



本节知识点概述

爬取淘宝商品数据

全能的Selenium

Selenium的环境搭建
Selenium的简单使用

实战环节

使用Selenium爬取淘宝
思路讲解+代码分享

说在最后

经验分享
回顾课程



DC学院
class.pkbigdata.com

造数



Selenium的环境搭建

1. 安装：在终端输入 `pip install selenium`
2. 下载：下载chromedriver，解压后放在...\Google\Chrome\Application\
3. 环境变量：将改目录添加至环境变量

```
%SystemRoot%\system32
%SystemRoot%
%SystemRoot%\System32\Wbem
%SYSTEMROOT%\System32\WindowsPowerShell\v1.0\
\软件\cmdr_mini
D:\nodejs\
C:\Program Files (x86)\Google\Chrome\Application
```





Selenium的简单使用

1. 导入Selenium包
2. 选择Chrome浏览器
3. 打开百度首页
4. 找到搜索框，输入关键词
5. 打印源码

```
# -*- coding:utf-8 -*-  
#!/usr/bin/env python  
  
from selenium import webdriver  
  
driver = webdriver.Chrome()  
driver.get("http://www.baidu.com")
```

```
# -*- coding:utf-8 -*-  
#!/usr/bin/env python  
  
from selenium import webdriver  
from selenium.webdriver.common.keys import Keys  
  
driver = webdriver.Chrome()  
driver.get("http://www.baidu.com")  
  
elem = driver.find_element_by_xpath('//*[@id="kw"]')  
elem.send_keys("Python selenium", Keys.ENTER)  
print(driver.page_source)
```





使用Selenium爬取淘宝

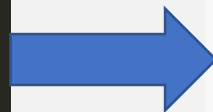
```
#!/usr/bin/env python
# coding:utf-8 -*-

from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.common.exceptions import TimeoutException
from pyquery import PyQuery as pq
from pymongo import MongoClient
import re

browser = webdriver.Chrome()
wait = WebDriverWait(browser, 10)

client = MongoClient()
db = client.taobao
data = db.data

def search(kd):
    try:
        browser.get('https://www.taobao.com/')
        input = wait.until(EC.presence_of_element_located((By.CSS_SELECTOR, "#q")))
        submit = wait.until(EC.element_to_be_clickable((By.CSS_SELECTOR, '#J_TSearchForm > div.search-button > button')))
        input.send_keys(kd)
        submit.click()
        total = wait.until(EC.presence_of_element_located((By.CSS_SELECTOR, '#main-srp-pager > div > div > div > div.total')))
        get_products()
        return total.text
    except TimeoutException:
        return search()
```



Robo 3T - 1.1

File View Options Window Help

New Connection (4)

- System
- taobao
 - Collections (1)
 - data
 - Functions
 - Users
 - test

db.getCollection('...')

New Connection localhost:27017 taobao

db.getCollection('data').find({})

data 0.002 sec.

Key	Value	Type
(1) ObjectId("59e6c162b385c725cc683...")	(7 fields)	Object
_id	ObjectId("59e6c162b385c725cc68333d")	ObjectId
image	//g-search1.alicdn.com/img/bao/uploaded/i4/...	String
price	¥ 188.00	String
deal	52	String
title	毛呢外套女中长款韩版2017新款秋冬款长袖学生宽...	String
shop	清扬时尚旗舰店	String
location	浙江 杭州	String
(2) ObjectId("59e6c162b385c725cc683...")	(7 fields)	Object
(3) ObjectId("59e6c162b385c725cc683...")	(7 fields)	Object
(4) ObjectId("59e6c162b385c725cc683...")	(7 fields)	Object
(5) ObjectId("59e6c162b385c725cc683...")	(7 fields)	Object
(6) ObjectId("59e6c162b385c725cc683...")	(7 fields)	Object
(7) ObjectId("59e6c162b385c725cc683...")	(7 fields)	Object
(8) ObjectId("59e6c162b385c725cc683...")	(7 fields)	Object
(9) ObjectId("59e6c162b385c725cc683...")	(7 fields)	Object
(10) ObjectId("59e6c162b385c725cc683...")	(7 fields)	Object
(11) ObjectId("59e6c162b385c725cc683...")	(7 fields)	Object
(12) ObjectId("59e6c162b385c725cc683...")	(7 fields)	Object
(13) ObjectId("59e6c162b385c725cc683...")	(7 fields)	Object
(14) ObjectId("59e6c162b385c725cc683...")	(7 fields)	Object
(15) ObjectId("59e6c162b385c725cc683...")	(7 fields)	Object
(16) ObjectId("59e6c162b385c725cc683...")	(7 fields)	Object
(17) ObjectId("59e6c162b385c725cc683...")	(7 fields)	Object
(18) ObjectId("59e6c162b385c725cc683...")	(7 fields)	Object
(19) ObjectId("59e6c162b385c725cc683...")	(7 fields)	Object

