公文知识图谱系统使用说明

2022年 08 月 25日

**文档历史**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **版本** | **作者** | | **修改内容** | | | **更新时间** | |
| V0.1 | 周文博 | | 新建 | | | 2022/08/10 | |
| V0.2 | 周文博 | | 完善配置环境和使用说明 | | | 2022/08/23 | |
| V1.0 | 李炜卓 | | 完善系统使用说明 | | | 2022/08/25 | |
|  |  | |  | | |  | |
|  |  | |  | | |  | |
|  |  | |  | | |  | |
|  |  | |  | | |  | |
|  |  | |  | | |  | |
|  |  | |  | | |  | |
|  |  | |  | | |  | |
|  |  | |  | | |  | |
|  |  | |  | | |  | |
|  |  | |  | | |  | |
|  |  | |  | | |  | |
|  |  | |  | | |  | |
|  |  | |  | | |  | |
|  |  | |  | | |  | |
|  |  | |  | | |  | |
|  |  | |  | | |  | |
|  |  | |  | | |  | |
|  |  | |  | | |  | |
|  |  | |  | | |  | |
| **版本号** | | | **更新内容** | | **编制人** | **确认人** | | **日期** | |
| V0.1 | | | 起草 | | 周文博 | 李炜卓 | | 2022/08/10 | |
| V0.2 | | | 修改 | | 周文博 | 李炜卓 | | 2022/08/23 | |
| V1.0 | | | 修改 | | 周文博 | 李炜卓 | | 2022/08/25 | |
|  | | |  | |  |  | |  | |
|  | | |  | |  |  | |  | |
|  | | |  | |  |  | |  | |
|  | | |  | |  |  | |  | |
|  | | |  | |  |  | |  | |
|  | | |  | |  |  | |  | |
|  | | |  | |  |  | |  | |
|  | | |  | |  |  | |  | |
|  | | |  | |  |  | |  | |
|  | | |  | |  |  | |  | |
|  | | |  | |  |  | |  | |
|  | | |  | |  |  | |  | |
|  | | |  | |  |  | |  | |
|  | | |  | |  |  | |  | |
|  | | |  | |  |  | |  | |

目 录

[1引言 1](#_Toc112318666)

[1.1系统总体需求 1](#_Toc112318667)

[1.2范围 1](#_Toc112318668)

[2系统架构 1](#_Toc112318669)

[3主要功能 3](#_Toc112318670)

[4运行条件 4](#_Toc112318671)

[4.1硬件运行条件 4](#_Toc112318672)

[4.2软件运行条件 4](#_Toc112318673)

[4.3部分软件安装说明 4](#_Toc112318674)

[4.4安装项目环境 5](#_Toc112318675)

[4.5测试 5](#_Toc112318676)

[5系统主要功能 6](#_Toc112318677)

[5.1系统功能概述 6](#_Toc112318678)

[5.2系统模块设计 6](#_Toc112318679)

[5.2.1归档模块 6](#_Toc112318680)

[5.2.2 通用要素抽取模块 7](#_Toc112318681)

[5.2.3 业务要素抽取模块 8](#_Toc112318682)

[5.2.4 文件操作模块 9](#_Toc112318683)

[5.2.5 通用模块 9](#_Toc112318684)

[6系统使用说明 10](#_Toc112318685)

[6.1智能归档操作流程 10](#_Toc112318686)

[6.2通用要素抽取操作流程 12](#_Toc112318687)

[6.3 业务要素抽取操作流程 13](#_Toc112318688)

[6.4 文件操作流程 14](#_Toc112318689)

[6.5 其他操作流程 15](#_Toc112318690)

# 

# 1引言

## 1.1系统总体需求

系统围绕海量电子公文中提取公文要素以及高质量的结构化信息，准确分析、挖掘电子公文的关系，对电子公文进行智能化归档，进一步提升实体信息抽取、关系信息抽取等中文自然语言处理技术在电子公文快速分析和归档管理的应用水平，推动人工智能技术在电子公文处理和管理的应用。

用户可以使用公文归档、通用要素抽取、业务要素抽取等功能来快捷提取结构化信息，获取归档及抽取的结果并进行可视化展示，并且以文件的方式储存结果。

## 1.2范围

表1 使用读者

|  |  |
| --- | --- |
| **预期读者** | **阅读建议** |
| 软件部分测试人员 | 仔细阅读全部内容，并履行开发规范。 |

# 2系统架构

公文智能抽取归档系统以电子公文知识信息抽取作为切入点，拟采用基于深度学习的命名实体识别技术、关系抽取技术以及机器学习等公文归档，对公文知识图谱进行构建，并为联合语义查询、公文关联推送、实时决策等上层核心业务提供支撑，系统的总体架构如图1所示：

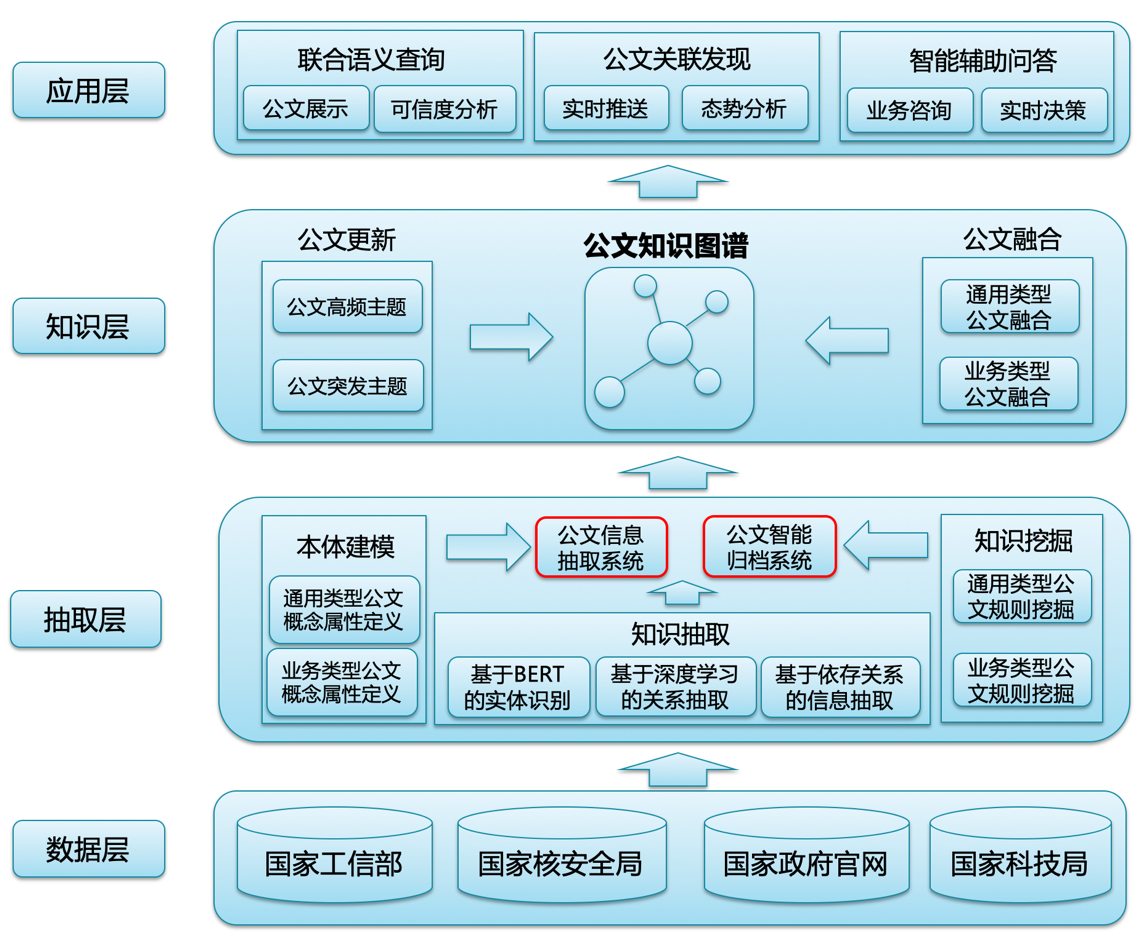
**

图1 公文知识图谱的总体架构

岗位能力知识图谱的构建与应用架构主要包含四层：数据层、抽取层、知识层以及应用层，其中**红色框**中的公文信息抽取系统与公文智能归档系统当下需求需研发的系统，它们是能力知识库构建的核心。接下来，将对这四层进行介绍：

