

**学习强国数据抓取评估**

|  |  |  |  |
| --- | --- | --- | --- |
| 文档版本号： | v 1.0 | 文档编号： |  |
| 文档密级： | 保密 | 归属部门/项目： | 项目开发 |
| 产品名： | XX | 子系统名： |  |
| 编写人： | XX | 编写日期： | 2019-05-06 |

**深圳市茁壮网络股份有限公司 版权所有**

**内部资料注意保密**

**修订记录：**

|  |  |  |  |
| --- | --- | --- | --- |
| **文档版本** | **修订人** | **修订日期** | **修订描述** |
| V 1.0 | hezy | 2019-05-06 | 初稿 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**目录**

[一、 需求说明 4](#_Toc29393)

[二、 素材来源 4](#_Toc18456)

[1、 学习强国所有栏目 4](#_Toc25372)

[三、 网站包含资源类型 5](#_Toc13244)

[1、 文字类资源 5](#_Toc17993)

[2、 视频音频类资源 6](#_Toc1290)

[3、 网页内canvas图片 6](#_Toc6781)

[四、 抓取方案（预定）： 6](#_Toc1583)

# 需求说明

学习强国全网数据爬取评估。

功能1：将爬取到的素材进行归类存储；

功能2：提供数据的接口供终端调用。

# 素材来源

## 学习强国所有栏目

截取网址：<https://www.xuexi.cn/>（学习强国网）

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **栏目：** | **各个子栏目：** | | | | | | | | | | |
| **首页** | 头条 | 重要活动 | 学习新视界 | 人民眼中的习近平 | 世界眼中的习近平 | 习近平同期声 | 习近平足迹 | 习近平日志 | 重要新闻 | 身边的感动 | 时播平台 |
| 宪法法律 | 党规党章 | 学习文件 | 中央政治局集体学习 | 学习时评 | 综合新闻 | 中宣部发布 | 你问我答 | 学习实践 | 新闻发步厅 | 文化进万家 |
| **学习新思想** | 八个明确 | 十四个坚持 | 重要活动 | 重要会议 | 重要讲话 | 重要文章 | 出国访问 | 指示批示 | 函电贺电 | 新时代纪实 |  |
| **十九大时间** | 十九大报告 | 十九大文献 | 十九大报告（分段） | 十九届中央全会 | 十九届中央纪委全会 | 延伸阅读 | 你问我答 | 学习面对面 |  |  |  |
| **学习理论** | 学习理论 |  |  |  |  |  |  |  |  |  |  |
| **红色中国** | 永远的丰碑 | 经典著作 | 红色记忆  （外网） | 文献记录片 | 历届党代会简介 | 党史故事 | 党史知识 | 党史研究 | 中国精神研究 |  |  |
| **学习科学** | 奖励名录 | 两院院士名录 | 科学思想研究 | 科学精神谈 | 国家工程 | 5G | 科技前沿 | 科技奖项 | 当代科学家的故事 | 大国工匠 | 一线风采 |
| 科普视界 | 政策解读 | 科普知识 | 中国科技史 | 中国历代科学家 | 世界科技博览 | 世界科技史 | 科学著作 | 心理辅导 |  |  |
| **环球视野** | 习近平外交思想研究 | 习近平出访足迹 | 世界眼中的习近平 | 一带一路新闻馆 | 一带一路博物馆 | 一带一路国情馆 | 一带一路文明馆 | 世界问题研究 | 世界文明研究 | 看世界 | 经典阅读 |
| **习近平文汇** | 习近平文汇 |  |  |  |  |  |  |  |  |  |  |
| **学习电视台** | 频道、分类 |  |  |  |  |  |  |  |  |  |  |
| **学习慕课** | 分类、课程 |  |  |  |  |  |  |  |  |  |  |
| **学习文化** | 中华文博 | 中华人物 | 中华古籍（外网） | 中国文学 | 中华诗词 | 中国戏曲 | 中国音乐 | 中国美术 | 中国建筑 | 中华武术 | 中华医药 |
| 中国成语 | 中国楹联 | 中国灯谜 |  | （备注：此栏目下各个栏目下还另有多个子栏目） | | | | | | |
| **强军新军** | 习近平强军思想研究 | 强军时评 | 统帅与官兵 | 三军风采 | 军中英模 | 学习军史 | 兵器大观 | 军史著作 | 中国军视（学习慕课中军事频道分类） | | |
| **美丽中国** | 生态文明建设思想研究 | 生态文明建设实践 | 摄影作品选 | 走遍中国 | 记录片 | 景区欣赏 | 历史文化名城 | 历史文化名镇（村） | 历史文化名街区 |  |  |

（注：淡蓝色标注表示可通过https://www.xuexi.cn/a3410a359fd312ba6c84eb77ffc872e8/datae43e220633a65f9b6d8b53712cba9caa.js类似js文件获取数据源；深蓝色标注可通过https://boot-source.xuexi.cn/data/app/11326597159201056043.js类似路径js文件获取数据源；

黄色标注表示为canvas构建图表，一般为pdf样式图）

# 网站包含资源类型

抓取到的素材根据上一章表格中的栏目结构进行分类存储；

## 文字类资源

文字类新闻需要抓取的信息包括：

文章标题（必选）

文章副标题（可选）

正文（必选）

图片链接（可选）

图片说明（可选）

视频链接（可选）

视频说明（可选）

新闻来源（可选）

新闻发表时间（可选）

第几届委员会（可选）

所属专题名称（可选）

## 视频音频类资源

视频类资源需要抓取的信息包括：

视频标题（必选）

视频文件下载（必选）

视频发表时间（必选）

## 网页内canvas图片

一般为pdf文件，例：

https://preview-pdf.xuexi.cn/preview/index.html?url=https%3A%2F%2Fpreview-pdf.xuexi.cn%2F5c248bd0923afa470533632a%2F63et3mvs6wo.pdf&amp;copy=1

# 抓取方案（预）：

该网站中一般网页为div列表标签的无跳转链接，可通过htmlUnit获取标签执行点击操作来获取跳转页面。

1. **图文新闻类资源：**

通过分析强国网各个栏目网页源码可知：图文新闻类资源网页数据源都在js文件上可以获取，

例如：

**例一：**

网页链接：

<https://www.xuexi.cn/4f91dddd8a2e71cee3fa490f5be2dcb3/e43e220633a65f9b6d8b53712cba9caa.html>

相对应的数据源js文件链接：

<https://www.xuexi.cn/4f91dddd8a2e71cee3fa490f5be2dcb3/datae43e220633a65f9b6d8b53712cba9caa.js>

**例二：**

网页链接：

<https://www.xuexi.cn/lgpage/detail/index.html?id=11326597159201056043>

相对应的数据源js文件链接：

<https://boot-source.xuexi.cn/data/app/11326597159201056043.js>

图文类资源数据源只有上面两种类型，各种栏目数据类项都在素材来源栏目中颜色标明了。

1. **视频音频文件：**

视频音频文件都为MP3、MP4后缀的url，都在网页相应数据源js文件中包含。

1. pdf书籍类文件：

此种资源为网页canvas生成的图片，暂没想到方法抓取