《用 Python 玩转数据》爬虫小项目（2项）

1. “3.2 爬虫小实验”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。

   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
        if count > 50:
            break
        print(count, item.string)
    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)
```