



本课件仅用于教学使用。未经许可，任何单位、组织和个人不得将课件用于该课程教学之外的用途(包括但不限于盈利等)，也不得上传至可公开访问的网络环境

1

数据科学导论

Introduction to Data Science

第二章 数据分析基础

黄振亚，陈恩红，刘淇

Email: cheneh@ustc.edu.cn, huangzhy@ustc.edu.cn

课程主页:

<http://staff.ustc.edu.cn/~huangzhy/Course/DS2021.html>



数据分析基础

2

□ 数据采集

Data Collection

□ 数据存储

Data Storage

□ 数据预处理

Data Preprocessing

□ 特征工程

Feature Engineering





数据采集

3

□ 无时无刻产生数据，获得数据的方式多种多样



网页



测量



数据库



监控



传统媒体



数据采集

4

- 数据检索
- 公开数据
- 批量数据获取
 - 网络爬虫
- 数据筛选



数据采集：数据检索

5

- 最简单、最灵活的数据获取方式就是依靠检索
- 学会使用搜索引擎
 - 百度：适合于搜索中文信息
 - Google：更适合搜索英文信息



11/10/2021



数据采集：数据检索

- 最简单、最灵活的数据获取方式就是依靠检索
- 学会使用搜索引擎
 - Google: 更适合搜索英文信息
 - Google Scholar, DBLP



文章 找到约 82 条结果 (用时0.04秒)

时间不限

2021以来

2020以来

2017以来

自定义范围...

按相关性排序

按日期排序

不限语言

中文网页

简体中文网页

Ekt: Exercise-aware knowledge tracing for student performance prediction
[Q.Liu, Z.Huang, Y.Yin, E.Chen, H.Xiong...](#) - IEEE Transactions on ..., 2019 - ieeexplore.ieee.org
 For offering proactive services (eg, personalized exercise recommendation) to the students in computer supported intelligent education, one of the fundamental tasks is predicting student performance (eg, scores) on future exercises, where it is necessary to track the ...
 ☆ 99 被引用次数: 77 相关文章 所有 12 个版本

[HTML] **Exercise Hierarchical Feature Enhanced Knowledge Tracing**
[H.Tong, Y.Zhou, Z.Wang](#) - International Conference on Artificial ..., 2020 - Springer
 ... Eur. J. Psychol. Assess. 16(1), 3 (2000)CrossRefGoogle Scholar. 5. Huang, Z., et al.:
Ekt:Exercise-aware knowledge tracing for student performance prediction. IEEE Trans. Knowl. Data Eng. (2019) Google Scholar. 6. Johnson, SC: Hierarchical clustering schemes ...
 ☆ 99 被引用次数: 4 相关文章 所有 6 个版本

mpg Context aware Knowledge Tracing Integrated with The Exercise

Index of /xml

Name	Last modified	Size	Description
Parent Directory			-
CHANGES.txt	2019-11-22 21:20	3.5K	
README.txt	2019-11-22 21:20	3.5K	
dblp.dtd	2019-11-22 21:20	12K	
dblp.xml.gz	2021-09-14 02:20	624M	
dblp.xml.gz.md5	2021-09-14 02:20	46	
docu/	2018-03-01 16:43	-	
osd.xml	2020-12-18 16:26	1.5K	
release/	2019-08-20 15:57	-	



数据采集：数据检索

7

- 最简单、最灵活的数据获取方式就是依靠检索
- 学会使用搜索引擎
 - Google: 更适合搜索英文信息
 - 2018.9, Google Dataset Search (Google 数据集搜索)
网址: <https://toolbox.google.com/datasetsearch>



目前仍处于测试阶段，支持中文搜索，但中国大陆的用户想要使用依然需要“梯子”

11/10/2021



数据采集：数据检索

- 最简单、最灵活的数据获取方式就是依靠检索
- 学会使用搜索引擎

找到 100 多个数据集

搜索

The New York Times
Coronavirus (Covid-19) Data in the United States
github.com
www.nytimes.com
csv

