# <newTermCollector>开发文档

版本修订日志

|  |  |  |
| --- | --- | --- |
| 修订时间 | 修订人 | 修订内容 |
| 2016年10月18日 | 杨恬 | 创建 |
| 2016年10月20日 | 杨恬 | 修改，抛弃了时间戳验证模块 |
| 2016年10月26日 | 杨恬 | 修改，增加了邮件提醒出错功能 |
|  |  |  |

## 1．1 newTermCollector 最新电影电视剧音乐节目爬取程序

1. 程序编号： newTermCollector
2. 程序名称：最新电影电视剧音乐节目爬取程序
3. 程序说明：本程序为最新电影电视剧音乐节目爬取程序，目前是为抓取网络上最新上映的电影，最新发布的歌曲，最新的电视连续剧以及电视节目功能研发，主要研发思路：选取几个权威的娱乐发布网站：时光网，百度排名，QQ音乐，网易音乐，抓取其中最新的电影电视剧音乐以及电视节目名称，包含的程序模块如下：

模块A：抓取网易音乐，QQ音乐最新音乐的功能

模块B：抓取时光上映电影以及上映场次功能

模块C：抓取百度排行榜上电视剧以及节目的排名

1. 程序功能可视化说明：
   1. 对于上映电影，程序访问以下网页：



抓取网页中的电影名称以及上映场数，返回一个json字符串如：

../../屏幕快照%202016-10-18%2014.16.01.png

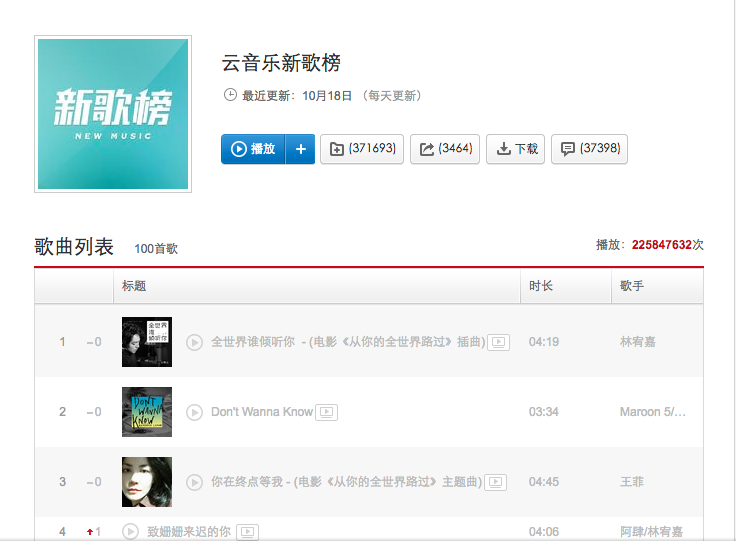
* 1. 对于QQ音乐，程序访问如下网页：



程序抓取网页上各大排行榜的音乐名称，存储于数组之中。

../../屏幕快照%202016-10-18%2014.22.22.png

* 1. 对于网易云音乐，程序访问以下网页：



程序抓取网易云音乐新歌榜的歌曲名称并返回名称数组：

../../屏幕快照%202016-10-18%2014.27.16.png

* 1. 对于电视剧节目，程序访问以下网页：



程序抓取排行榜中的电视剧以及综艺节目名称，返回如下格式的json字符串：

../../屏幕快照%202016-10-18%2014.33.23.png

1. 程序输入与输出说明：

输入：无

输出：输出json格式的结果字符串。如：

音乐：["不息之河 ","A Hard Rain's A-Gonna Fall","呵护","Freal Luv","Don't Wanna Know","PPAP ","Knockin' On Heaven's Door","董小姐","Million Reasons","你的美丽","安和桥","斑马，斑马","Faded","悟空 ","你会爱我到什么时候","球球","Somewhere Now","漂洋过海来看你","可惜没如果 ","丑八怪 ","刚刚好","Home","从你的全世界路过","演员","离开前叫醒我","牛仔很忙 ","彩虹 ","绅士","Make You Feel My Love","Sugar "]

电视剧&节目：{"keywords" : ["诛仙青云志","麻雀","胭脂","老九门","微微一笑很倾城","择天记","步步惊心：丽","笑傲江湖","欢喜密探","幻城","琅琊榜","画江湖之不良人","好家伙","小别离","长征","双刺","湄公河大案","新白娘子传奇","中国式关系","使徒行者","THE K2","甄嬛传","盗墓笔记","兰陵王妃","伪装者","武林外传","平凡的世界","不良人","请回答1988","遥远的距离","白蛇传","世界上的另一个我","芈月传","欢乐颂","旋风少女第二季","青青草","STB超级教师","锻刀","余罪","定制幸福","灰姑娘与四骑士","嫉妒的化身","仙剑奇侠传3","百思不得姐","太太万岁","花千骨","秦时明月","越狱","养个孩子不容易","大嫁风尚"],"nature":"电视剧名"}