**数据层**：知识库构建的首要问题是选取所采用的数据源。这里的知识库构建的目标是抽取并整合多源异构数据。项目中的数据源包含国家工信部、国家核安全局、国家政府官网、国家科技局等公开的海量公文数据。

**抽取层**：抽取层主要包含本体构建、知识抽取以及知识挖掘三大模块。本体构建主要根据通用类型公文与业务类型公文的要素与属性进行建模，其中通用类型包括请示、报告、意见、函、通知、纪要、命令、决定、公告、通告、通报、批复、决议、公报共15类，业务类型公文则聚焦军事、安全、科技三大领域。知识抽取则是对半结构化的表格以及无结构化的进行解析和抽取，主要采用基于BERT的实体识别技术、基于深度学习的关系抽取技术以及基于依存关系的信息抽取技术。知识挖掘则是通过频繁项挖掘与模版挖掘的技术来对不同类型的公文进行潜在的规则挖掘。

**知识层**：知识层则是基于公文信息抽取系统与公文智能归档系统将公文转为可靠的结构化知识的基础上，形成公文知识图谱，表达的形式以(主语、谓语、宾语)三元组为主。此外，公文知识图谱会进一步将公文周边数据（如：新闻报道、百科等）进行融合，并对于新增公文信息进行实时捕获、准确归档。以此来快速发现公文的一些突发主题和高频主题，服务于上层的应用。

**应用层**：应用层是系统和用户的接口，主要通过构建的公文知识图谱来支撑特定领域的服务效果，具体包括联合语义查询、公文关联发现、智能辅助问答。其中联合语义查询可以进行公文信息的全面展示、公文周边信息的可信度判断，公文关联发现则是助力于公文的实时推送与特定业务公文的态势分析，而智能辅助问答则是为使用者提供更快捷、准确的业务咨询与实时决策。

# 3主要功能

公文智能抽取归档系统聚焦在军事、安全、科技等前沿领域场景中的海量电子公文数据，采用知识图谱、深度学习等前沿人工智能技术，重点解决长度、内容不确定的电子公文中要素抽取、智能归档问题，并将语义信息与结构化知识予以应用，应用场景如下：

1．**通用公文要素的抽取**：从公文的通用类型角度出发，采用基于依存分析的模板技术对新增电子公文中的要素进行实时抽取，其类型包括报告、意见、函、通知、命令、决定、公告、通告、通报、批复10种不同类型的公文，以实现公文实体的精准识别。

2．**业务公文要素的抽取**：从公文的业务类型角度出发，采用深度学习的少样本抽取模型对新增电子公文中的要素进行实时抽取，其业务类型覆盖军事、安全、科技等前沿领域，实现对公文内容的结构解析。

3．**新增公文的智能归档**：从公文的业务类型角度出发，采用机器学习、深度学习等技术，实现新增电子公文对于军事、安全、科技等领域的二级文档分类，以完成对公文智能归档。

# 4运行条件

## 4.1硬件运行条件

系统测试硬件主机为高性能台式主机，CPU为虑酷睿i7-12700KF处理器显卡为RTX3080显卡(10GB显存)，内存为DDR5容量为32GB 。

## 4.2软件运行条件

软件操作系统为如下表所示，其中安装的操作系统为Linux Ubuntu 18.4版本

表2 软件运行环境

|  |  |
| --- | --- |
| **部署软件** | **最低版本** |
| Python | 3.7以上版本 |
| JDK | 1.8以上版本 |
| MySQL | 5.7以上版本 |
| Node.js | V16以上版本 |
| nginx | 1.20.1以上版本 |
| docker | docker\_1.5以上社区版本 |

## 4.3部分软件安装说明

本项目采用python 3.7作为编程语言。为方便起见，建议使用miniconda作为python管理工具。

1. Miniconda下载

登陆<https://docs.conda.io/en/latest/miniconda.html>，下载对应系统的miniconda软件，并安装。

1. 下载完成后，在终端内新建python环境，输入如下命令：

conda create -n env python==3.7

1. 输入conda activate env，激活python环境

## 4.4安装项目环境

在项目文件夹内，提供了所有所需的项目依赖包requirements.txt

在成功激活python环境后，输入

pip install -r requirements.txt

确保安装完毕所需依赖。

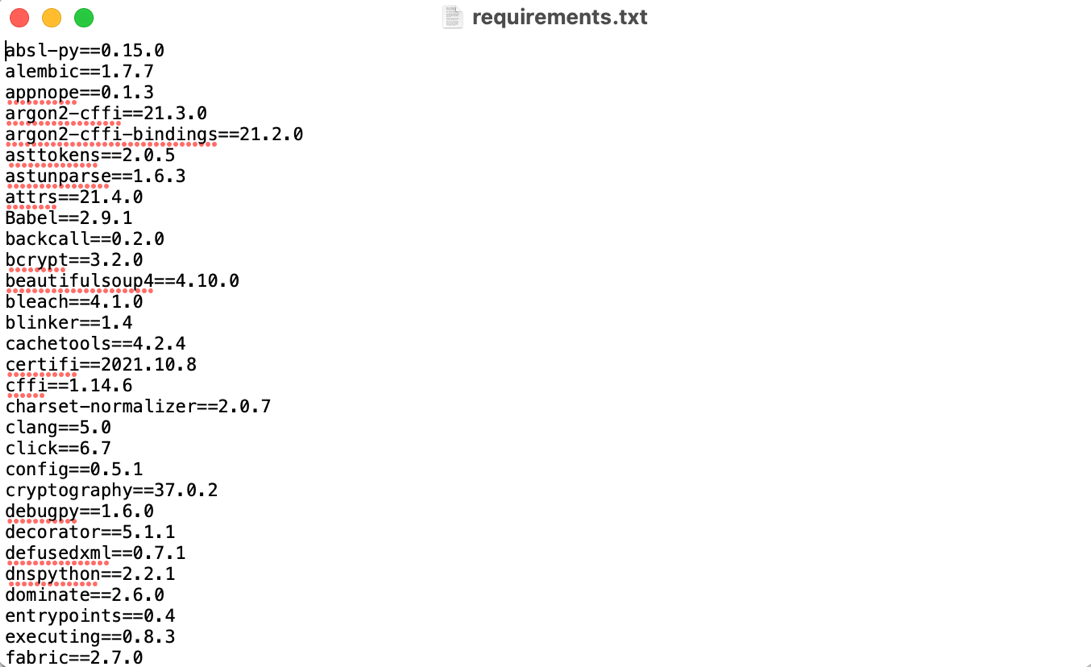


图2 requirements.txt 示例

## 4.5测试

（1）最后命令行执行python，看到以下截图信息，则代表python安装成功。

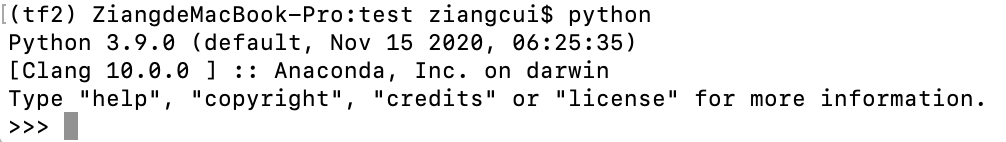


图3 python安装成功示例

（2）依赖包若安装成功，在执行完此前的pip install命令后，则会显示成功安装的信息，如下图所示：

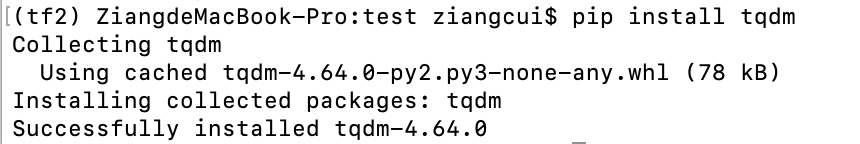


图4 依赖安装成功示例

**以上步骤都完成后，环境即完全配置完毕。**

# 5系统主要功能

## 5.1系统功能概述

|  |  |
| --- | --- |
| 模块名称 | 功能简述 |
| 公文归档模块 | 对上传的公文通过已有模型进行预测，获得归档结果并在前端进行展示 |
| 通用要素抽取模块 | 对上传的公文进行通用要素模板抽取，提取相对的实体与公文要素，并将抽取结果在前端进行展示 |
| 业务要素抽取模块 | 对上传的公文进行业务要素模板抽取，提取相对的实体与公文要素，并将抽取结果在前端进行展示 |
| 文件操作模块 | 提供上传文件与下载文件功能，将前端上传的文件保存至指定路径，并且可以把所需文件返回给前端进行下载操作 |
| 通用模块 | 各部分通用的操作，包括获取任务信息，删除任务等 |