Our World in Data
Coronavirus Disease (COVID-19) - the data
ourworldindata.org
csv

The New York Times
Coronavirus (Covid-19) Data in the United States
访问 github.com 访问 www.nytimes.com
csv
数据集提供者
The New York Times
许可

阶段，支
中国大陆
依然需要



数据采集：公开数据

- 国内常见公开数据渠道
 - 国家相关部门统计信息
 - 中国银行业监督管理委员会
 - 中国国家统计局

数据解读

更多>>

- 2015年8月社会融资规模增量统计数据报告
- 2015年8月金融统计数据报告
- 2015年7月社会融资规模增量统计数据报告
- 2015年7月金融统计数据报告
- 2015年上半年地区社会融资规模增量统计数据

统计信息	
• 2015年总资产、总负债 (月度)	2015-08-25
• 2015年总资产、总负债 (季度)	2015-08-10
• 2015年银行业金融机构用于小微企业的贷款情况表 (季度)	2015-08-10
• 2015年商业银行主要指标分机构情况表 (季度)	2015-08-10
• 2015年商业银行主要监管指标情况表 (季度)	2015-08-10
• 2015年银监会监管统计信息发布日程表	2015-02-13
• 2014年总资产、总负债 (季度)	2015-02-13
• 2014年商业银行主要监管指标情况表 (季度)	2015-02-13
• 2014年商业银行主要指标分机构情况表 (季度)	2015-02-13

□ 代表性公开数据集

- 1400万的图像数据
 - <http://www.image-net.org/>
- Amazon从2008年开始就为开发者提供几十TB的开发数据
 - <http://aws.amazon.com/datasets>
- YouTube视频的统计与社交网络数据
 - <http://netsg.cs.sfu.ca/youtubedata/>

统计公报
| 更多

- 年度统计公报
- 经济普查公报
- 人口普查公报
- 农业普查公报
- R&D普查公报
- 其他统计公报
- 基本单位普查公报
- 工业普查公报
- 三产普查公报



数据采集：公开数据

10

□ 代表性公开数据集

- 用户评分MovieLens: <https://grouplens.org/datasets/movielens/>
- 文本数据-头条: <https://github.com/aceimnorstuvwxyz/toutiao-text-classification-dataset>
- 金融数据-股票: <https://github.com/asxinyu/Stock>
- 网络数据-Large scale network: <https://snap.stanford.edu/data/>
- 教育数据:
 - ASSISTmentsData-学业: <https://sites.google.com/site/assistmentsdata/home/>
 - BASEGroup: <https://github.com/bigdata-ustc/EduData>
- 阿里天池数据-数据平台: <https://tianchi.aliyun.com/dataset/>
- 公开大数据竞赛的数据: KDDCup, NeurIPS Challenge



数据采集：批量数据获取

- 大量数据的获取难以手动实现，需借助**爬虫程序**
 - 也有可能通过交易（购买）“数据”而得
- 网络爬虫是一个自动在网上抓取数据的程序
 - 爬虫本质上就是**下载**特定网站网页的HTML/JSON/XML数据，并对数据进行**解析、提取与存储**
 - 通常先定义一组**入口URL**，根据页面中的其他URL，**深度优先**或**广度优先**的遍历访问，逐一抓取数据

The screenshot shows a news website interface with a search bar at the top. Below the search bar, there are several news articles and a 'Hot News' section. The articles include titles like '习近平对智利进行国事访问 发表署名文章 专题', '李克强：打造改革创新开放的新标杆', '纸币要消失了？央行筹备数字货币', '国家卫计委：全国性公共场所控烟条例有望今年出台', '韩日三审“气象灾害预防条约’ 维护“输入法’', '韩媒：韩日签署《空情协定》为深化双边军事合作起步', '父亲半道弃女儿遗体于垃圾箱内 当事人接受调查', and '联想A320T手机系统升级失败开不了机怎样重新刷机'. The 'Hot News' section features a large image of Xi Jinping at a podium and a table of related news items.



