# 表格去重助手说明书

v1.0（2024年8月）

## 1. 概述

### 1.1 软件名称

表格去重助手 (Duplicate Data Identifier)

### 1.2 软件简介

表格去重助手是一款基于Python Pandas数据库开发的桌面应用程序，用于识别和处理Excel文件中的重复数据。它允许用户处理单个文件或批量文件，并输出包含重复和非重复数据的整合后文件。

该软件适用于数据清理和数据管理的场景，特别是在需要对大量数据进行去重处理时。

### 1.3 软件功能

* 处理单个Excel文件中的重复数据
* 批量处理多个Excel文件中的重复数据
* 根据用户指定的唯一标识符列名，自动分类并输出重复和非重复数据
* 支持对处理结果进行自定义命名

## 2. 功能说明

### 2.1 主菜单界面

**功能**：提供用户选择处理单个文件或批量文件的入口。

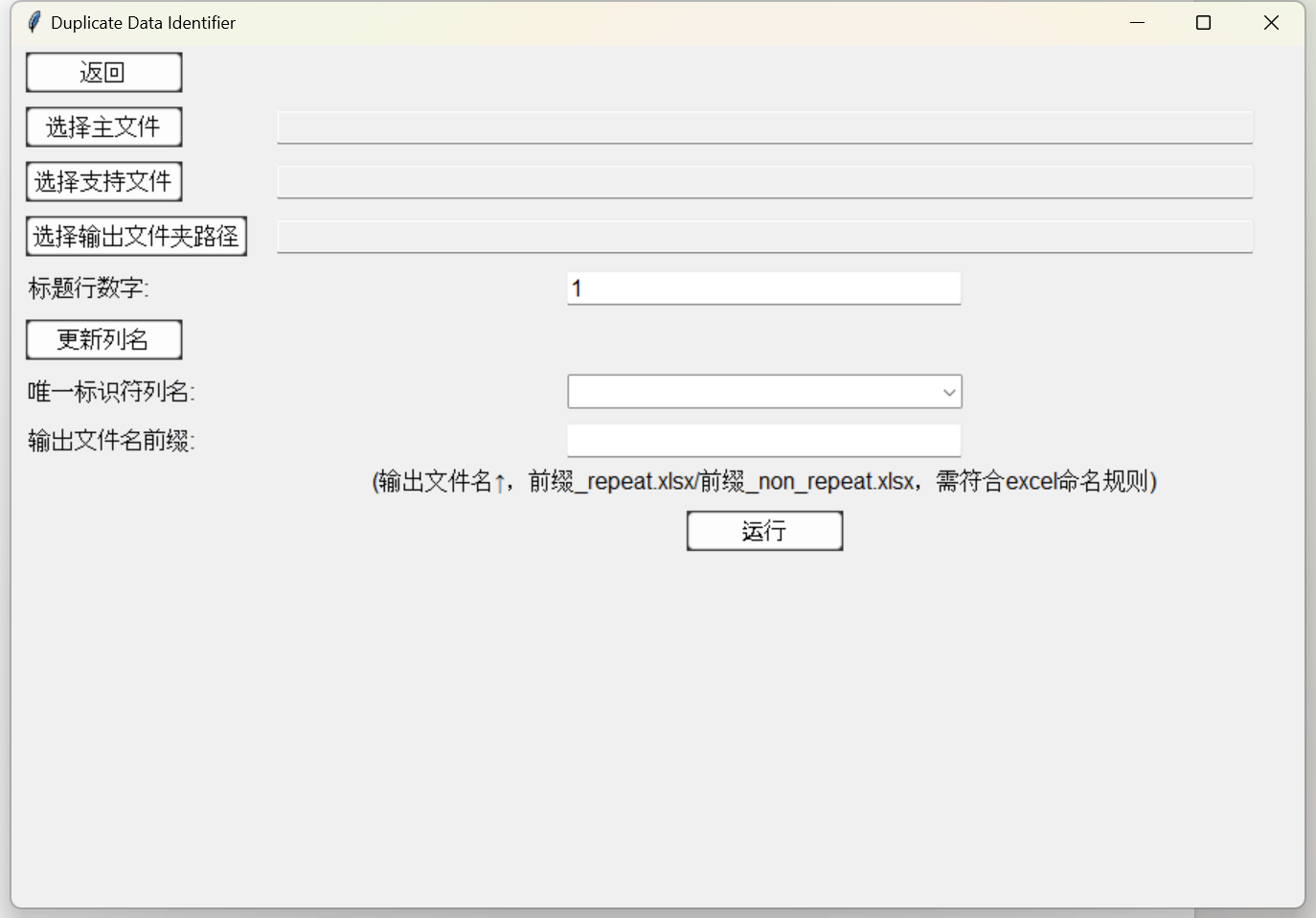
* **处理单个文件**：进入单个文件处理界面。
* **处理批量文件**：进入批量文件处理界面。



### 2.2 单个文件处理模块界面介绍

**功能**：用户选择主文件、支持文件，并配置相关参数后，程序自动识别主文件中的重复数据并生成处理结果。

* **返回**：返回主菜单。
* **选择主文件**：用户选择需要处理的主Excel文件。
* **选择支持文件**：用户选择用于比较的支持Excel文件。
* **选择输出文件夹路径**：用户选择输出结果的文件夹路径。
* **标题行数字**：用户输入Excel文件中标题行所在的行号，默认为1。
* **更新列名**：提取主文件中的所有列名，供用户选择唯一标识符列。
* **唯一标识符列名**：用户选择一个列作为唯一标识符，用于识别重复数据。
* **输出文件名前缀**：用户为生成的结果文件设置前缀，程序会生成带有"\_repeat.xlsx"和"\_non\_repeat.xlsx"后缀的文件。
* **运行**：执行重复数据的识别与分类，生成包含重复数据和非重复数据的Excel文件。



