数据科学导论

Introduction to Data Science

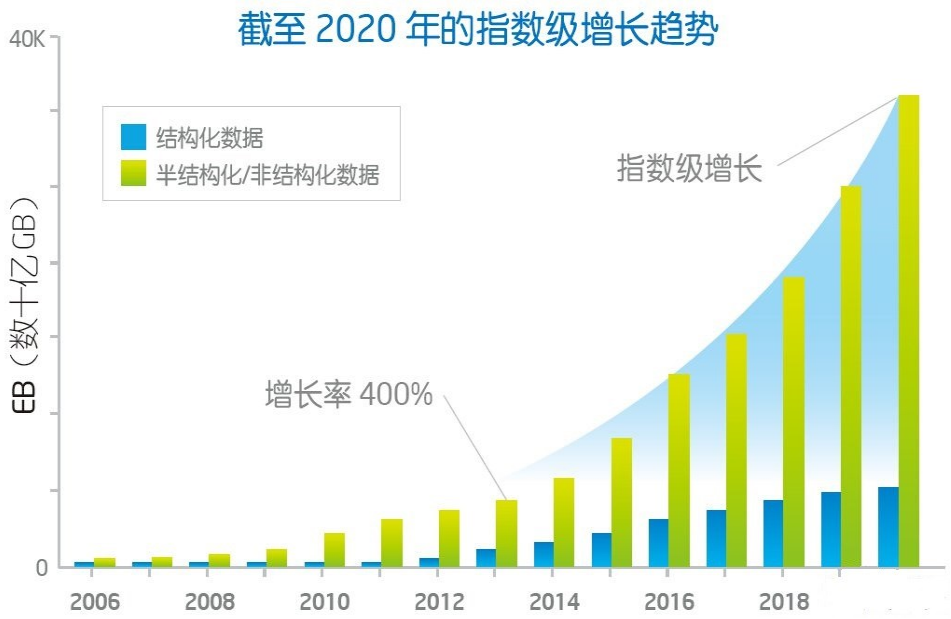
第二章 数据分析基础

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

- 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

- 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

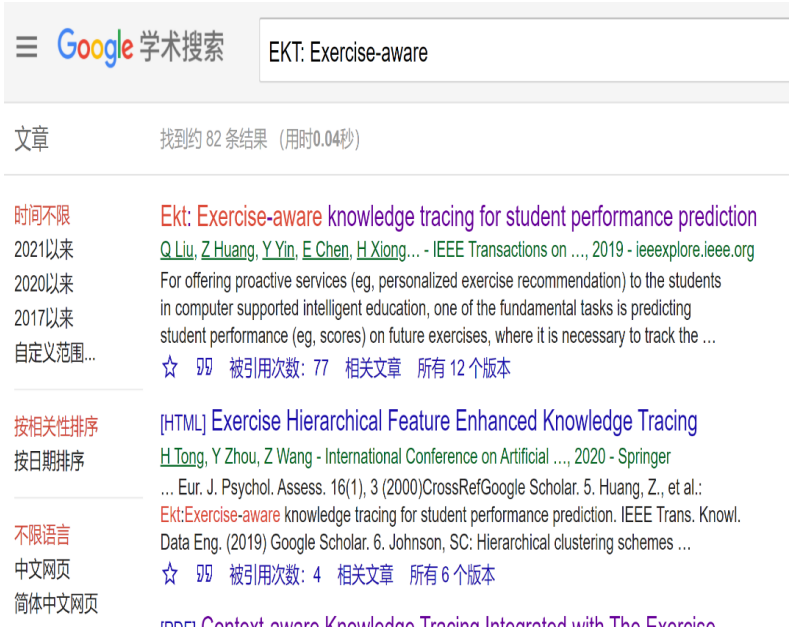
- 数据检索
- 公开数据
- 批量数据获取
  - 网络爬虫
