极客学院
jikexueyuan.com

定向爬虫
MongoDB与Scrapy

定向爬虫：MongoDB与Scrapy — **课程概要**

- MongoDB介绍与安装
- Python与MongoDB
- Scrapy应用MongoDB
- 实战——小说爬虫

MongoDB介绍与安装

MongoDB介绍与安装

- MongoDB的介绍
- MongoDB的安装
- MongoDB可视化

MongoDB介绍与安装 — **MongoDB的介绍**

MongoDB是一个跨平台的NoSQL，基于Key-Value形式保存数据。其储存格式非常类似于Python的字典，因此用Python操作MongoDB会非常的容易。

MongoDB is an open source, document database designed for ease of development and scaling.

——MongoDB Home Page

MongoDB介绍与安装 — **MongoDB的安装**

- 下载文件：

  https://www.mongodb.org/downloads

- 创建文件夹：

  mkdir data

- 执行命令：

  mongod --dbpath ./data

- 搞不定？

  打开极客学院，搜索“MongoDB”

MongoDB介绍与安装 — **MongoDB可视化**

- 打开网址：

  http://www.mongovue.com/

- 下载MongoVUE
- 安装MongoVUE
- 运行MongoVUE

Python与MongoDB

Python与MongoDB

- pymongo的安装
- Python 操作MongoDB

Python与MongoDB— **pymongo的安装**

核心命令：

```
pip install pymongo
easy_install pymongo
```

Python与MongoDB — **Python操作MongoDB**

```
import pymongo
connection = pymongo.MongoClient()
tdb = connection.Jikexueyuan
post_info = tdb.test
post_info.insert(xxx)
post_info.remove(xxx)
```

Scrapy应用MongoDB

Scrapy应用MongoDB

- 配置文件的编写
- pipelines的编写

Scrapy应用MongoDB — **配置文件的编写**

在settings.py中配置MongoDB的IP地址、端口号、数据记录名称，可以实现方便的更换MongoDB的数据库信息。

在settings.py中引用pipelines.py从而使pipelines生效。

Scrapy应用MongoDB — **pipelines的编写**

在pipelines中可以像普通Python文件操作MongoDB一样编写代码处理需要保存到MongoDB的数据。然而不同的是这里的数据来自items。这样做的好处是将数据的抓取和处理分开。

实战——小说爬虫

实战——小说爬虫

目标网站：盗墓笔记小说网站

目标网址：http://www.daomubiji.com/

目标内容：

盗墓笔记小说的信息，具体内容包括：

- 书标题
- 章数
- 章标题

输出结果保存到MongoDB中。

定向爬虫：MongoDB与Scrapy

本套课程中我们学习了MongoDB在定向爬虫中的应用，你应当掌握以下知识：

- 安装和使用MongoDB
- 能在可视化界面中查看MongoDB的信息
- 能够操作使用Python读写MongoDB
- Scrapy中将爬取结果保存到MongoDB中

你可以使用MongoDB保存一些其他内容。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