### 2.3 批量文件处理模块界面介绍

**功能**：用户选择多个文件夹，并配置相关参数后，程序自动批量识别文件中的重复数据并生成处理结果。

* **返回**：返回主菜单。
* **选择主文件夹**：用户选择存放主Excel文件的文件夹。
* **选择支持文件夹**：用户选择存放支持Excel文件的文件夹。
* **选择输出文件夹路径**：用户选择输出结果的文件夹路径。
* **选择示例文件**：用户选择一个示例文件用于提取列名。
* **标题行数字**：用户输入标题行所在的行号，默认为1。
* **更新列名**：提取示例文件中的列名，供用户选择唯一标识符列。
* **文件名列表**：用户输入需要处理的文件名列表，文件名之间用逗号分隔。
* **重复文件名前缀/后缀**：用户为处理结果的重复数据文件设置命名前缀和后缀。
* **非重复文件名前缀/后缀**：用户为处理结果的非重复数据文件设置命名前缀和后缀。
* **重复数据汇总文件名**：用户设置汇总所有重复数据的Excel文件名。
* **非重复数据汇总文件名**：用户设置汇总所有非重复数据的Excel文件名。
* **运行**：执行批量文件的重复数据识别与分类，生成各文件的处理结果和汇总文件。



## 3. 操作指南

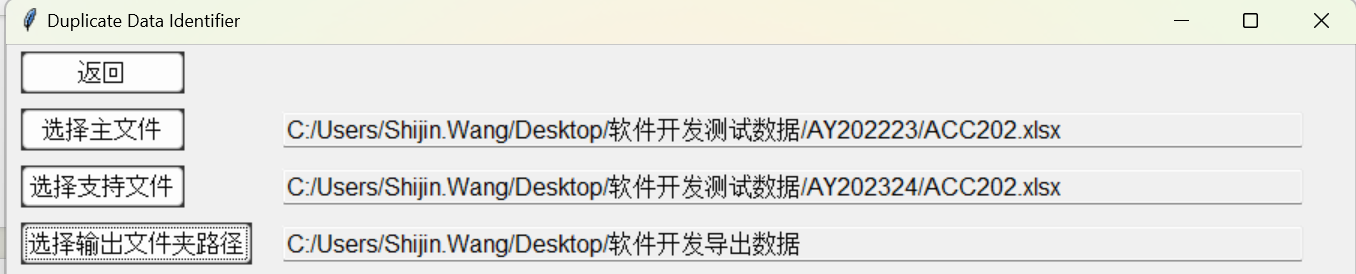