- 数据筛选

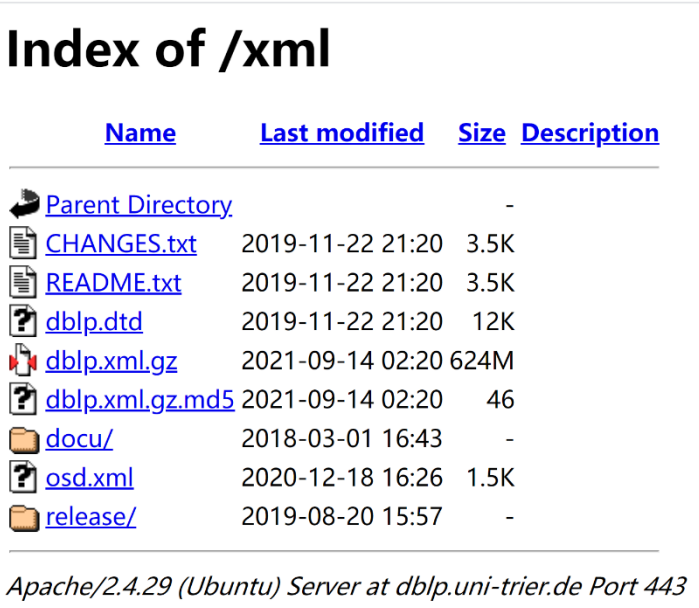
数据采集：数据检索

- 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

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





数据采集：数据检索

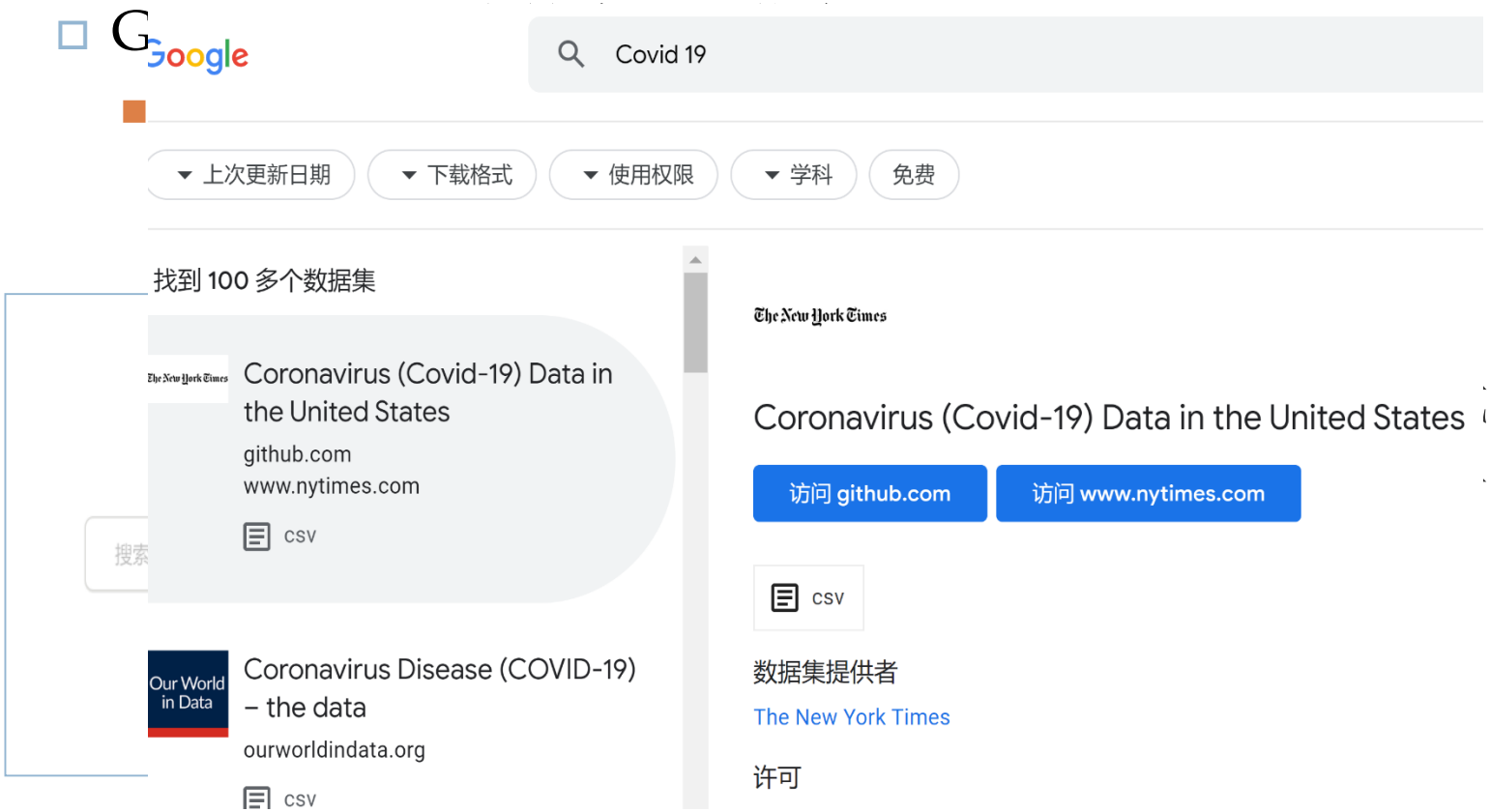
- 最简单、最灵活的数据获取方式就是依靠检索
- 学会使用搜索引擎
  - Google：更适合搜索英文信息
    - 2018.9，Google Dataset Search（Google 数据集搜索）

