[教学贴]dlsite相关脚本（python，powershell）

内容关键词： 乱码问题解决，一行代码获取你硬盘上所有音声的rj编号（powershell，win8以上自带），爬取dlsite作品信息程序（python） 适合有较多日本音声资源的人观看

最近从网上下载了很多同人音声相关的合集。 加上我以前下载的一些，总共1tb。整理这些资源，解压然后重新打包存到自己的硬盘花了好几天时间。

网上的资源比较杂乱，存在以下几个点：

1. 文件命名，个人比较喜欢 rjid+ 空格 + 作品标题的形式，最好再加上cv，这样自己用的时候，rjid可以用来排序，也可以上dlsite查询相关作品，是作品的唯一标识，作品标题可以帮我们了解作品内容，cv:作为音声作品，cv是很重要的一个标识。
2. 压缩格式，很多zip文件打开文件名是乱码的，这是因为在日本打包成zip的音声作品，使用的是日文字符编码，我们如果在国内的系统上打开就会乱码。使用le之类的转区工具打开可以解决这个问题

   另外rar，7z格式支持utf8编码，所以不会有乱码问题。7z打包太慢，所以我一般不使用，rar除了打包更快一些，还支持添加一部分冗余信息，用做恢复记录，比较适合在网络中传输。网上有大量的资源都是rar压缩，

   比如说大部分gal的放流

   所以我压缩统一使用rar压缩，而且默认添加3%冗余信息

   另外文件过大的时候最好分卷压缩：

   1. 百度云由于很大一部分流量靠p2p下载，分成多个文件可以提高整体的下载速度
   2. 多个文件下载的包出错，可以重新下载对应的part，单个大文件你就只能重新下载

   网上下载的很多都是文件夹套文件夹的套娃形式，比如外层一个rj编号，里面是作品名的一个文件夹，这样我觉得阅览不太方便，我通常会把里面那层拿出来再把标题复制到rj编号后面，搞了100个左右文件我就放弃了。

   乱码问题解决的方法是，使用ConvertZ这个工具，可以对文件名，或者mp3的tap标签转码

   日文编码通常是shift-jis，我通常把它转成utf8

   最后我在整理文件的过程中放弃了，文件名能有个RJ编号，就满意了。

   其实还有个严重的问题，这就是我接下来要说的了

   很多文件都找不到标题，通常我会翻里面的readme文件找标题，

   还有些里面没有readme文件——通过rjid去dlsite上搜索。

   其实光有标题也是不够了，还有声优，以及作品描述的信息是比较关键的。

所以我就写了一个python脚本，爬取dlsite上的相关信息。因为我在分析这个站点的过程中没找到什么可以请求的jsonp或者ajax的接口，信息貌似是后端直接渲染上去的，而且还用的是.net后端

所以这个爬虫是直接下载页面再解析的。

通过一个rjid的列表，构造请求把每个商品详情页下载下来，然后解析每个页面，把所需的信息分别保存到**3个文件**

- catalog.txt文件，简单的目录，仅包含rj编号和所有文件标题
- details.txt文件，除了标题和编号，还有dlsite商品右侧的所有信息cv，厂商，剧本插画之类，还包括商品下面的描述
- infoList.json文件，一般人用不到，保存了列表数据结构，需要的时候可以用程序从里面读入，因为python保存的json里面文字默认是unicode编码的，所以阅览也不方便
- html 目录，一般人用不到，包含所有下载下来的页面，需要时可以重新解析其他数据，以后就不用重复下载了

这个程序需要一个每行一个RJ编号的列表，而且**字母统一需要大写**:

```
PS B:\Downloads\onsen\同人音声\助眠治愈> Get-ChildItem -Recurse | foreach{if($_.Name.C
RJ113566
RJ121353
RJ160464
RJ179870
RJ201564
RJ206132
RJ216099
RJ220854
RJ225200
RJ232583
RJ235435
RJ237824
RJ238860
RJ239241
RJ239461
RJ239824
RJ239937
RJ235431
RJ236747
RJ205446
RJ221743
RJ240488
RJ205233
RJ226622
RJ238959
RJ240010
RJ211552
RJ229357
RJ238940
```

