《统计资料指标数据抽取工具操作手册》

# 1. 简介‌

## 1.1 工具概述

对包括统计公报、政府工作报告在内的统计资料进行解析。输入统计资料文件，得到其中包含的各类指标形成的指标表。

## 1.2 适用场景

项目前期整理基础数据时应用该工具，例如可对历年统计公报进行数据抽取，并对比历年各项指标，以对现状分析工作提供支撑。

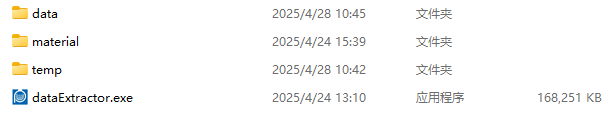
# 2. 系统要求‌

支持windows7及更高平台运行，内存需预留500MB以上，需保证稳定网络连接（无丢包、无卡顿）。

# 3. 安装与配置‌

解压“数据提取工具.zip”，得到以下4个文件，即完成安装。

（下文中“根目录”即指此层文件夹目录）



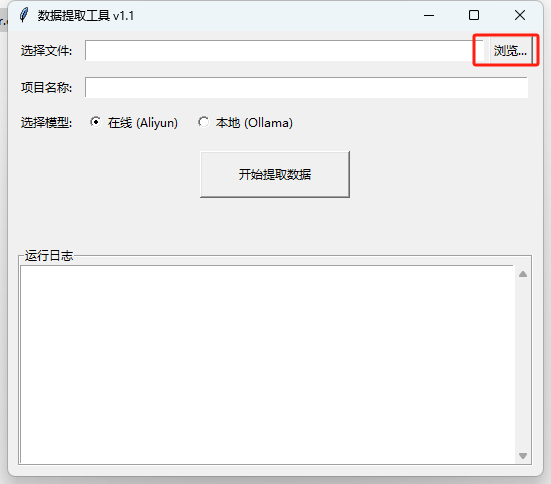
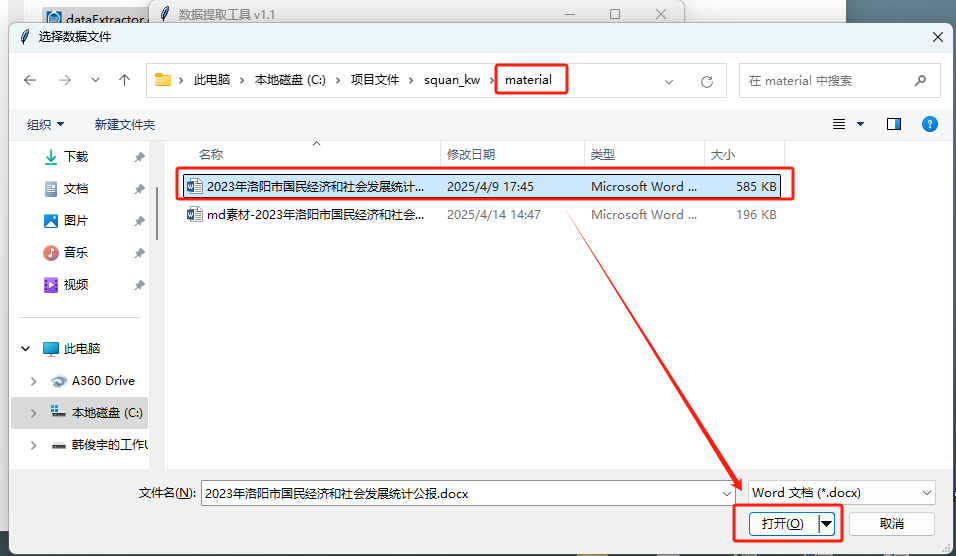
# 4. 工具使用说明‌

