1. 数据获取方式

1、企业生产的用户数据：大型互联网公司有海量用户，所有他们积累数据有天然的优势

有意识的中小型企业，也开始积累的数据。

2、数据管理资讯公司：通常这样的公司有很庞大的数据采集团队，一般会通过市场调研、问卷调查、固定的验本检测 和各行各业的公司进行合作、专家对话

3、政府、机构提供的公开数据：政府通过各级政府上报的数据进行合并、机构都是权威的第三方网站

4、第三方数据平台购买的数据： 通过各个数据交易平台来购买各行各业的数据，根据获取的难度不同，价格也不同。

5、爬虫爬取数据:如果市场上没有，或者价格太高、互联网定向爬取数据

二、爬虫：抓取网页数据的程序。

三、爬虫如何提取网页数据：

1.网页都有自己唯一的URL（统一资源定位符）来进行定位。

2.网页都使用HTML（超文本标记语言）来描述页面信息。

3.网页都使用HTTP/HTTPS（超文本传输协议）协议来传输HTML数据。

爬虫的设计思路：

1. 确定需要爬取的网页URL地址
2. 通过HTTP/HTTPS协议来获取对应的HTML页面
3. 提取HTML页面有用的数据：
4. 如果需要的数据，就保存起来。
5. 如果页面的其他URL，那就继续执行第二步

四 选择Python爬虫的原因，PHP、java、C/C++、Python等等都可以。

PHP虽然是世界最好的语言（适合写网页，对多线程支持、异步支持不好，并发处理能力很弱）爬虫是工具性程序，对速度和效率要求比较高。

Java的网络爬虫生态圈已很完善，是Python爬虫最大的对手。但Java语言本身很笨重，代码量很大。重构成本比较高，任何修改都会导致代码的大量变动。爬虫经常需要修改采集代码。

C/C++运行效率和性能几乎最高，但学习成本很高，代码成型比较慢。不是正确的选择。

Python 语法优美、代码简洁、开发效率高、支持的模块多，相关的HTTP请求模块和HTML解析模块非常丰富。还有强大的爬虫Scrapy，以及成熟高效的scrapy-redis分布式策略。

