# 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

- 1 通用爬虫设计
  - 爬虫主要流程:请求,解析,储存
- ② 爬虫实例: JD评论爬虫
  - 静态爬虫: 以京东的

- 动态爬虫: 以京东的 商品评论为例
  - 常见问题
  - 翻页策略: 构造url
  - 常见错误
  - 解决新思路
  - 爬虫的修改

- 1 通用爬虫设计
  - 爬虫主要流程:请求,解析,储存
- ② 爬虫实例: JD评论爬虫
  - 静态爬虫: 以京东的 产品价格为例

- 动态爬虫:以京东的 商品评论为例
  - 常见问题
  - 翻页策略: 构造url
  - 常见错误
  - 解决新思路
  - 爬虫的修改

### 请求

- 判断静态还是动态页面。方法:审查元素(查看源代码)看我们所需的信息是否出现。
- 库的选择: 通常请求用requests库。
- 具体请求方式: get或者post(视具体情况而定)

## 解析

• 库的选择: bs4或者json

• 方法: 静态——BeautifulSoup, 动态——json

## 储存

- 储存格式: csv或者json
- 库的选择: Python的csv模块

- 1 通用爬虫设计
  - 爬虫主要流程:请求,解析,储存
- ② 爬虫实例: JD评论爬虫
  - 静态爬虫: 以京东的 产品价格为例

商品评论为例

动态爬虫:以京东的

- 常见问题
- 翻页策略: 构造url
- 常见错误
- 解决新思路
- 爬虫的修改

静态爬虫:以京东的产品价格为例 动态爬虫:以京东的商品评论为例 帮政策略:构造url 常见策略:构造url 常见新退路

#### 步骤

- 判断出它是静态内容
- ② 利用requests库请求html
  - 反爬
  - 具体症状: 返回的html错误或者出现403等报错数字
  - 原因: 一般,浏览器在向服务器发送请求的时候,会有一个请求头——User-Agent,它用来标识浏览器的类型.当我们使用requests来发送请求的时候,默认的User-Agent是python-requests/2.\*\*。
  - 反爬策略: 修改User-Agent如修改为'Mozilla/5.0 (Macintosh; Intel Mac OS X 10\_11\_2) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/47.0.2526.80 Safari/537.36'

静态爬虫:以京东的产品价格为例 动态爬虫:以京东的商品评论为例 常见问题 藉页策略:构造url 常见错误 解决新思路

### 目录

- 1 通用爬虫设计
  - 爬虫主要流程:请求,解析,储存
- ② 爬虫实例: JD评论爬虫
  - 静态爬虫: 以京东的 产品价格为例

- 动态爬虫:以京东的 商品评论为例
  - 常见问题
  - 翻页策略: 构造url
  - 常见错误
  - 解决新思路
  - 爬虫的修改

静态爬虫:以京东的产品价格为例 动态爬虫:以京东的商品评论为例 常见问题 帮贝策略:构造url 常见错误 解决新思路 哪里也的表表

#### 难点:

- 如何获取url
- 如何请求数据
- 如何处理返回的html

动态爬虫: 以京东的商品评论为例

## 步骤

● 审查元素(建议使用谷歌浏览器)

注意:刷新,浏览整个页面 观察内容: 一般在xhr或者js中,获取正确的url

② 获取内容

使用库:requests

具体方法:

- 观察headers,决定使用get或post以及对应的请求头 (一般来讲,其中我们需要关注的是Cookie / Host / Origin / Referer / User-Agent / X-Requested-With等参数。)
- 好消息是,这样的请求往往得到的内容是json格式的,所以 我们非但不会加重爬中的任务, 反而可能会省去解

i态爬虫:以京东的产品价格为例 j态爬虫:以京东的商品评论为例 :见问题 即页策略:构造url :见错误 :汉错误 □ to ke:n

#### 目录

- 1 通用爬虫设计
  - 爬虫主要流程:请 求,解析,储存
- ② 爬虫实例: JD评论爬虫
  - 静态爬虫: 以京东的 产品价格为例

● 动态爬虫:以京东的 商品评论为例

- 常见问题
- 翻页策略: 构造url
- 常见错误
- 解决新思路
- 爬虫的修改

① 获取html时出现错误

解决方案:观察headers

判断是否是反爬(print出来状态码)解决方法:修改请求头

- ② 解析html的坑:
  - 返回的不是正常的json格式 方法: import re 用正则表达式洗字符串
  - 编码错误

查看编码: 观察headers或print出来

方法:转化成unicode.