### 3.1 单个文件处理操作流程

在处理单个文件时，

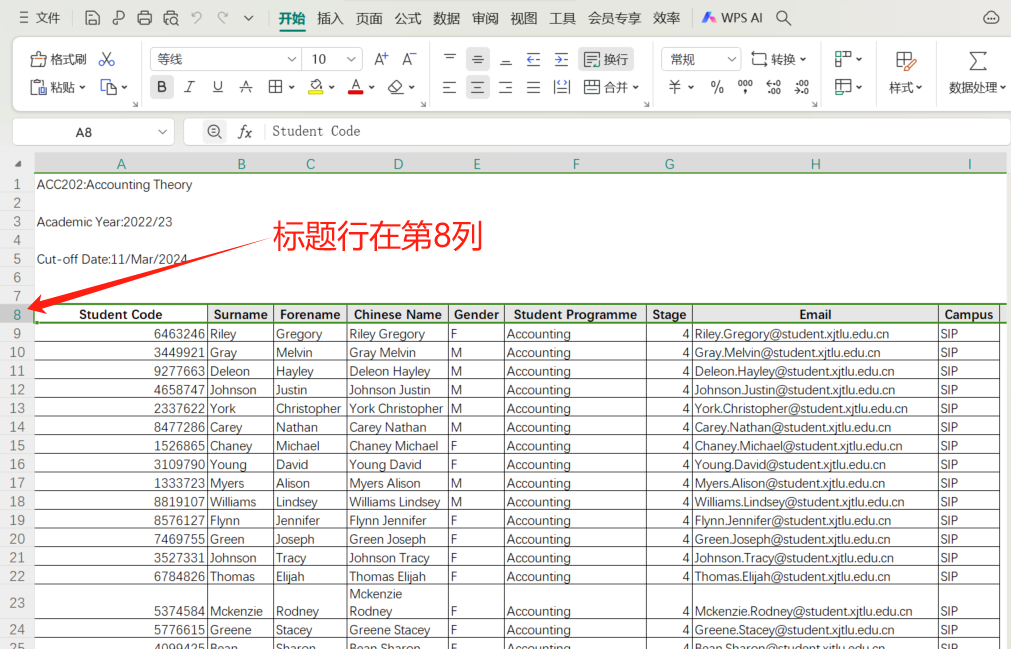
首先点击“选择主文件”按钮，选择需要被处理的Excel文件路径地址。

在选择支持文件时，点击“选择支持文件”按钮，选择用以辅助识别重复/非重复内容的支持文件路径。

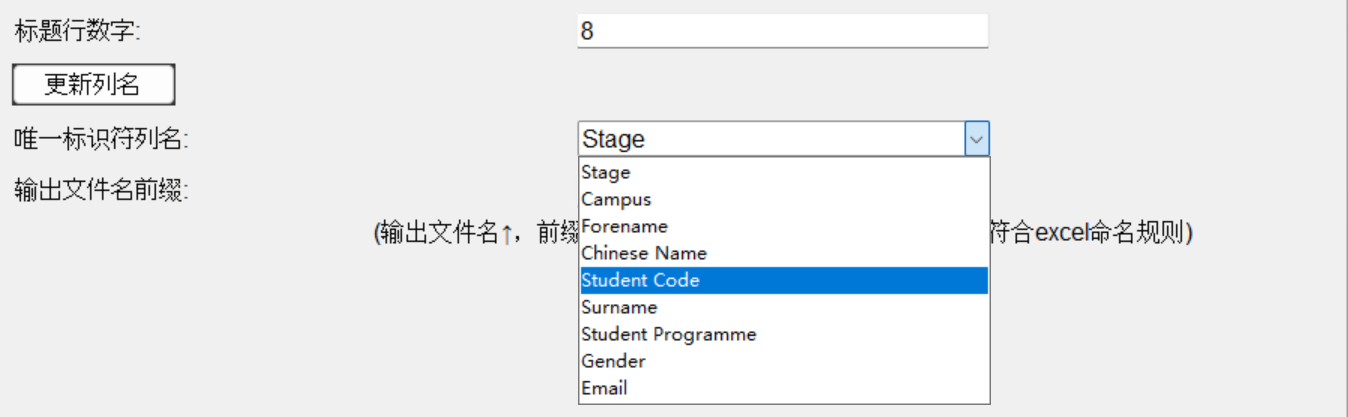
点击“选择输出文件夹路径”按钮，选择最终生成的重复/非重复文件保存的路径位置。



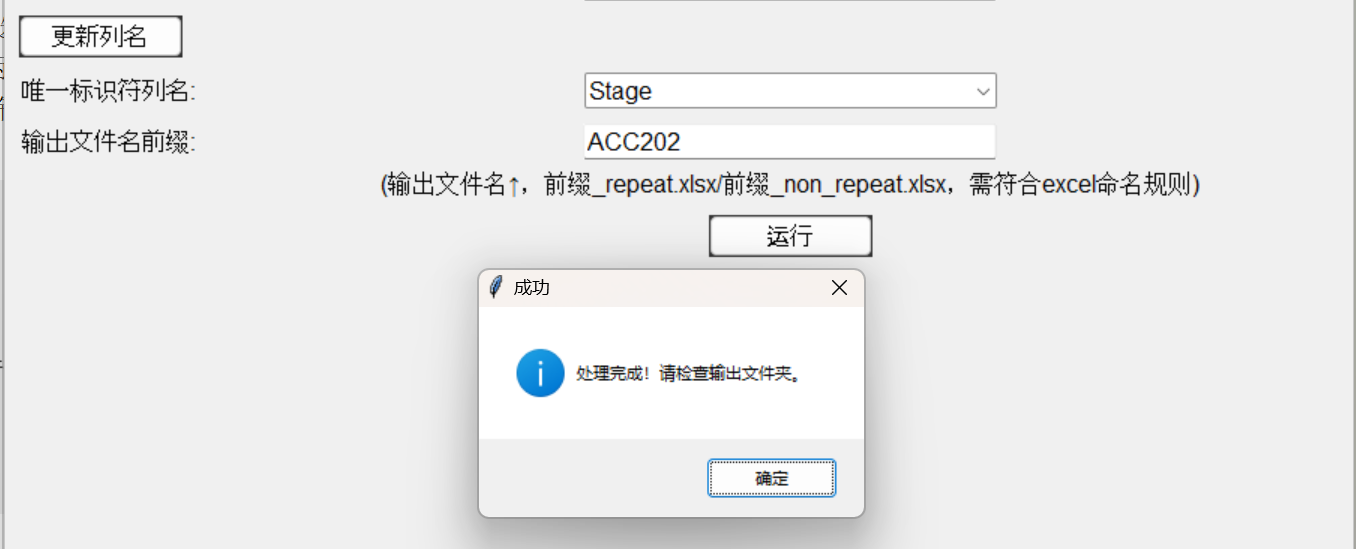
如下图Excel文件，由于该表格含有信息的标题行位于第8行，因此在“标题行数字”框内输入“8”，并点击，更新列名。



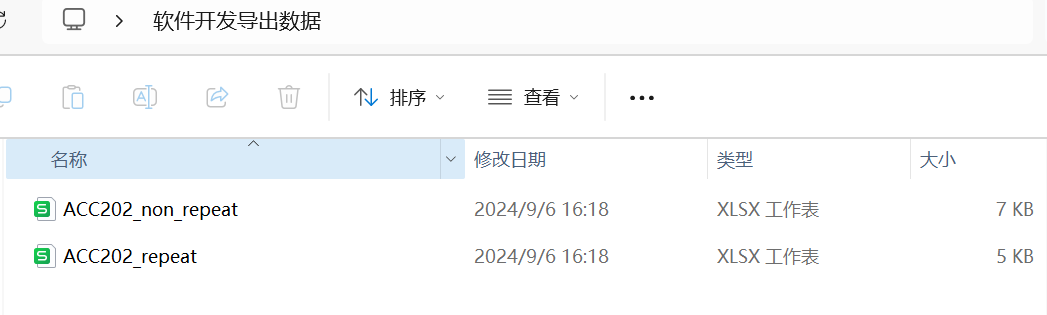
此操作将会读取选择的行（此处为第8行）每一列的名称，并在“唯一标识符列名”的下拉框中显示，用户需要挑选用以筛选的唯一标识符列名的选项。在此这个案例中，选择“Student Code”也就是学号，作为筛选两个表格的唯一标识符列名，也就是说，软件会根据两个文件中的Student Code（学号）来筛选出2022年和2023年ACC202课程的重复和非重复学生。



