# 1.爬虫基础知识学习

## 1.1 beautifulsoup

Beautiful Soup是python的一个库，最主要的功能是从网页抓取数据。官方解释如下：

Beautiful Soup提供一些简单的、python式的函数用来处理导航、搜索、修改分析树等功能。它是一个工具箱，通过解析文档为用户提供需要抓取的数据，因为简单，所以不需要多少代码就可以写出一个完整的应用程序。

Beautiful Soup自动将输入文档转换为Unicode编码，输出文档转换为utf-8编码。你不需要考虑编码方式，除非文档没有指定一个编码方式，这时，Beautiful Soup就不能自动识别编码方式了。然后，你仅仅需要说明一下原始编码方式就可以了。

Beautiful Soup已成为和lxml、html6lib一样出色的python解释器，为用户灵活地提供不同的解析策略或强劲的速度

http://cuiqingcai.com/1319.html

## 1.2 flask

Flask 是一个 Python 实现的 Web 开发微框架，其 WSGI 工具箱采用 Werkzeug ，模板引擎则使用 Jinja2。Flask也被称为 “microframework” ，因为它使用简单的核心，用 extension 增加其他功能。Flask没有默认使用的数据库、窗体验证工具。然而，Flask保留了扩增的弹性，可以用Flask-extension加入这些功能：ORM、窗体验证工具、文件上传、各种开放式身份验证技术。

http://docs.jinkan.org/docs/flask/

http://docs.jinkan.org/docs/flask/quickstart.html

相关知识

1）jinja模板

http://docs.jinkan.org/docs/jinja2/

2）蓝图

http://docs.jinkan.org/docs/flask/blueprints.html

3）表单

http://docs.jinkan.org/docs/flask-wtf/quickstart.html

https://segmentfault.com/a/1190000000997917

https://segmentfault.com/a/1190000002172627

## 1.3 bootstrap

Bootstrap，来自 Twitter，是目前最受欢迎的前端框架。Bootstrap 是基于 HTML、CSS、JAVASCRIPT 的，它简洁灵活，使得 Web 开发更加快捷

http://www.runoob.com/bootstrap/bootstrap-tutorial.html

## 1.4 Flask-SQLAlchemy

数据库操作

http://www.pythondoc.com/flask-sqlalchemy/quickstart.html

https://segmentfault.com/a/1190000004618621

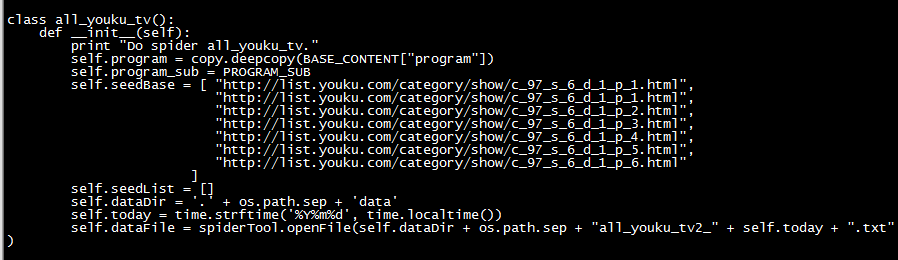
# 2.爬虫模板学习

## 2.1 临时/定时爬取

日常工作维护最多就是临时爬取了，视频网站的结构发生变化或者新增加cp的爬取，首先要做的就是临时爬取和定时爬取了。爬取内容的分类与各视频网站的分类有关，常见的电影、电视剧、综艺、动漫、记录片等。临时爬取基本与定时爬取爬取采用相同模板（部分改变），下面主要模板以定时爬取介绍（临时爬取的差异性再另加说明），核心是三重爬取的过程。