## 5.2系统模块设计

### 5.2.1归档模块

#### 5.2.1.1功能

对上传的公文通过已有模型进行预测，获得归档结果并在前端进行展示，提供归档接口、查询归档状态接口、查询归档结果接口三个接口。

#### 5.2.1.2接口

|  |  |
| --- | --- |
| 名称 | 归档接口 |
| 接口地址 | /classification |
| 请求方式 | POST |
| 参数类型 | Body/json |
| 请求参数 | task\_id, file\_path |

|  |  |
| --- | --- |
| 名称 | 查询归档状态接口 |
| 接口地址 | /classification/status |
| 请求方式 | GET |
| 参数类型 | Params |
| 请求参数 | task\_id |

|  |  |
| --- | --- |
| 名称 | 查询归档结果接口 |
| 接口地址 | /classification/result |
| 请求方式 | POST |
| 参数类型 | Body/json |
| 请求参数 | task\_id, batch\_size, batch\_num |

### 5.2.2 通用要素抽取模块

#### 5.2.2.1功能

对上传的公文进行通用要素模板抽取，提取相对的实体与公文要素，并将抽取结果在前端进行展示，提供抽取接口、查询抽取结果接口。

#### 5.2.2.2接口

|  |  |
| --- | --- |
| 名称 | 通用要素抽取接口 |
| 接口地址 | /general/extract |
| 请求方式 | POST |
| 参数类型 | Body/json |
| 请求参数 | task\_id, file\_path |

|  |  |
| --- | --- |
| 名称 | 查询通用要素抽取结果接口 |
| 接口地址 | /general/result |
| 请求方式 | POST |
| 参数类型 | Body/json |
| 请求参数 | task\_id, batch\_size, batch\_num |

### 5.2.3 业务要素抽取模块

#### 5.2.3.1功能

对上传的公文进行业务要素模板抽取，提取相对的实体与公文要素，并将抽取结果在前端进行展示，提供抽取接口、查询抽取结果接口。

#### 5.2.3.2接口

|  |  |
| --- | --- |
| 名称 | 业务要素抽取接口 |
| 接口地址 | /business/extract |
| 请求方式 | POST |
| 参数类型 | Body/json |
| 请求参数 | task\_id, file\_path |

|  |  |
| --- | --- |
| 名称 | 查询业务要素抽取结果接口 |
| 接口地址 | /business/result |
| 请求方式 | POST |
| 参数类型 | Body/json |
| 请求参数 | task\_id, batch\_size, batch\_num |

### 5.2.4 文件操作模块

#### 5.2.4.1功能

提供上传文件与下载文件功能，将前端上传的文件保存至指定路径，并且可以把所需文件返回给前端进行下载操作，提供上传文件接口、下载文件接口。

#### 5.2.4.2接口

|  |  |
| --- | --- |
| 名称 | 上传文件接口 |
| 接口地址 | /upload |
| 请求方式 | POST |
| 参数类型 | Body/form-data |
| 请求参数 | file |

|  |  |
| --- | --- |
| 名称 | 下载文件接口 |
| 接口地址 | /download |
| 请求方式 | GET |
| 参数类型 | Params |
| 请求参数 | task\_id, task\_type |

### 5.2.5 通用模块

#### 5.2.5.1功能

各部分通用的操作，包括获取任务信息，删除任务等，提供获取任务信息接口、删除任务接口。

#### 5.2.5.2接口

|  |  |
| --- | --- |
| 名称 | 获取任务信息接口 |
| 接口地址 | /get\_all |
| 请求方式 | GET |
| 参数类型 | Params |
| 请求参数 | task\_type |

|  |  |
| --- | --- |
| 名称 | 删除任务接口 |
| 接口地址 | /delete |
| 请求方式 | POST |
| 参数类型 | Body/form-data |
| 请求参数 | task\_id, task\_type |

# 6系统使用说明

## 6.1智能归档操作流程

在前端上传文件后，后端生成task\_id和file\_path并返回给前端，前端暂时储存并生成任务。

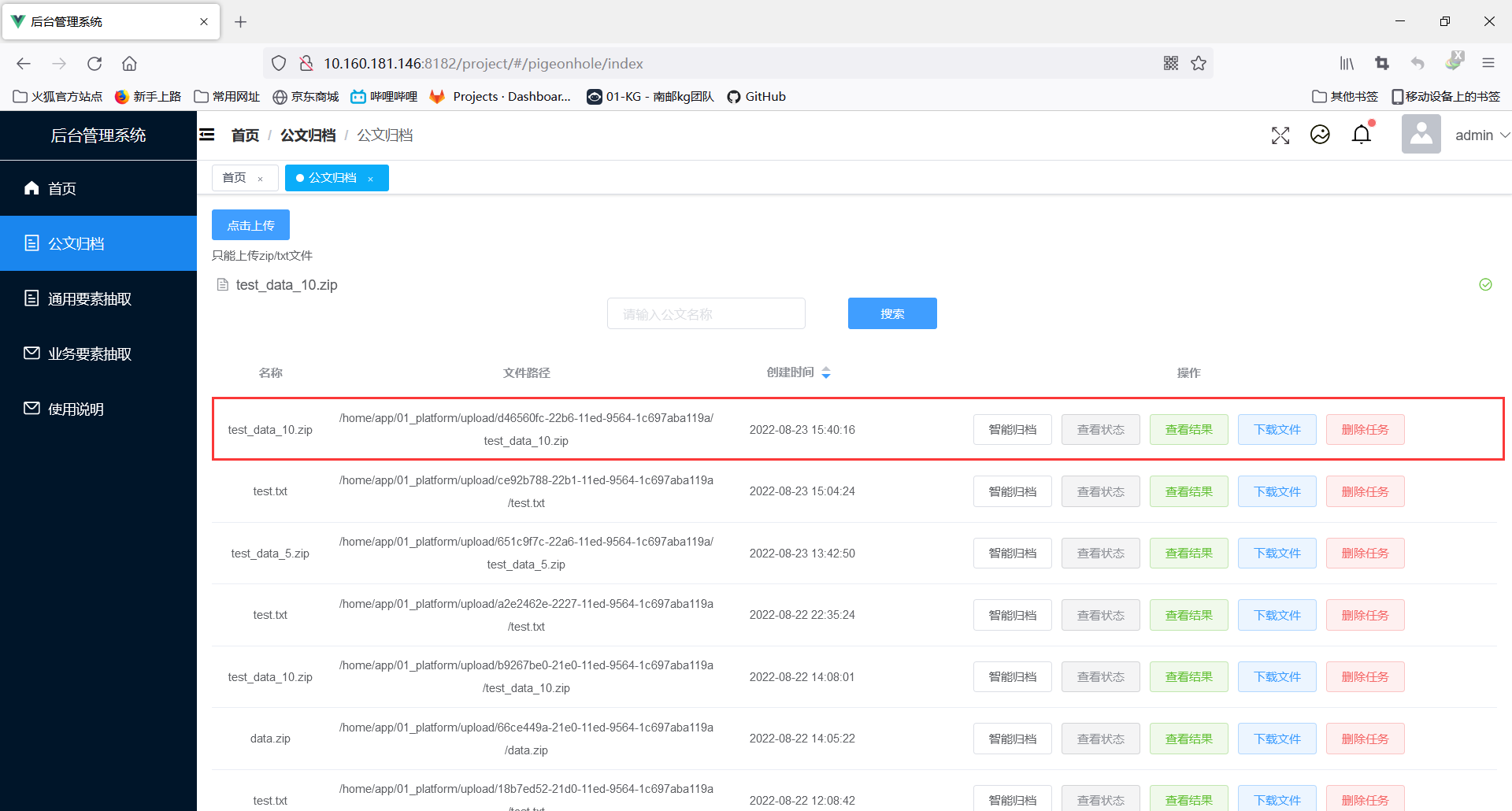


图5 归档任务创建界面

用户可以点击智能归档按钮开始归档，此时前端将调用归档接口，后端对该任务对应的task\_id和file\_path进行查找和解析，将对应文本进行预处理操作，保证其格式及编码正确。接下来输入文本并调用已有模型进行预测，并返回给前端已经开始归档。