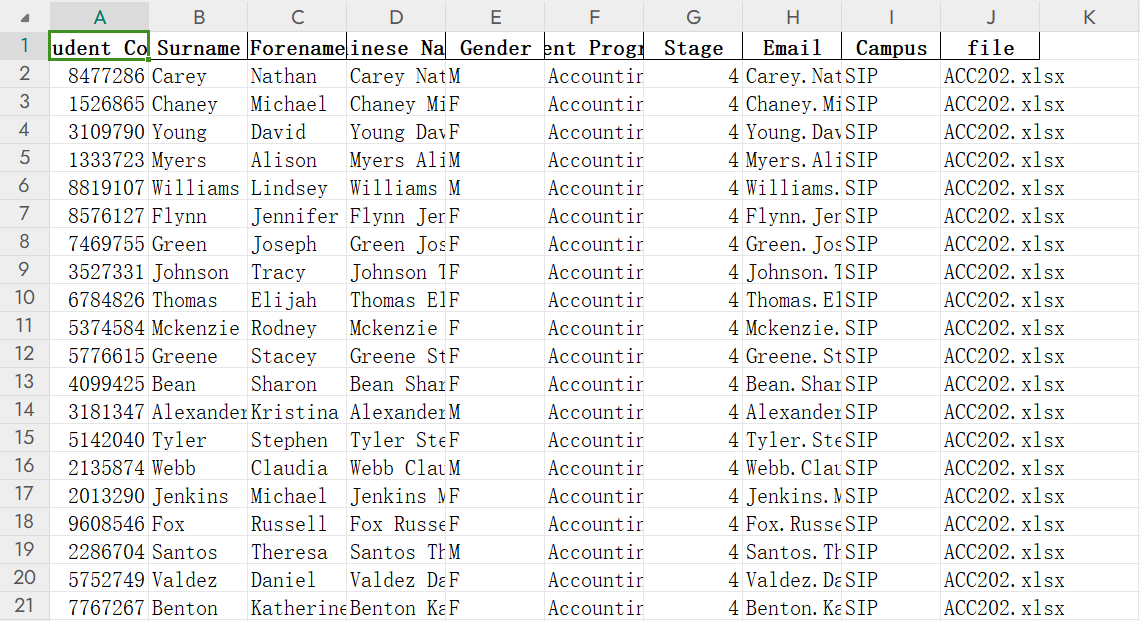
接下来，在“输出文件名前缀”框内输入你想自定义的内容，如图所示，如果输入前缀为“ACC202”，那么得到的文件名如图所示。