### 2.1.1定义爬取类

主要定义视频网址或视频资源接口网址，以及爬取结果存放的目录文件



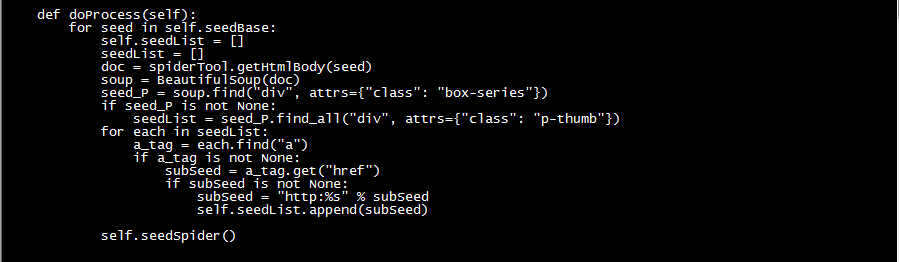
定义爬虫模块类



视频网址示例

### 2.1.2爬取处理函数

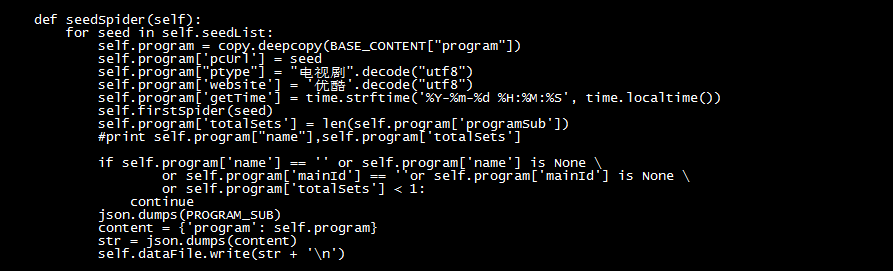
爬取的主处理函数，遍历所有的种子视频网址列表，依次爬取所有的节目集，主要在定时爬取中用到