数据采集：网络爬虫

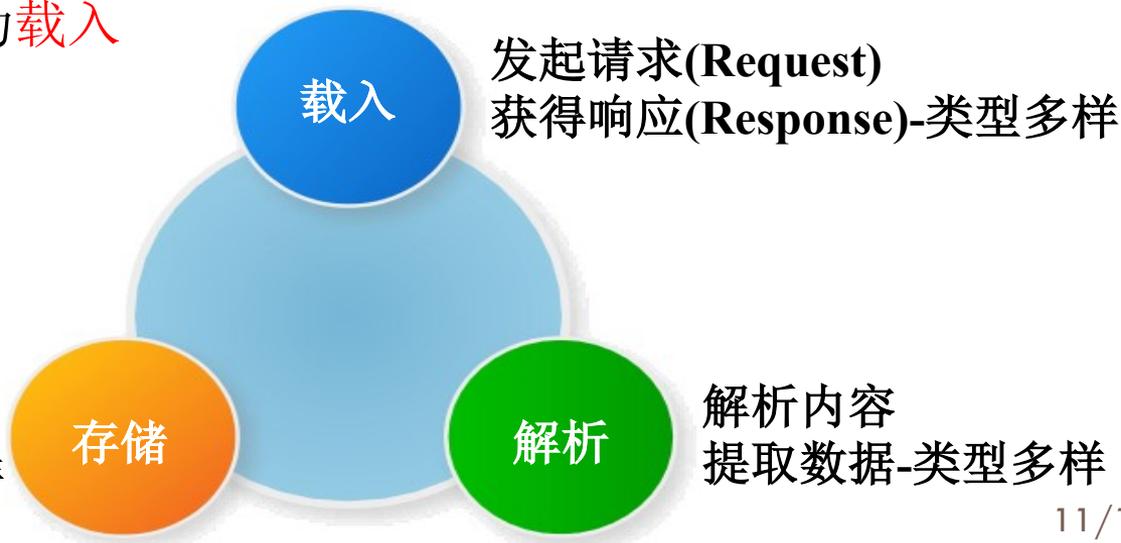
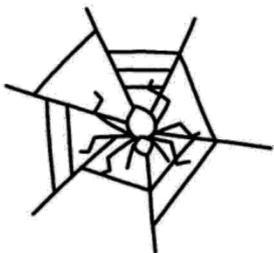
□ 网络爬虫是什么？

□ 网络爬虫（又被称为网页蜘蛛，网络机器人，网页追逐者），是一种按照一定的规则，自动的抓取万维网信息的程序或者脚本。

■ 请求网站并提取数据的自动化程序

□ 爬虫的行为可以划分为：载入、解析、存储，

■ 最复杂的部分为载入

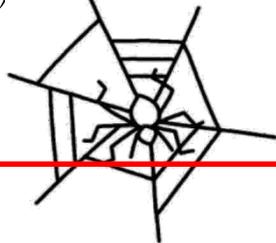




数据采集：网站数据

访问网页示例

- 网站数据主要依托于网页（html, 超文本标记语言）展示
- 用户Request服务器，服务器response信息（html等）



存储数据





数据采集：网站数据

14

- 网页示例
 - 网站数据主要依托于网页（html, 超文本标记语言）展示
 - 用户Request服务器，服务器response信息（html等）
 - 课程主页<http://staff.ustc.edu.cn/~huangzhy/Course/DS2021.html>
 - 右键“检查”查看网页源代码

Introduction to Data Science 数据科学导论

课程代码：CS1503

学院：011计算机科学与技术系

教师：[陈恩红](#), [黄振亚](#), [刘淇](#)

