基于 Python 的 RSSHub 实现

cscnk52, 2024/12/13

何为 RSS

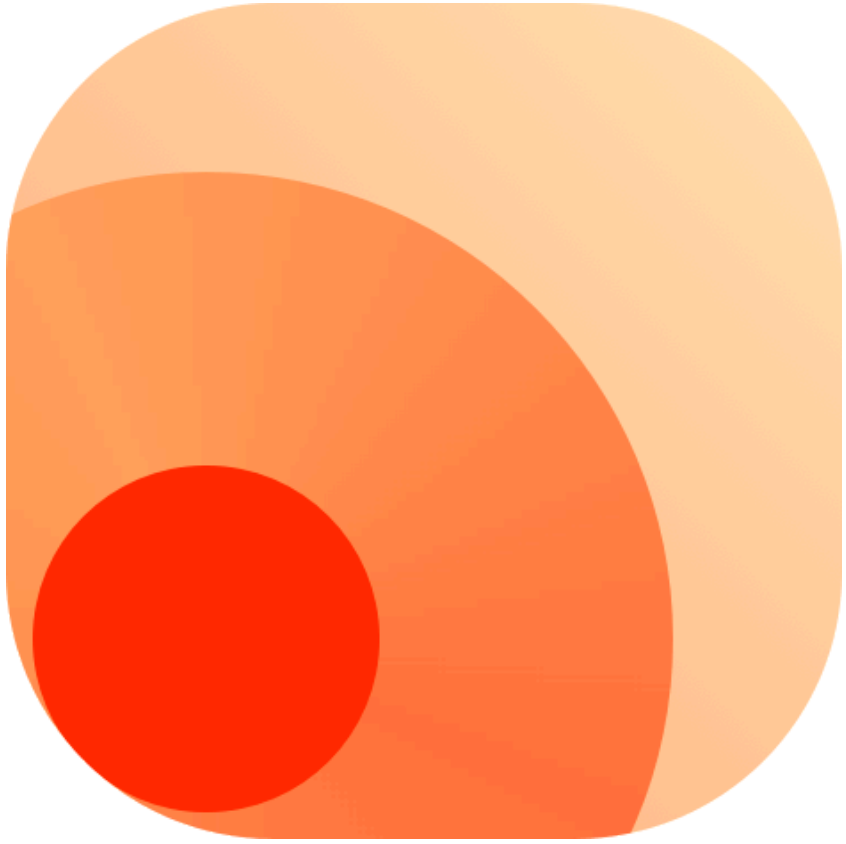
RSS[1]（英文全称：RDF Site Summary 或 Really Simple Syndication），中文译作简易信息聚合，也称聚合内容，是一种消息来源格式规范，用以聚合多个网站更新的内容并自动通知网站订阅者。

使用 RSS 后，网站订阅者便无需再手动查看网站是否有新的内容，同时 RSS 可将多个网站更新的内容进行整合，以摘要的形式呈现，有助于订阅者快速获取重要信息，并选择性地点阅查看。

[1] RSS - 维基百科

何为 RSSHub

RSSHub[1] 是一个开源、简单易用、易于扩展的 RSS 生成器，可以给任何奇奇怪怪的内容生成 RSS 订阅源。RSSHub 借助于开源社区的力量快速发展中，目前已适配数百家网站的上千项内容。



[1] RSSHub - GitHub

代码构成

```
.
├── config.py            # 存储配置
├── database.py          # 数据库操作相关
├── feed_generator.py    # 生成 RSS 文件
├── fetch.py             # 抓取网页内容
├── main.py              # 入口文件
├── test.py              # 测试文件
└── utils.py             # 实用函数
```

以抓取哈理工新闻网下 `理工要闻` 栏目为例，抓取所有新闻条目，然后输出成规范的 RSS 文件，并存入 SQLite 数据库中。

RSS 文件频道规范[1]

Element	Description	Example
title	The name of the channel. It's how people refer to your service. If you have an HTML website that contains the same information as your RSS file, the title of your channel should be the same as the title of your website.	GoUpstate.com News Headlines
link	The URL to the HTML website corresponding to the channel.	http://www.goupstate.com/
description	Phrase or sentence describing the channel.	The latest news from GoUpstate.com, a Spartanburg Herald-Journal Web site.

[1] RSS 2.0 Specification - Required channel elements

RSS 文件条目规范[1]

Element	Description	Example
title	The title of the item.	Venice Film Festival Tries to Quit Sinking
link	The URL of the item.	http://nytimes.com/2004/12/07FEST.html
description	The item synopsis.	<description>Some of the most heated chatter at the Venice Film...</description>
guid	A string that uniquely identifies the item. More	
pubDate	Indicates when the item was published. More.	

[1] RSS 2.0 Specification - Elements of <item>

使用到的第三方库

- black - 用于格式化 Python 代码
- bs4[1] - 用于解析爬取的 HTML 数据
- concurrent - 用于多线程爬取数据
- datetime - 用于处理时间
- feedgen - 生成规范的 RSS 文件
- feedparser - 解析 RSS 文件
- re - 正则表达式用于提取字符串
- sqlite3 - 用于 SQLite 数据库操作
- unittest - 用于单元测试 命令行

[1]Dummy package for Beautiful Soup (beautifulsoup4)

Reference:

- DIYgod. (2024, June 19). The crash and rebirth of a six-year-old open source project. https://diygod.cc/6-year-of-rsshub
- DIYgod. (n.d.). DIYgod/rsshub: ❤️ everything is RSSible. GitHub. https://github.com/DIYgod/RSSHub
- RSS Advisory Board. RSS icon. (n.d.). https://www.rssboard.org/rss-specification
- Wikimedia Foundation. (2024, December 12). RSS. Wikipedia. https://zh.wikipedia.org/wiki/RSS