《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。

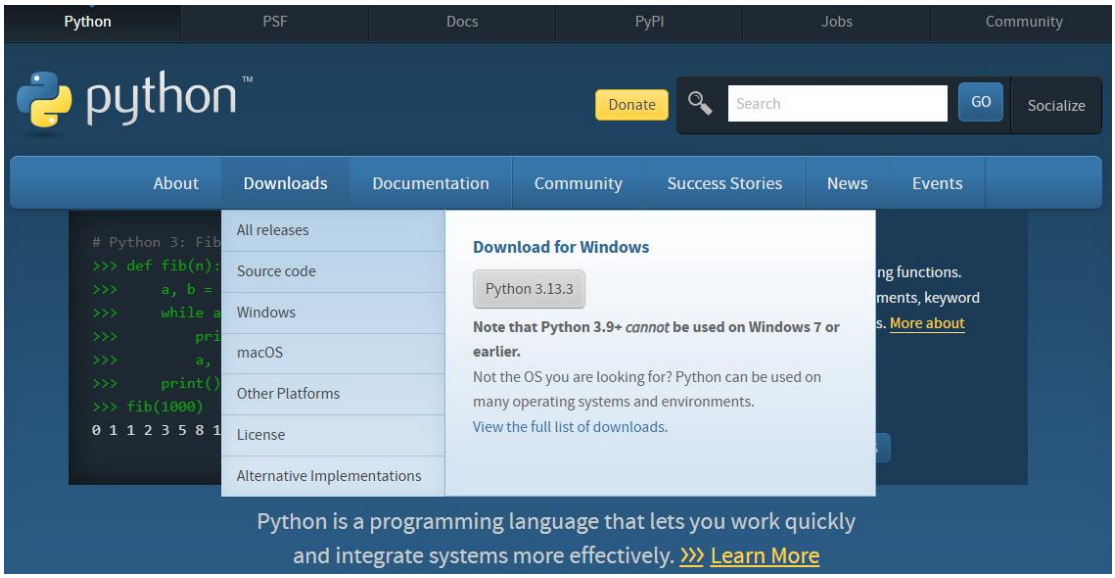
2. 系统要求

支持 windows7 及更高平台运行，支持 3.12.6 及更高 Python 版本，第三方库详见同目录下 requirements.txt 文件。

3. 安装与配置

3.1 安装 Python 环境

前往 https://www.python.org/下载 Python 安装包并安装。



3.2 安装运行程序必要的第三方库

在程序根目录下，通过命令提示符（win+R 调出运行窗口，并输入 cmd 回

车启动命令提示符），运行如下命令：

```
pip install -r requirements.txt
```

4. 工具使用说明

4.1 输入参数详解

4.1.1 输入文件

统计资料文件放置于根目录下的 material 文件夹内。

输入文件相对路径，格式例如“material/2023 年南平市水资源公报.pdf”。需要保留最前面的“material/”

4.1.2 项目名称

自定义名称，最后生成的 EXCEL 将以此输入进行命名。

4.1.3 搜索模式

模式可选择“local”或“online”。

Local 模式下将使用公司本地部署的 qwq32b 模型，适用于不可暴露于网络的甲方直接提供资料；Online 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。

一般来说 Online 模式效果远好于 Local 本地模式。

4.2 配置文件说明

配置文件包括和 config.config 和 local_ollama_ip.config 共计 2 份文件。config.config 为阿里云验证授权文件，存储了阿里云大模型调用所使用的 api key；local_ollama_ip.config 为公司本地部署的大模型的 ollama 接口地址。

4.3 输出结果说明

输出文件默认保存在 data 文件夹中。

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，并准备使用本工具进行数据抽取。

5.2 输入数据准备

将通过附件下载的公报保存至 material 文件夹内，如为在线公报则保存 URL 网址信息。

以记事本打开 data_extract.py，将第 205-207 行（即最后一行）修改为项目对应信息。（如下所示，标黄为需要修改的内容。其中第一行 file_address 所填写的文件名需与实际文件名完全一致，并包含文件扩展名）

```
'file_address':u'material/XXXX 年 A 市国民经济和社会发展统计公报.docx',
'proj_name': u'A 市',
'model': 'online'
```

5.3 执行步骤

在根目录下使用命令提示符（文件夹内空白处右键，点击“在终端中打开”），输入以下命令开始运行程序。

```
python data_extract.py
```

据经验推算，1 份统计公报约耗时 1.5 小时左右。过程中无需其他操作。

5.4 输出结果展示（附 Excel 截图）

最终在 data 文件夹内形成的 EXCEL 如下图所示。

	A	B	C
1	序号	指标名称	指标数值
2	1	地区生产总值（亿元）	939.05
3	2	第一产业增加值（亿元）	184.11
4	3	第二产业增加值（亿元）	261.36
5	4	第三产业增加值（亿元）	493.58
6	5	人均地区生产总值（元）	26858
7	6	居民消费价格总指数（比上年上涨%）	1.5
8	7	食品烟酒（比上年上涨%）	2.6
9	8	粮食（比上年上涨%）	1.2
10	9	菜（比上年上涨%）	9.4
11	10	畜肉类（比上年上涨%）	11.5
12	11	衣着（比上年上涨%）	1
13	12	居住（比上年上涨%）	2.3
14	13	生活用品及服务（比上年上涨%）	0.2
15	14	交通和通信（比上年上涨%）	-2.5
16	15	教育文化和娱乐（比上年上涨%）	1.9
17	16	医疗保健（比上年上涨%）	3.3
18	17	其他用品和服务（比上年上涨%）	0.6
19	18	全部财税收入（亿元）	156.4
20	19	财政总收入（亿元）	148.87

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

Local 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。

Online 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，则本工具不可用。

7. 常见问题解答（FAQ）

Q1: 依赖库安装总是超时如何处理？

在命令提示符下运行如下命令：

```
python -m pip install --upgrade pip
pip config set global.index-url https://mirrors.tuna.tsinghua.edu.cn/pypi/web/simple
```

将使用清华大学的国内镜像安装包进行安装，省去连接海外 Python 服务器

的步骤。

Q2: 输出 Excel 内容为空的可能原因？

local 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 online 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工具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。