极客学院
jikexueyuan.com

定向爬虫
Scrapy初探

定向爬虫：Scrapy初探 — **课程概要**

- Scrapy介绍与安装
- Scrapy爬取网页
- Scrapy文件结构
- 实战——豆瓣爬虫

Scrapy介绍与安装

Scrapy介绍与安装

- Scrapy介绍
- Scrapy安装

Scrapy介绍与安装 — **Scrapy介绍**

Scrapy是Python开发的一个快速 web爬虫抓取框架，用于抓取web站点并从页面中提取结构化的数据。Scrapy用途广泛，可以用于数据挖掘、监测和自动化测试。

An open source and collaborative framework for extracting the data you need from websites. In a fast, simple, yet extensible way.

——Scrapy Home Page

Scrapy介绍与安装 — Scrapy安装

- 安装lxml(第三课已经安装)
- 安装zope.interface

  https://pypi.python.org/pypi/zope.interface/4.1.2
- 安装Twisted

  https://pypi.python.org/pypi/Twisted
- 安装pyOpenSSL

  https://pypi.python.org/pypi/pyOpenSSL
- 安装pywin32

  http://sourceforge.net/projects/pywin32/files/pywin32/
- 安装Scrapy（pip install scrapy）

Scrapy爬取网页

Scrapy爬取网页

- Scrapy生成Project
- Scrapy爬取网页

Scrapy爬取网页 — **Scrapy生成Project**

核心代码：

```
scrapy startproject xxx
```

Scrapy爬取网页 — **Scrapy爬取网页**

```
import scrapy
from scrapy.contrib.spiders import CrawlSpider
from scrapy.http import Request
from scrapy.selector import Selector


Xxx = selector.xpath(xxxxx).extract()
```

Scrapy文件结构

Scrapy文件结构

Project中的文件包括:

- items.py
- settings.py
- pipelines.py

Scrapy文件结构 — **items.py**

Items.py定义需要抓取并需要后期处理的数据。

Item objects are simple containers used to collect the scraped data. They provide a dictionary-like API with a convenient syntax for declaring their available fields.

——Scrapy官方手册

Scrapy文件结构 — **settings.py**

settings.py文件配置Scrapy，从而修改user-agent，设定爬取时间间隔，设置代理，配置各种中间件等等。

The Scrapy settings allows you to customize the behaviour of all Scrapy components, including the core, extensions, pipelines and spiders themselves.

——Scrapy官方手册

Scrapy文件结构 — **pipelines.py**

pipeline.py用于存放执行后期数据处理的功能，从而使得数据的爬取和处理分开。

After an item has been scraped by a spider, it is sent to the Item Pipeline which process it through several components that are executed sequentially.

——Scrapy官方手册

实战——豆瓣爬虫

实战——豆瓣爬虫

目标网站：豆瓣电影TOP250

目标网址：http://movie.douban.com/top250

目标内容：

豆瓣电影TOP250中的250部电影，具体内容包括：

- 电影名称
- 电影信息
- 电影评分

输出结果：生成csv文件

定向爬虫：Scrapy初探

本套课程中我们学习了Scrapy最简单的一些操作，你应当掌握以下知识：

- 配置Scrapy环境
- 生成Scrapy的Project
- 使用Scrapy爬取信息
- 将爬取结果保存为csv文件

你可以使用Scrapy爬取一些网页。如果想继续提高，你可以继续在极客学院学习《定向爬虫入门》课程。

极客学院

jikexueyuan.com

中国最大的IT职业在线教育平台