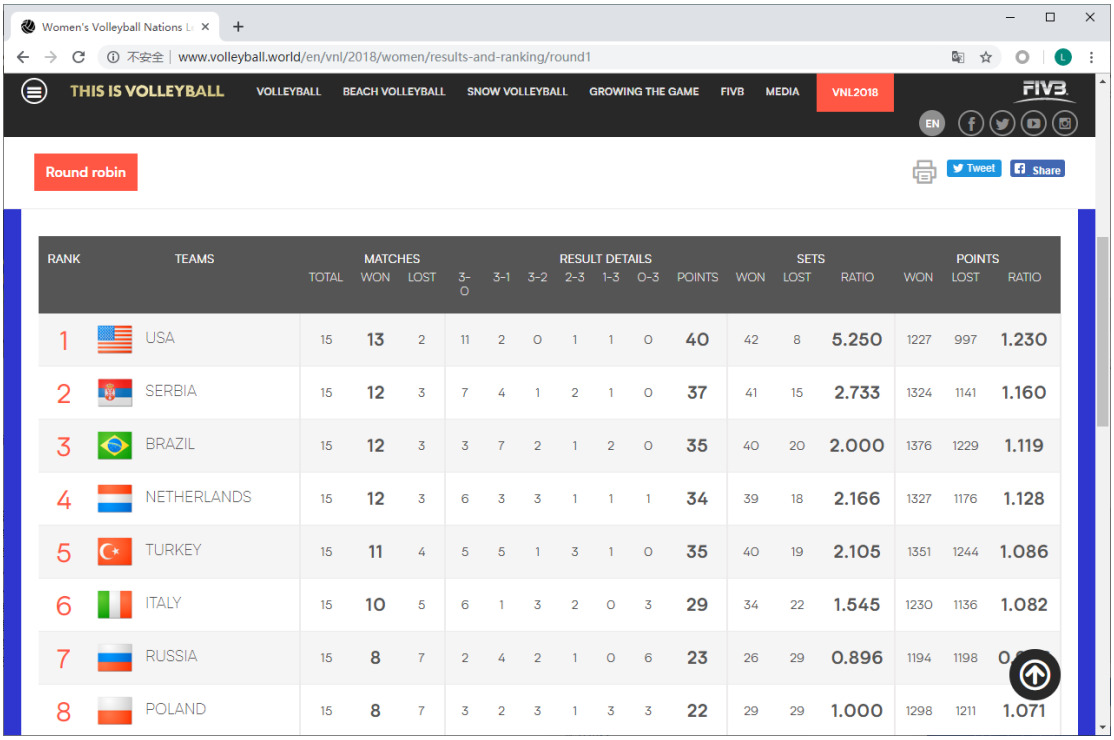
《用 Python 玩转数据》爬虫小项目（3 项）

Dazhuang@NJU

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(star)的平均值。提示：有的评论中并不包含评分。

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://www.volleyball.world/en/vnl/2018/women/results-and-ranking/round1）上的数据（包括 TEAMS and TOTAL, WON, LOST of MATCHES）



RANK	TEAMS	MATCHES TOTAL	MATCHES WON	MATCHES LOST	RESULT DETAILS 3-0	3-1	3-2	2-3	1-3	0-3	POINTS	SETS WON	SETS LOST	SETS RATIO	POINTS WON	POINTS LOST	POINTS RATIO
1	USA	15	13	2	11	2	0	1	1	0	40	42	8	5.250	1227	997	1.230
2	SERBIA	15	12	3	7	4	1	2	1	0	37	41	15	2.733	1324	1141	1.160
3	BRAZIL	15	12	3	3	7	2	1	2	0	35	40	20	2.000	1376	1229	1.119
4	NETHERLANDS	15	12	3	6	3	3	1	1	1	34	39	18	2.166	1327	1176	1.128
5	TURKEY	15	11	4	5	5	1	3	1	0	35	40	19	2.105	1351	1244	1.086
6	ITALY	15	10	5	6	1	3	2	0	3	29	34	22	1.545	1230	1136	1.082
7	RUSSIA	15	8	7	2	4	2	1	0	6	23	26	29	0.896	1194	1198	0.9
8	POLAND	15	8	7	3	2	3	1	3	3	22	29	29	1.000	1298	1211	1.071

提示：在处理时可以用已学的方法将每一项需要的内容（如 USA 和 15）单独解析出来，但这种做法将有联系的数据打散了，较好的做法是将每个 TEAM 的相关数据按组解析出来。但是由于包含这 4 项信息的源代码（请自行观察）分在多行并且行首有多个空格，因此在处理时在构造正则表达式时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，包括换行符和空格）。

【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，例如 *1084336*】

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

"""

Comments    parsing

@author: Dazhuang

"""

import requests, re, time

from bs4 import BeautifulSoup



count = 0

i = 0

s, count_s, count_del = 0, 0, 0

lst_stars = []

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('span', 'short')

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

    # Other way: we can use a whole regular expression to pattern comments and rangking
stars
```

```
        p = re.findall(pattern, r.text)
        for item in comments:
            count += 1
            if count > 50:
                # count the number of comments more than 50 of the page
                count_del += 1
            else:
                print(count, item.string)
        for star in p:
            lst_stars.append(int(star))
        time.sleep(5)           # delay request from douban's robots.txt
        i += 1
        for star in lst_stars[:-count_del]:       # calculate the rating star of 50 comments
            s += int(star)
if count >= 50:
        print(s // (len(lst_stars)-count_del))
```

【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/')

    # put the re expression on one line and pay attention to the '\n'

    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】

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

"""

Crawler

@author: Dazhuang

"""

import re

import requests

def crawler(url):

    try:
```

```
            r = requests.get(url)
        except requests.exceptions.RequestException as err:
            return err
        r.encoding = r.apparent_encoding
        # put the re expression on one line
        pattern =
    re.compile('href="/en/vnl/2018/women/teams/.*?">(.*?)</a></figcaption>\s+</figure>\
    s+</td>\s+<td>(.*?)</td>\s+<td class="table-td-bold">(.*?)</td>\s+<td class="table-td-
    rightborder">(.*?)</td>')
        p = re.findall(pattern, r.text)
        return p


    if __name__ == "__main__":
        url = 'http://www.volleyball.world/en/vnl/2018/women/results-and-
ranking/round1'
        result = crawler(url)
        print(result)
```