      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”

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



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

- 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/

数据采集：公开数据

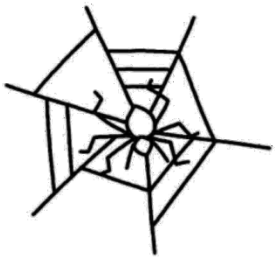
- 代表性公开数据集
  - 用户评分MovieLens：https://grouplens.org/datasets/movielens/
  - 文本数据-头条：https://github.com/aceimnorstuvwxz/toutiao-text-classfication-dataset
  - 金融数据-股票：https://github.com/asxinyu/Stock
  - 网络数据-Large scale network：https://snap.stanford.edu/data/
  - 教育数据:
    - ASSISTmentsData-学业: https://sites.google.com/site/assistmentsdata/home/
    - 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

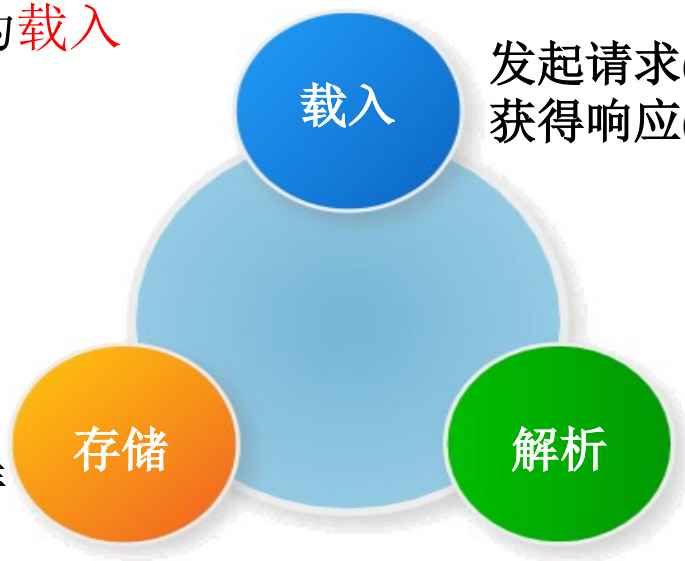
数据采集：批量数据获取

- 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

数据采集：网络爬虫

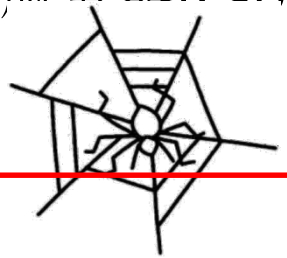
- 网络爬虫是什么？
  - 网络爬虫（又被称为网页蜘蛛，网络机器人，网页追逐者），是一种按照一定的规则，自动的抓取万维网信息的程序或者脚本。
    - 请求网站并提取数据的自动化程序
  - 爬虫的行为可以划分为：载入、解析、存储，
    - 最复杂的部分为载入





数据采集：网站数据

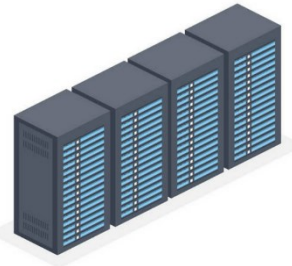
- 访问网页示例
  - 网站数据主要依托于网页（html, 超文本标记语言）展示
  - 用户Request服务器，服务器response信息（html等）