上课时间：每周二下午第8、9节，教室：3B201

11/10/2021



数据采集：网站数据

15

```
□ <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html>
<head>

</head>
<body>
<font size="8px"><strong> <center>Introduction to Data Science </center></strong> <font>
<font size="8px"><strong> <center>数据科学导论 </center></strong> <font>
<br>
<font size="6px"><strong> <center>课程代码：CS1503 </center></strong> <font>
<font size="6px"><strong> <center>学院：011计算机科学与技术系 </center></strong> <font>
<font size="6px"><strong> <center>教师：<a href="http://staff.ustc.edu.cn/~cheneh/" target="-parent">陈恩红</a>
<br><font size="6px"><strong> <center>上课时间：每周二下午第8、9节，教室：3B201</center></strong> <font>

<hr size="1px" noshade>
助教（含作业提交）：
<br><dir><font size="5px">
<li>刘嘉聿 jy251198@mail.ustc.edu.cn</li><br>
<li>QQ群： 697196774 </li></font>
```



网络爬虫：载入

16

- 载入：将目标网站数据下载到本地
 - Html, HyperText Markup Language
 - 爬虫程序向服务器发送网络请求 Request，获取相应的网页
 - 网站常用网络协议：http, https
 - 数据常用请求方式：get, post
 - get: 参数常放置在URL中
 - `http://www.adc.com?p=1&q=2&r=3`,
 - 问号后为参数
 - post: 参数常放置在一个表单中（报文头（header））
 - 在向目标URL发送请求时，将参数放置在一个网络请求的报文头中
 - 更安全



网络爬虫：载入

17

□ 载入：将目标网站数据下载到本地

□ 数据常用请求方式：get, post

■ **get**：参数常放置在URL中

■ http://www.adc.com?p=1&q=2&r=3，问号后为参数

■ 例如，https://www.baidu.com/s?wd=图片



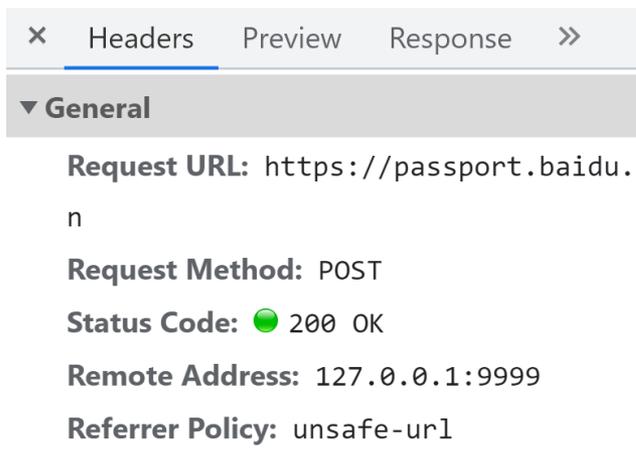
请求头 Accept: text/html.applicati



网络爬虫：载入

18

- 载入：将目标网站数据下载到本地
 - 数据常用请求方式：get, post
 - **post**：参数常放置在一个表单中
 - 在向目标URL发送请求时，将参数放置在一个网络请求的报文头中
 - 相比于Get，多了Form Data部分（请求体）
 - 更安全：登录操作常用（不会放在URL后面）





网络爬虫：载入

19

- 载入：将目标网站数据下载到本地
 - 数据常用请求方式：get, post
 - 获得服务器的响应：Response，即获取网页源代码

响应头

响应体，即网页源代码

▼ Response Headers View source

Bdpagetype: 2

Bdqid: 0xdabb3fd4000073b3

响应状态：200，404等

HTTP状态码

```

<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html>
<head>

</head>
<body>
<font size="8px"><strong> <center>Introduction to Data Science </center></strong> <font>
<font size="8px"><strong> <center>数据科学导论 </center></strong> <font>
<br>
<font size="6px"><strong> <center>课程代码: CS1503 </center></strong> <font>
<font size="6px"><strong> <center>学院: 011计算机科学与技术系 </center></strong> <font>
<font size="6px"><strong> <center>教师: <a href="http://staff.ustc.edu.cn/~cheneh/" target="-parent">陈恩红</a>
<br><font size="6px"><strong> <center>上课时间: 每周二下午第8、9节, 教室: 3B201</center></strong> <font>

<hr size="1px" noshade>
助教(含作业提交):
<br><dir><font size="5px">
<li>刘嘉聿    jy251198@mail.ustc.edu.cn</li><br>
<li>QQ群: 697196774 </li></font>