最初我是用python产生当前目录的这种列表，用来做合集其实也够用了。因为那个目录命名比较规律，直接都是rj名。

但是呢，python脚本运行每次要打开命令行敲太麻烦，而且想想写递归和字符串处理都挺麻烦的，干脆就用powershell脚本，获取rj编号的列表，然后再打开python脚本执行。

powershell脚本当时也不太会，所以赶快找点资料学一下。微软一个ai妹子的视频不错，看完第一课（20分钟）我就会打印当前目录的RJ列表了，后面看了powershell中文博客，知道怎么递归查找，这个程序也算是完善了

接下来，教大家的是，一行脚本递归获取当前目录下所有RJ编号的列表

首先要确认powershell的环境是否安装

在win8 以上都会自带powershell3.0以上的版本，linux和其他版本的windows也可以下载。

1. 在当前目录下打开powershell命令行

按住shift+鼠标右键（在你想打开的目录窗口中），就会有相应的选项

2.输入命令

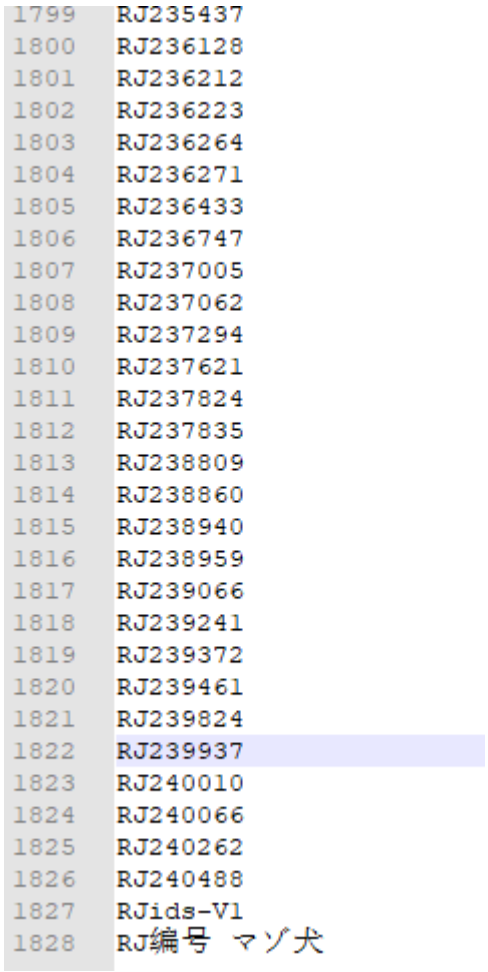
```
# 递归搜索当前目录下所有包含RJ的文件名，去除重复文件名后，保存到rjidList.txt文件中

Get-ChildItem -Recurse | foreach{if($_.Name.Contains("RJ"))
{$_.Name.Substring($_.Name.IndexOf('RJ'),8)}}|select -Unique| Out-File rjidList.txt
```

就可以的到所有RJ编号分行排的txt文件了，而且已经经过去重处理

你也可以修改相关两个"RJ" 字符为其他的字符，统计包含其他字符的列表

把末尾的 rjidList.txt 换成其他名字也可以

去掉`|select -Unique`，再换一个文件名，就可以得到一个不去重的列表，两个对比一下，就能知道你重复了哪些文件



如图，可以知道我总共的音声，有1828个，另一个不去重处理的有2300多个，可知重复了600多个

这里推荐使用notepad++，功能比较强大，这里显示行数，还有我用了文本按行排序的功能。

python脚本可以通过这个列表下载相关信息。已经用pyinstaller打包成exe文件，这样没有python环境的电脑也可以运行。引入了多线程模块，1800多个页面信息总共大小接近200m，2分钟就下载完了

下面放一下这个工具的链接，powershell脚本，和python脚本的exe文件，还有ConvertZ转码工具这三个

使用需求：

1. 有powershell命令行
2. 能翻墙

其余未知，只在自己win10电脑上测过。

使用方式： 1.右键点击powershell脚本，选择运行powershell。

会开启ie代理（命令行或者其他应用程序的流量默认不走ss，需要修改成ie代理），产生列表文件（固定文件名，python脚本要调用），启动dlsiteSpider.exe脚本。

2. 等待执行完成。