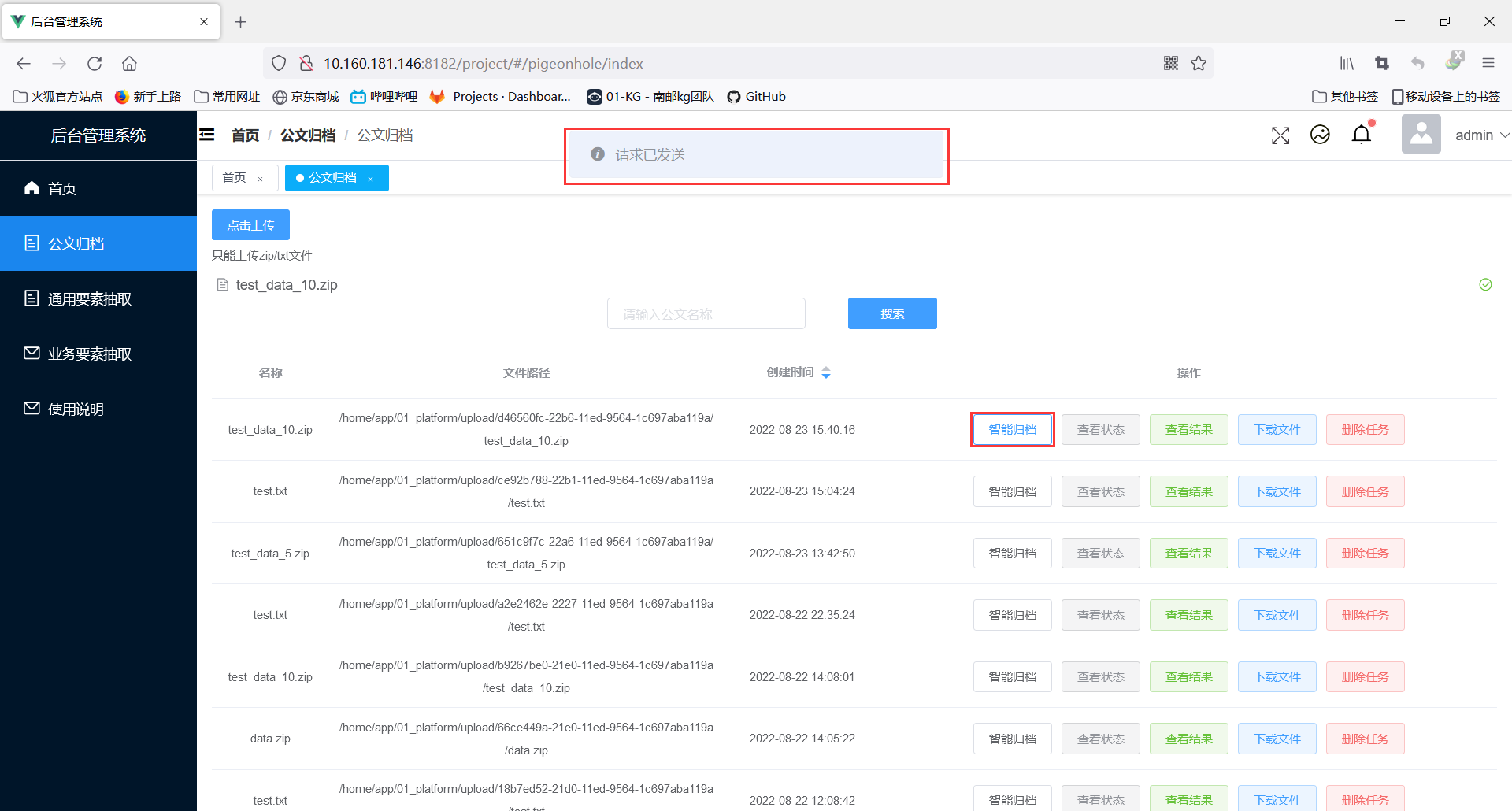


图6 开始归档界面

在进行智能归档后，用户可以相应的点击查看状态和查看结果的按钮，调用后端相应的接口，后端将从mysql数据库中查找相应结果，并将结果处理成相应格式返回给前端。

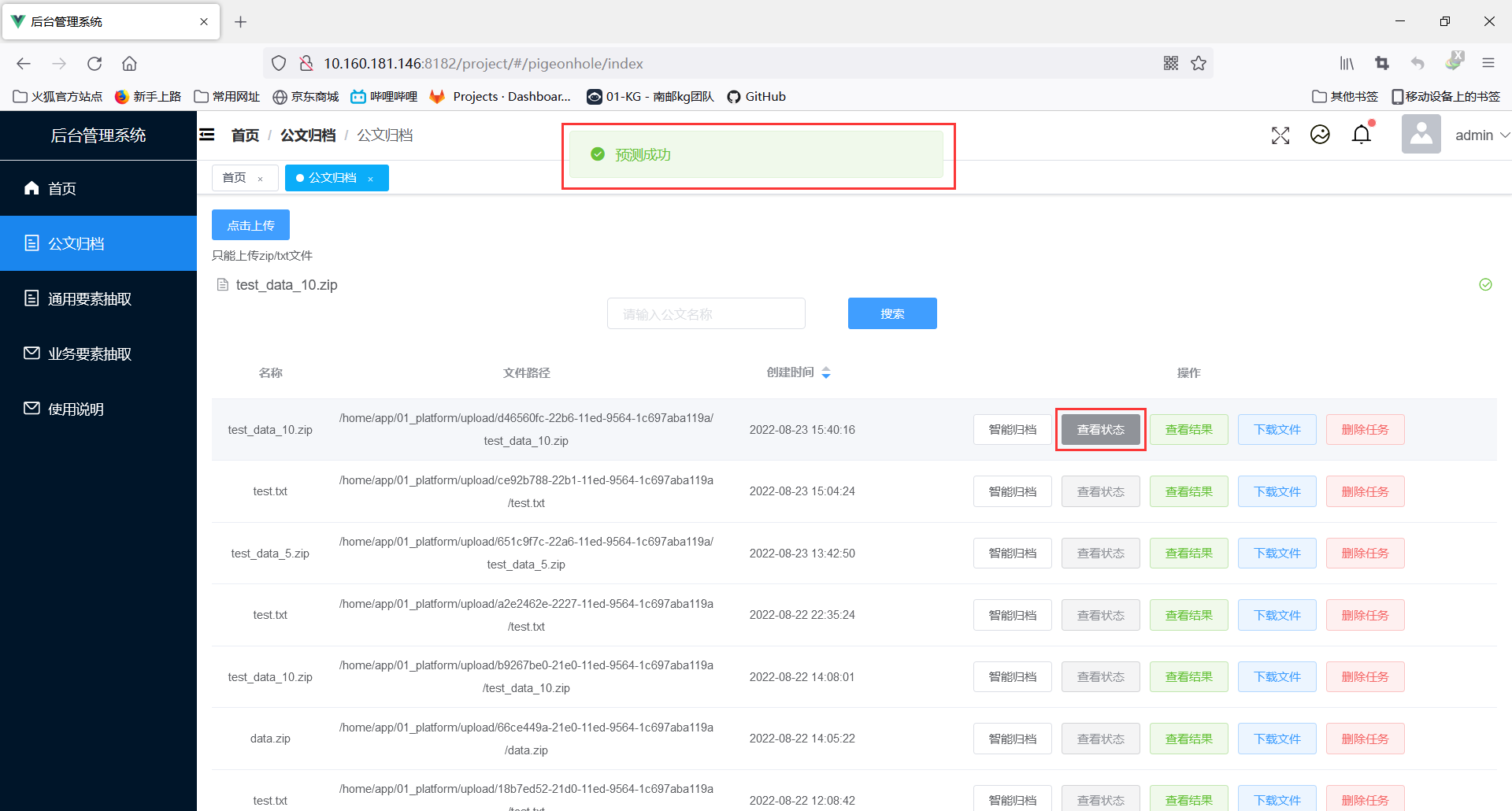


图7 查看归档状态界面

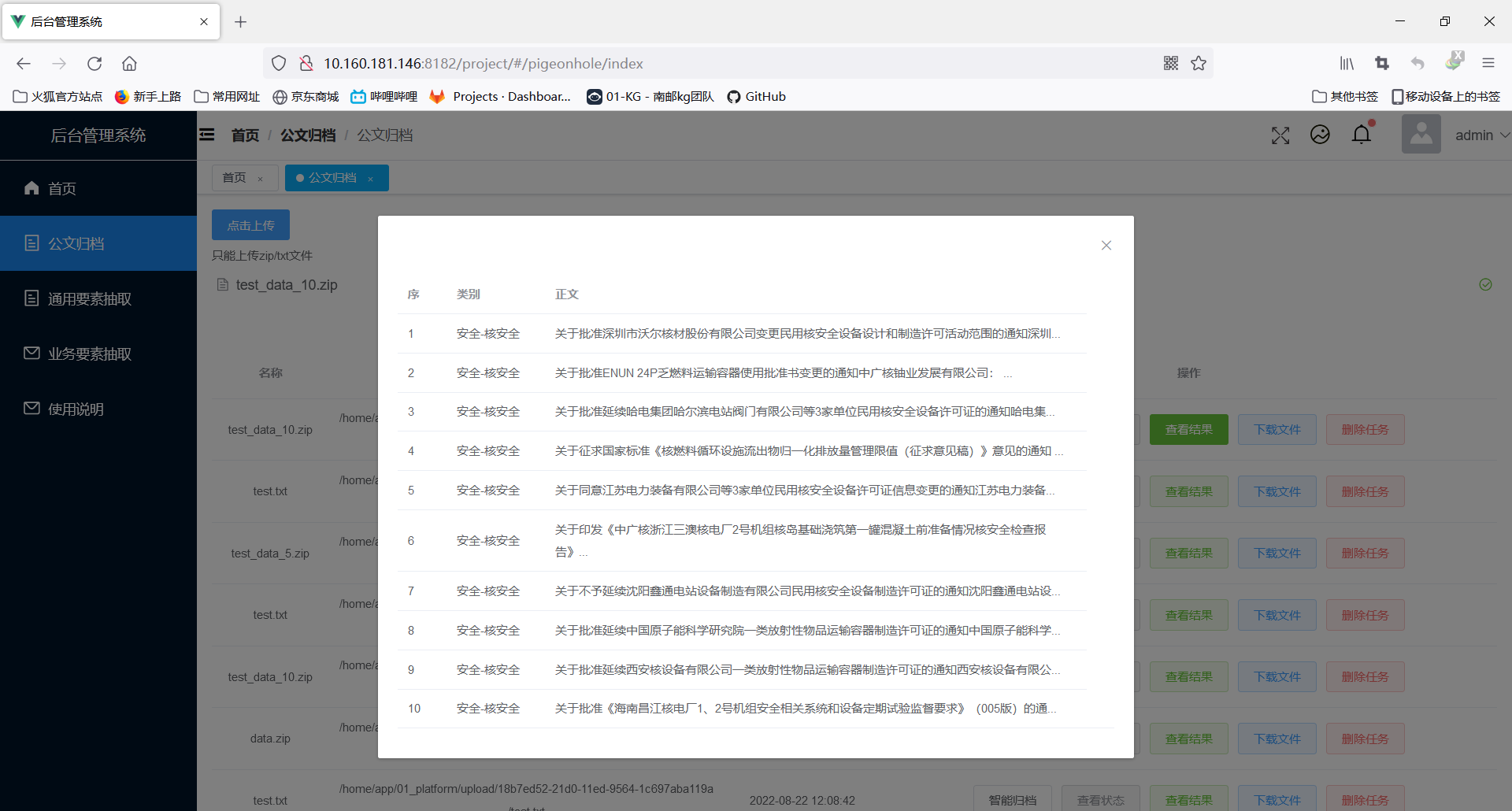


图8 查看归档结果界面