存储数据

浏览器解析
展示数据

发起请求(Request)



获得响应(Response)
网页数据

数据采集：网站数据

- 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

数据采集：网站数据

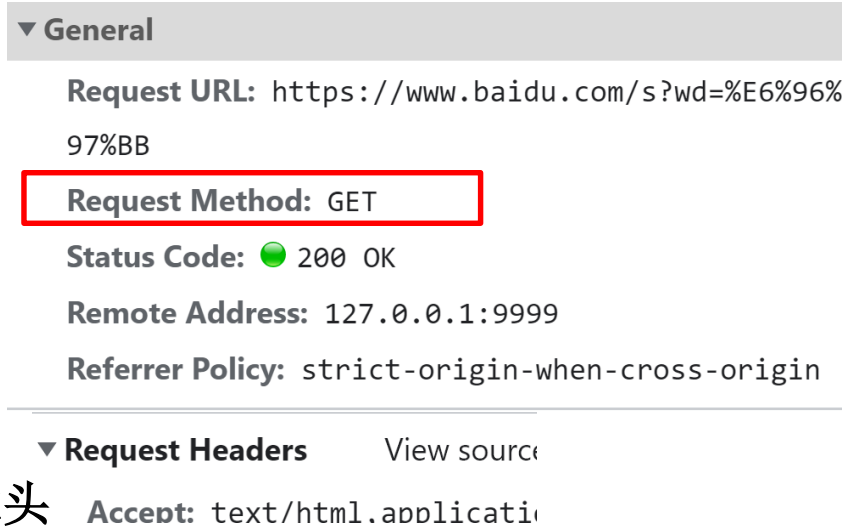

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html>
<head>

</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

- 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

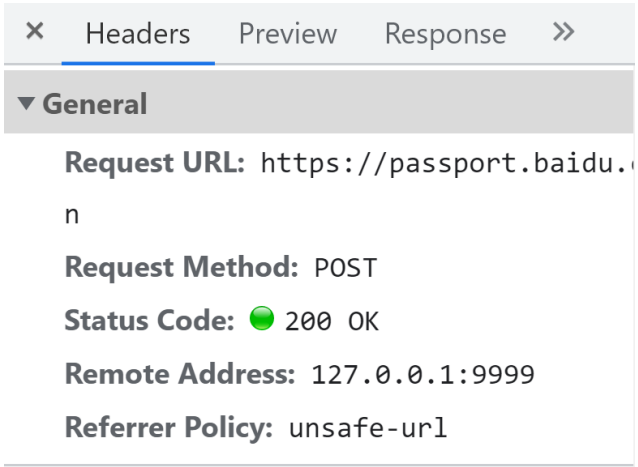
- 载入：将目标网站数据下载到本地
  - 数据常用请求方式：get，post
    - **get：** 参数常放置在**URL**中
      - http://www.adc.com?p=1&q=2&r=3，问号后为参数
      - 例如，https://www.baidu.com/s?wd=图片



