# 课程设计实验报告

241300009王琳睿

1. **选题描述：**
2. **⽅案设计：**

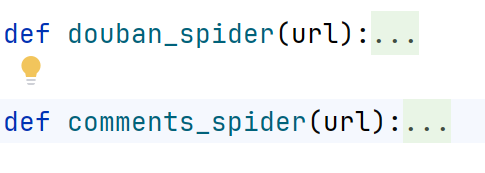
整个功能的实现分成两步，首先获取Top10电影的基本信息，将其存储进SQL数据表movie\_info，再获取这十部电影各自的前60条短评,将其存储进SQL数据表short\_comments。

其中movie\_info的结构为：

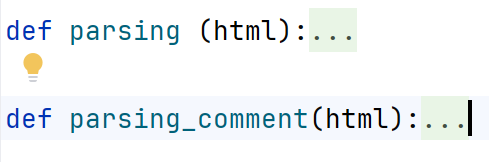
**三、代码模块的功能划分与描述：**

代码分为四个功能模块：

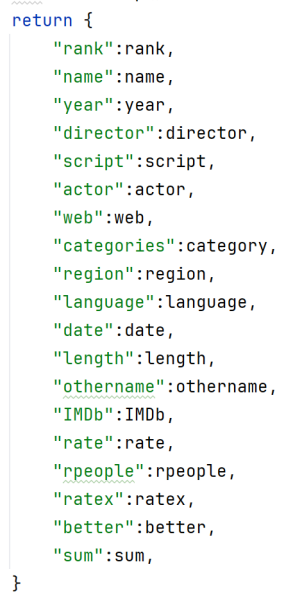
1. 获取页面HTML，2.解析并获取目标信息，3.存储信息，4.主代码（main）
2. 获取页面HTML部分，使用爬虫功能，并增加了反—反爬机制，将获取的HTML转换成字符串类型存储在一个string变量中。



1. 用正则表达式将该string变量的内容进行匹配，提取出需要的信息。



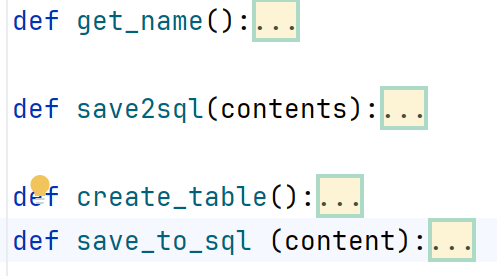
其中parsing函数解析电影信息，返回一个字典



parsing\_comments函数先将20条短评信息分割成列表存储，再依次解析，返回一个字典，字典的values（除name外）为列表：



1. 将获取到的数据存入MySQL



get\_name函数用于获取movie\_info表里的电影名称，并判断当前短评信息属于哪一部电影。

save2sql函数创建movie\_info表并将电影基本信息存入movie\_info表中

create\_table函数用于创建short\_comments表

save\_to\_sql 函数将短评及相关信息存入short\_comments表中

（该数据需多次存入，所以创建表和存数据分开操作）

1. main模块共有两个.py文件，分别为main\_get\_info,用于获取基本信息并存储，遍历10个url，将返回的10个字典存入一个列表，最后一起存入数据表；main\_get\_short\_comments,用于获取短评信息并存储，先创建数据表再遍历30个url,依次将返回的字典存入数据表。
2. **实现效果：**

第一张表中的部分内容



第二张表中的部分内容：