Logs

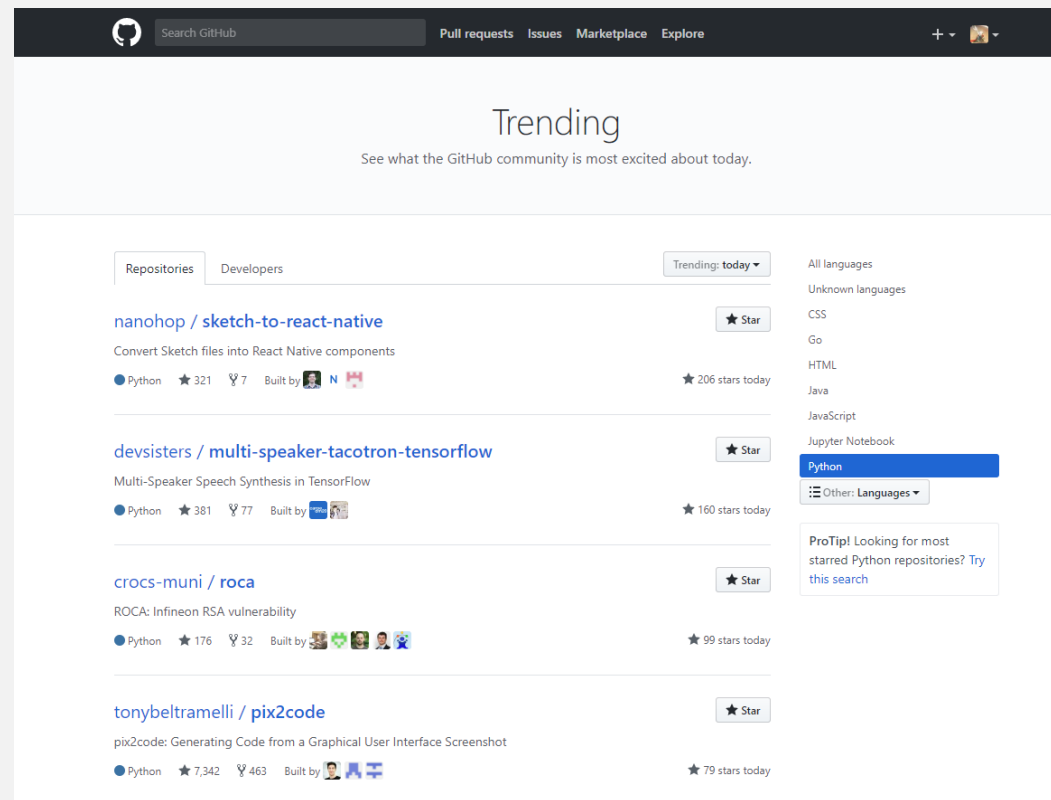


说在最后

常用的学习方法

经常浏览的网站

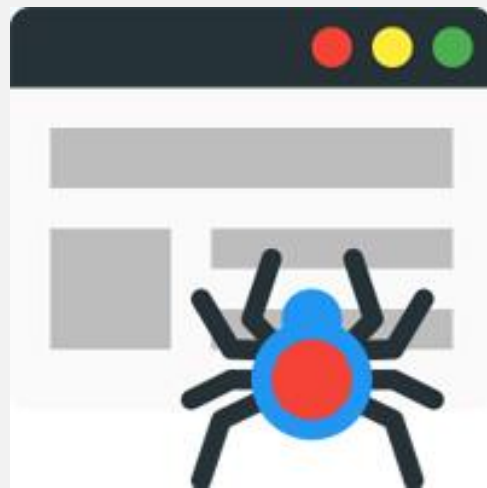
一年的学习经历





说在最后

- 第一课_什么是爬虫
- 第二课_初识Python爬虫
- 第三课_使用Requests爬取豆瓣短评
- 第四课_使用xpath解析豆瓣短评
- 第五课_使用pandas保存豆瓣短评数据
- 第六课_爬取知乎用户数据
- 第七课_爬取拉勾职位信息
- 第八课_爬取淘宝商品价格





完成作业

- 了解Selenium是什么
- 用Selenium去爬取更多网站
- 回顾入门课程，迎接进阶部分



DC学院
class.pkbigdata.com

造数

爬虫工程师

更多数据科学课程，上DC学院：class.pkbigdata.com



关注 DataCastle



关注造数



造数

