数据采集与智能问答系统

本次数据采集课程设计基于数据采集课程大作业，要求结合大语言模型完成自选数据采集与问答系统任务。课设规定1人1组，独立完成一个数据采集与智能问答系统。

# 一、开发工具与环境

* 编程语言：Python3.x
* 数据采集框架：Scrapy（2.x以上）
* 并发优化：Scrapy内置并发机制（Twisted异步网络库）及Scrapy-Redis分布式架构
* 数据存储与管理：MongoDB（模拟分布式大数据存储环境）或ElasticSearch（实现类搜索引擎实时查询）
* 大语言模型（LLM）接口：OpenAI API（如GPT系列）、LangChain框架辅助实现复杂链式问答功能
* 数据分析与可视化：Pandas、Numpy、Matplotlib、Seaborn、Plotly

**三、系统目标**  
搭建一个完善的自动化、智能化数据处理系统，实现“一键启动”，自动化爬取目标网站或数据源信息，采集、清洗与存储数据，融入大语言模型（LLM）完成智能分析与问答交互，提供数据结果的可视化展示，并以用户友好的交互式界面呈现。

**四、本次课程设计相比之前大作业的细化改进点**

* 融入大语言模型（如OpenAI GPT等）实现智能数据分析和自然语言问答功能；
* 利用Scrapy框架实现高效数据采集，并融入并发分布式技术（如Scrapy-Redis）；
* 改进数据存储结构，使用分布式存储系统模拟（MongoDB集群或Elasticsearch分片集群）；
* 增加智能可视化分析模块，数据可视化更加具体丰富；
* 增强系统的一键启动，一次性完成数据采集、分析、存储、问答任务；

**五、具体实现细节**  
（一）一键启动机制

* 编写统一的入口脚本（[例如run.py](http://xn--run-uc0ep96b.py/)），实现系统启动后，自动运行以下步骤： step1: 启动Scrapy爬虫采集数据； step2: 数据处理：数据清洗与字段提取、结构化存储； step3: 启动数据分析及可视化模块； step4: 启动LLM 模块与用户交互问答接口；
* 使用logging统一日志记录系统启动后的每个步骤运行状态。

（二）数据爬取与并发机制

* 使用Scrapy框架快速高效爬取目标网站数据；
* 设置合理rules规则，完成自动翻页与深度爬取；
* 使用Scrapy的内置并发设置（例如CONCURRENT\_REQUESTS）优化并发；
* 可选扩展（优秀加分）：搭载Scrapy-Redis实现分布式爬虫，模拟分布式集群爬取。

（三）大数据存储方案（模拟大型存储系统）

* 模式一：MongoDB 分布式存储  
  配置MongoDB副本集或者分片（sharding）模拟分布式存储环境，存储Scrapy采集的数据。
* 模式二（推荐）：Elasticsearch分布式实时存储与检索  
  利用Elasticsearch的索引和分片功能，实现稳定、高效的实时存储、检索与分析查询，尤其适合后续问答查询与交互分析功能实现。

（四）接入大语言模型融合智能分析与问答模块

* 使用OpenAI API或本地大语言模型接口（如Llama2、ChatGPT API）；
* 构建LangChain框架作为智能问答的辅助工具链，实现数据检索与智能回答的高效整合；
* 结合爬取的数据，通过Embedding技术生成向量存储；构建向量数据库（例如FAISS或ChromaDB），实现高效的数据召回和语义搜索；
* 用户可通过命令行或Web API向系统提出自然语言问题，系统快速查找并且调用LLM进行智能分析，返回自然语言答案。

（五）数据分析与智能可视化

* 统计分析模块：使用Pandas库清洗和分析数据，提供数据统计分析；
* 数据分析报告生成：自动生成基本统计描述、趋势分析、数据分布等；
* 数据可视化模块： |- Matplotlib & Seaborn实现静态图表（直方图、折线图、散点图） |- Plotly实现交互式动态数据可视化，展示结果的Web Dashboard页面

**六、系统整体流程框架示意图**

数据采集与分析问答智能系统 ├─数据采集（Scrapy框架 + Scrapy-Redis并发分布式） │ ├─爬虫启动 │ ├─网页抓取 │ └─数据抽取（解析） │ ├─数据清洗与预处理 │ ├─格式转换与清洗（Pandas/Numpy） │ └─数据去重与存储优化 │ ├─数据存储（模拟大数据存储） │ ├─MongoDB分布式存储 或 Elasticsearch │ └─向量数据库（FAISS/ChromaDB） │ ├─智能数据分析模块 (大语言模型LLM) │ ├─数据语义检索与知识库构建（LangChain） │ ├─问答交互接口 (LLM API) │ └─自然语言查询与智能应答 │ └─数据分析与可视化展示 ├─数据统计报表 ├─交互式图表 (Matplotlib/Seaborn/Plotly) └─Web Dashboard可视化界面展示

**七、交付物与考核点（答辩与评估）**  
本课设的交付内容包括：

* 系统设计报告（需求分析、方案细化、系统架构、关键技术说明等）
* 源代码与项目说明文档（含数据采集、存储、分析问答功能代码）
* 智能分析与交互问答系统演示
* 数据可视化结果展示

考核与评分依据：

* 一键启动正常运行（基本要求，必要条件）；
* Scrapy框架实现高效爬取，并发性能达到良好；
* 模拟分布式存储结构明确，可扩展性强；
* LLM模块自然语言理解与数据分析能力强，交互友好且准确；
* 数据分析与可视化效果明确丰富，界面美观实用；
* 文档提交全面、规范，现场答辩时能详细说明问题。