概述

本章就要就urllib中parse、request模块的重点API进行说明，也是以后大家最常用到的API。

- 本文不会列举所有的API。
- 本文以实例方式演示说明API，即直接上代码

实例

- 先看一个基本的实例：

```
#-*- coding:utf-8 -*-


__author__ = "苦叶子"


import urllib.parse
import urllib.request


if __name__ == "__main__":
    print("urllib API实例演示说明")

    # 访问百度首页
    response=urllib.request.urlopen('http://www.baidu.com')

    # 打印下首页是html源码
    # 获取完整的响应内容，便于断言其中的特定值
    html=response.read()
    print(html)

    # 打印下http header信息
    # 有时候我们需要提前header值来用于下一个请求
    header = response.info()
    print(header)

    # 获取下状态码 http响应的status code
    # 接口测试的一个断言，就是断言状态码
    status_code = response.getcode()
    print(status_code)

    # 打印下本次请求的目标url
    url = response.geturl()
    print(url)
```

- 下面我们基本的爬虫实例

我们尝试爬取下博客园首页的一些链接。

注意：需要用到前基础篇html.parser模块相关是技术

```
#-*- coding:utf-8 -*-


__author__ = "苦叶子"


import urllib.parse
import urllib.request
from html.parser import HTMLParser


class BlogHTMLParser(HTMLParser):
    data = []
    data_key = ""

    def __init__(self):
        HTMLParser.__init__(self)
        self.is_a = False

    def handle_starttag(self, tag, attrs):
        # 处理开始为a的标签
        if tag == "a":
            self.is_a = True
            for name,value in attrs:
                if name == "href":
                    # 提取a的href属性值
                    self.data_key = value

    def handle_data(self, data):
        # 处理结束为a的标签
        if self.is_a and self.lasttag == "a":
            # 将a标签的href属性值作为key， a的文本作为data构建字典
            self.data.append({self.data_key : data})

    def handle_endtag(self, tag):
        # 处理a结束标签
        if self.is_a and self.lasttag == "a":
            self.is_a = False

    def get_data(self):
        # 返回所有从a中提取到的目标数据
        return self.data


if __name__ == "__main__":
    print("urllib爬取博客园首页实例演示说明")
```

```
    url = "https://www.cnblogs.com/"

    # 访问首页
    response = urllib.request.urlopen(url)

    # 获取首页的html
    data = response.read().decode(encoding="utf-8")

    # 提取所有的链接
    blogHtmlParser = BlogHTMLParser()
    blogHtmlParser.feed(data)
    links = blogHtmlParser.get_data()
    print(links)
```

小结

在做爬虫的一些基础研究、学习时，建议能多多使用urllib，加深、加强对http的理解和掌握。

扫一扫关注微信公众号：