解决了这些问题,我们就应该能顺利地写出一篇单面爬虫了,那么如何翻页呢?

静态爬虫: 以京东的产品价格为例 动态爬虫: 以京东的商品评论为例 常见问题 翻页策略: 构造url 常见错误 级独等里路

#### 目录

- 1 通用爬虫设计
  - 爬虫主要流程:请求,解析,储存
- ② 爬虫实例: JD评论爬虫
  - 静态爬虫: 以京东的 产品价格为例

● 动态爬虫:以京东的 商品评论为例

- 常见问题
- 翻页策略: 构造url
- 常见错误
- 解决新思路
- 爬虫的修改

静态爬虫; 以京东的产品价格为例 动态爬虫; 以京东的商品评论为例 常见问题 翻页策略; 构造url 常见错误 解决新思路

- 观察url,修改其中的参数来重新运行观察有什么效果
- ② 找到最大面数,然后利用字符串格式化来修改url 用迭代的方法写成循环。
- ③ 观察运行效果

- 1 通用爬虫设计
  - 爬虫主要流程:请 求,解析,储存
- ② 爬虫实例: JD评论爬虫
  - 静态爬虫: 以京东的

产品价格为例

- 动态爬虫:以京东的 商品评论为例
  - 常见问题
  - 翻页策略: 构造url
  - 常见错误
  - 解决新思路
  - 爬虫的修改

静态爬虫:以京东的产品价格为例 动态问题 朝灭策略:以京东的商品评论为例 朝灭策略:构造url **常见**策略:构造url 解决新退路

- 返回的是空list现在猜测原因是正则表达式的问题)
- ❷ 编码错误(解决思路:从一开始请求时就将其转化为unicode)
- ValueError

- 1 通用爬虫设计
  - 爬虫主要流程:请 求,解析,储存
- ② 爬虫实例: JD评论爬虫
  - 静态爬虫: 以京东的 产品价格为例

● 动态爬虫:以京东的 商品评论为例

- 常见问题
- 翻页策略: 构造url
- 常见错误
- 解决新思路
- 爬虫的修改

通用爬虫设计 爬虫实例: JD评论爬虫 静态爬虫,以京东的产品价格为例 动态爬虫,以京东的商品评论为例 常见问题 翻页策略:构造url 需见籍误 肥新退路 爬虫的修改

静态爬虫:以京东的产品价格为例 动态爬虫:以京东的商品评论为例 常见问题 翻页策略:构造url 常见错误 解识进路路 爬虫的修改

## 目录

- 1 通用爬虫设计
  - 爬虫主要流程:请求,解析,储存
- ② 爬虫实例: JD评论爬虫
  - 静态爬虫: 以京东的 产品价格为例

- 动态爬虫:以京东的 商品评论为例
  - 常见问题
  - 翻页策略: 构造url
  - 常见错误
  - 解决新思路
  - 爬虫的修改

静态爬虫:以京东的产品价格为例 动态爬虫:以京东的商品评论为例 需见问题 翻页策略:构造url 常见错误 解决新思路 爬虫的修改

- 函数式编程
- ② 循环的写法
- 3 报错处理

通用爬虫设计 爬虫实例: JD评论爬虫 静态爬虫,以京东的产品价格为例 动态爬虫;以京东的商品评论为例 常见问题 翻页策略:构造url 常见销误 限见的起路 作用现额 形成的

