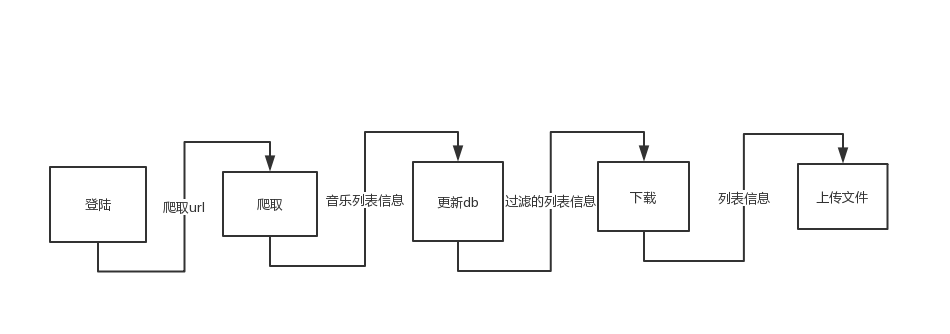
## 猜歌歌曲爬虫

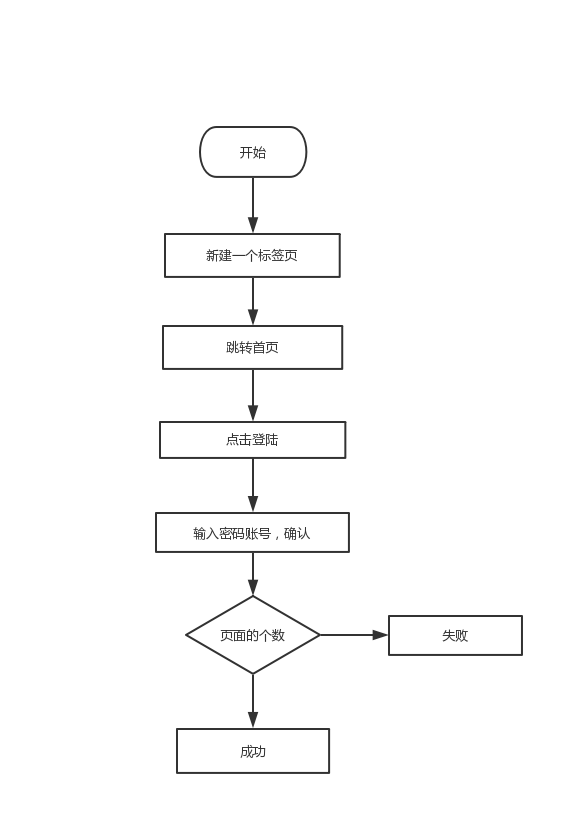
1. 流程控制



分为 登陆,爬取,更新数据库,下载,上传模块5个模块。

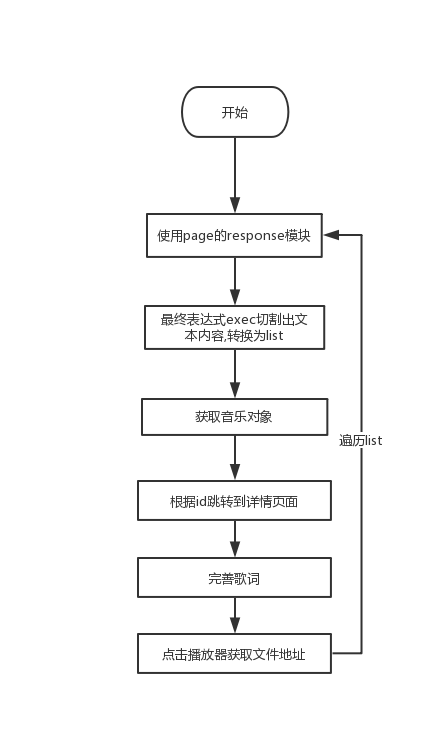
1. 模块设计

1.登陆模块，使用浏览器模拟操作借助于puppeteer模块。



登陆模块可能存在的问题是在登陆次数较多时会出现二维码问题

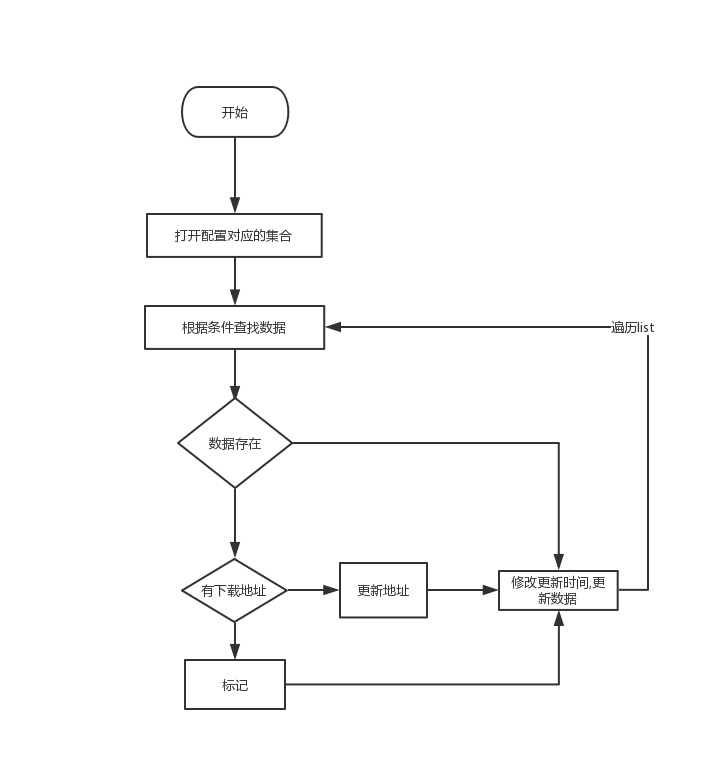
2.爬取模块



1.输入为页面的url 通过爬取list的数据获取id 再根据id 跳转详情页面获取歌唱和文件的真实下载地址，需要注意的是在这里地址有一个过期时间,需要根据id获取文件的真实地址为函数,在下载失败的时候,重新获取文件的地址更新音乐对象的地址。

