

# 实例1：中国大学排名定向爬虫

WS06

---



嵩天

[www.python123.org](http://www.python123.org)

# The Website is the API ...



## Requests

自动爬取HTML页面  
自动网络请求提交

## robots.txt

网络爬虫排除标准



## Beautiful Soup

解析HTML页面



## Projects

实战项目A



掌握定向网络数据爬取和网页解析的基本能力

# Python网络爬虫与信息提取

**python**  
弹指之间 · 享受创新

04X -Tian



# “中国大学排名定向爬虫”实例介绍

http://www.zuihaodaxue.cn/zuihaodaxuepaiming2016.html

2016中国最好大学排名. × +

zuihaodaxue.cn/zuihaodaxuepaiming2016.html

English | 订阅我们

网站首页

中国大学排名

世界大学排名

原创分析

要闻资讯

院校信息

联系我们

首页 / 中国大学排名 / 软科中国最好大学排名2016

软科中国最好大学排名2016

Powered by Scopus

2016

“中国最好大学排名”的排名范围是教育部公布的全国普通高等学校名单（2015年5月21日公布）中，1216所办学层次为本科的大学。这其中公办大学793所、民办大学140所、独立学院283所。  
[查看排名方法](#)

排名	学校名称	省市	总分	指标得分
				生源质量（新生高考成绩得分）
1	清华大学	北京市	95.9	100.0
2	北京大学	北京市	82.6	98.9
3	浙江大学	浙江省	80	88.8
4	上海交通大学	上海市	78.7	90.6
5	复旦大学	上海市	70.9	90.4

相关文章

中国最好大学排名的特点

中国最好大学排名-排名方法

2016中国最好大学排名

2016生源质量排名

2016科研规模排名

2016科研质量排名

2016顶尖成果排名

2016顶尖人才排名

2016科技服务排名

2016产学研合作排名

2016成果转化排名

报告下载

# 功能描述

输入：大学排名URL链接

输出：大学排名信息的屏幕输出（排名，大学名称，总分）

技术路线：requests-bs4

定向爬虫：仅对输入URL进行爬取，不扩展爬取

# 定向爬虫可行性

<http://www.zuihaodaxue.cn/zuihaodaxuepaiming2016.html>

```
<tbody class="hidden_zhpm" style="text-align:center;">
  <tr class="alt">
    <td>1</td><td><div align="left">清华大学</div></td><td>北京市</td><td>95.9</td><td class="hidden-xs need-hidden indicator5">100.0</td><td class="hidden-xs need-hidden indicator6" style="display:none;">97.90%</td><td class="hidden-xs need-hidden indicator7" style="display:none;">37342</td><td class="hidden-xs need-hidden indicator8" style="display:none;">1.298</td><td class="hidden-xs need-hidden indicator9" style="display:none;">1177</td><td class="hidden-xs need-hidden indicator10" style="display:none;">109</td><td class="hidden-xs need-hidden indicator11" style="display:none;">1137711</td><td class="hidden-xs need-hidden indicator12" style="display:none;">1187</td><td class="hidden-xs need-hidden indicator13" style="display:none;">593522</td>
  </tr>
  <tr>
    <td>2</td><td><div align="left">北京大学</div></td><td>北京市</td><td>82.6</td><td class="hidden-xs need-hidden indicator5">98.9</td><td class="hidden-xs need-hidden indicator6" style="display:none;">95.96%</td><td class="hidden-xs need-hidden indicator7" style="display:none;">36137</td><td class="hidden-xs need-hidden indicator8" style="display:none;">1.294</td><td class="hidden-xs need-hidden indicator9" style="display:none;">986</td><td class="hidden-xs need-hidden indicator10" style="display:none;">87</td><td class="hidden-xs need-hidden indicator11" style="display:none;">439403</td><td class="hidden-xs need-hidden indicator12" style="display:none;">799</td><td class="hidden-xs need-hidden indicator13" style="display:none;">7343</td>
  </tr>
  <tr class="alt">
    <td>3</td><td><div align="left">浙江大学</div></td><td>浙江省</td><td>80</td><td class="hidden-xs need-hidden indicator5">88.8</td><td class="hidden-xs need-hidden indicator6" style="display:none;">96.46%</td><td class="hidden-xs need-hidden indicator7" style="display:none;">41188</td><td class="hidden-xs need-hidden indicator8" style="display:none;">1.059</td><td class="hidden-xs need-hidden indicator9" style="display:none;">803</td><td class="hidden-xs need-hidden indicator10" style="display:none;">86</td><td class="hidden-xs need-hidden indicator11" style="display:none;">959511</td><td class="hidden-xs need-hidden indicator12" style="display:none;">833</td><td class="hidden-xs need-hidden indicator13" style="display:none;">64392</td>
  </tr>
```

