China's Prices Project at Xiamen University(CPP@XMU)

爬虫技术简介

岳忠信

厦门大学　经济学院

2016年12月10日

目录

请求

- 判断网页是静态还是动态：查看源代码
  - 在HTML里找到数据:静态网页找不到数据：动态网页
- 寻找入口(URL)
  - 静态网页：直接请求网页URL
  - 动态网页：通过审查元素寻找数据API接口(XHR或者JS)
- 判断请求方式：查看网页请求类型
  - Get请求：无数据上传，通常数据通过构造URL上传
  - Post请求：需要上传请求数据
- 工具：requests[1]

[1] http://docs.python-requests.org/en/master/

解析

- 静态HTML：bs4[2]
- 动态：
  - 标准JSON格式：json
  - 字符串格式：re(正则表达式)等

[2] https://www.crummy.com/software/BeautifulSoup/

储存

- CSV格式：csv
- JSON格式：json
- SQL：python-sql
- MongoDB：pymongo
- ...

动态网页寻找入口

- 审查元素-Network：抓包
- 在XHR里面找JSON格式的数据
- 如果XHR没有，在JS里面找

反爬及应对策略

- 症状：
  - 4XX：无法访问，通常是无法请求
  - 3XX：重定向，通常是需要输入验证码或者需要登陆
- 可能原因：
  - 检查请求头
  - 控制访问速率
  - 需要用户登陆

反爬及应对策略

- 初级解决方案：
  - 代理UA
  - 控制速率
  - 换IP
  - 识别验证码
  - 模拟登录

- 主要问题：数据格式混乱
- 解决方案：正则表达式+技巧

- 构造URL
- 匹配"下一页"
- 判断是否重复爬取

- 解决方案：
  - 基本原则：输入程序解码成unicode，输出程序编码成其他编码
  - 技巧：反复检查编码


CPP