**1. 项目概述**

* **项目名称**：Scrapy Dushu 爬虫项目
* **项目简介**：该项目是一个基于 Scrapy 的爬虫，用于从 www.dushu.com 网站上爬取书籍信息，并将数据存储到 MySQL 数据库中。项目旨在收集特定分类的书籍数据，包括书名、作者和封面图片等信息。

**2. 功能说明**

* **爬取书籍信息**：爬虫从指定的分类页面开始，逐页抓取书籍列表中的信息。
* **数据存储**：抓取到的数据通过 MysqlPipeline 被存储到 MySQL 数据库中的 books 表。
* **并发请求控制**：通过 Scrapy 的配置文件可以设置并发请求数和下载延迟，以避免过度请求服务器。

**3. 依赖说明**

* **Python 版本**：建议使用 Python 3.7 及以上版本
* **依赖库**：
  + Scrapy：用于构建爬虫框架
  + PyMySQL：用于与 MySQL 数据库交互
* **安装依赖**： 使用 pip 安装依赖：

pip install scrapy pymysql

**4. 安装说明**

* **步骤 1**：克隆项目代码到本地

git clone https://github.com/yourusername/scrapy\_dushu.git

cd scrapy\_dushu

**步骤 2**：安装项目依赖

pip install -r requirements.txt

**步骤 3**：配置 MySQL 数据库

确保 MySQL 服务已启动，并创建数据库 spider：

CREATE DATABASE spider;

在项目的 settings.py 文件中，设置数据库连接信息：

DB\_HOST = 'localhost'

DB\_PORT = 3306

DB\_USER = 'root'

DB\_PASSWORD = 'leo'

DB\_NAME = 'spider'

DB\_CHARSET = 'utf8'

创建 books 表：

CREATE TABLE books (

id INT AUTO\_INCREMENT PRIMARY KEY,

title VARCHAR(255),

author VARCHAR(255),

src TEXT

);

5. **使用说明**

**运行爬虫**： 在项目根目录下执行以下命令启动爬虫：

scrapy crawl read

**查看抓取数据**： 爬虫运行后，可以通过 MySQL 客户端查看抓取到的数据：

SELECT \* FROM books;

**6. 常见问题**

* **问题 1**：爬虫无法启动或连接数据库
  + **解决方案**：检查 MySQL 服务是否启动，确保 settings.py 中的数据库配置正确。
* **问题 2**：抓取到的数据为空
  + **解决方案**：检查爬虫的 start\_urls 和 rules 是否正确配置，确保目标页面有可抓取的数据。