爬取主处理函数

### 2.1.3第一重爬取

也可叫种子爬取，根据爬取到的种子列表逐次爬取节目

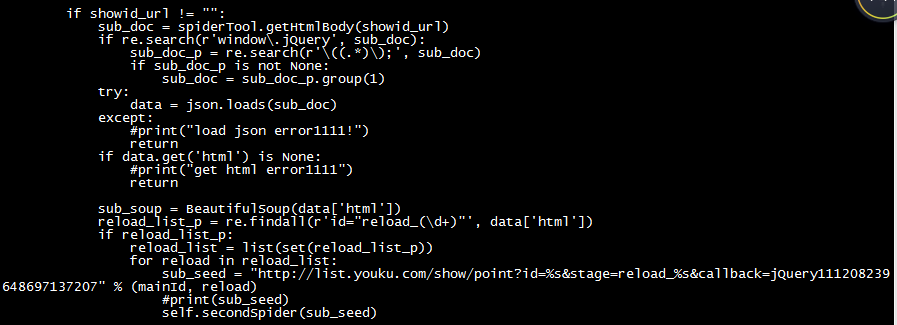


种子爬取

### 2.1.4第二重爬取

爬取节目集，即是种子对应的节目集的爬取，爬取节目对应的证评分、海报、节目名、发行年代、别名、地区、演员、导演、类型、播放次数、简介、mainID等信息，

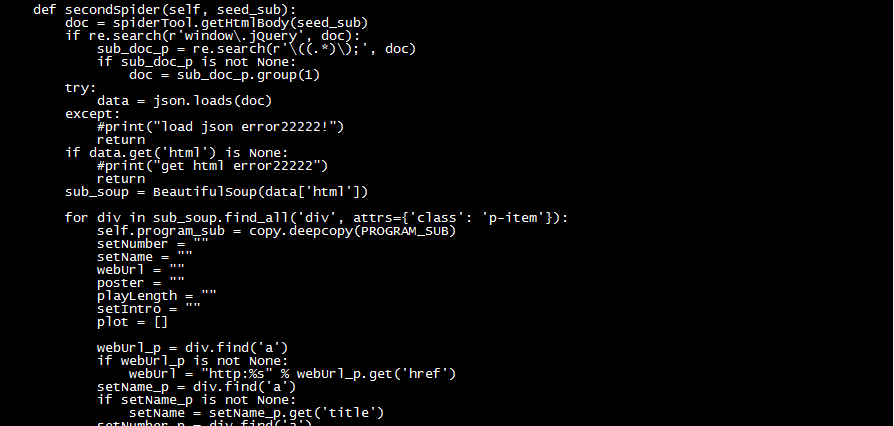




节目爬取

### 2.1.5第三重爬取

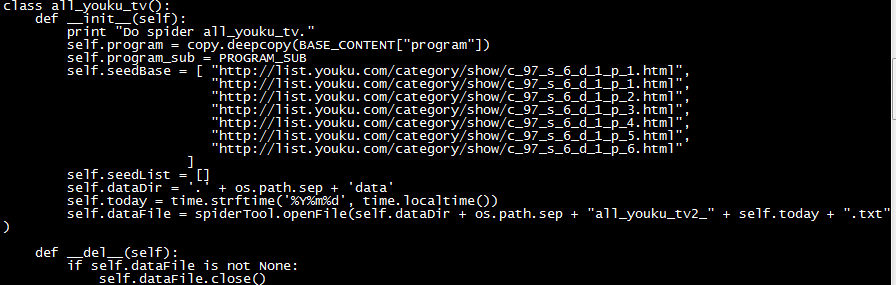
爬取节目集下的所有节目信息，比如爬取电视剧下的所有的集，爬取信息如，集号、节目名、播放地址、海报、播放时长、简介等，其中集号、节目名、播放地址是必不可少的，集号必须是数字，不支持字符

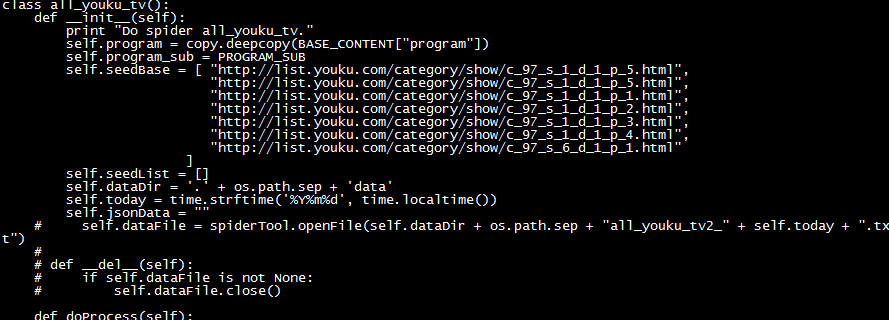


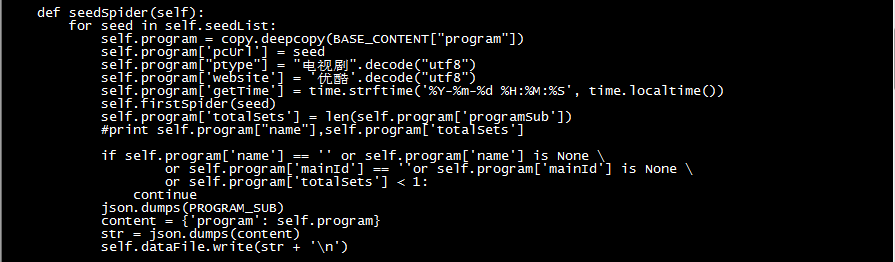
节目剧集爬取

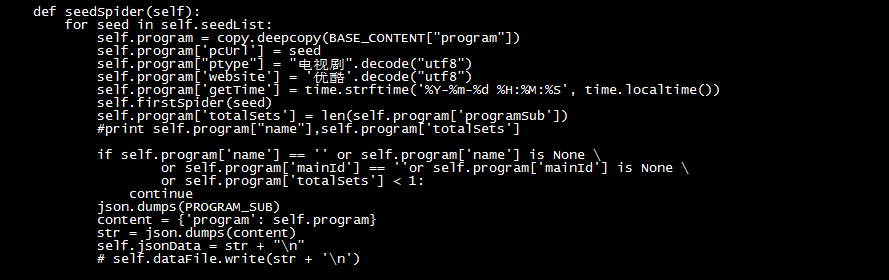
### 2.1.6临时爬取与定时爬取的差异

下图对比中，前面的是定时爬取的，后面是临时爬取的









临时爬取的示例如下：





