孢子V2.3抓取政采招标项目需求文档

**文档版本记录**

|  |  |  |  |
| --- | --- | --- | --- |
| **文档版本号** | **时间** | **修改内容** | **备注** |
| V0.9 | 2018.7.23 | 创建文档 |  |
| V0.9.1 | 20180.9.25 | 修改文档 |  |
|  |  |  |  |
|  |  |  |  |

# 1.项目背景

随着孢子平台不断发展壮大，开发商不断入驻，但项目远远供不应求。

开源节流思想，聚合政府采购项目需求信息。项目需求扩量，满足更多平台用户。（**政采-政府采购简称）**

# 2.项目目标：

1.**自动化抓取政采项目**

政采项目来源包括：国家级网站、地方级网站、垂直行业网站、平台型网站等几十个来源网站。

计划一期完成5个站点抓取：国家级、地方级站点；

2.**版本计划时间**：2018.9.25-2018.9.26

# 3.业务流程图



# 4.产品功能矩阵

|  |  |  |  |
| --- | --- | --- | --- |
| **模块** | **功能点** | **简要描述** | **备注** |
| 爬虫设计 | 爬虫功能设计 | 自动化抓取、关键词搜索、防屏蔽等 | 1.技术方案看是否有现成的参考，功能一开始不需要完美，都可以简化，跑通流程即可； |
| 项目去重 | 抓取项目去重 | 抓起来源较多，需要对抓取的项目进行去重 |  |
| 项目抓取 | 抓取项目字段信息 | 抓取项目核心字段信息 |  |
| 项目入库 | 项目入库功能 | 抓取的项目，直接通过模拟用户账户，发布到孢子后台，方便运营对项目进行审核审核 | 所有项目，都使用账户- Gauin 模拟发布到该账户下。 |

# 5.需求说明

## 5.1目标网站

### 5.1.1 抓取目标网站汇总

|  |  |  |  |
| --- | --- | --- | --- |
| **网站类型** | **网站名称** | **网站地址** | **备注** |
| 国家招标网站 | 中国政府采购网 | http://search.ccgp.gov.cn/bxsearch | 1.项目列表有类型字段：我们只抓取两个类型下的项目：公开招标、询价公告  2.该网站支持搜索关键词全文匹配搜项目 |
| 地方招标网站 | 浙江政府采购网 | http://www.zjzfcg.gov.cn/  purchaseNotice/index.html?\_=1532361561939 | 1.项目左侧列表有公告类型，我们只抓取的类型：采购公告 |
| 河北省政府采购网 | <http://www.ccgp-hebei.gov.cn/>  province/cggg/zbgg/ | 1.项目左侧列表有采购公告，我们只抓取的类型：招标公告  2.该网站没有标题搜索功能，只能通过翻页抓取，匹配标题； |
|  | 湖南省政府采购网 | <http://www.ccgp-hunan.gov.cn>  /page/notice/more.jsp# | 1.项目左侧列表有公告类型，我们只抓取的类型：采购公告  2.该网站可以检索公告标题； |
| 广东政府采购网 | <http://www.gpcgd.com/gpcgd/portal/>  portal-news!list?portalNews.typeId=40011 | 1.项目左侧列表有采购信息，我们只抓取的类型：采购信息公告  2.该网站可以检索公告标题； |

### 5.1.2 爬虫设计要点

|  |  |  |
| --- | --- | --- |
| **设计要点** | **说明** | **备注** |
| 抓取频率 | 30分钟队列抓取一次 | 具体抓取频率根据机器性能等确定 |
| 抓取队列 | 根据频率每次重新抓取，需要记录上次抓取到的记录，以此确定这次抓取的终止点 |  |

## 5.2 关键词捞项目

### 5.2.1关键词

**需求说明：**

1. 抓取关键词，需要是一张表配置，方便后续增加词库；

2.根据项目关键词抓取到的项目，在入库的时候，自动归到对应所属领域分类中；

|  |  |
| --- | --- |
| **项目关键词** | **所属领域** |
| VR、虚拟现实、虚拟  航拍、全景、摄影、实景、漫游、无人机  公关传媒、媒体公关、活动策划、公关营销  3D、模型、建模、CG、Maya、3dMax、动画、美术、大空间行走、行为捕捉、动作捕捉、体感互动、Kinect  U3D、Unity、UE4、虚幻引擎  AR、增强现实、增强  OpenCV、SLAM、ARKit、ARCore、Metaio美桃、vision based、Vuforia、EasyAR、ARToolKit、QQ AR、Wikitude、HiAR、VoidAR  MR、混合现实、全息、投影  HTC Vive、Oculus Rift、Gear、PSVR、体验店、体验馆、DK2、大朋、Matterport  Hololens、联想MR、Magic leap、Google Glass | ARVR |
| 人工智能、智能、AI、机器视觉、语音识别、深度学习、神经网络、自然语音处理、智慧旅游、智能音箱  TensorFlow、PaddlePaddle  智慧城市、特色小镇、物联网、IOT、智能家居、智慧家居、智能家电  工业制造、工业4.0、智能制造  智慧教育、智慧金融  新能源、能源矿产  智能汽车、无人驾驶、无人车  新零售、智能营销 | 人工智能 |
| 可穿戴设备、虚拟助手、智能医疗、智慧医疗、智能健康、医学影像  智能医疗、智能血糖仪、智能血压计、智能心电仪、智能温度计、智能秤、智能按摩器  智能手表、智能手环 | 智能医疗 |
| 机器人、工业机器人、服务机器人、教育机器人、水下机器人、建筑机器人、安全营救机器人、消防机器人  工业自动化、发那科 | 机器人 |
| 3D打印 | 3D打印 |

### 5.2.2 抓取去重设计

**☆需求说明：**

1.抓取的项目，在入库的时候，需要和已抓取项目标题进行比对，完全一样则抛弃；

2.该功能如果开发成本较高，则一期不做此功能；

## 5.3 抓取核心信息和孢子项目入库对应字段

|  |  |  |
| --- | --- | --- |
| **抓取字段** | **孢子项目库字段** | **备注说明** |
|  | 项目类型 | 默认对应-政采项目 |
| 项目标题 | 项目标题 |  |
|  | 所属领域 | 根据抓取用到的关键词-  对应的所属领域关系归类 |
| 项目描述 | 项目描述 |  |
|  | 项目预算 | 默认预设：100000，运营同学会自己修改 |
|  | 完成时间 | 和发布时间一致，待运营同学自己修改 |
|  | 交付方式 | 默认预设为： 沟通确定 |
|  | 发布人 | 默认显示 xx先生，待运营同学自己修改  所有项目，都使用账户- Gauin 模拟发布到该账户下。 |
|  | 联系方式 | 默认预设 微信-xxxx。注意这里是两个字段的值。待运营同学修改 |
| 发布时间 | 发布时间 |  |
| 来源链接 | 来源链接 | 该字段为新增字段  只针对 政采项目增加该字段，前端不展示 |