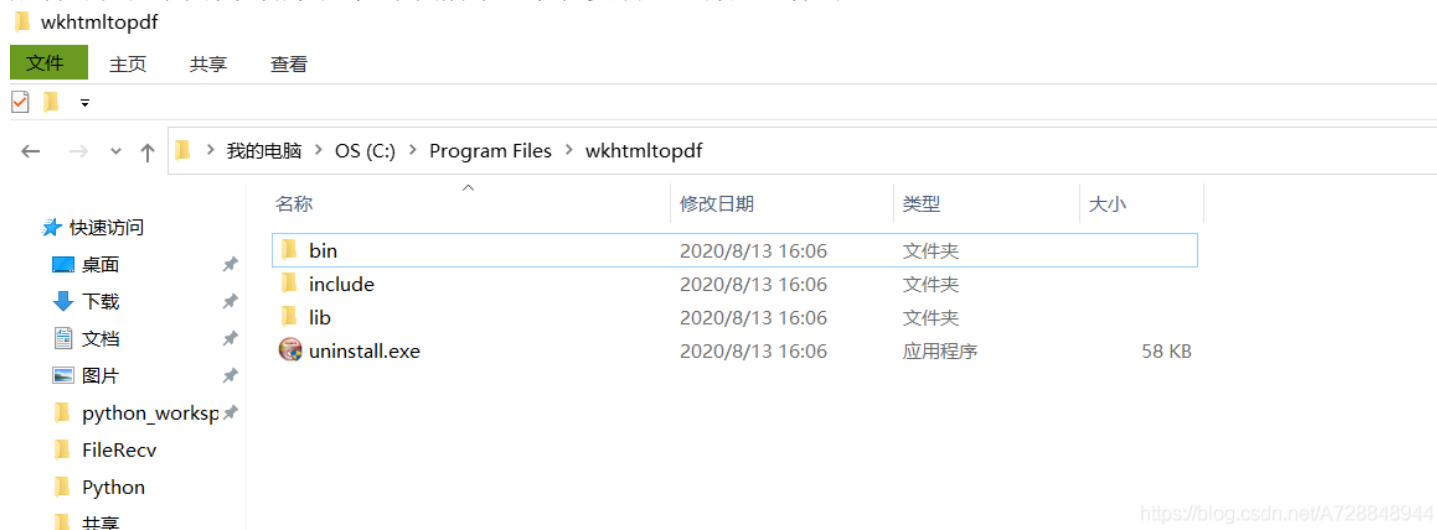


是不是相当简洁, 没有任何多余的

## 3. 将获得的html转成pdf

### 3.1 检查工具是否成功安装

成功可以在下图找到相关包如下图所示 (默认安装地址都是一样的)



### 3.2 使用pdfkit模块

#### 3.2.1 先进行简单的配置

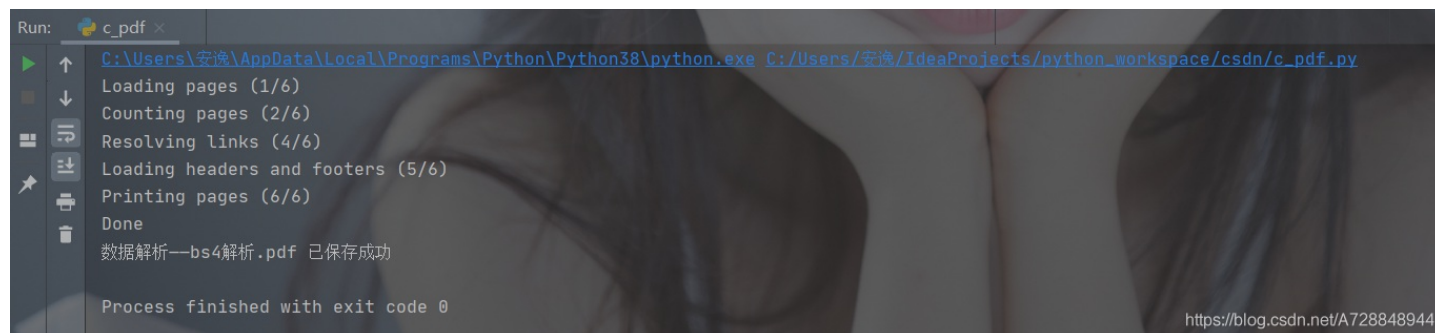
```
config = pdfkit.configuration(wkhtmltopdf=r'C:\Program Files\wkhtmltopdf\bin\wkhtmltopdf.exe')
```

这就是上面路径下bin的wkhtmltopdf.exe一个软件

#### 3.2.2 开始转换

```
pdfkit.from_file(title+'.html', title+'.pdf', configuration=config)
print(title+'.pdf', '已保存成功')
```