## 6.2通用要素抽取操作流程

用户首先通过前端上传需要进行抽取的文件，前端根据后端的返回信息创建一个新的任务供用户操作。任务创建好后，用户可以选择快速抽取功能对上传的文本进行要素抽取。

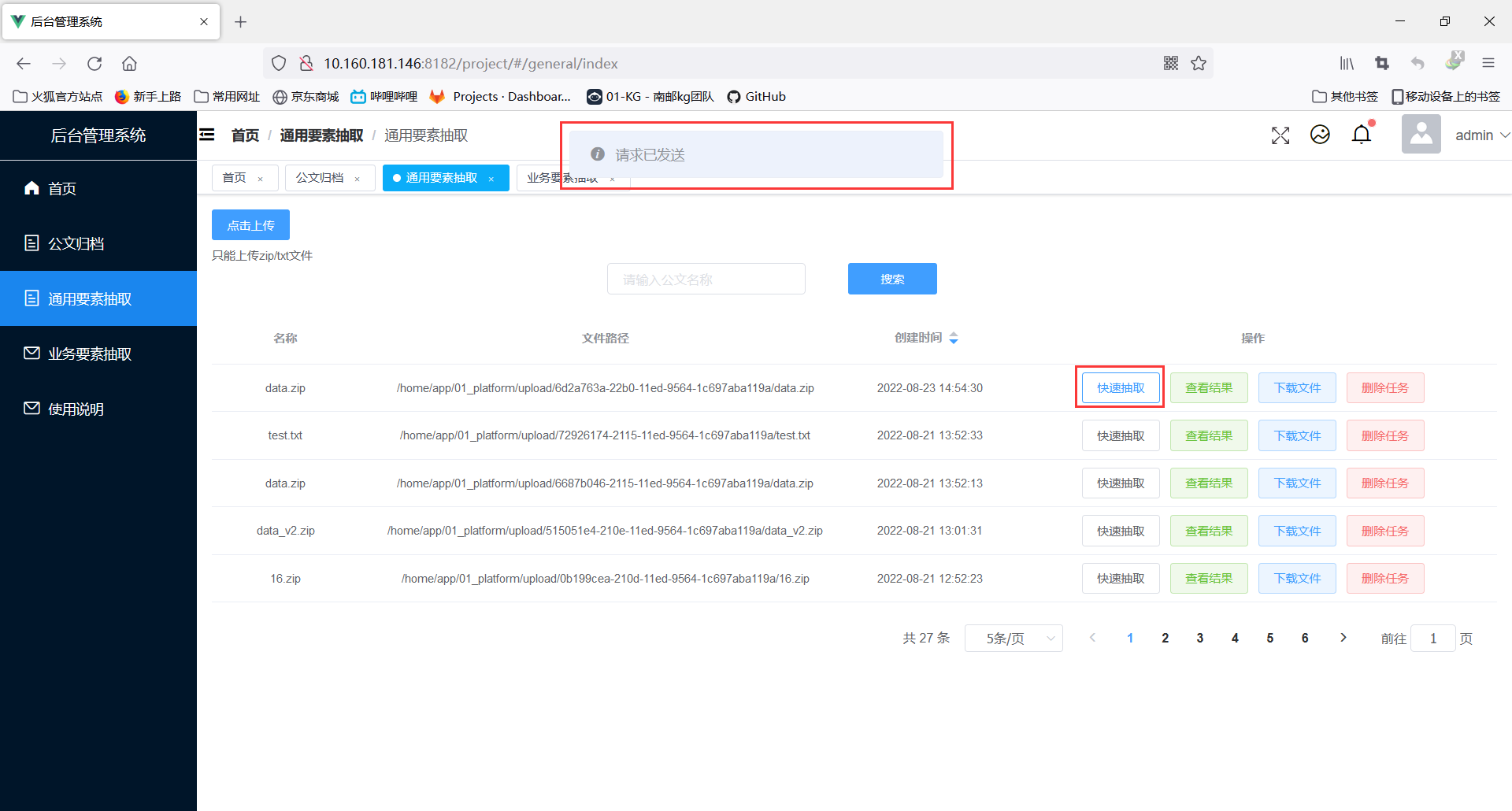


图9 通用要素快速抽取界面

在进行快速抽取后，用户可以通过查看结果来获取可视化的抽取结果，展示所有的抽取字段。

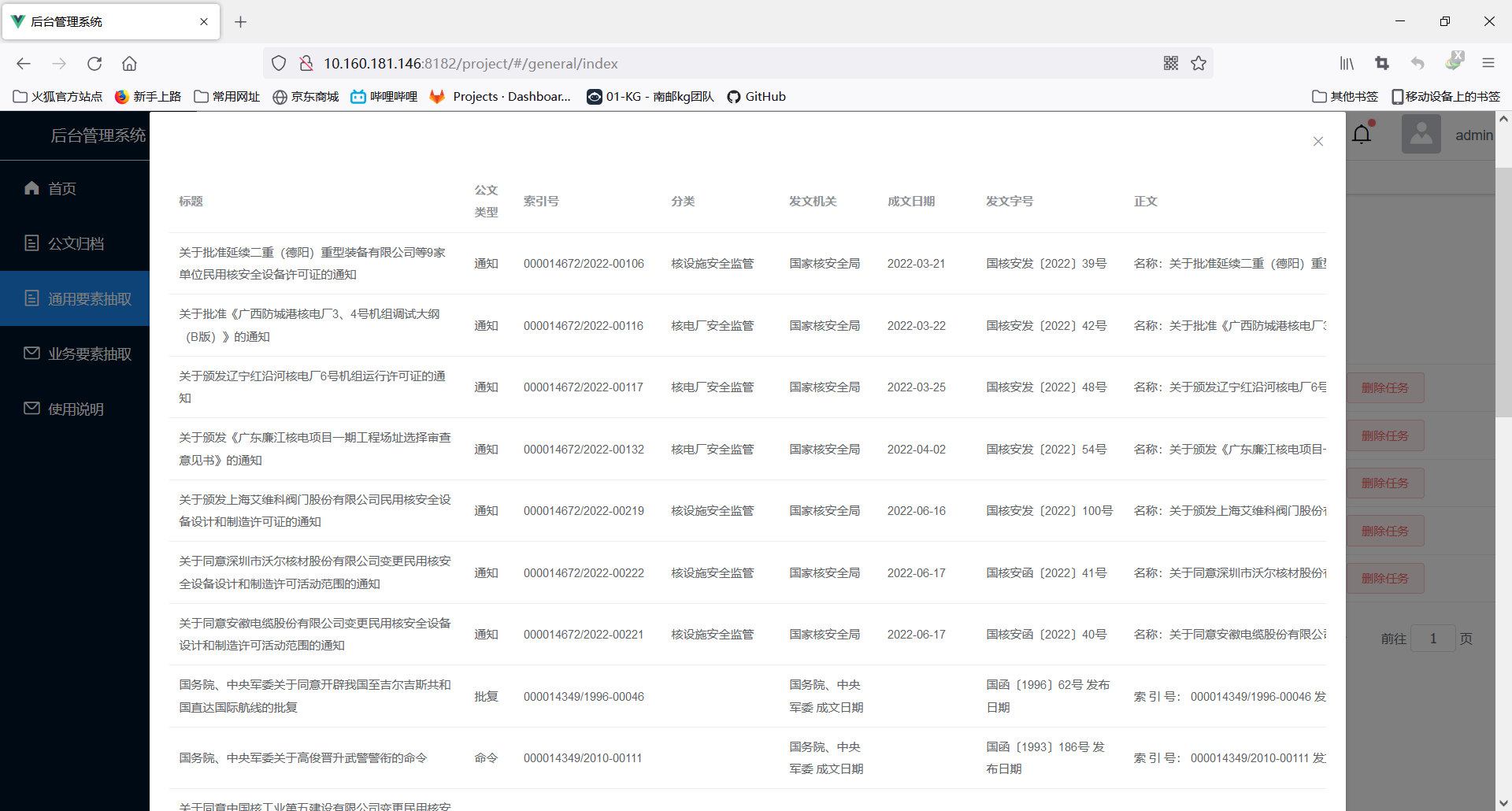


图10 查看通用要素抽取结果界面