# 定向爬虫可行性

<http://www.zuihaodaxue.cn/robots.txt>

# 功能描述

排名	学校名称	总分
1	清华大学	95.9
2	北京大学	82.6
3	浙江大学	80
4	上海交通大学	78.7
5	复旦大学	70.9
6	南京大学	66.1
7	中国科学技术大学	65.5
8	哈尔滨工业大学	63.5
9	华中科技大学	62.9
10	中山大学	62.1
11	东南大学	61.4
12	天津大学	60.8
13	同济大学	59.8
14	北京航空航天大学	59.6
15	四川大学	59.4
16	武汉大学	59.1



# 程序的结构设计

排名	学校名称	省市	总分	推荐得分
				生源质量（新生高考成绩得分）
1	清华大学	北京市	95.9	100.0
2	北京大学	北京市	82.6	98.9
3	浙江大学	浙江省	80	88.8
4	上海交通大学	上海市	78.7	90.6
5	复旦大学	上海市	70.9	90.4
6	南京大学	江苏省	66.1	90.7
7	中国科学技术大学	安徽省	65.5	90.1
8	哈尔滨工业大学	黑龙江省	63.5	80.9
9	华中科技大学	湖北省	62.9	83.5
10	中山大学	广东省	62.1	81.8

二维数据

步骤1：从网络上获取大学排名网页内容

步骤2：提取网页内容中信息到合适的数据结构

步骤3：利用数据结构展示并输出结果

# 程序的结构设计

- 步骤1：从网络上获取大学排名网页内容 `getHTMLText()`
- 步骤2：提取网页内容中信息到合适的数据结构 `fillUnivList()`
- 步骤3：利用数据结构展示并输出结果 `printUnivList()`

# “中国大学排名定向爬虫”实例编写



# main()

```
import requests
from bs4 import BeautifulSoup

def getHTMLText(url):
    return ""

def fillUnivList(ulist, html):
    pass

def printUnivList(ulist, num):
    print("Suc" + str(num))

def main():
    uinfo = []
    url = 'http://www.zuihaodaxue.cn/zuihaodaxuepaiming2016.html'
    html = getHTMLText(url)
    fillUnivList(uinfo, html)
    printUnivList(uinfo, 20) # 20 univs
main()
```

# getHTMLText()

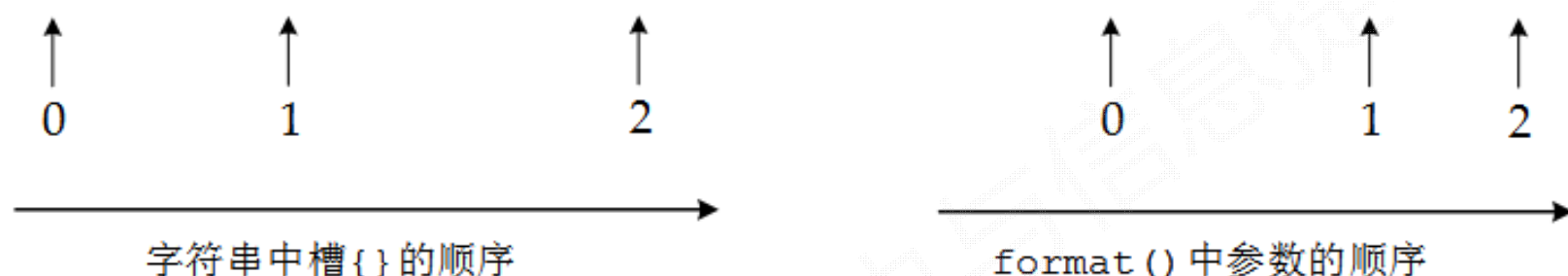
```
def getHTMLText(url):  
    try:  
        r = requests.get(url, timeout=30)  
        r.raise_for_status()  
        r.encoding = r.apparent_encoding  
        return r.text  
    except:  
        return ""
```

# fillUnivList()

```
import bs4
```

```
def fillUnivList(ulist, html):  
    soup = BeautifulSoup(html, "html.parser")  
    for tr in soup.find('tbody').children:  
        if isinstance(tr, bs4.element.Tag):  
            tds = tr('td')  
            ulist.append([tds[0].string, tds[1].string, tds[3].string])
```

"{ }：计算机{ }的CPU占用率为{ }%。".format("2016-12-31","PYTHON",10)



:	<填充>	<对齐>	<宽度>	,	<.精度>	<类型>
引导 符号	用于填充的 单个字符	< 左对齐 > 右对齐 ^ 居中对齐	槽的设定输 出宽度	数字的千位 分隔符 适用于整数 和浮点数	浮点数小数 部分的精度 或 字符串的最 大输出长度	整数类型 b, c, d, o, x, X 浮点数类型 e, E, f, %

# printUnivList()

```
def printUnivList(ulist, num):  
    print("{:^10}\t{:^6}\t{:^10}".format("排名", "学校名称", "总分"))  
    for i in range(num):  
        u=ulist[i]  
        print("{:^10}\t{:^6}\t{:^10}".format(u[0], u[1], u[2]))
```