而且，调用其他语言接口非常方便（胶水语言）

五、课程介绍

1如何抓取HTML页面

HTTP请求的处理，urllib、urllib2、requests

处理后的请求可以模拟浏览器发送请求，获取服务器响应的文件

2 解析服务器响应的内容 re、xpath、BeautifulSoup4(bs4)/jsonpath、pyquery等

使用某种描述性的语言来给需要提取的数据定义一个匹配规则，符合这个规则的数据就会被匹配。

3如何采集动态HTML、验证码的处理

通用动态网页采集：Selenium + PhantomJS（无界面）：模拟真实浏览器加载js、ajax等非静态的数据

Tesseract：机器学习库，机器图像识别系统，可以处理简单的验证码，复杂的可以通过手动输入 或者专门的打码平台

4 Scrapy框架：（Scrapy，Pyspider）

高定制性高性能（异步网络框架twisted），所有数据下载速度非常快，提供了数据存储、数据下载、提取规则等组件

5 分布式策略 加分项不是必备 中小型企业没必要

Scrappy redis，在scrapy的基础上添加了一套以redis数据库为核心的组件。

让scrapy框架支持分布式的功能，主要在Redis里做请求指纹去重、请求分配、数据临时存储。Set集合去重

6 爬虫 反爬虫 反反爬虫之间的斗争：

User-Agent、代理、验证码、动态数据加载、加密数据。

数据价值，是否值得去费劲做反爬虫。

1 机器成本+人力成本 》数据价值 就不反了 一般做到封IP就结束

2 面子战争

爬虫和反爬虫之间的斗争，最后一定是爬虫获胜。

只要是真实用户可以浏览的网页数据，爬虫就一定能爬下来。

六 通用爬虫 聚焦爬虫

通用爬虫 搜索引擎用的爬虫系统 只能爬文本 Word等文字相关 图片电影不能

目标：尽可能的把互联网上所有网页下载下来，放到本地服务器里形成备份，

再对这些网页做相关处理（提取关键字、去掉广告），最后提供一个用户检索接口。

抓取流程：

A 首选选取一部分已有的URL，把这些URL放到待爬取队列。

B 从队列里取出这些URL，然后解析DNS得到主机IP，然后去这个对应的服务器里下载HTML页面，保存到本地服务器。之后把这个爬取的URL放到已爬取队列。

C 分析这些网页内容，找出网页里其他的URL连接，继续执行第二步，直到爬取结束。

搜索引擎如何获取一个新网站的URL：

1 主动向搜索引擎提交网址

2 在其他网址里设置网站的外链

3 搜索引擎和DNS服务商进行合作，可以快速收录新的网站

通用爬虫并不是万物皆可爬，需要遵守

Robots协议：协议指明了通用爬虫可以爬取网页的权限。

Robot.txt（只是一个建议） 并不是所有爬虫都遵守，一般只有大型的搜索引擎爬虫才会遵守。 个人写爬虫，就不管了。

通用爬虫工作流程： 爬取网页 存储数据 内容处理 提供检索/排名服务

搜索引擎排名

1 PageRank值： 根据网站的流量（点击量/浏览量/人气）统计，流量越高，越值钱，排名越靠前。

2 竞价排名：

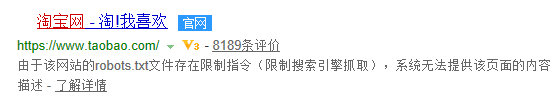
通用爬虫缺点

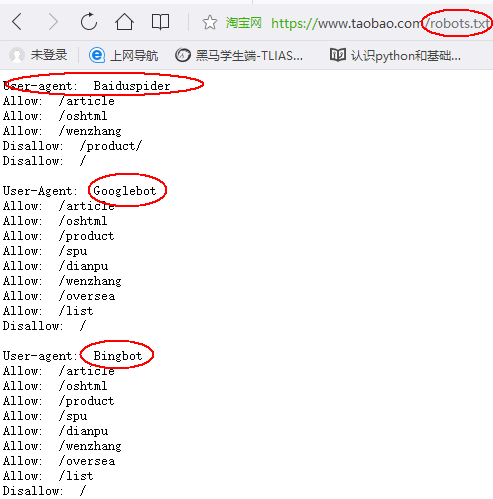
1 只能提供和文本相关的内容（HTML。Word、PDF）等等，但是不能提供多媒体文件（音乐、图片、视频）和二进制文件（程序、脚本）

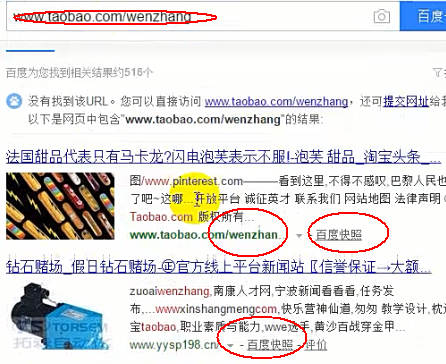
2 提供的结果千篇一律，不能针对不同领域的人提供不同的搜索结果。

3 不能理解人类语义上的检索。

从而聚焦爬虫出现了。面向主题爬虫，面向需求爬虫：会针对某种特定的内容去爬取信息，而且信息和需求尽可能相关。







DNS 就是把域名解析成IP的一种技术。

Windows对话框中 ping [www.baidu.com](http://www.baidu.com) –t 可以获得IP，直接访问服务器。

访问百度 https：//www.baidu.com 先从域名解析到DNS，通过DNS解析给浏览器一个IP地址，再把IP地址发送到指定的服务器（每个服务器都有一个IP地址）上面

聚焦爬虫 爬虫程序员写的再对某种内容爬虫

**HTTP协议 （HyperText Transfer Protocol，超文本传输协议）：是⼀种发布**

**和接收 HTML页面的方法。**

**HTTPS （Hypertext Transfer Protocol over Secure Socket Layer）简单讲是**

**HTTP的安全版，在HTTP下加入SSL层。**

**SSL （Secure Sockets Layer 安全套接层）主要用于Web的安全传输协议，**

**在传输层对网络连接进行加密，保障在Internet上数据传输的安全。**

**HTTP 的端口号为 80 ，**

**HTTPS 的端口号为 443**

**网络爬虫抓取过程可以理解为模拟浏览器操作的过程 。**

**浏览器的主要功能是向服务器发出请求，在浏览器窗口中展示您选择的网络**

**资源，HTTP是⼀套计算机通过网络进行通信的规则。**

**HTTP通信由两部分组成： 客户端请求消息与服务器响应消息**

**1. 当用户在浏览器的地址栏中输入⼀个URL并按回车键之后，浏览器会向**

**HTTP服务器发送HTTP请求。HTTP请求主要分为“Get”和“Post”两种方**

**法。Get是所有长型参数都会在URL上显示出来，Post参数隐藏**

**2. 当我们在浏览器输入URL http://www.baidu.com 的时候，浏览器发送⼀**

**个Request请求去获取 http://www.baidu.com 的html文件，服务器把**

**Response文件对象发送回给浏览器。**

**3. 浏览器分析Response中的 HTML，发现其中引用了很多其他文件，比如**

**Images⽂件，CSS⽂件，JS⽂件。 浏览器会⾃动再次发送Request去获**

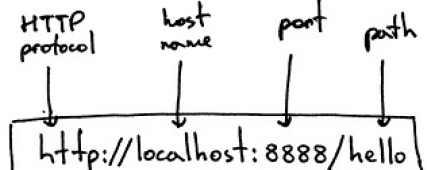
**取图⽚，CSS⽂件，或者JS⽂件。**

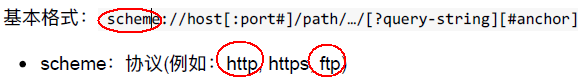
**4. 当所有的⽂件都下载成功后，⽹⻚会根据HTML语法结构，完整的显示**

**出来了。**

**URL（Uniform / Universal Resource Locator的缩写）：统⼀资源定位符，**

**是⽤于完整地描述Internet上⽹⻚和其他资源的地址的⼀种标识⽅法。**



****

**host：服务器的IP地址或者域名**

**port：服务器的端口（如果是走协议默认端口，缺省端口80）**

**path：访问资源的路径**

**query-string：参数，发送给http服务器的数据**

**anchor：锚（跳转到网页的指定锚点位置）**

**例如：**

**ftp://192.168.0.116:8080/index**

**http://www.baidu.com**

****