## 4.1 输入参数详解

### 4.1.1 选择文件

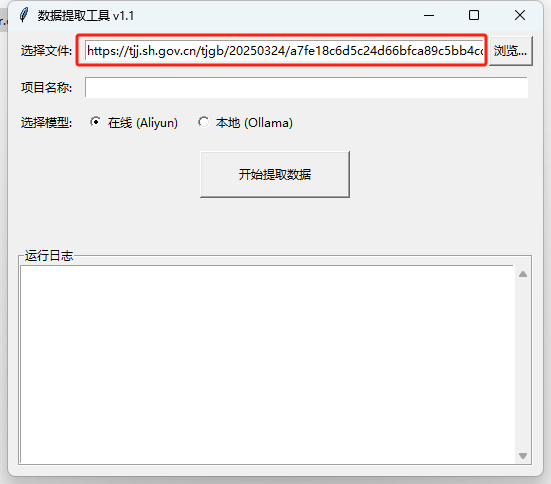
复制统计资料文件，并将复制的副本放置于根目录下的material文件夹内。

对于已下载文件：左键单击“浏览”，选择material文件夹内需要进行数据抽取的文件。选择完毕后单击“打开”。



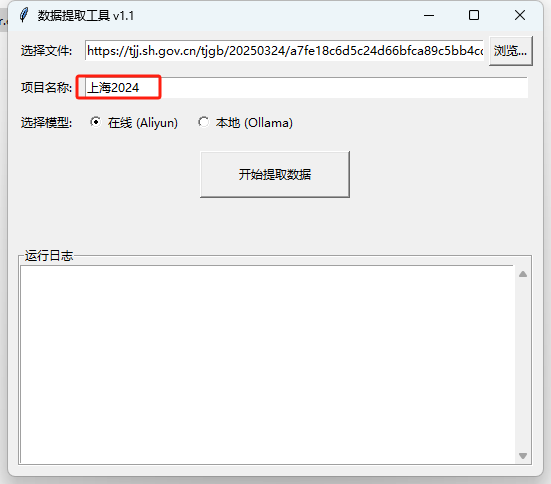
对于网页文件：直接在“选择文件”右侧的空白栏中填入网页地址（如https://tjj.sh.gov.cn/tjgb/20250324/a7fe18c6d5c24d66bfca89c5bb4cdcfb.html）。



### 4.1.2 项目名称

自行输入自定义名称，最后生成的EXCEL将以此处输入的参数进行命名。建议命名中包含地名、年份（如“上海2024”）。

（由于网页文件难以准确抽取项目名称，所以这里需要自己输入一个便于区分的自定义参数）。



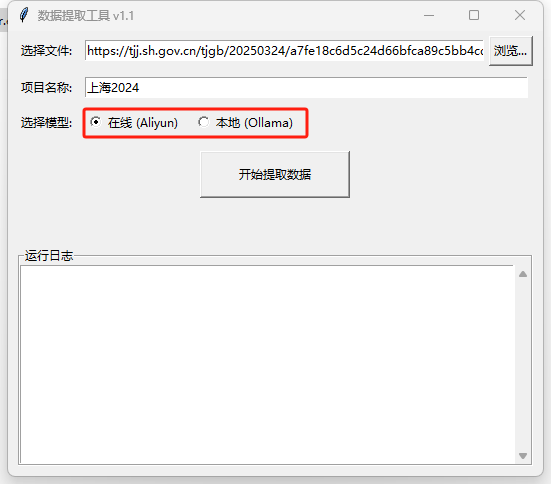
### 4.1.3 选择模式

模式可选择“在线”或“本地”。

在线模式下，将调用阿里云在线部署的qwen-plus模型，适用于各类公开发布或已公示的统计资料；

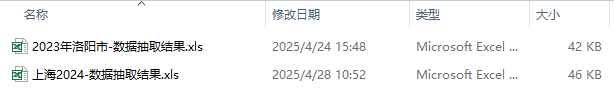
本地模式下，将使用公司本地部署的qwq32b模型，适用于不宜暴露于网络的甲方直接提供的资料。

如无特殊要求，十分推荐使用在线模式，不受算力限制下本工具将发挥更好性能，在线模式的数据抽取结果也远好于本地模式。



## 4.2 输出结果说明

输出文件默认保存在data文件夹中，文件名为“项目名称-数据抽取结果.xls”。



输出EXCEL包含3列，分别为序号（从1开始的简单数字序号）、指标名称（含单位）和指标具体数值。



# 5. 示例演示‌

## 5.1 示例场景描述

假设需要分析A市的城镇发展情况，需要通过统计公报强化数据支撑。

按要求下载A市的统计公报（或记录在线统计公报的网址），并准备使用本工具进行数据抽取。

## 5.2 输入数据准备

将通过附件下载的A市公报保存至material文件夹内，如为在线公报则复制URL网址。

1.在“选择文件”栏中通过单击“浏览”按钮选择文件，或将URL网址复制到空白栏中。

2.“项目名称”栏中，检查或自行填入“A市XXXX年”。

3.“选择模型”中，根据实际情况选择合适的模式。

## 5.3 执行步骤

点击“开始提取数据”按钮，开始运行本程序。据经验推算，1份统计公报约耗时10分钟左右。过程中无需其他操作。

程序完成后，将自动弹框提醒，并说明程序运行时长、数据抽取结果保存路径等信息。可点击“确定”关闭弹窗。

## 5.4 输出结果

最终在data文件夹内形成数据抽取结果EXCEL。

# 6. 注意事项‌

## 6.1 输入文件格式限制

现已完成对word文档（docx、doc）、txt文件、网页文件（html、shtml）、pdf文件的支持，其中pdf文件暂不支持照片扫描。

对其他类型的文件不保证输出结果准确性。

## 6.2 网络连接问题

本地模式下，需时刻保证与公司TJUPDI-WORK无线网络保持连接。否则返回结果将存在异常。

在线模式下，需时刻保证对外网络畅通。如网络存在丢包或不稳定情况，则本工具不保证数据抽取质量。

# 7. 常见问题解答（FAQ）‌

## Q1: 输出Excel内容为空的可能原因？

本地模式下，由于公司硬件限制，无法同时供应多个用户运行程序，因此会导致数据空白。建议切换至在线模式，或等待片刻再行尝试。

# 8. 技术支持与反馈‌

## 8.1 联系方式

可联系院内创研中心提交使用问题反馈与优化建议。

## 8.2 版本更新计划

计划增加对照片扫描类型pdf文件的支持。

计划增加对统计资料内图表图片的数据抽取功能。

# 附录‌

## A. 版本更新日志

2025年4月24日星期四：发布第一版完整程序。

2025年4月28日星期一：完成程序GUI界面设计。