3.db模块

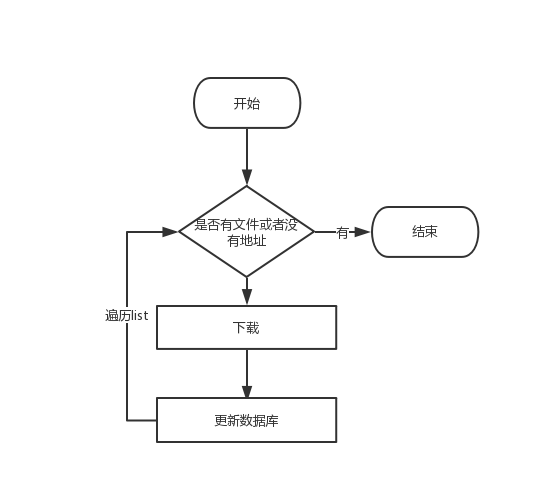
在链接db的时候根据歌曲不同,分为了不同的集合



这里文件检查的顺序是本地,服务器上。如果存在就标示。使用的地方在三个地方,第一个是在爬取完数据的时候需要对数据进行合并,下载完成后需要对数据的本地路径进行更新。首先数据爬取完毕的时候，需要关心的是是否有数据的真实地址。数据下载关心的是数据是否有本地数据的路径。数据上传完成更新的时候需要关心数据库是否有本地数据。

4.数据下载模块。

使用download下载发现问题,链接超时和打开403code。超时的原因是内部代理拦截了请求,使用非公司网络解决,或设置代理。403是由于文件有一个过期时间,那么解决方案是在下载的时候需要一个下载队列,再次请求和更新数据库.



5.数据上传模块