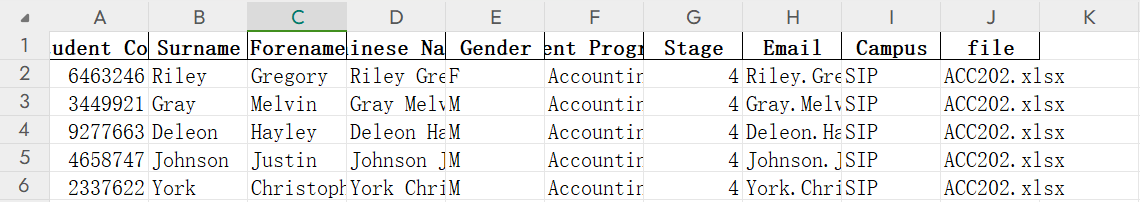
处理成功，下图为输出的文件名和表格内容



某对文件（ACC202）非重复内容Excel：



某对文件（ACC202）重复内容Excel：



### 3.2 批量文件处理操作流程

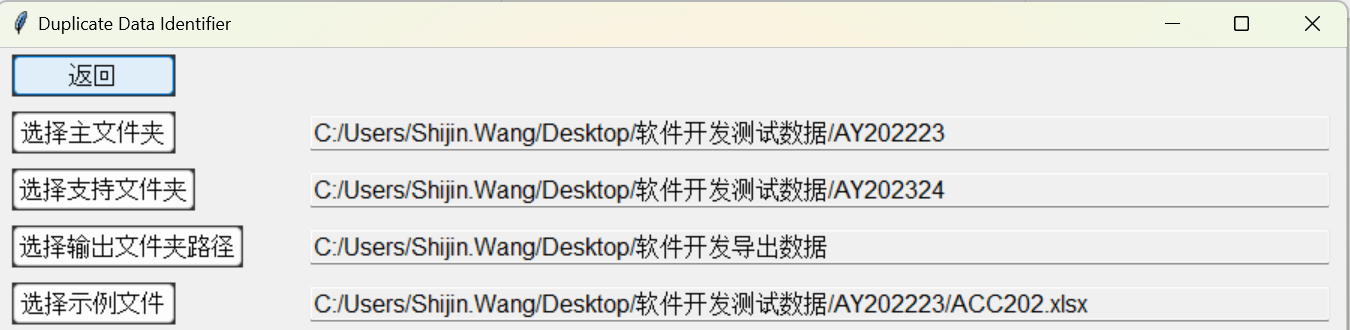
在处理多个文件时，

首先点击“选择主文件夹”按钮，选择需要被处理的Excel文件路径地址。

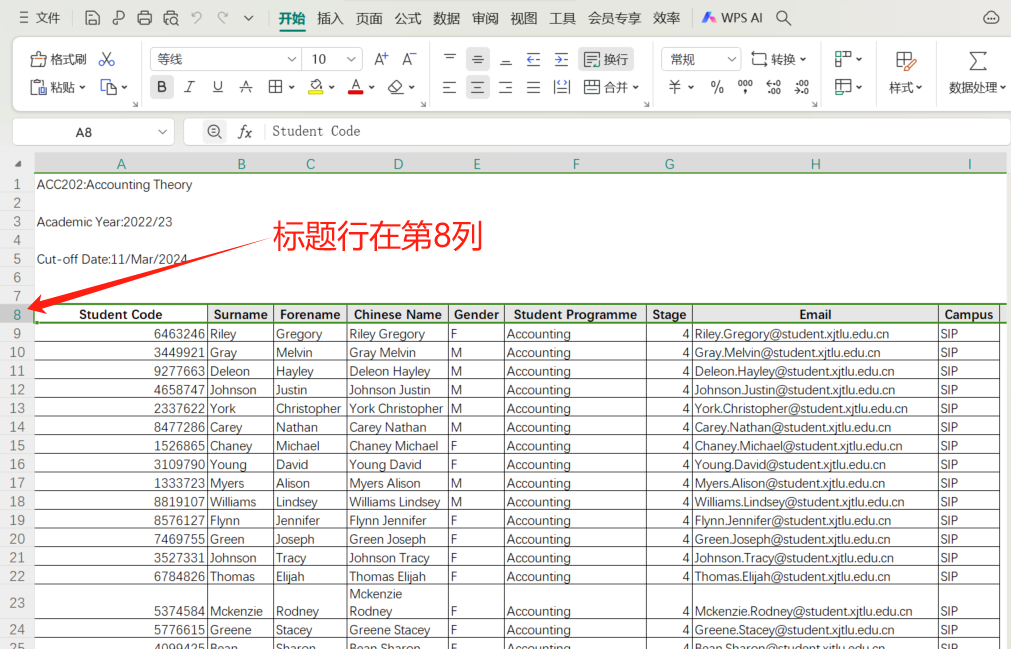
在选择支持文件时，点击“选择支持文件夹”按钮，选择用以辅助识别重复/非重复内容的支持文件路径。

