项目简介-爬虫组

爬虫实例：JD评论爬虫

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
- MongoDB
- . . .


CPP