## 6.3 业务要素抽取操作流程

用户首先通过前端上传需要进行抽取的文件，前端根据后端的返回信息创建一个新的任务供用户操作。任务创建好后，用户可以选择快速抽取功能对上传的文本进行要素抽取。

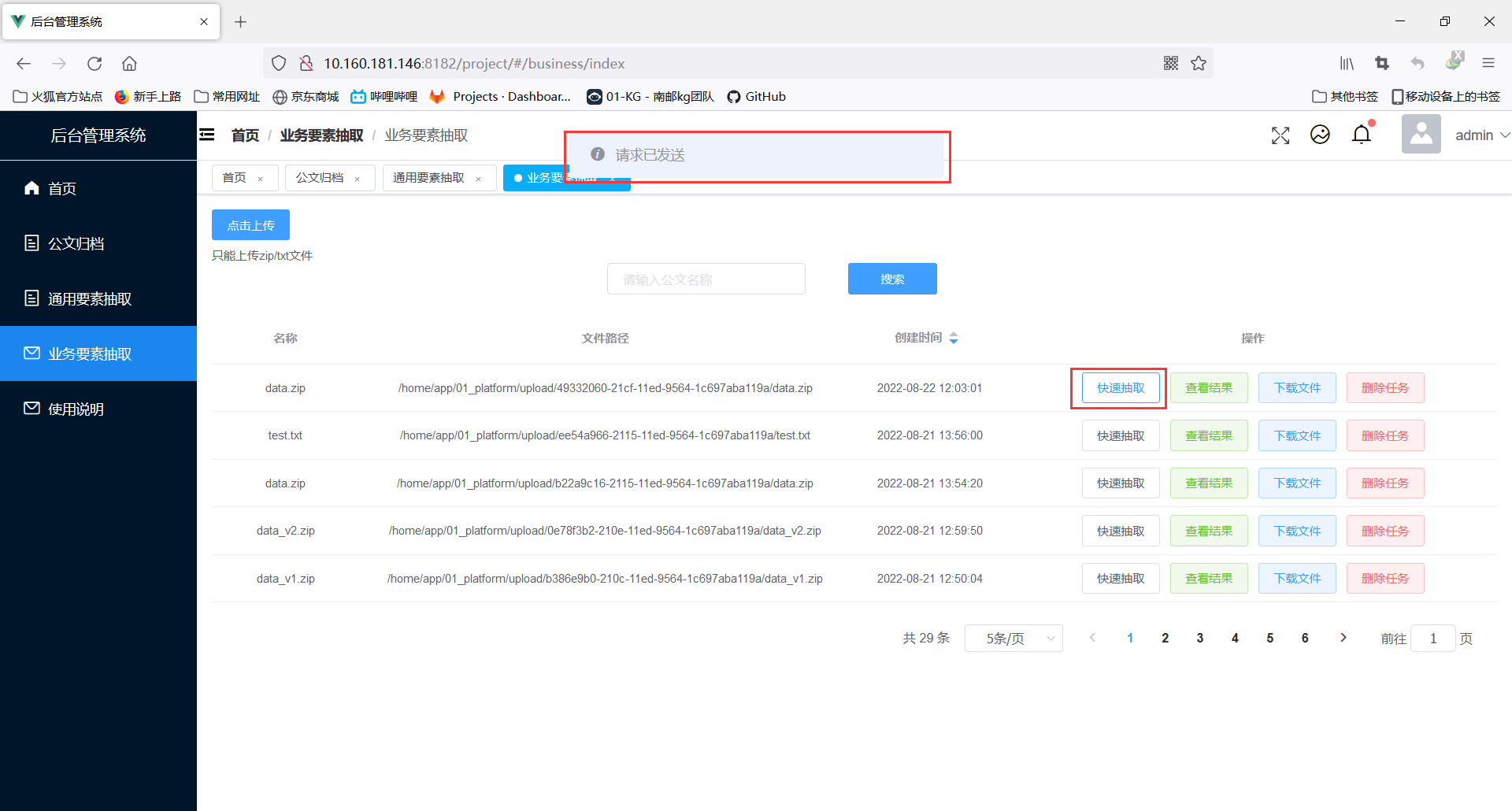


图11 业务要素快速抽取界面

在进行快速抽取后，用户可以通过查看结果来获取可视化的抽取结果，展示所有的抽取字段。

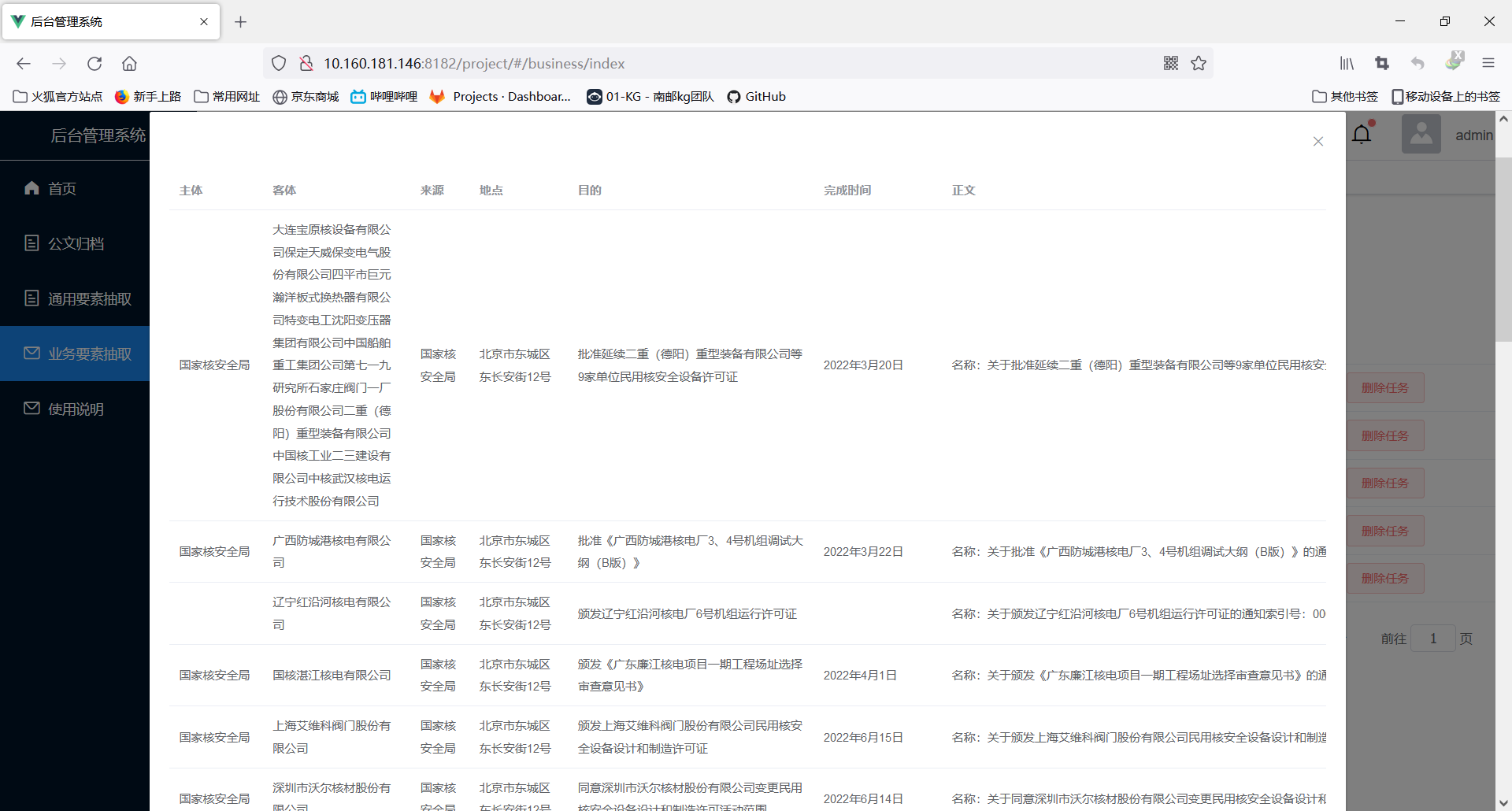


图12 查看业务要素抽取结果界面