请求头

网络爬虫：载入

- 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）



请求体

网络爬虫：载入

- 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

响应状态：200， 404等

HTTP状态码

响应体，即网页源代码

```
<!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

- 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

- 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

- 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）

网络爬虫：解析

- 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

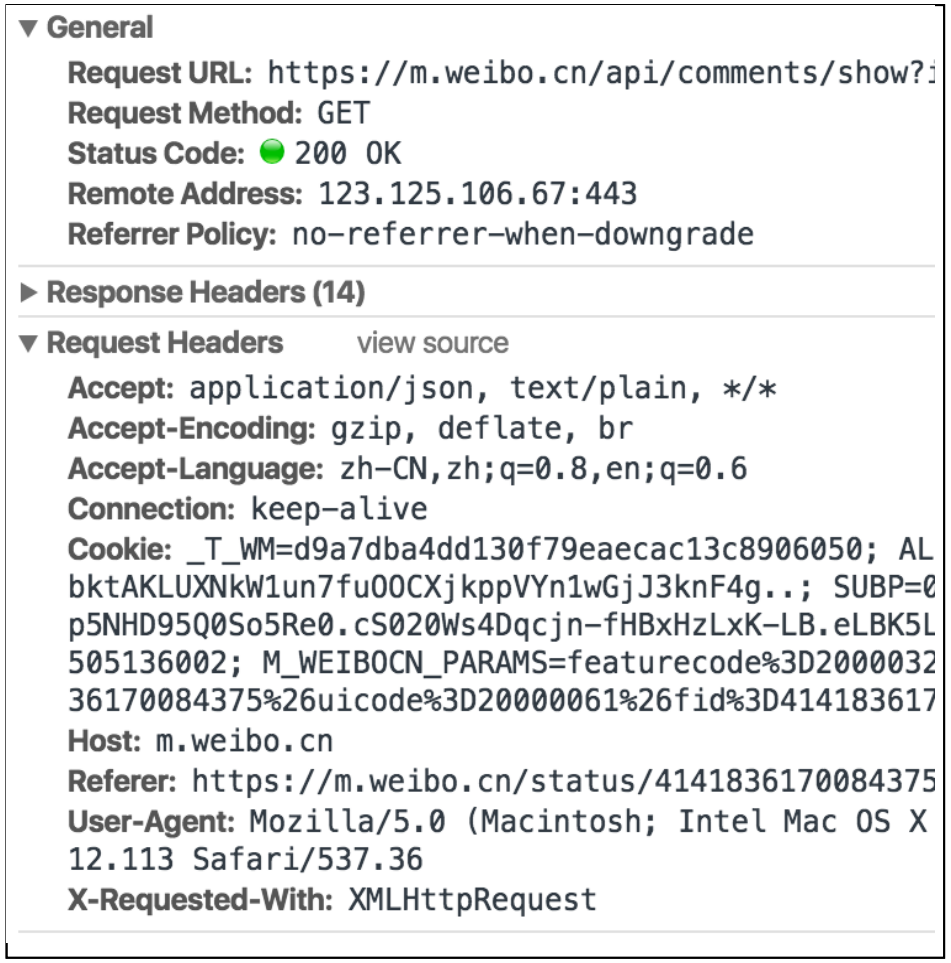
```
<?xml version="1.0" encoding="utf-8"?>
<country>
  <name>中国</name>
  <province>
    <name>黑龙江</name>
    <cities>
      <city>哈尔滨</city>
      <city>大庆</city>
    </cities>
  </province>
  <province>
    <name>广东</name>
    <cities>
      <city>广州</city>
      <city>深圳</city>
      <city>珠海</city>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

网络爬虫：解析(对比JSON与XML)

