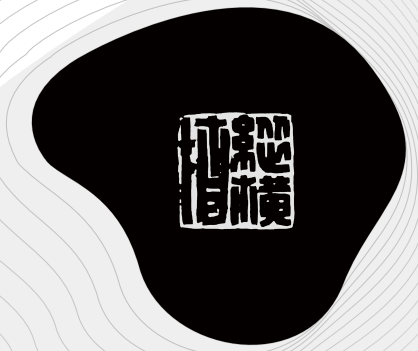


2019

网络数据采集与解析

目 录

CONTENT

1.1 什么是爬虫

- 严格意义上讲，我们不讲也用不到爬虫
  - 真正的爬虫是搜索引擎使用的
  - 重点在于算法效率、覆盖度
- 我们实际上用的是数据采集器和解析器
  - 找到数据
  - 下载数据
  - 解析数据
  - 存储数据

1.2 能不能找计算机的人写爬虫

libpku - 北京大学课程资料整理

Fork this on GitHub

/ .

.travis.yml README.md

// 专业课

/// web技术概论

Web技术概论2016秋期末1.jpg Web技术概论2016秋期末2.jpg web技术概论.docx web技术概论考题.doc

/// 人工智能

林作铨AI2013春期末.pdf

/// 人工智能概论

人工智能概论知识点整理by高飙.pdf

// toc

- .
  - 专业课
    - web技术概论
    - 人工智能
    - 人工智能概论
    - 人群与网络
      - 人群与网络2017秋期末
    - 信号与系统
    - 信息论与编码基础
    - 偏微分方程

1.3

为什么爬虫资源这么少

- 某种意义上说具有法律风险
- 爬虫是一个斗智斗勇的过程
- 爬虫技术更新太快
  - 反爬虫策略
  - 前端工程师
  - 爬虫工具

2.1

爬虫是否违法

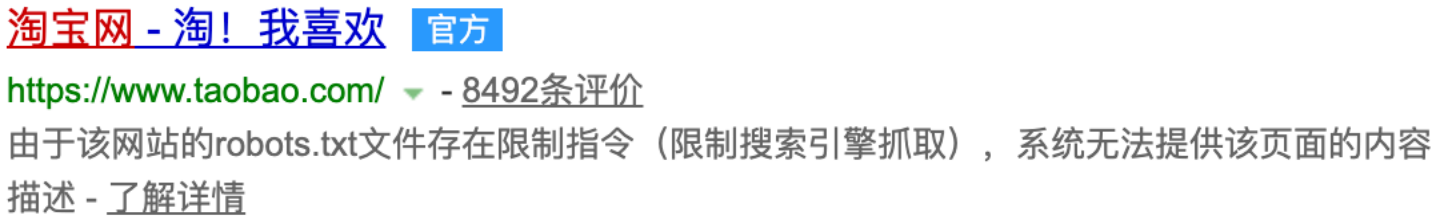
公网上一般没有问题

09.重点：什么样的爬虫是违法?

如果爬虫程序采集到**公民的姓名、身份证件号码、通信通讯联系方式、住址、账号密码、财产状况、行踪轨迹等个人信息，并将之用于非法途径的，则肯定构成非法获取公民个人信息的违法行为**。除此之外，根据相关规定，对于违反国家有关规定，向他人出售或者提供公民个人信息，情节严重的，窃取或者以其他方法非法获取公民个人信息的，均可构成成“侵犯公民个人信息罪”，处三年以下有期徒刑或者拘役，并处或者单处罚金;情节特别严重的，处三年以上七年以下有期徒刑，并处罚金。重点关注：下列情况下，爬虫有可能违法，严重的甚至构成犯罪。

1. 爬虫程序规避网站经营者设置的反爬虫措施或者破解服务器防抓取措施，非法获取相关信息，情节严重的，有可能构成“非法获取计算机信息系统数据罪”。
2. 爬虫程序干扰被访问的网站或系统正常运营，后果严重的，触犯刑法，构成“破坏计算机信息系统罪”
3. 爬虫采集的信息属于公民个人信息的，有可能构成非法获取公民个人信息的违法行为，情节严重的，有可能构成“侵犯公民个人信息罪”。

2.2 爬虫是否压力过大



- Robot.txt针对的是搜索引擎爬虫
- 大多数情况下，咱们只是遍历一次、单线程
- 不存在服务器压力的问题

2.3

反爬虫策略

- 限制频率
  - 等一等，问题不大
- 限制登陆
  - Cookie信息是基本的操作
  - 模拟登陆，用户vip还是买的
- 限制次数
  - 代理ip风险可能比较大，但是并不直接违法
- 数据结果
  - 我们又不转卖数据

3.1

爬虫框架

- 把网页抓下来（crawler）
  - 寻找一个方式，找到所有网页的url
  - 通过手段，把网页保存在本地
- 把网页解析出来（parser）
  - 正确解析内容
  - 合理存储方式

4.1 实操建议

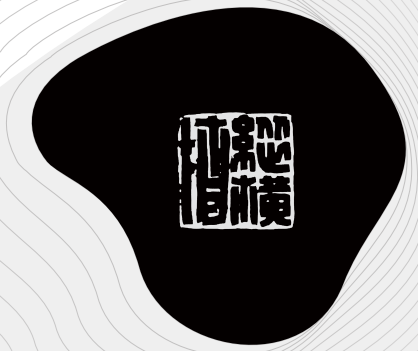
目前，我校管院的两位老师均出价2万元购买了我的股吧数据。和讯网数据的抓取正在进行中。

我还抓取过佰腾网、中国商品网、企查查等网站，项目经验丰富，数据质量有保证。

- 东方财富网股吧是国内最大、最有代表性的股票论坛，已经有使用该网站数据的文章发表在《金融研究》、《管理世界》等期刊上。我的数据是沪深两市上市公司的所有帖子的原始数据。数据跨度长达十年，数据量达到约8亿条，包含详细的帖子信息和用户信息。更详细的说明请参见“说明书-东方财富股吧数据”，样本数据请参见“数据样本-东方财富网股吧数据”。

- 和讯网是是我国知名的财经资讯网站，其新闻覆盖广度、报道专业性和数据收集处理可行性高。北京大学数字金融研究中心使用来自和讯网的新闻数据编撰了互联网金融情绪指数。我的数据预期跨度长达13年，预期数据量为9千万条。更详细的说明同样请参见附件。

感谢您能在百忙之中抽空阅读我的邮件。如果您对这两份数据感兴趣，想做进一步了解或是有其他数据想要抓取的话，请与我联系。



2019

THANKS