# 全代码

请阅读全代码

Python网络爬虫与信息提取

# “中国大学排名定向爬虫”实例优化



# 输出结果的中文对齐问题

排名	学校名称	总分
1	清华大学	95.9
2	北京大学	82.6
3	浙江大学	80
4	上海交通大学	78.7
5	复旦大学	70.9
6	南京大学	66.1
7	中国科学技术大学	65.5
8	哈尔滨工业大学	63.5
9	华中科技大学	62.9
10	中山大学	62.1
11	东南大学	61.4
12	天津大学	60.8
13	同济大学	59.8
14	北京航空航天大学	59.6
15	四川大学	59.4
16	武汉大学	59.1

# printUnivList()

```
def printUnivList(ulist, num):  
    print("{:^10}\t{:^6}\t{:^10}".format("排名", "学校名称", "总分"))  
    for i in range(num):  
        u=ulist[i]  
        print("{:^10}\t{:^6}\t{:^10}".format(u[0], u[1], u[2]))
```

# 中文对齐问题的原因

:	<填充>	<对齐>	<宽度>	,	<.精度>	<类型>
引导 符号	用于填充的 单个字符	< 左对齐 > 右对齐 ^ 居中对齐	槽的设定输出 宽度	数字的千位 分隔符 适用于整数 和浮点数	浮点数小数 部分的精度 或 字符串的最 大输出长度	整数类型 b, c, d, o, x, X 浮点数类型 e, E, f, %

当中文字符宽度不够时，采用西文字符填充；中西文字符占用宽度不同

# 中文对齐问题的解决

当中文字符宽度不够时，采用西文字符填充；中西文字符占用宽度不同

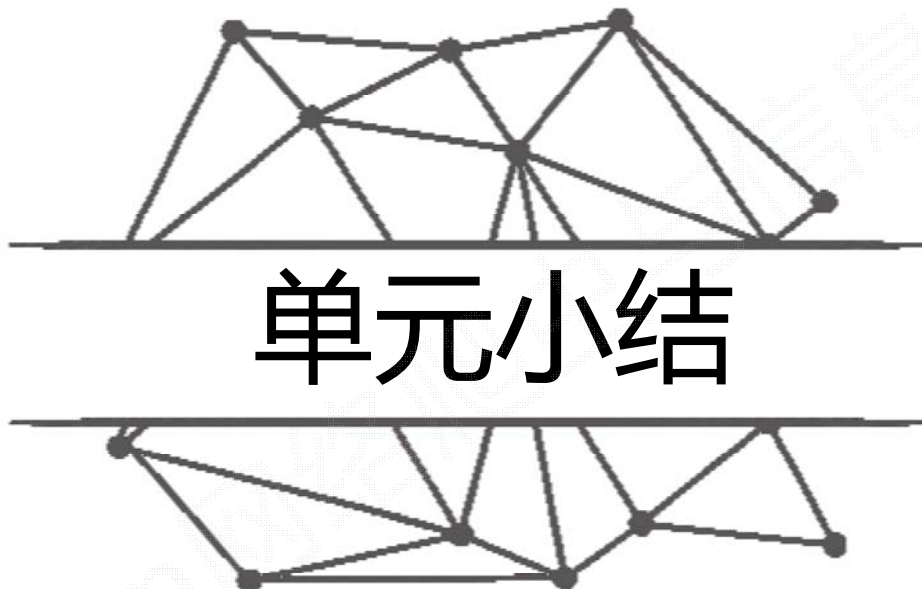
采用中文字符的空格填充 `chr(12288)`

# printUnivList()

```
def printUnivList(ulist, num):  
    tplt = "{0:^10}\t{1:{3}^10}\t{2:^10}"  
    print(tplt.format("排名", "学校名称", "总分", chr(12288)))  
    for i in range(num):  
        u=ulist[i]  
        print(tplt.format(u[0], u[1], u[2], chr(12288)))
```

排名	学校名称	总分
1	清华大学	95.9
2	北京大学	82.6
3	浙江大学	80
4	上海交通大学	78.7
5	复旦大学	70.9
6	南京大学	66.1
7	中国科学技术大学	65.5
8	哈尔滨工业大学	63.5
9	华中科技大学	62.9
10	中山大学	62.1
11	东南大学	61.4
12	天津大学	60.8
13	同济大学	59.8
14	北京航空航天大学	59.6
15	四川大学	59.4
16	武汉大学	59.1
17	西安交通大学	58.9



The diagram consists of two symmetrical, interconnected node-link structures. The top structure is positioned above a horizontal line, and the bottom structure is positioned below it. Each structure features a series of nodes (represented by small black dots) connected by straight lines (edges). The connections form a complex, web-like pattern that resembles a stylized, abstract representation of a network or a biological structure. The overall layout is centered on the page, with the text '单元小结' (Unit Summary) acting as a focal point between the two graphical elements.

## 单元小结

# 实例1：中国大学排名定向爬虫

采用requests-bs4路线实现了中国大学排名定向爬虫

对中英文混排输出问题进行优化