

# 智能金融科技技术

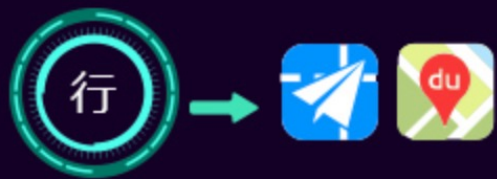
## 数据爬取

李晨辉 副教授 博士

华东师范大学计算机学院

# 用户画像

## 用户数据



## 基于大数据的客户画像算法

Indexing

Count

String Convert

Extract

sequence

Join

String Truncate

Filter

Sort

Aggregate

## 客户全景画像

性别 / 年龄  
收入 / 职业

汽车 / 金融  
娱乐 / 社交



消费地点 /  
金额 / 频率

分期类型 /  
金额

1

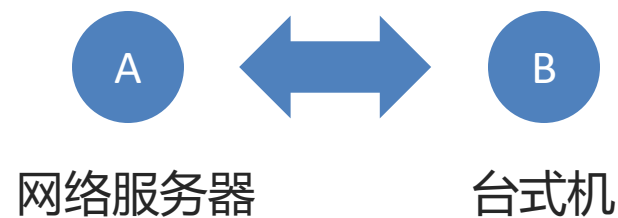
# 爬虫基本操作



# 浏览器访问网站流程

假设A有一个网络服务器。而B想用一台台式机通过浏览器访问A的服务器上运行的某个网站。整个访问过程归纳如下：

- B输入访问网站的地址后，B的电脑传输一段二进制的数数据，这些数据包含**数据头**和**数据内容**
  - **数据头**包含发送方B的mac地址和目的地的ip地址
  - **数据内容**包含了针对A网络服务器的请求，例如，获得某个网页页面。
- B的本地网络路由器将数据打包传输到A的ip地址
- B的数据最后通过物理电缆进行传输
- A的服务器接受到了B的数据包
- A的服务器识别存于**数据头**的端口号，发现是80，意味着这是一个网页请求，于是调用网页服务器相关的程序
- 网页服务器程序接受到如下信息：
  - This is a GET request
  - The following file is requested: index.html
- 网页服务器程序载入正确的HTML 文件，并打包通过本地路由发送给B的电脑



## Python模拟页面访问

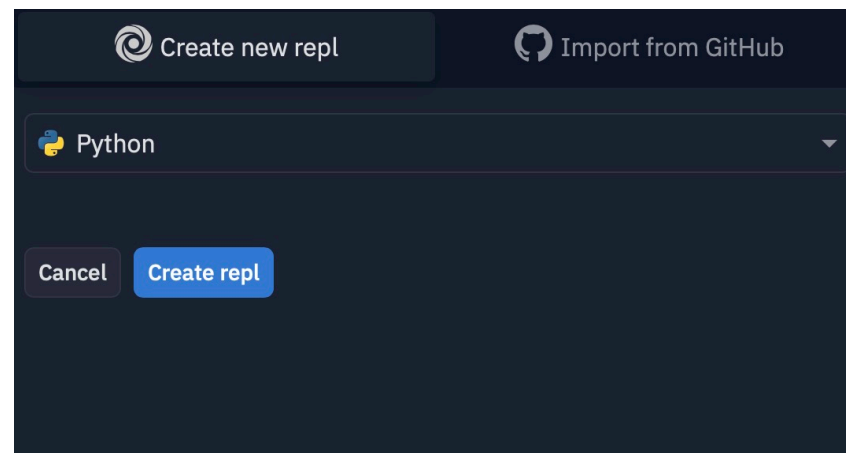
执行以下代码，获得页面内容

```
from urllib.request import urlopen  
html = urlopen("http://pythonscraping.com/pages/page1.html")  
print(html.read())
```

**Python3的在线编辑器，使用前须创建一个repl**

**使用前需要注册登陆，网络环境合适**

<https://repl.it/languages/python3>



# BeautifulSoup

1. BeautifulSoup将HTML的内容组织成了Python可以识别的对象格式
2. 因为BeautifulSoup不是Python默认的库，需要手动安装（或者通过Anaconda安装）
3. 手动安装方式：pip install beautifulsoup4

获得网页页面某字段的内容

```
from urllib.request import urlopen
from bs4 import BeautifulSoup
html = urlopen("http://www.pythonscraping.com/exercises/exercise1.html")
bsObj = BeautifulSoup(html.read(), "html.parser");
print(bsObj.h1)
```

这段代码解析了exercise1.html这个HTML文件，并输出了h1这个字段的内容：  
<h1>An Interesting Title<h1>

# BeautifulSoup

## 获得网页页面某字段的内容

```
from urllib.request import urlopen
from bs4 import BeautifulSoup
import ssl
ssl._create_default_https_context = ssl._create_unverified_context
html = urlopen("https://www.ecnu.edu.cn/info/1094/59213.htm")
bsObj = BeautifulSoup(html.read(), "html.parser");
rs = bsObj.find_all(attrs={"name": "description"});
print ( rs[0]['content'] );
print ( rs[1]['content'] );
```

