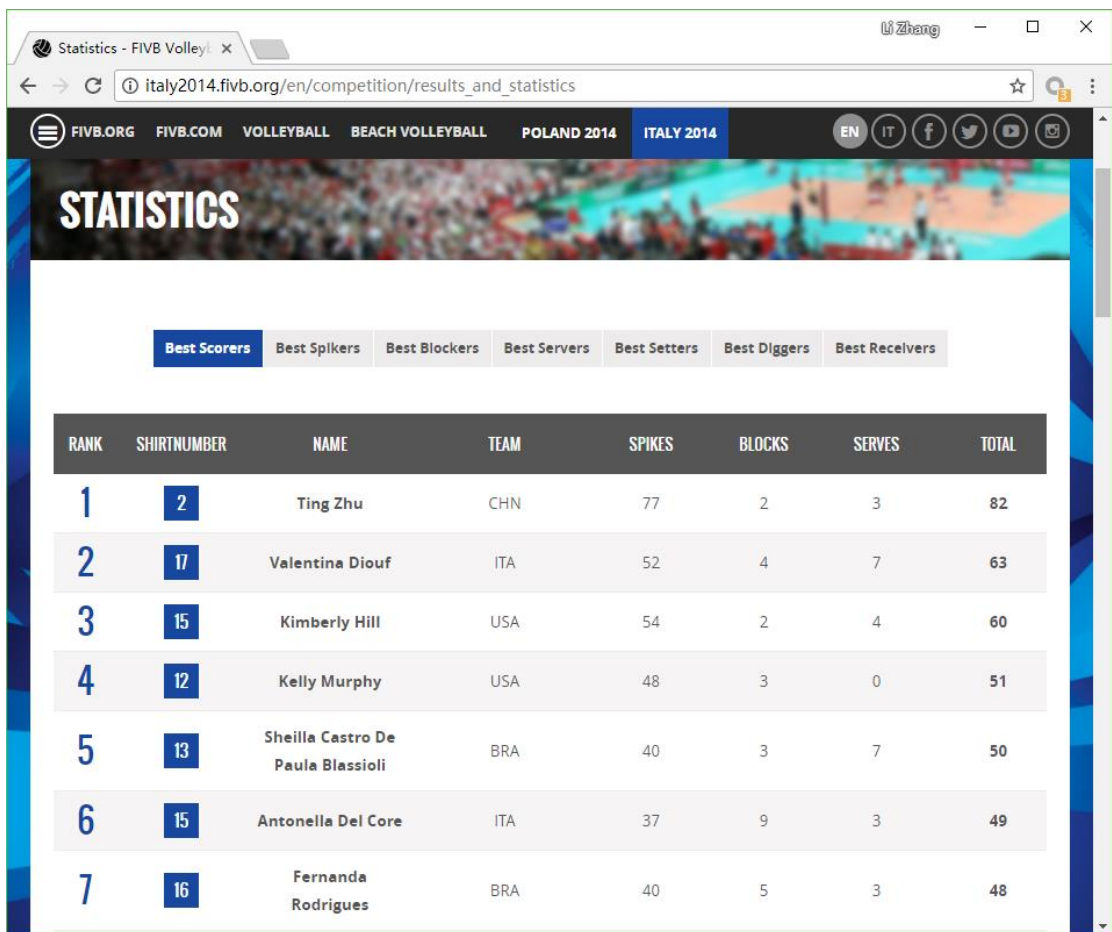


## 《用 Python 玩转数据》爬虫小项目（3 项）

1. “迷你爬虫编程小练习”进阶：抽取某本书的前 50 条短评内容并计算评分的平均值。
2. 在 “<http://money.cnn.com/data/dow30/>” 上抓取道指成分股数据并将 30 家公司的代码、公司名称和最近一次成交价放到一个列表中输出。
3. 请爬取网页 ([http://italy2014.fivb.org/en/competition/results\\_and\\_statistics](http://italy2014.fivb.org/en/competition/results_and_statistics)) 上的数据



The screenshot shows a web browser window displaying the FIVB Volleyball Italy 2014 statistics page. The page has a dark header with navigation links for FIVB.ORG, FIVB.COM, VOLLEYBALL, BEACH VOLLEYBALL, POLAND 2014, and ITALY 2014. Below the header, there's a large banner with the word "STATISTICS" in white. Underneath the banner, there are several tabs: "Best Scorers", "Best Spikers", "Best Blockers", "Best Servers", "Best Setters", "Best Diggers", and "Best Receivers". The "Best Scorers" tab is selected. Below the tabs is a table with the following columns: RANK, SHIRTNUMBER, NAME, TEAM, SPIKES, BLOCKS, SERVES, and TOTAL. The table lists the top 7 scorers.