- 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

抓包工具
获取请求



网络爬虫：抓取微博评论

获得评论的json格式

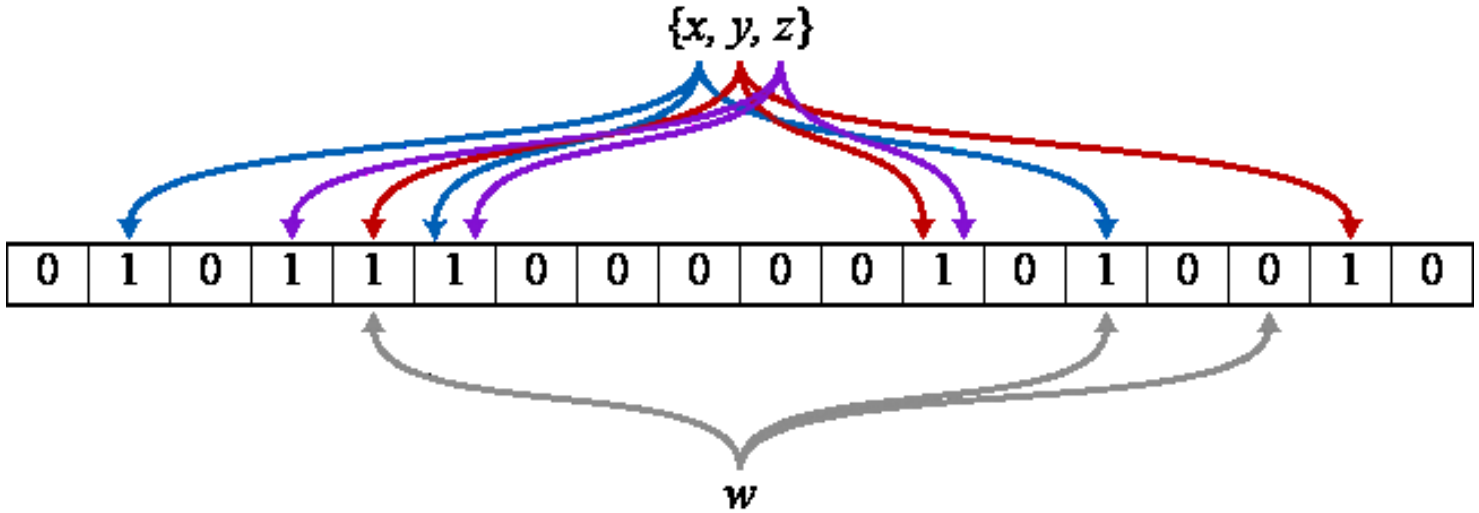

```
{
    "mod_type":"mod/pagelist",
    "previous_cursor":"",
    "next_cursor":"",
    "card_group":[
        {
            "id":4142016554789113,
            "created_at":"08-18 08:46",
            "source":"柔光自拍vivo X7",
            "user":Object{...},
            "text":"回复<a
href="/n/%E9%82%93%E8%B6%85">@邓超</a>:不管是谁,
请大家记住陈赫的话，他们很好，感情都很好。恳请各家粉丝不要
戏太多就好<i class="face face_1 icon_1">[微笑]</i>
没准你们那么嫌弃骂的那么难听，人家正主还是感情好的时不时去
吃火锅呢，你们不累吗？别用自己对他的爱去给他造成困扰",
            "reply_id":4142015488402958,
            "reply_text":"<a
href='/u/'5187664653>@邓超</a>：我也不知道<i
class="face face_1 icon_20">[doge]</i>",
            "like_counts":10811,
            "liked":false,
            "mod_type":"mod/single/infobox"
```


解析出需要的字段

用户id	时间	内容
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
……	……	……

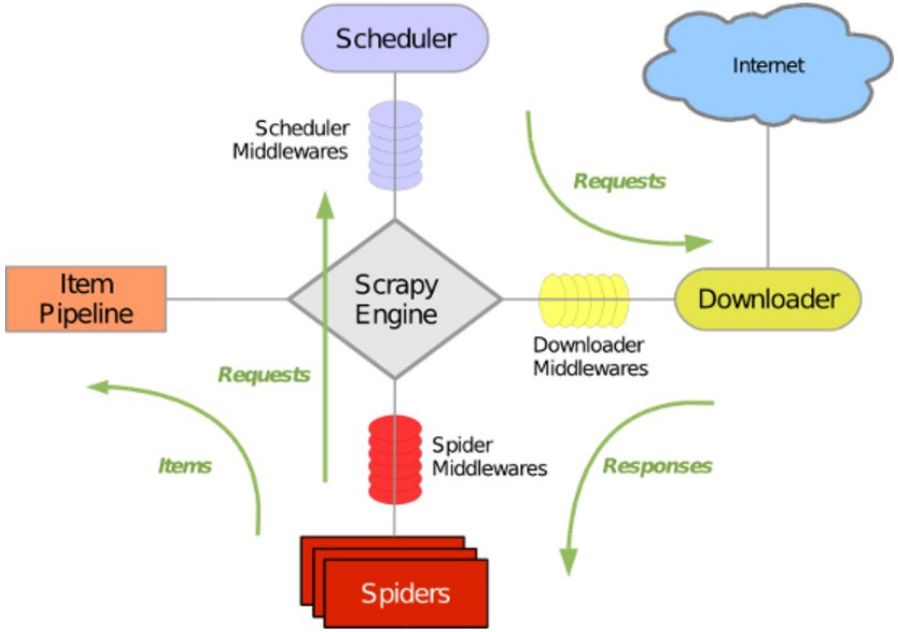
网络爬虫：去重服务

- 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

- 基于Java的工具
  - HttpClient
  - Jsoup
- 基于Python的工具
  - **Scrapy**
  - Beautiful-soup



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

- 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

- 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

- 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

- 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