点击“选择输出文件夹路径”按钮，选择最终生成的重复/非重复文件保存的路径位置。

再点击“选择示例文件”按钮，选择用于作为样本的示例文件。

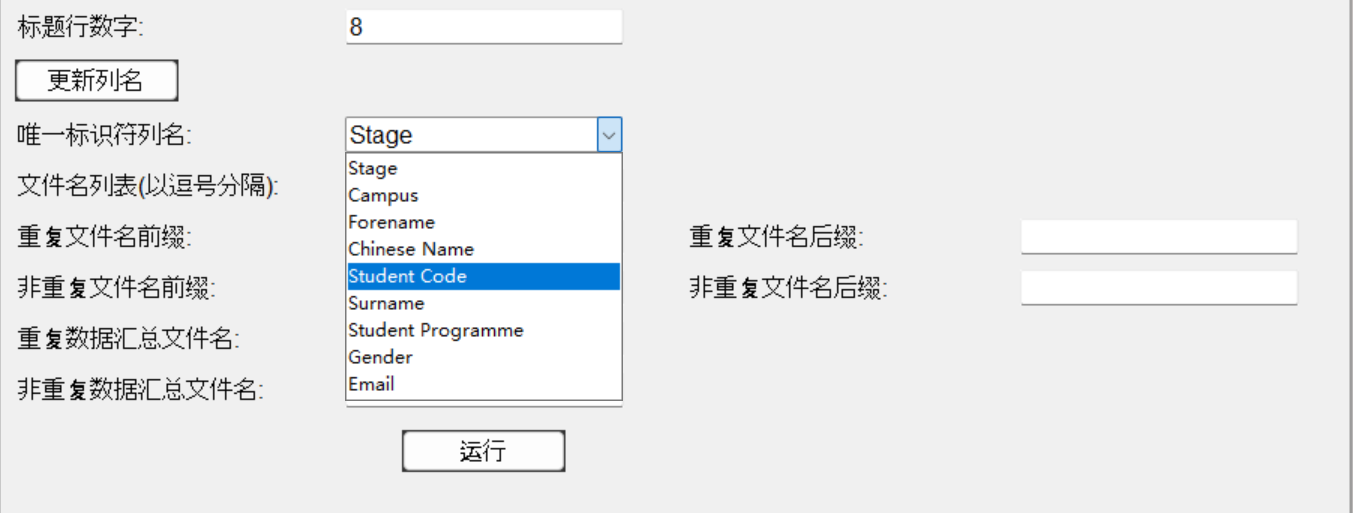


示例文件如下图Excel文件，由于该表格含有信息的标题行位于第8行，因此在“标题行数字”框内输入“8”，并点击，更新列名。



此操作将会读取选择的行（此处为第8行）每一列的名称，并在“唯一标识符列名”的下拉框中显示，用户需要挑选用以筛选的唯一标识符列名的选项。

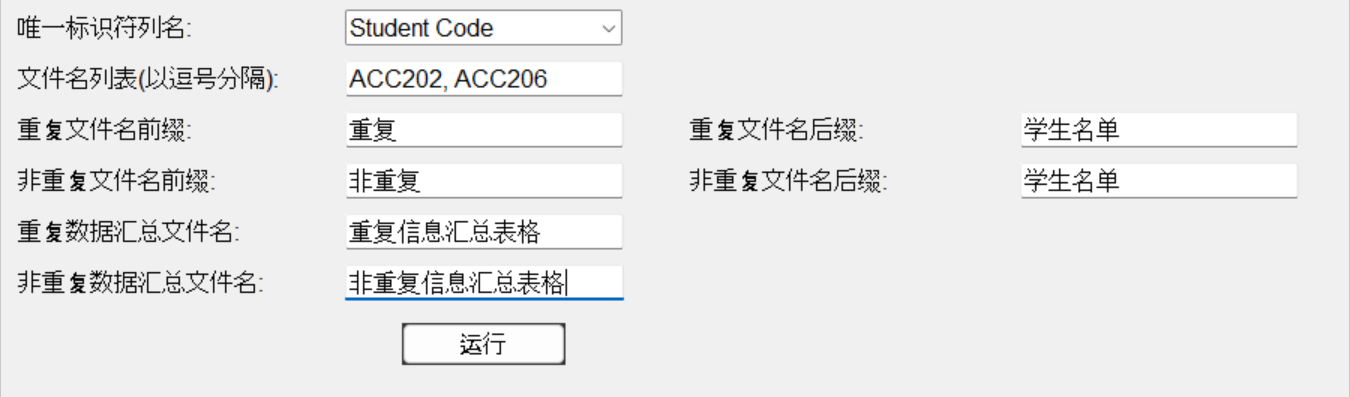
在此这个案例中，选择“Student Code”也就是学号，作为筛选两个表格的唯一标识符列名.也就是说，软件会根据两个文件夹中的Excel文件中的Student Code（学号）来筛选出2022年和2023年ACC202, ACC206课程的重复和非重复学生。



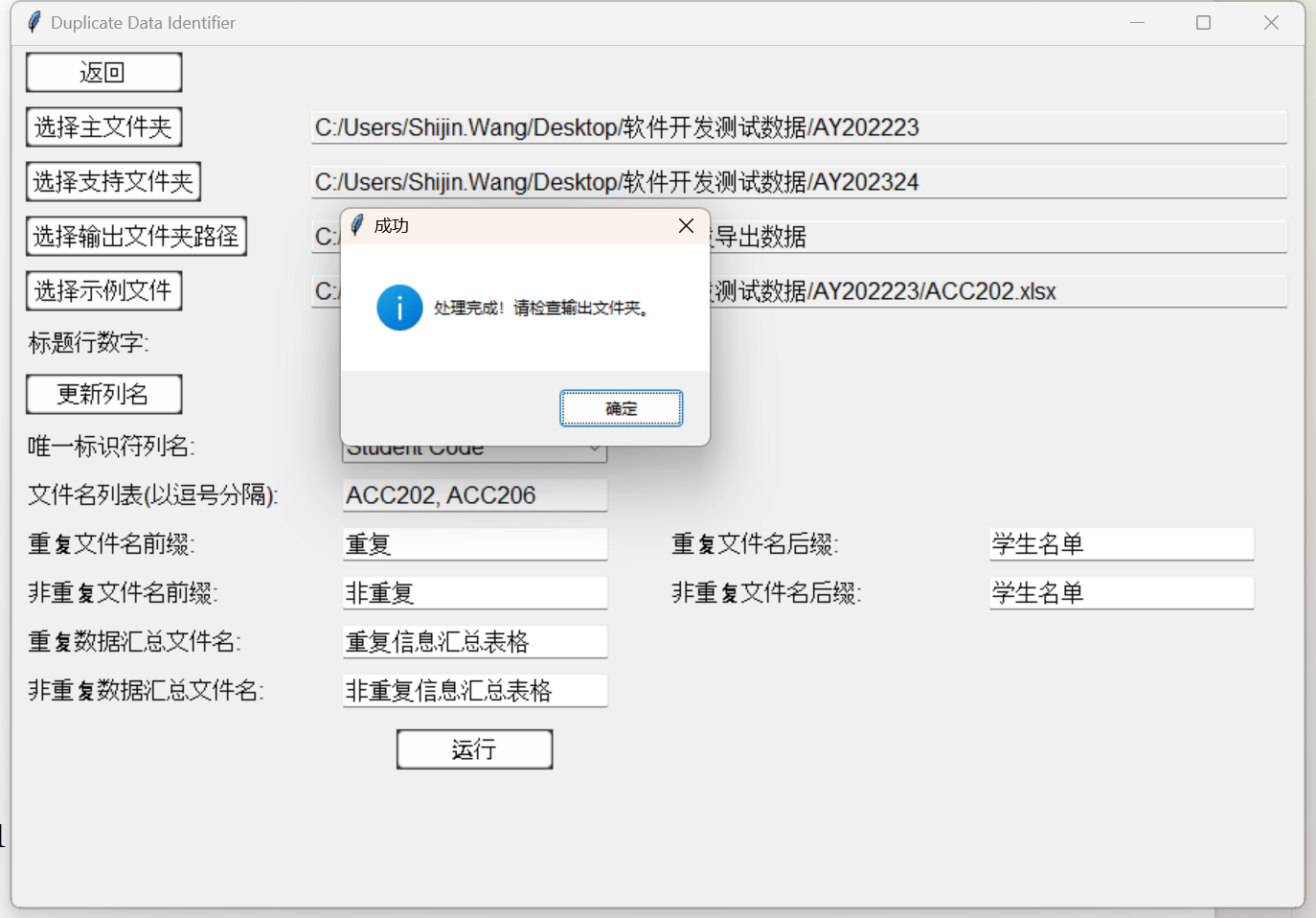
在选择唯一标识符列名后，由于本次是处理批量（多个文件对），因此需要在“文件名列表”右侧方框中，输入文件名列表，例如在本次处理中，需要输入“ACC202, ACC206”。

