**《文档查重》操作手册**

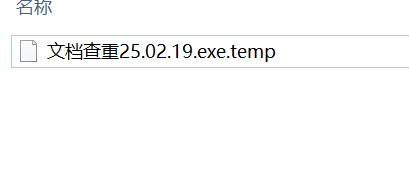
亲爱的朋友，当你看到这个文档的时候，恭喜你，你已经对文档之间重复内容无所畏惧了，你可以轻松使用程序查找出来多个文档之间的重复文本、重复表格以及重复图片。尤其在招投标工作中，当你上传了两个投标文件和一个招标文件时，就可以找到同时存在于两个投标文件中的文本而不存在于招标文件的文本（因为我们在制作投标文件时，常常会复制招标文件的某些文字），这种坚定又轻松的感觉，相信只有你经历过人工查重才能有所体会。

下面我简单介绍一下操作方法，如有不妥，请指正！

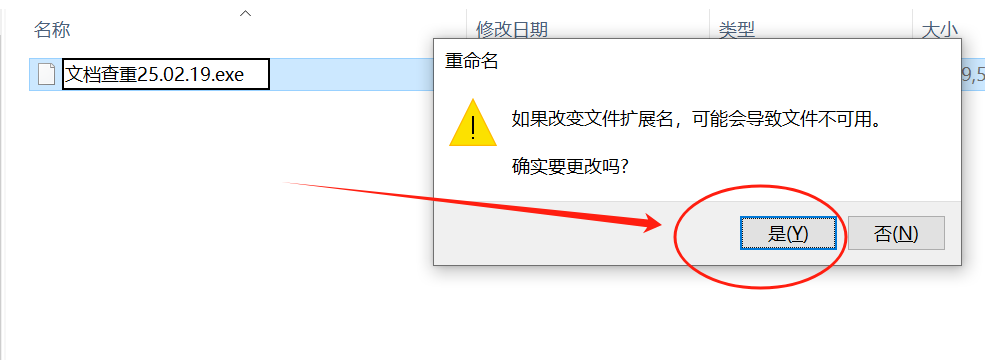
**软件环境要求：win10以上系统。**

**对比的文件格式：.docx文件。如果是.doc文件需要先将文件另存为.docx文件，再进行导入比对。**

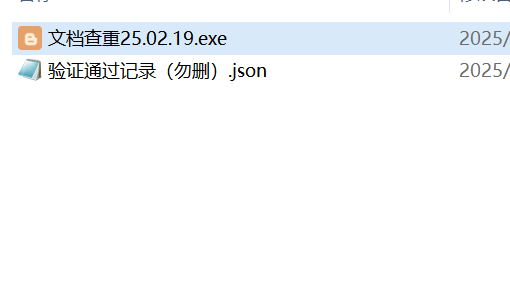
