

**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](#_Toc23154)

[二、 素材来源 4](#_Toc13517)

[1、 学习强国所有栏目 4](#_Toc25381)

[2、 自治区人大栏目对接 5](#_Toc3668)

[3、 自治区政协栏目对接 5](#_Toc24278)

[4、 智慧党建栏目对接 6](#_Toc22802)

[三、 需要抓取的内容 6](#_Toc16715)

[1、 文字类资源 6](#_Toc30023)

[2、 视频类资源 7](#_Toc32523)

[3、 领导介绍 7](#_Toc28032)

[4、 信件 7](#_Toc18449)

[5、 提案、建议 8](#_Toc11012)

[四、 接口说明 8](#_Toc15596)

# 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：** |  |  |  |  |  |  |  |  |  |  |  |
| **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** |  |  |  |  |  |  |  |  |  |  |  |
| **习近平汇** |  |  |  |  |  |  |  |  |  |  |  |
| **学习电视台** |  |  |  |  |  |  |  |  |  |  |  |
| **学习慕课** |  |  |  |  |  |  |  |  |  |  |  |
| **学习文化** |  |  |  |  |  |  |  |  |  |  |  |
| **强军新军** |  |  |  |  |  |  |  |  |  |  |  |
| **美丽中国** |  |  |  |  |  |  |  |  |  |  |  |

（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，暂无法获取数据源）

## 自治区人大栏目对接

截取网址：<http://www.nxrd.gov.cn/index.html>（[宁夏回族自治区人民代表大会常务委员会](http://www.so.com/link?m=aU5GjZukskJk2r8o0YTVk7Ir%2BjvqALlpgJCEuP%2BLMsvAXtZ92fC1ZG1d7s114KRn3JsVwkU95EfTO8nWa4JYTN%2FZolAhKFb6Kp3TMPyZIwjTIBAhLz9%2Fkcgnmmqs%3D" \t "_blank) ）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **人大概况** | **人大新闻** | **常委会工作** | **工作动态** | **人大专题** |
| 组织机构 |  | 工作计划(ok) | 通知公告（ok） | 全国人代会宁夏代表团(ok) |
| 领导视窗(简历图片问题) |  | 立法动态(ok) | 自身建设(ok) | 自治区人民代表大会会议(ok) |
| 领导讲话（TRS\_Editor） |  | 监督视窗（ok） | 代表工作(ok) | 自治区人大常委会会议 |
|  |  | 重大事项决定（ok） | 市县区人大(ok) | 自治区人大常委会主任会议(ok) |
|  |  | 人事任免（ok） |  | 机关主题教育实践活动 |

## 自治区政协栏目对接

截取网址：<http://www.nxzx.gov.cn/>（宁夏政协网）

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **政协概况** | **政协会议（2017）** | **委员会工作** | **政协工作动态** | **政协新闻** | **政策法规** |
| 组织机构 | 政协全体会议(ok) | 提案委员会(ok) | 工作动态 | 热点专题（ok） | 规章制度 |
| 政协简介（ok） | 政协常委会议（ok） | 经济委员会(ok) | 市县政协 | 各地政协新闻（ok） | 理论研究 |
| 领导讲话（ok） | 政协主席会议（ok） | 人口资源环境委员会(ok) | 委员风采 | 图片动态（ok） | 文史资料 |
| 党派团体（ok） |  | 科教文卫体委员会(ok) | 社情民意 | 书画摄影 |  |
|  |  | 社会和法制委员会(ok) |  |  |  |
|  |  | 民族和宗教委员会(ok) |  |  |  |
|  |  | 文史和学习委员会(ok) |  |  |  |
|  |  | 港澳台侨联络委员会(ok) |  |  |  |

## 智慧党建栏目对接

截取网址：http://www.nxdjw.gov.cn/（宁夏党建网）

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **党建动态** | **公示公告** | **远程教育** | **要闻动态** | **党建工作** | **专题网页** |
| 基层党建 | 公示公告 | 党员教育 | 党建要闻 | 干部工作 | 12380举报网站-去掉 |
| 党建研究 | 任前公示 | 共产党员电视栏目 | 高层声音 | 人才工作 | 组工资讯 |
|  | 人事任免 | 资源中心 | 党建新闻 | 资料下载 | 党史教育网 |
|  |  | 专题专栏 | 图片新闻 | 政策法规 | 反邪教警示教育 |
|  |  | 全国资源 |  | 廉政建设 | 宁夏干部教育培训网络学院 |
|  |  |  |  |  | 三农呼叫中心-去掉 |

# 需要抓取的内容

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）

视频链接（必选）

视频发表时间（必选）

## 领导介绍

领导介绍需要抓取的信息有：

领导姓名（必选）

领导职务（可选）

领导头像（必选）

工作分工（可选）

个人简介（必选）

## 信件

信件类新闻需要抓取的数据有：

来信人（必选）

来信时间（必选）

诉求类型（必选）

所属区域（必选）

来信标题（必选）

来信内容（必选）

回复部门（必选）

回复内容（必选）

回复时间（必选）

## 提案、建议

议案、建议类新闻需要抓取的数据有：

编号（必选）

届次（必选）

类别（必选）

主办（必选）

协办单位（必选）

案由（必选）

正文（必选）

办理情况（必选）

# 接口说明

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 功能 | 接口说明 | 备注 |
|  | 获取素材 | 根据栏目结构、媒资类型返回素材 |  |