RANK	SHIRTNUMBER	NAME	TEAM	SPIKES	BLOCKS	SERVES	TOTAL
1	2	Ting Zhu	CHN	77	2	3	82
2	17	Valentina Diouf	ITA	52	4	7	63
3	15	Kimberly Hill	USA	54	2	4	60
4	12	Kelly Murphy	USA	48	3	0	51
5	13	Sheilla Castro De Paula Blassioli	BRA	40	3	7	50
6	15	Antonella Del Core	ITA	37	9	3	49
7	16	Fernanda Rodrigues	BRA	40	5	3	48

参考程序见下一页

【参考代码：将 url 中的 **bookid** 换成自己想查看的书的 id】

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

```
"""
```

```
Comments parsing
```

```
@author: Dazhuang
```

```
"""
```

```
import requests, re, time
```

```
from bs4 import BeautifulSoup
```

```
count = 0
```

```
i = 0
```

```
sum, count_s = 0, 0
```

```
while(count < 50):
```

```
    try:
```

```
        r = requests.get('https://book.douban.com/subject/bookid/comments/hot?p=' + str(i+1))
```

```
    except Exception as err:
```

```
        print(err)
```

```
        break
```

```
    soup = BeautifulSoup(r.text, 'lxml')
```

```
    comments = soup.find_all('p', 'comment-content')
```

```
    for item in comments:
```

```
        count = count + 1
```

```
        print(count, item.string)
```

```
    if count == 50:
```

```

        break

pattern = re.compile('<span class="user-stars allstar(.*) rating"')

p = re.findall(pattern, r.text)

for star in p:

    count_s = count_s + 1

    sum += int(star)

time.sleep(5)    # delay request from douban's robots.txt

i += 1

if count == 50:

    print(sum / count_s)

```

#### 【参考代码】

```

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

"""

Get dji stock data

@author: Dazhuang

"""

import requests

import re

def retrieve_dji_list():

    r = requests.get('http://money.cnn.com/data/dow30/')

```

```

search_pattern =
re.compile('class="wsod_symbol">(.*?)</a>.*?<span.*?>(.*?)</span>.*?

\n.*?class="wsod_stream">(.*?)</span>')

dji_list_in_text = re.findall(search_pattern, r.text)

return dji_list_in_text

```

```

dji_list = retrieve_dji_list()

print(dji_list)

```

### 【参考代码】

提示：由于包含信息的源代码分在 3 行，所以在处理时要使用多行模式（用 flags=re.M 表示），并且要把换行时的空白字符表示出来（用\s+可表示多个空白字符）。

```

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

"""

Crawler

@author: Dazhuang

"""

import requests

import re

r = requests.get('http://italy2014.fivb.org/en/competition/results_and_statistics')

r.encoding = r.apparent_encoding

pattern = re.compile('<td><a

id="wcbbody_o_wcgridpade50e7ca82ec64ee2b91ea4cc6c4e00c6_1_PlayerStatisticsTabl

e_BestScorers_Name_*?>'

```

```
href="/en/competition/teams/.*/players/.*/id=.*?">(.*?)</a></td>|s+<td
id="wcbbody_o_wcgridpade50e7ca82ec64ee2b91ea4cc6c4e00c6_1_PlayerStatisticsTabl
e_BestScorers_TeamCell_.*?"><a
id="wcbbody_o_wcgridpade50e7ca82ec64ee2b91ea4cc6c4e00c6_1_PlayerStatisticsTabl
e_BestScorers_Team_.*?"
href="/en/competition/teams/.*/">(.*?)</a></td>|s+<td>(.*?)</td>|s+<td>(.*?)</td>|s+
<td>(.*?)</td>|s+<td>(.*?)</td>',flags=re.M)
```

```
p = re.findall(pattern, r.text)
```

```
print(p)
```