文本关键词抽取项目技术方案

一、需求

1. 搭建数据平台，处理&储存订单中的文本数据
2. 知识图谱搭建，分层抽取知识框架中的各级关键词，构建知识图谱

二、 技术选型

1. 数据平台

订单数据以订单号+子编码-文本内容的形式储存，数据索引方式简单，数据内容为长文本。综合考虑数据模型和性能，采用MongoDB储存订单的文本数据。

2. 知识图谱

2.1. 抽取模型

由于无人工标注的数据，前期的抽取使用大模型（LLM）直接做推理。结合成本和后期调整的潜力，使用LLAMA/LLAMA2模型+LORA微调的形式直接做抽取，后续若有来自使用方的反馈输入，可以微调LORA权重提升抽取信息的质量

2.2. 图谱储存

考虑到图谱的应用场景不涉及复杂查询及兼容原始数据，图谱储存同样使用MongoDB

3. 系统服务

服务无高并发等性能要求，使用bottle/Flask等python HTTP框架

三、技术方案

1. 接口设计

1.1. 文档管理

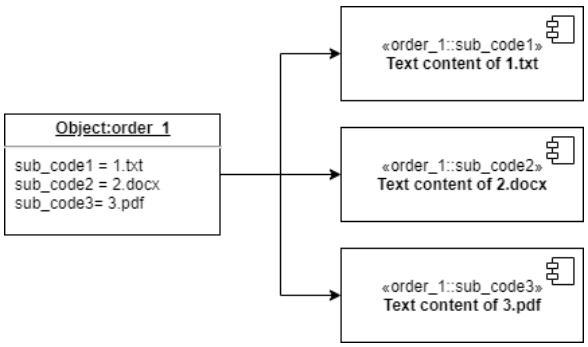
1. 新增订单 `[POST] /api/dm/order/new`
2. 更新订单 `[POST] /api/dm/order/update`
3. 查询订单任务状态 `[GET] /api/dm/order/status/<order_id>`

1.2. 算法

1. 关键词抽取 `[POST] /api/algo/extract`
2. 图谱查询 `[GET] /api/algo/query_kg`

2. 数据库设计

2.1. 原始数据



数据schema

```
{
    "order_id": "string",
    "sub_code": "string",
    "file_name": "string",
    "text": "string"
}
```