最后，在处理批量文件时，软件还支持自定义重复文件名，非重复文件名的前后缀，并会自动根据各个重复/非重复文件，合并生成汇总文件，同样地，汇总文件的名称也支持自定义。

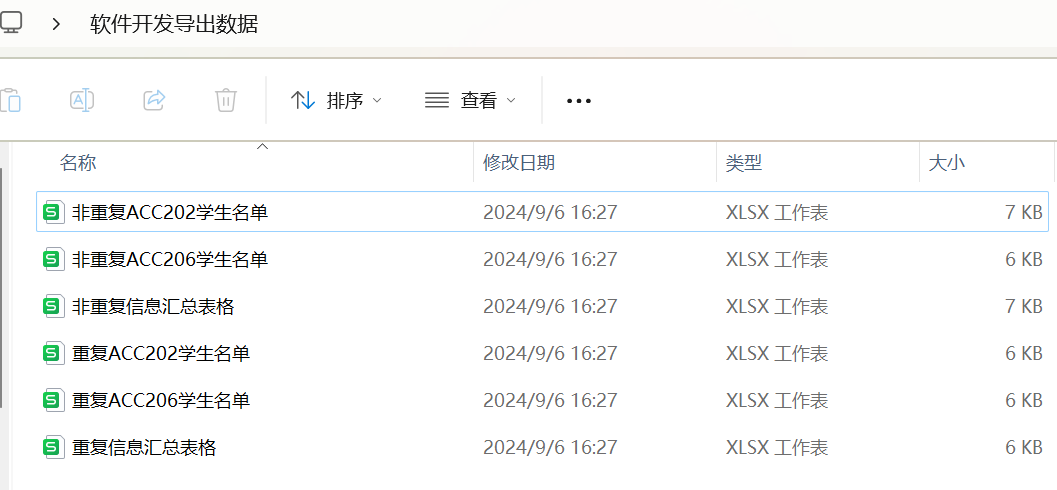
如下图所示



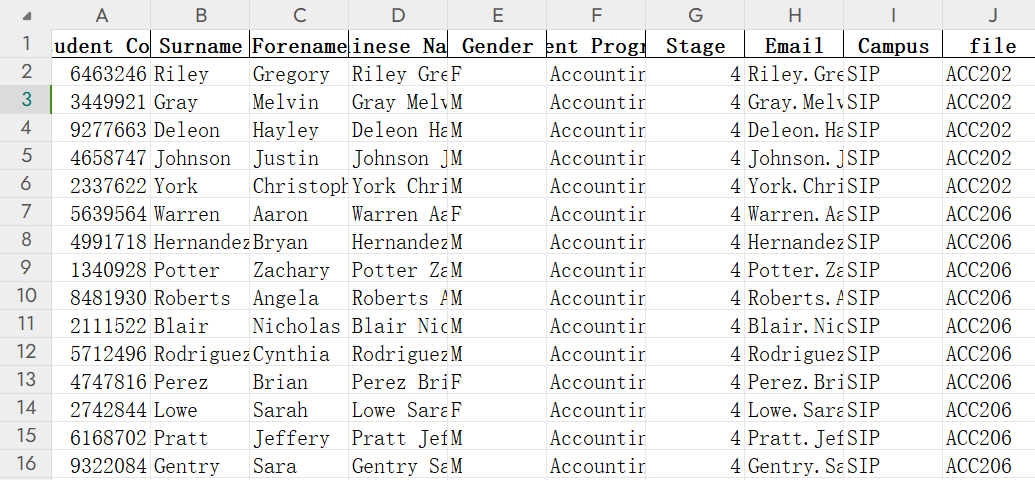
点击运行，显示处理完成。



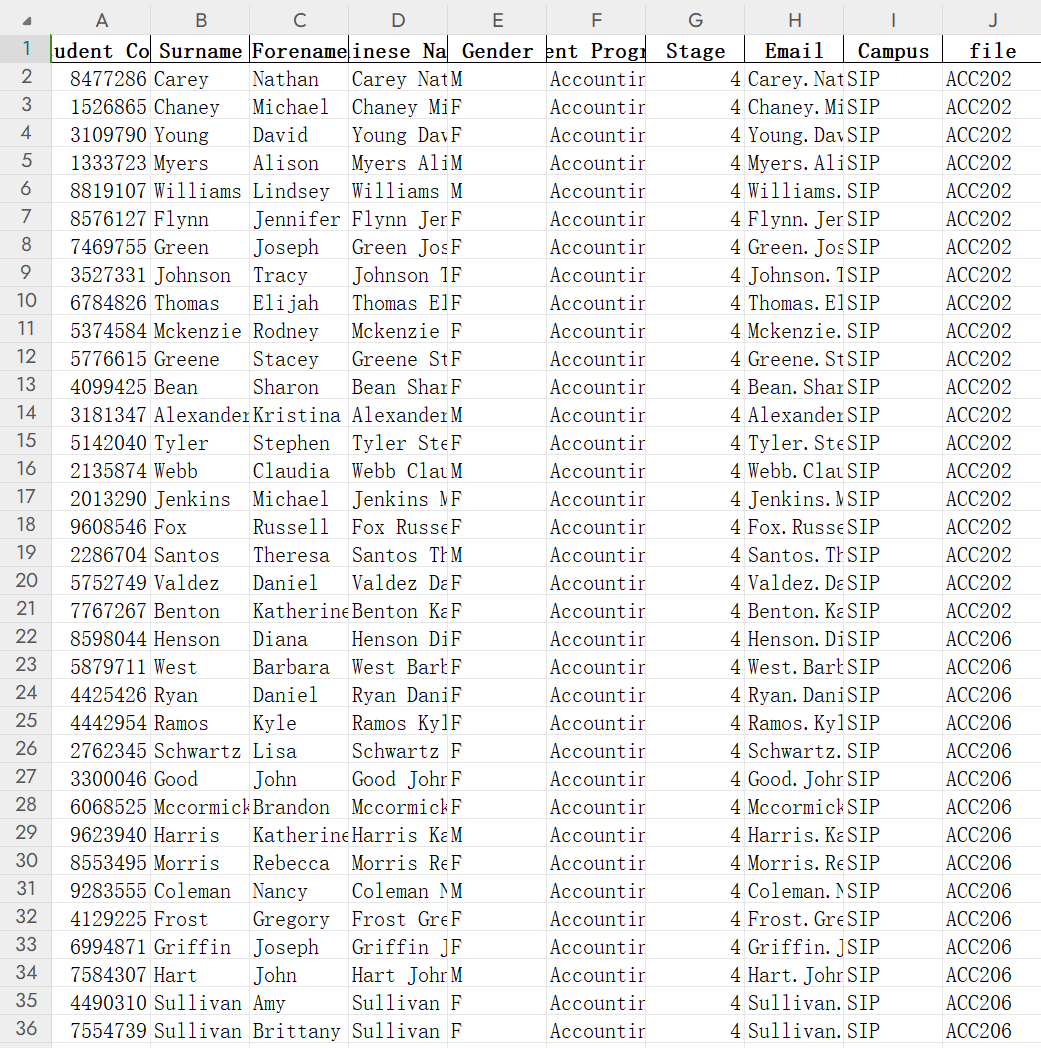
处理成功，下图为输出的文件名和表格内容。

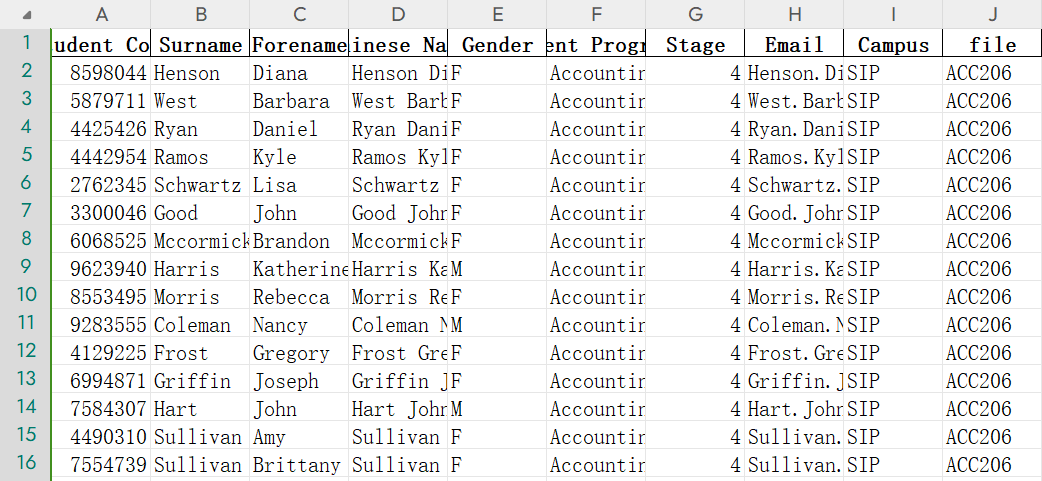


ACC202, ACC206重复信息汇总表格。将两个不同的文件对中重复的信息合并进入了一个表格，并在最右边列加入了“file”列，显示信息的来源文件名称。



ACC202，ACC206不重复信息汇总表格。将两个不同的文件对中不重复的信息合并进入了一个表格，并在最右边列加入了“file”列，显示信息的来源文件名称。



某对Excel文件（ACC206）的非重复内容Excel：

至此，处理单个/批量文件的具体步骤已完全呈现

## 4. 错误处理

### 4.1 输入验证

* 在运行处理之前，系统会检查所有必填字段是否已经填写，若有缺失则会提示错误信息。

### 4.2 文件与路径选择

* 系统要求用户选择有效的文件路径和输出文件夹路径，如选择错误，系统将弹出错误提示。

### 4.3 数据处理异常

* 在处理过程中，如出现任何异常（例如文件格式问题、列名问题等），系统将捕捉错误并弹出提示信息，指导用户进行修正。

## 5. 数据结构与存储

### 5.1 数据结构

* 主文件和支持文件：包含Excel文件的路径和数据表中的列。
* 处理结果：两个输出Excel文件，分别包含重复数据和非重复数据。

### 5.2 存储

* 处理结果会根据用户设置的文件名前缀和后缀保存在指定的输出文件夹中。