## 6.4 文件操作流程

用户可以点击上传文件按钮，选择本地的zip文件或者txt文件进行上传。后端会将表单中的file数据进行解析，拿到文件名并从配置文件中生成存储路径file\_path，将文件保存至相应路径，并生成一个随机的task\_id返回给前端。

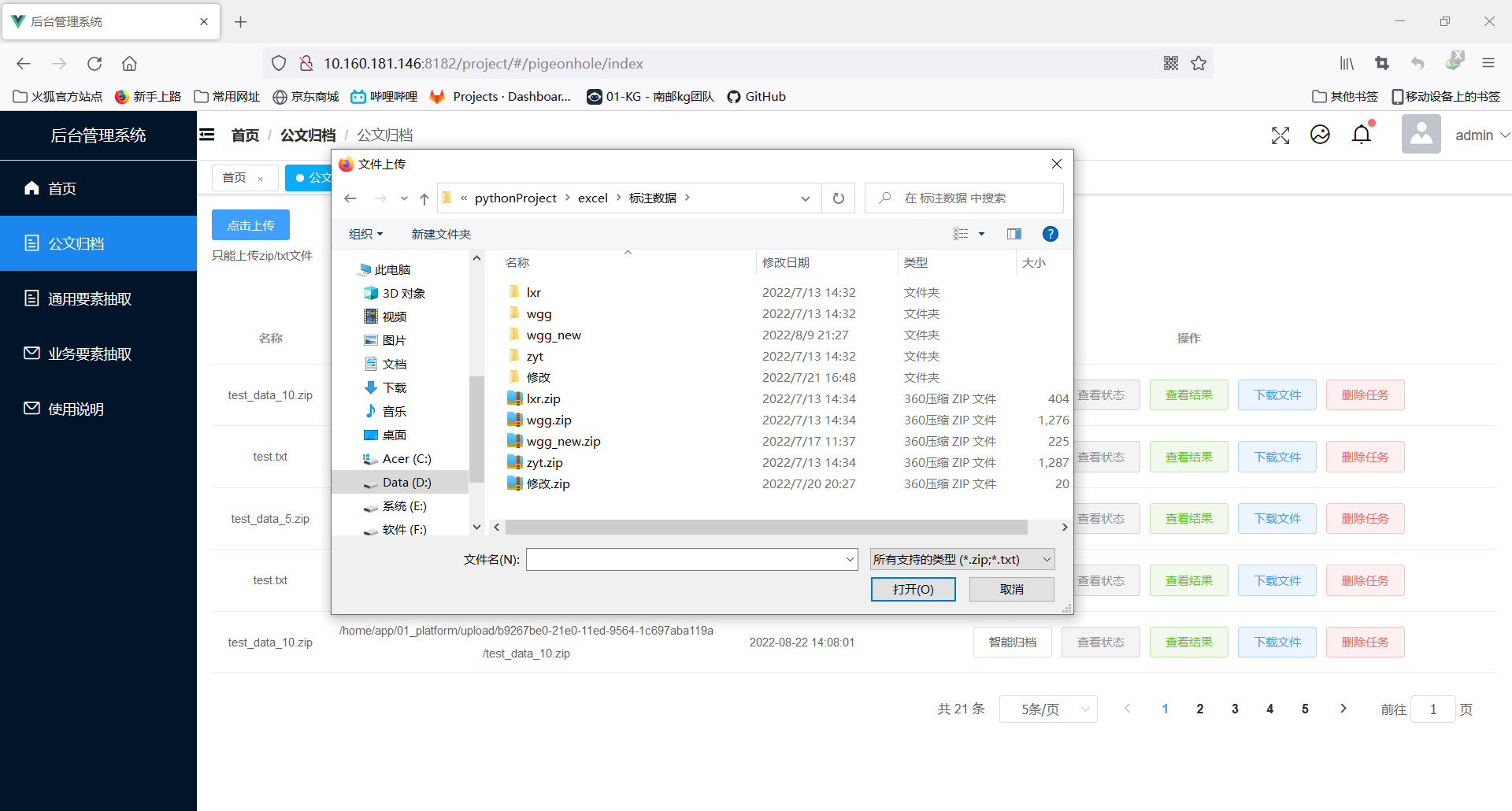


图13 上传文件界面

当需要下载文件时，用户点击下载文件按钮，前端将task\_id传至后端，后端通过task\_id在数据库中查找相对应的文本信息与结果信息，将其进行整合并生成csv文件返回至前端进行下载。