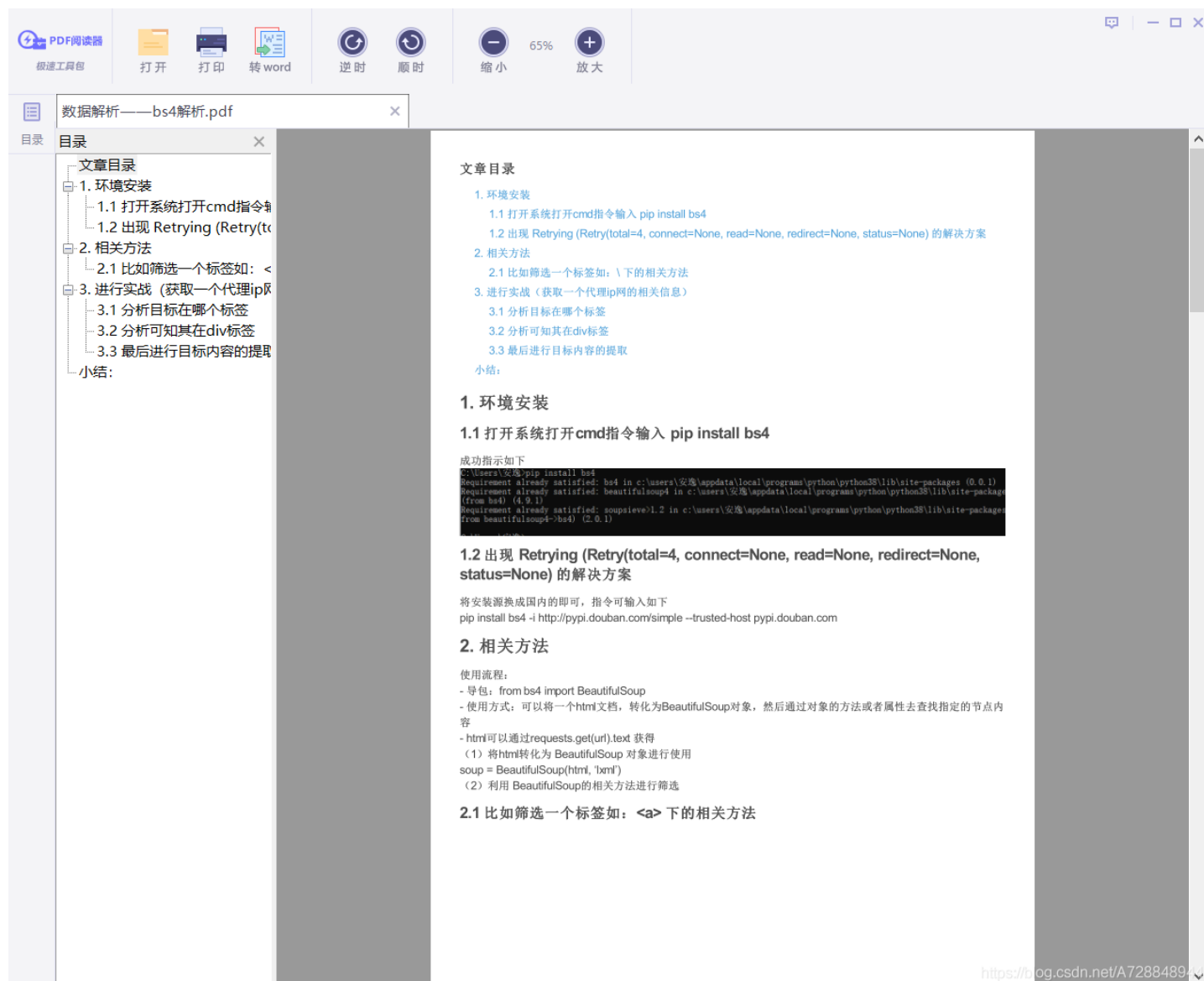
输出如下：



观察是否成功打印 如下图所示：



观察是否可用：



== 大功完成了！ ==

最后当然是全部代码拉

```
import pdfkit
import requests
import parsel

url = 'https://blog.csdn.net/A728848944/article/details/108026415'
headers = {
    'user-agent' : 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/84.0.4147.125 Safari/537.36'
}

response = requests.get(url=url, headers=headers)
html = response.text
selector = parsel.Selector(html)
title = selector.css('.title-article::text').get()
article = selector.css('article').get() # 提取标签为article 的内容

src_html = '''
<!DOCTYPE html>
<html>
<head>
    <meta charset="UTF-8">
    <title>Document</title>

</head>
<body>
    {content}
</body>
</html>
'''

with open(title+'.html', mode='w+', encoding='utf-8') as f:
    f.write(src_html.format(content=article))
    print(title+'=====保存成功')

config = pdfkit.configuration(wkhtmltopdf=r'C:\Program Files\wkhtmltopdf\bin\wkhtmltopdf.exe')
pdfkit.from_file(title+'.html', title+'.pdf', configuration=config)

print(title+'.pdf', '已保存成功')
```

## 小福利

获取一个博主所有文章pdf代码

```

import pdfkit
import requests
import parsel
import os
import time

src_html = '''
<!DOCTYPE html>
<html>
<head>
    <meta charset="UTF-8">
    <title>Document</title>

</head>
<body>
    {content}
</body>
</html>
'''

headers = {
    'user-agent' : 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/84.0.4147.125 Safari/537.36'
}

def download_one_page(page_url):
    response = requests.get(url=page_url, headers=headers)
    html = response.text
    selector = parsel.Selector(html)
    title = selector.css('.title-article::text').get()
    article = selector.css('article').get() # 提取标签为article 的内容

    with open(title+'.html', mode='w+', encoding='utf-8') as f:
        f.write(src_html.format(content=article))

    config = pdfkit.configuration(wkhtmltopdf=r'C:\Program Files\wkhtmltopdf\bin\wkhtmltopdf.exe')
    pdfkit.from_file(title+'.html', title+'.pdf', configuration=config)
    print(title+'.pdf', '=====已保存成功')

def down_all_url(index_url):
    index_response = requests.get(url=index_url, headers=headers)
    index_selector = parsel.Selector(index_response.text)
    urls = index_selector.css('.article-list h4 a::attr(href)').getall()
    for url in urls:
        download_one_page(url)
        time.sleep(2)

if __name__ == '__main__':
    down_all_url('https://blog.csdn.net/A728848944')

```