2023年春季学期 北京大学通选课

计算机科学与编程入门

第10讲 网络爬虫（3）

主讲教师：陆俊林 纪晓璐

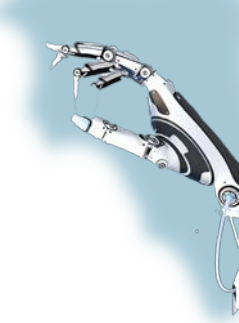
2023年4月24日

主要内容

- 01. requests获取网页内容
- 02. pyquery解析网页内容
- 03. bs4解析网页内容
- 04. requests+bs4解析网页实例

01 requests获取网页内容

requests-第三方库



pip install requests

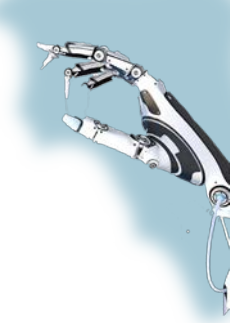
import requests

requests.get(url, params={ }, headers={ })

- 返回一个响应对象requests.Response
  - encoding属性可设置编码
  - text属性可获得网页文件
  - content属性可获得二进制信息

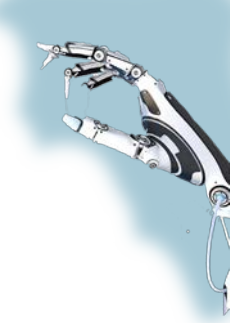
02 pyquery解析网页内容

pyquery-第三方库



```
pip install pyquery

from pyquery import PyQuery as pq

pq_doc = pq(filename='./data/css_example.html')


pq_items = pq_doc('#Tom p')

pq_ps = pq_items.find('p')
```

pyquery库的使用



① 读取本地文件，生成pyquery对象

② 选取节点

```
pq_items = pq_doc('p a')#选取所有p节点中的a节点
```

③ 用items函数，转换成“生成器”

```
item_list = pq_items.items() #转换成“生成器”
```

④ 遍历“生成器”中的所有节点元素，获取节点的属性

```
for item in item_list:
    print(item)
    print(item.attr('href'))
    print(item.attr('target'))
    print(item.text())
```

03 bs4解析网页内容

使用bs4解析网页

- 主要类：BeautifulSoup
- 对一段HTML文本进行解析，包括以下步骤：

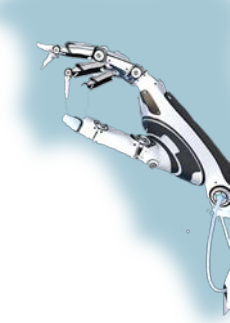
①根据HTML文本，得到一个BeautifulSoup对象:

```
from bs4 import BeautifulSoup
soup = BeautifulSoup(html,” html.parser” )
```

② 在BeautifulSoup对象中查找，得到Tag对象或Tag对象列表：

```
.div                                    找到其中第一个<div>对象
.find( “div” )                          找到其中第一个<div>对象
.find_all ( “div” )                     找到其中所有的<div>对象
.find( “p” ,{ “class” :” news” })       找到其中属性class为news的<p>对象
.find_all ( “p” ,{ “class” :” news” }) 找到其中所有属性class为news的<p>对象
```

使用bs4解析网页



③ 对于得到的tag对象，可以得到其属性：

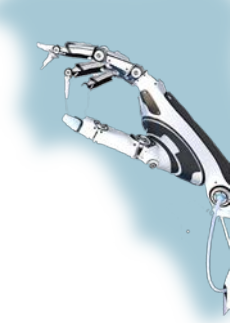
tag.text	得到对象内部的文字，不包含尖括号和尖括号中的内容
tag.[“属性名”]	得到对象的属性值，一个字符串或者列表
tag.attrs	得到对象的所有属性，一个字典

④ 由于网页的嵌套对象的特点，可以用tag对象进一步查找或遍历：

tag.div	得到对象内部的div对象
tag.find() 或tag.find_all()	在内部进一步查找
tag.parent	得到对象的父对象
tag.contents	得到对象的所有子对象

04 requests+bs4解析网页实例

作业（详见教学网-课程作业）



- 1. 用手工登录配合的方法，获取登录知乎的 cookies，供后续使用
  - cookies 保存在“./data/ my_cookies.json”，后续程序使用时也指向这个位置，以便核查时更换其他 cookies
  - 这是一个单独的程序，和后续程序可以各自独立运行
- 2. 使用已有的 cookies，用 selenium 库登录知乎，抓取热榜话题和对应的热度值，抓取结果存成CSV格式的文件

本讲到此结束，谢谢！

计算机科学与编程入门