```



网络爬虫：载入

20

- 实际操作：抓取一个静态网页步骤
 - 首先确定URL，例如：<http://www.baidu.com>
 - 其次确定请求的方式以及相关参数：
 - 直接用浏览器实现：chrome, firefox浏览器抓包工具，详见
 - <http://jingyan.baidu.com/article/3c343ff703fee20d377963e7.html>
 - 或者抓包工具：charles等，详见
 - <http://blog.csdn.net/jiangwei0910410003/article/details/41620363/>
 - 最后在代码中按照特定的请求方式（get, post）向URL发送参数，即可收到网页的结果



网络爬虫：载入

21



□ 但部分页面的数据是**动态加载的**

□ Ajax异步请求

- 网页中的部分数据需要**浏览器渲染** (JavaScript调用接口获取数据)
- 用户的某些点击、下拉的**操作**触发才能获得

□ 解决方案：

- 借助抓包工具，分析Ajax某次操作所触发的请求，通过代码实现相应的请求
 - 有技术难度，但抓取速度快。
- 利用智能化的工具：selenium webdriver
 - 用**程序控制驱动浏览器**，模拟浏览器
 - 可以**模拟实现人的所有操作**
 - 操作简单，但是速度慢
 - 因为爬虫需要启动浏览器，浏览器需要渲染页面，所以速度比较慢
- 其他：Splash, Pyv8等



网络爬虫：载入

22

- **反爬虫**：随着网络爬虫对目标网站访问频率的加大，网站禁止爬虫程序继续访问
- 常见反爬手段：
 - 出现用户登录界面，需要验证码
 - 禁止某个固定帐号或ip一段时间内访问网站
 - 更有甚者，直接返回错误的无用数据
- 应对措施：
 - 优化爬虫程序，尽量减少访问次数，尽量不抓取重复内容
 - 使用多个cookie（网站用来识别用户的手段，每个用户登录会生成一个cookie）
 - 使用多个ip（可以用代理实现）

安全验证 ×

您的帐号可能存在安全风险，为了确保为您本人操作，请先进行安全验证。

发送成功

验证方式

186*****23手机 ▼

请输入六位验证码 重新发送(57)

确定



网络爬虫：解析

23

- 解析：在载入的结果中**抽取特定的数据**，载入的结果主要分成三类html、json、xml
 - html
 - Java工具包：jsoup等
 - Python工具包：beautifulSoup等
 - json
 - Java工具包：json-lib、org-json、jackson等
 - Python工具包：json、demjson等
 - Xml
 - Java工具包：dom4j等
 - Java工具包：xml、libxml2等



解析内容
提取数据-
类型多样



网络爬虫：解析(对比JSON与XML)

```
{  
  "name": "中国",  
  "province": [{  
    "name": "黑龙江",  
    "cities": {  
      "city": ["哈尔滨", "大庆"]  
    }  
  },  
  {  
    "name": "广东",  
    "cities": {  
      "city": ["广州", "深圳", "珠海"]  
    }  
  },  
  .....  
}]
```

对象，成员：键值对

```
<?xml version="1.0" encoding="utf-8"?>  
<country>  
  <name>中国</name>  
  <province>  
    <name>黑龙江</name>  
    <cities>  
      <city>哈尔滨</city>  
      <city>大庆</city>  
    </cities>  
  </province>  
  <province>  
    <name>广东</name>  
    <cities>  
      <city>广州</city>  
      <city>深圳</city>  
      <city>珠海</city>  
    </cities>  
  </province>  
  .....  
</country>
```



网络爬虫：解析(对比JSON与XML)

25

- 可读性
 - Json简洁，XML规范，xml比较好
- 可扩展性
 - 均很好
- 数据体积
 - Json数据量少，传输快。Xml数据量大，传输慢
- 编码解码
 - Json容易，xml复杂（树结构，父子节点）
- 数据描述
 - Xml数据描述更好
- 数据交互
 - Json与JavaScript交互更方便，易于解析。XML更适合跨平台共享



网络爬虫： 抓取微博评论

 **邓超** 🏆
8-17 20:49 来自 iPhone 7 Plus

跑男最新名单.....