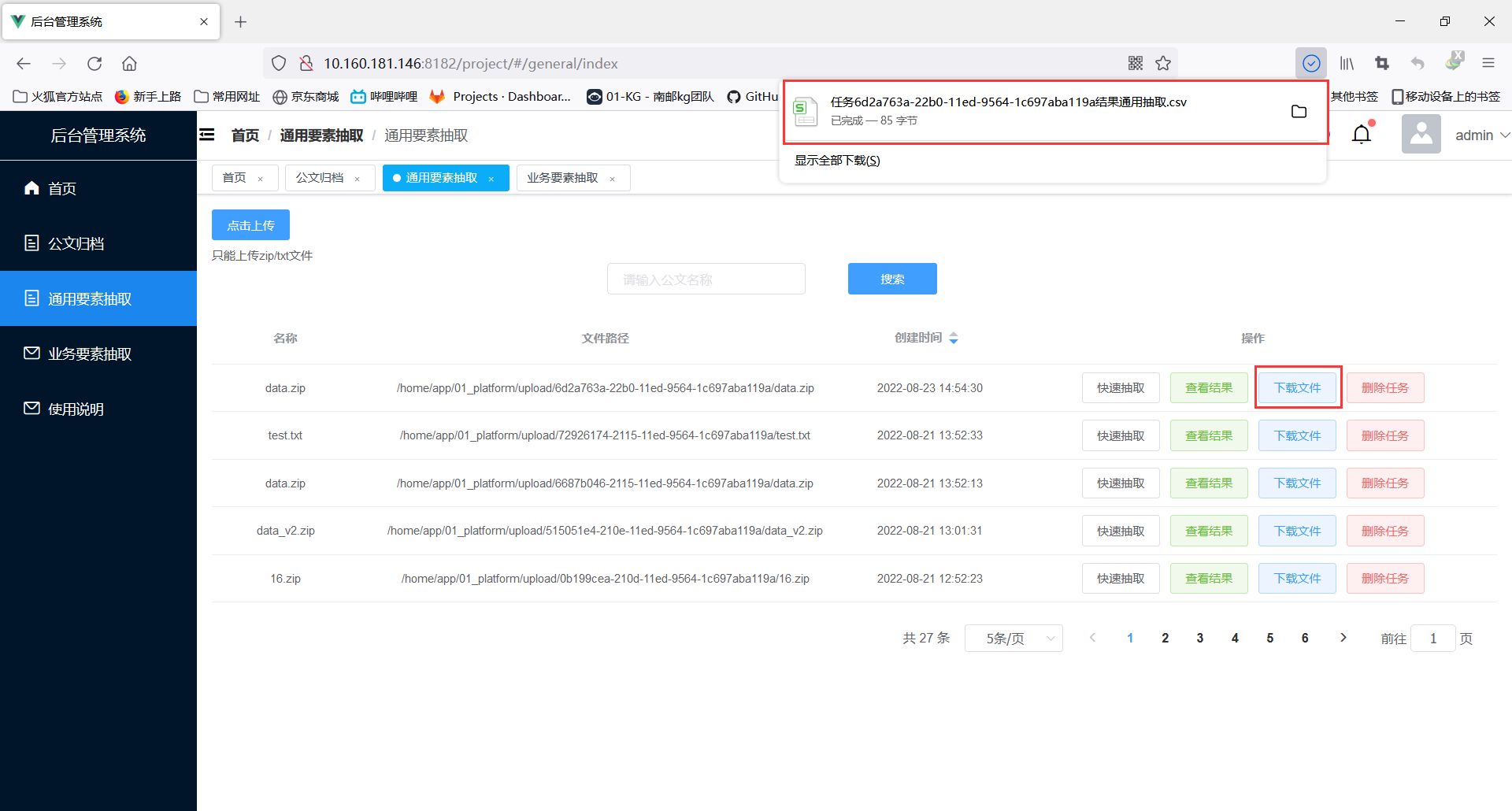


图14 下载文件界面

## 6.5 其他操作流程

用户每次切换前端页面时，都会自动调用获取任务信息接口，后端在数据库中查找所有不重复的相应功能的任务信息并返回给前端，前端根据这些信息加以展示。

当用户想要删除任务时，可以点击删除任务按钮，前端将调用删除任务接口，后端将对应的任务信息从数据库里删除。

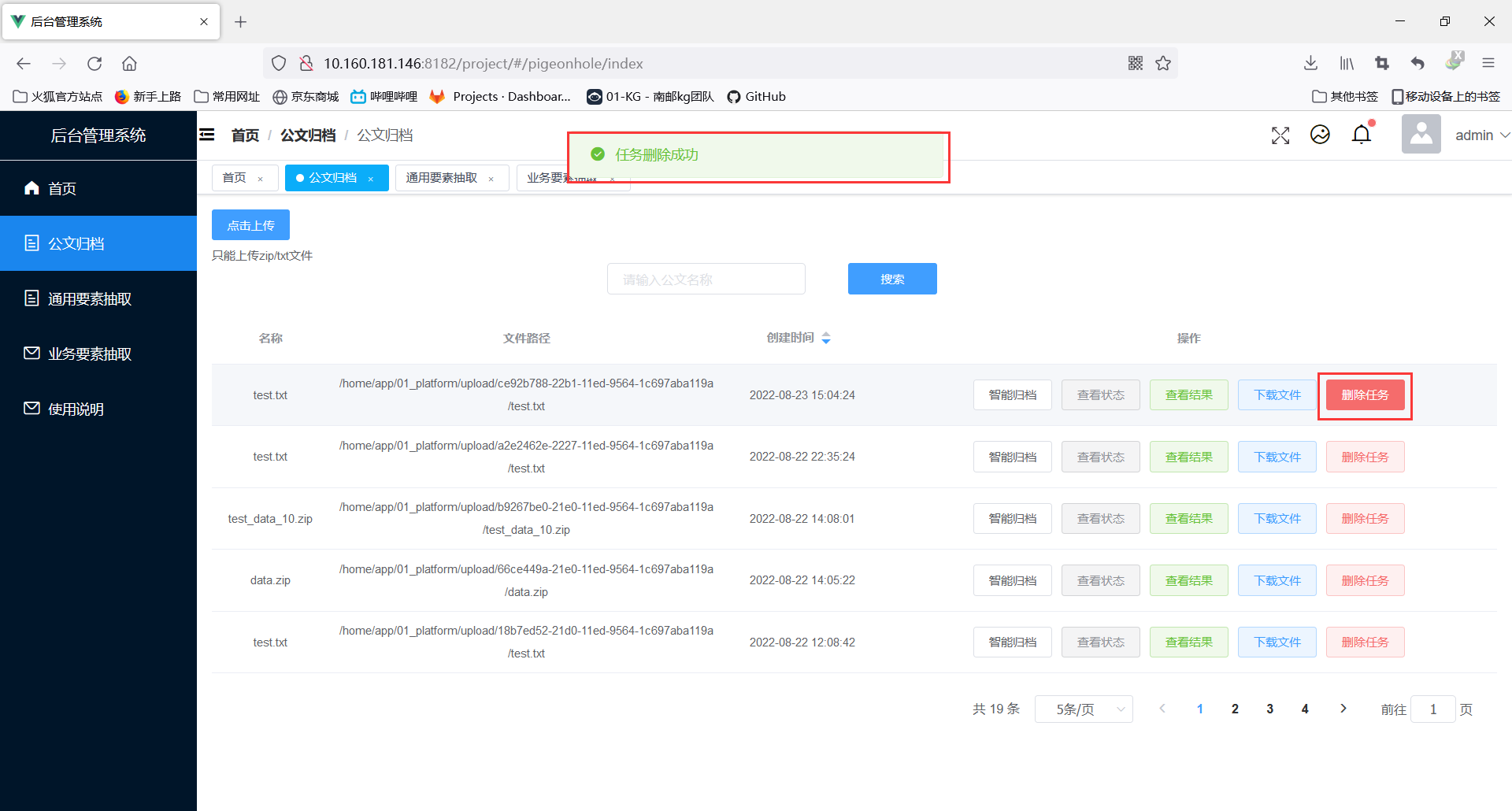


图15 删除任务界面