## 腾讯课堂爬虫程序设计

#### 需求背景

抓取ke.qq.com的数据，统计每天各个分类的课程数量，提供一个网址，可以查看历史数据

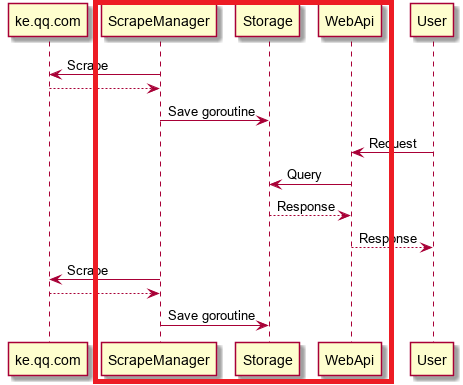
#### 设计思路

|  |  |
| --- | --- |
| 功能点 | 描述 |
| 抓取每日腾讯课堂的所有课程 | 观察腾讯课堂页面，使用go-colly框架实现页面数据抓取功能 |
| 提供数据查询功能，能够查询每天不同类型的课程数量 | 实现restful接口并提供不同类型课程的数量统计和某一类课程下所有课程详情的查询功能 |
| 保存每日腾讯课堂的所有课程，并能查询历史数据 | 通过不同的数据表来保存每日课程，表名为日期 |
| 提供网址能够让其他用户调用restful接口 | 需要提供公网ip或域名让用户访问我的个人电脑 |

#### 设计方案

###### 主要流程

红色方框内为本程序的主要组成部分，主要分为抓取模块，存储模块，Web模块三部分。抓取模块获取到课程信息后，将课程信息组装好放入管道中，数据写入协程监听到管道变化，将管道内容读出并写入数据库。Web模块提供查询接口，并在请求到来时查询数据库。



###### 数据库表设计

数据库选用mysql，设置CourseId为主键，添加CourseType，即课程类型为索引，方便后续查询。提供的信息包括课程Id、课程名称、老师、课程类型、价格

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| CourseId(pri key) | CourseName | Teachers | CourseType(index) | Price |
|  |  |  |  |  |
|  |  |  |  |  |

###### 抓取腾讯课堂页面

通过了解决定使用爬虫框架go-colly来抓取页面，主要爬取如下几个页面：

1. 从https://ke.qq.com/course/list页面开始，先爬取分类页面，观察发现即使点击所有课程中的全部课程，最多也只显示34页的课程，与实际课程数量不符。故决定先爬取所有大类，再进入大类获取大类下的所有小类，直至所有类别均已访问



1. 翻页页面，选取下一页的链接元素，直到无法翻页为止

IMG_256

1. 课程列表页面，获取该页面的所有课程链接，随后访问每一门课程



1. 课程信息页面，获取课程的价格，id，名称，教师等信息



###### 提供restful接口

1. 提供接口能够查询不同分类下的课程数量

可通过直接查询数据库实现每一类课程的数量统计，查询语句设计为：

select CourseType, count(CourseType) as TotalCount from ${tableName} group by CourseType;

通过查询并统计CourseType列的数据得出如下表格，解析后返回给用户

|  |  |
| --- | --- |
| CourseType | TotalCount |
|  |  |

1. 提供接口能够查询某一分类下的所有课程信息

通过如下查询语句实现该功能：

SELECT \* FROM ${tableName} WHERE CourseType=${courseType}

1. 将服务暴露给用户

由于家里的ip地址为非公网ip，需要对ip和端口进行映射，否则无法通过外网访问，需要在网上找一个地址映射工具，使用户可用域名访问我的服务

#### 接口设计

|  |  |  |
| --- | --- | --- |
| 接口功能 | 接口描述 | 备注 |
| 查询不同分类下的课程数量 | /api/v1/coursetype/:data | data为查询日期，也是数据库表名，示例为“2021\_4\_24” |
| 查询某一分类下的所有课程信息 | /api/v1/courseinfo/:data/:type | 1. data为查询日期，也是数据库表名，示例为“2021\_4\_24” 2. type为查询的课程类型，示例为“IT·互联网” |