📄 344900 | 💬 303031

转发 344900 评论 303031

 **陈赫**
08-18
天霸

 **邓超**
08-18
我们都很好，谢谢大家❤️

 **邓超**
08-18
我也不知道🐼

 **贼亮zl**
08-17
迪丽热巴💋💋

抓包工具
获取请求

▼ **General**

Request URL: https://m.weibo.cn/api/comments/show?i
Request Method: GET
Status Code: 🟢 200 OK
Remote Address: 123.125.106.67:443
Referrer Policy: no-referrer-when-downgrade

▶ **Response Headers (14)**

▼ **Request Headers** [view source](#)

Accept: application/json, text/plain, */*
Accept-Encoding: gzip, deflate, br
Accept-Language: zh-CN,zh;q=0.8,en;q=0.6
Connection: keep-alive
Cookie: _T_WM=d9a7dba4dd130f79eaecac13c8906050; ALbktAKLUXNkw1un7fu00CXjkppVYn1wGjJ3knF4g.; SUBP=0p5NHD95Q0So5Re0.cS020Ws4Dqcjn-fHBxHzLxK-LB.eLBK5L505136002; M_WEIBO_CN_PARAMS=featurecode%3D200003236170084375%26uicode%3D20000061%26fid%3D414183617
Host: m.weibo.cn
Referer: https://m.weibo.cn/status/4141836170084375
User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 12.113 Safari/537.36
X-Requested-With: XMLHttpRequest



网络爬虫： 抓取微博评论

获得评论的json格式

京ICP备15025187号-1 邮箱: service@json.cn

```

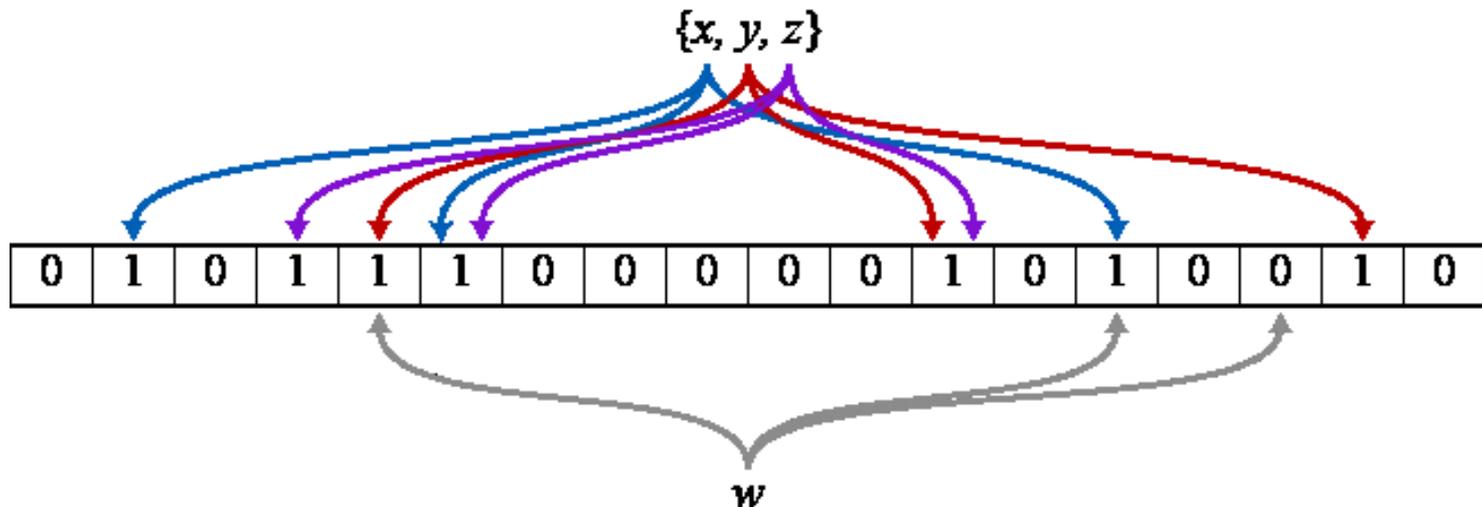
    "mod_type": "mod/pagelist",
    "previous_cursor": "",
    "next_cursor": "",
    "card_group": [
      {
        "id": "4142016554789113",
        "created_at": "08-18 08:46",
        "source": "柔光自拍vivo X7",
        "user": @Object{...},
        "text": "回复
```



