学号 专业 姓名

实验日期 教师签字 成绩

实验报告

【实验名称】 开发者社区数据采集

【实验目的】

掌握基于虚拟机的Python数据采集环境搭建

了解开发者社区Segmentfault主要功能

掌握HTML基础知识及XPath语法

掌握Python数据库操作方法

掌握Scrapy爬虫框架、爬取策略设计、数据存储表设计

【实验原理】

【实验内容】

**题目**：使用Scrapy实现一个开发者社区爬虫，爬取开发者社区用户、资源等信息，并存储到MySQL数据库中。

**要求**：1.搭建开发环境，包括安装Python解释器和Scrapy（推荐使用Anaconda[[1]](#footnote-1)）、MySQL数据库[[2]](#footnote-2)、可视化数据库管理工具（推荐使用HediSQL[[3]](#footnote-3)）；

2.分析Segmentfault[[4]](#footnote-4)社区的主要功能，为需要爬取的信息设计存储数据表；

3.设计爬取策略流程图，给出爬取的初始页面，待爬取URL抽取方式等信息；

4.开发Scrapy爬虫，爬取相关数据，存储到MySQL数据库。

**提高（选做）**：考虑多台机器并行爬取的设计策略，例如使用数据库集中管理，爬虫分布式部署，统一从数据库获取待爬取URL，数据统一存储。

**实验结果（实验步骤及相关核心代码）**：

1. https://www.anaconda.com/distribution/ [↑](#footnote-ref-1)
2. https://dev.mysql.com/downloads/installer/ [↑](#footnote-ref-2)
3. https://www.heidisql.com/download.php [↑](#footnote-ref-3)
4. https://segmentfault.com/ [↑](#footnote-ref-4)