电影：[{"freq":"9999","name":"圆梦巨人"},{"freq":"1727","name":"湄公河行动"},{"freq":"805","name":"从你的全世界路过"},{"freq":"451","name":"暗杀游戏"},{"freq":"291","name":"勇士"},{"freq":"220","name":"黑处有什么"},{"freq":"183","name":"王牌逗王牌"},{"freq":"150","name":"爵迹"},{"freq":"118","name":"宾虚"},{"freq":"90","name":"鲁滨逊漂流记"},{"freq":"37","name":"地球四季"},{"freq":"35","name":"逗鸟外传：萌宝满天飞"},{"freq":"23","name":"七月与安生"},{"freq":"22","name":"铠甲勇士捕王"},{"freq":"11","name":"星际迷航3：超越星辰"},{"freq":"11","name":"幽灵医院"},{"freq":"6","name":"凶手还未睡"},{"freq":"5","name":"爱神箭"},{"freq":"3","name":"疯狂丑小鸭"},{"freq":"2","name":"勇士"},{"freq":"2","name":"冰川时代5：星际碰撞"},{"freq":"2","name":"太阳河"},{"freq":"2","name":"四渡赤水"},{"freq":"2","name":"遵义会议"},{"freq":"2","name":"樱桃小丸子 来自意大利的少年"},{"freq":"2","name":"叶问2：宗师传奇"},{"freq":"1","name":"同桌的你"},{"freq":"1","name":"十万个冷笑话"},{"freq":"1","name":"小熊的夏天"},{"freq":"1","name":"心语阳光"},{"freq":"1","name":"HOLD住爱"},{"freq":"1","name":"龙骑侠"},{"freq":"1","name":"既然青春留不住"},{"freq":"1","name":"我是哪吒"},{"freq":"1","name":"爱上处女座"},{"freq":"1","name":"三个孬家伙"},{"freq":"1","name":"心灵解码"},{"freq":"1","name":"凤凰谷"},{"freq":"1","name":"美人鱼之海盗来袭"},{"freq":"1","name":"迁徙的鸟"},{"freq":"1","name":"那年我对你的承诺"},{"freq":"1","name":"惊天破"},{"freq":"1","name":"驴得水"},{"freq":"1","name":"机械师2：复活"},{"freq":"1","name":"侠探杰克：永不回头"},{"freq":"1","name":"龙珠Z：复活的弗利萨"},{"freq":"1","name":"热血雷锋侠之激情营救"},{"freq":"1","name":"路边野餐"},{"freq":"1","name":"盛先生的花儿"},{"freq":"1","name":"但丁密码"}]

1. 处理过程流程图：

../../../Documents/未命名.pdf

1. 开放接口（若被调用，需要提供开放的接口说明）：

无

1. 引入的包及接口：

bs4&BeautifulSoup: 实用的html解析模块

scrapy：成熟的爬虫模块，运用其中的Selector选择器定位目标元素

selenium：配合phantomjs实现伪浏览器操作

retrying: 出错重试模块

开发分支的地址:<https://123.56.224.148:8888>

测试分支的地址:<https://60.205.6.30:8888>

正式分支的地址:<https://60.205.146.52:8888>

1. 程序源码目录结构说明：

newTermCollector2/

├── \_\_configure\_\_

│   ├── ConfigureParser.py

│   ├── \_\_init\_\_.py

│   └── configure.xml

├── \_\_doc\_\_

│   └── newTermCollection开发文档.docx

├── \_\_init\_\_.py

├── \_\_spider\_\_

│   ├── \_\_init\_\_.py

│   ├── netEaseMusic.py

│   ├── spidermovie.py

│   ├── spidermusic.py

│   └── spiderseries.py

├── \_\_util\_\_

│   ├── \_\_init\_\_.py

│   └── library.py

└── main.py

│

└──run.sh

\_\_configure\_\_: 包含配置文件configure.xml以及配置文件解析类ConfigureParser

\_\_doc\_\_: 包含程序说明文档

\_\_spider\_\_：包含四个抓取程序的代码

\_\_util\_\_: 包含一些常用的函数如md5加密等

resource: 包含phantomjs可执行文件

run.sh 启动文件

1. 程序启动方式说明

执行：chmod +x run.sh; ./run.sh

1. 程序配置文件及说明

配置文件：configure.xml

post\_base: 数据传输的接口，以逗号分隔

update\_duration: 程序更新运行的周期（默认一天运行一次）

url：抓取数据的接口网址

element: 目标数据所在的元素，用于xpath定位

attribute：目标元素拥有的属性名，用于xpath定位

value: 该属性的值，用于xpath定位

post\_url: 和post\_base组成完整的数据传输接口

crawl\_type: 区分电视剧抓取和节目抓取

executable\_path: phantomjs存在路径

last\_timestamp:上次运行的时间戳，用于与接口对比验证删除操作是否合法（已舍弃）

admin: 管理员邮箱，用于出错时发邮件

1. 部署程序打包方式：

无

1. 程序部署说明：

将文件放置于服务器，执行run.sh即可

1. 程序潜在的问题，推荐解决方案及其他特殊说明

由于抓取量较小，程序没有考虑被封禁的问题。不过程序加入了retry模块可以在出错的时候重试。

Resource中的phantomjs可能会报wrong exec format错误，解决办法只有安装phantomjs然后在配置文件中加入phantomjs路径。

1. 版本号和CHANGES

newTermCollector 1.0