网络爬虫：去重服务

28

- 去重服务：避免信息的重复抓取，减小存储空间
 - Bloom过滤器**：由一个很长的二进制向量和一系列随机映射函数组成，通过多个hash函数将一个元素映射成一个位阵列（Bit Array）中的多个点
 - 只有多个hash结果都一样时，才说明数据是重复的

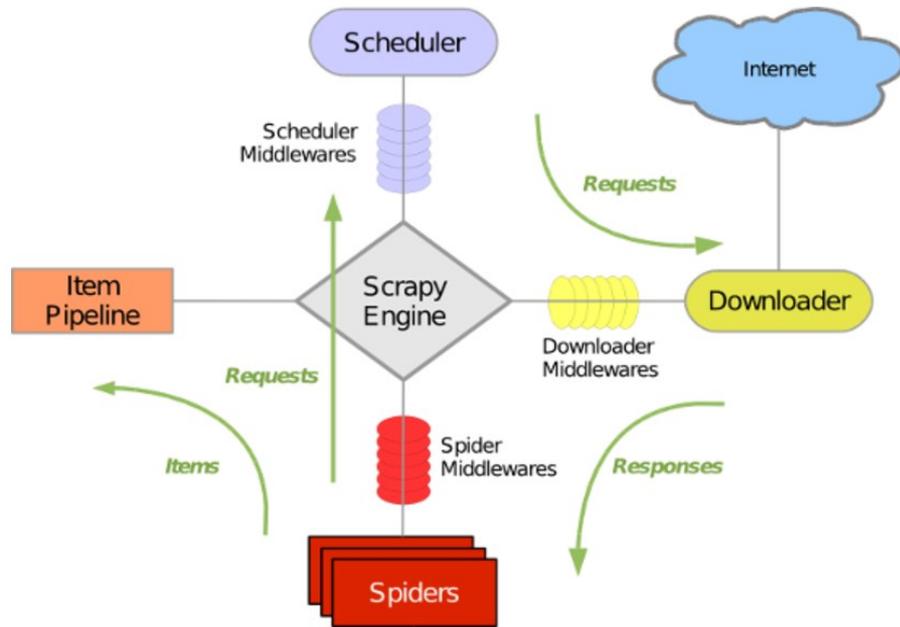




网络爬虫：现有技术

29

- 基于Java的工具
 - HttpClient
 - Jsoup
- 基于Python的工具
 - **Scrapy**
 - BeautifulSoup



现有的爬虫框架很成熟，能够合理的控制爬取的过程，并有效的处理爬取过程中出现的各种异常，推荐使用Scrapy



网络爬虫：现有技术

30

■ ItSucks工具

- 支持通过下载模板和**正则表达式**来定义下载规则
- 提供swing GUI操作界面

■ Spidernet工具

- 以**递归树**为模型的多线程web爬虫程序
- 存储于**sqlite**数据文件

■ 完整解决方案

- 基于用户浏览器的爬虫（插件）
- 八爪鱼
- 火车采集器



火车采集器
网页数据采集利器



数据采集

31

- 注意网站规定
- 注意法律规定
 - 2021年6月1日，《中华人民共和国数据安全法》
- 注意数据使用规范
- etc



数据采集

32

- 数据检索
- 公开数据
- 批量数据获取
 - 网络爬虫
- 数据筛选



数据分析基础

33

- 数据采集 Data Collection
- 数据存储 Data Storage
- 数据预处理 Data Preprocessing
- 特征工程 Feature engineering