当前位置: 首页 / 新闻栏目 / 新闻热点 / 正文

## 华东师大与中银金科、中国银行上海分行签约

🕒 2021年12月25日

幸福之花战略

12月22日下午，华东师范大学与中银金融科技有限公司、中国银行上海市分行签署战略合作协议，深入推进科技、金融、教育三项赋能，标志着银校合作进入崭新阶段。签约仪式在浦东新区中银大厦举行，华东师范大学校长钱旭红院士、副校长周傲英，中国银行上海市分行行长张守川，中银金融科技有限公司董事长邢桂伟出席。中国银行上海市普陀支行行长宋崇勇主持签约仪式。



2

# 词频词云





# 词频统计

1. 使用爬虫工具获得以下网页文本数据：

<https://news.ecnu.edu.cn/fc/98/c1833a261272/page.htm>

2. 使用词频工具进行词频统计

<http://ictclas.nlpir.org/>





3

**Selenium**



## 基本思路：

- 使用Selenium模拟浏览器行为

[Selenium的基本使用](#)

## 基本步骤：

- 下载Chrome驱动器，放到Python目录

<https://sites.google.com/a/chromium.org/chromedriver/home>

注意：下载的版本需与浏览器版本一致，Chrome中输入chrome://settings/help 确定版本

- 示例代码

```
from selenium import webdriver
```

```
browser = webdriver.Chrome()
```

```
browser.get('https://www.ecnu.edu.cn')
```

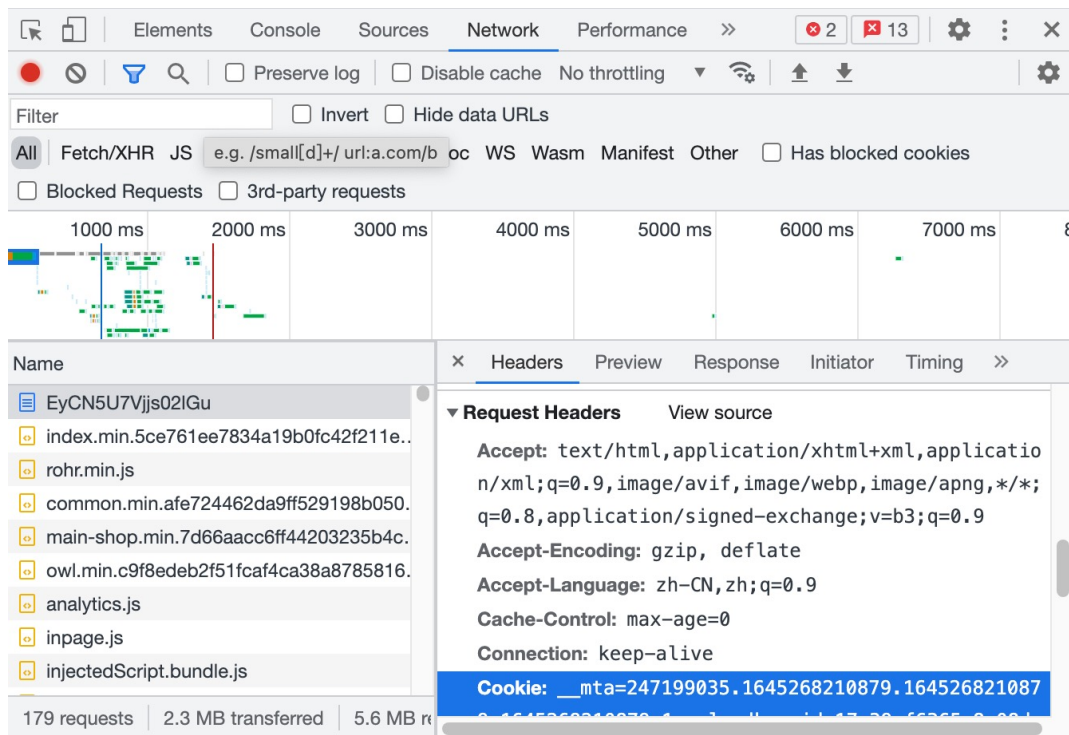
```
rs = browser.find_element_by_id('top-nav')
```



**Cookie**

# Cookie

- 浏览器登录账号后获取Cookie信息



- 示例代码

```
import requests
```

```
#要抓取的目标页码地址
```

```
url = "https://www.qcc.com/web/search/risk?key=东方财富证券股份有限公司"
```

```
#抓取页码内容，返回响应对象
```

```
headers = {"User_Agent": "Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/64.0.3282.186 Safari/537.36", "Cookie": "该值从浏览器获取", "Host": "www.qcc.com"}
```

```
response = requests.get(url, headers = headers )
```

```
with open("test.html", "w", encoding="utf-8") as f:
```

```
    f.write(response.content.decode());
```

- 拉勾网（招聘要求）：<https://www.lagou.com/>
- 企查查（公司信息）：<https://www.qcc.com/>
- 东方财富（上市公司）：<https://www.eastmoney.com/>
- **大众点评（餐饮消费）**：<https://www.dianping.com/>