

上市公司数据采集 - 任务解答

requests + re 抓静态, selenium 抓动态, 最后下载现金流量表。

```
In [1]: import os
import re
import csv
import time
import requests
import chardet
from bs4 import BeautifulSoup
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.chrome.options import Options
from selenium.webdriver.chrome.service import Service
```

1. 正则获取静态数据

```
In [2]: url = 'https://finance.sina.com.cn/'
headers = {'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537
          '(KHTML, like Gecko) Chrome/64.0.3282.186 Safari/537.36'}

r = requests.get(url, headers=headers)
r.encoding = chardet.detect(r.content)['encoding']
html = r.text
```

```
In [3]: nums = re.findall(r'\d+\.\d*', html)
nums[:20]
```

```
Out[3]: ['2026',
        '05',
        '13',
        '13',
        '51',
        '00',
        '8',
        '7',
        '24',
        '000089',
        '1',
        '1',
        '5',
        '5',
        '4',
        '7',
        '6',
        '5',
        '7',
        '24']
```

```
In [4]: html2 = ''.join(re.findall(re.compile(r'id="wmt_tabs".*?>(.*?)</div>', re.S), html)
titles = re.findall(r'<a.*?>(.*?)</a>', html2)
titles
```

```
Out[4]: ['沪深京', '亚太', '欧洲', '美国', '商品', '货币']
```

2. selenium 获取动态数据

```
In [5]: options = Options()
options.add_argument('--headless=new')
service = Service(r'D:\mypython\myprojects\env\chromedriver-win64\chromedriver.exe')
browser = webdriver.Chrome(service=service, options=options)

browser.get('http://vip.stock.finance.sina.com.cn/corp/view/iframe/vAK_NewStockIssu
time.sleep(2)

element = browser.find_element(By.CLASS_NAME, 'wrap')
lst = [td.text for td in element.find_elements(By.TAG_NAME, 'th')]
rows = [lst[i:i + 4] for i in range(0, len(lst), 4)]
browser.quit()
rows[:6]
```

```
Out[5]: [['日期', '股票名称', '申购代码', '申购价格'],
['05-18', '长进光子', '787635', '--'],
['05-13', '嘉德利', '732435', '15.76'],
['05-13', '朗信电气', '920220', '28.29'],
['05-13', '惠康科技', '001237', '53.26'],
['05-07', '天海电子', '001365', '27.19']]
```

3. 下载现金流量表

```
In [6]: save_dir = r'D:\datasets'
os.makedirs(save_dir, exist_ok=True)

def download(stockid):
    u = f'https://money.finance.sina.com.cn/corp/go.php/vFD_CashFlow/stockid/{stockid}'
    resp = requests.get(u, headers=headers, timeout=30)
    resp.encoding = 'gb2312'
    table = BeautifulSoup(resp.text, 'html.parser').find('table', id='ProfitStateme
    data = [[c.get_text(strip=True) for c in tr.find_all(['td', 'th'])] for tr in t
    data = [r for r in data if r]
    company = data[0][0].split('(')[0]
    path = os.path.join(save_dir, f'{company}({stockid}) 现金流量表.csv')
    with open(path, 'w', newline='', encoding='utf-8-sig') as f:
        csv.writer(f).writerows(data)
    print('保存: ', path)

for code in ['600357', '600358', '600359']:
    download(code)
```

保存: D:\datasets\承德钒钛(600357) 现金流量表.csv

保存: D:\datasets\国旅联合(600358) 现金流量表.csv

保存： D:\datasets\新农开发(600359) 现金流量表.csv