**一、下载**



当你下载“文档查重25.02.19.exe”程序文件时，很有可能保存到计算机上是带有后缀的文件，不用担心，这是计算机为了保护你免遭病毒侵害的措施。

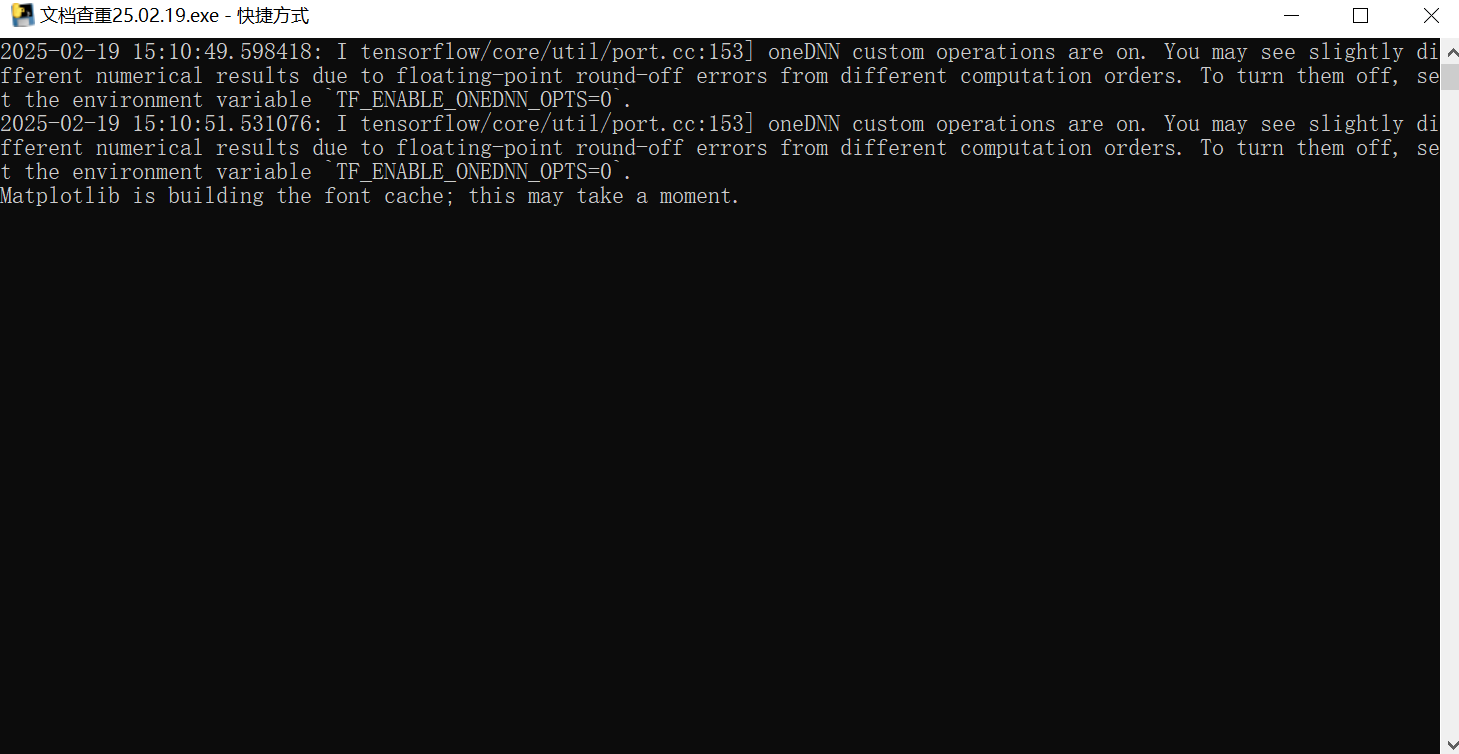


你可以直接修改文件名“文档查重25.02.19.exe.temp”为“文档查重25.02.19.exe”，选择“是”。

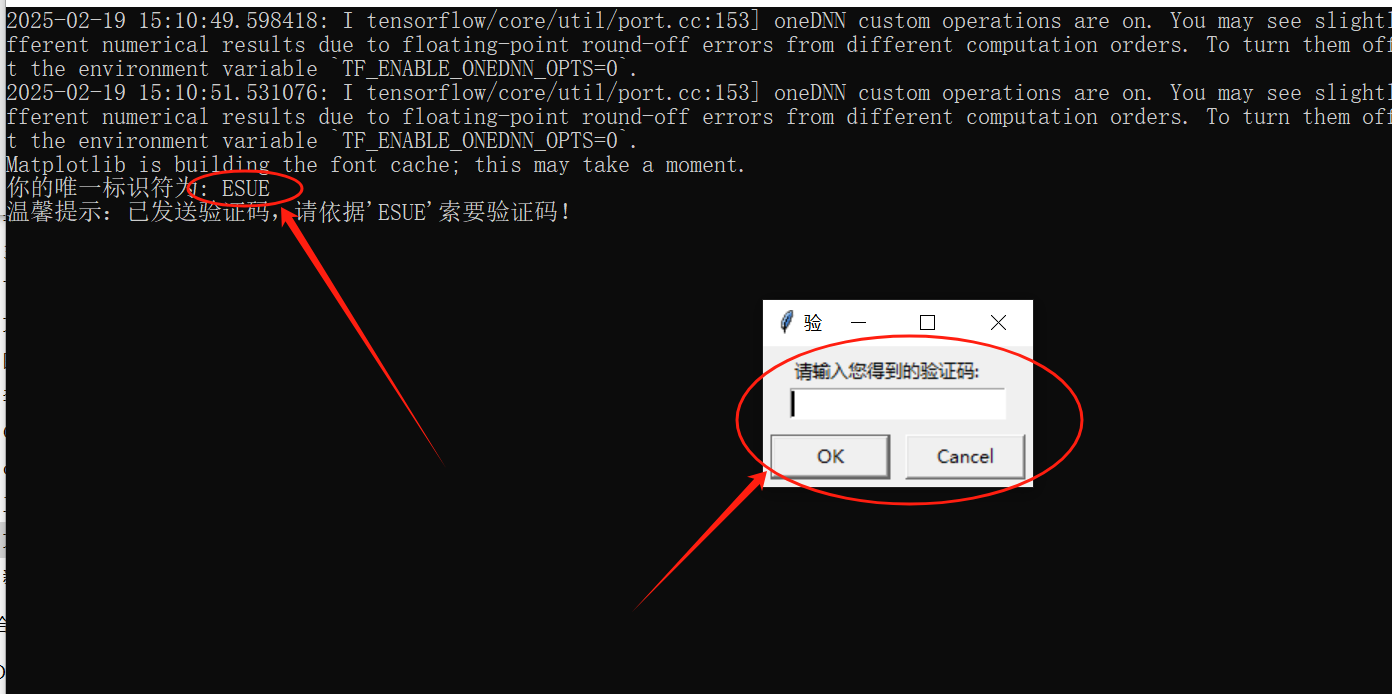


然后得到程序文件“文档查重25.02.19.exe”，请把这个文件保存在安全的地方，然后创建快捷方式，把快捷方式放在方便的地方进行使用，以免误删。

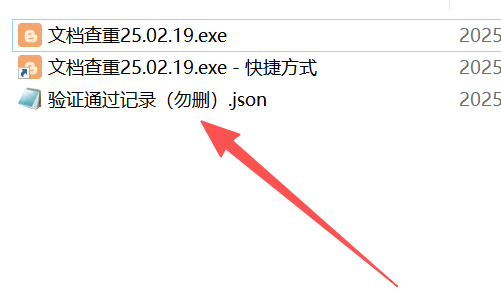
**二、验证**



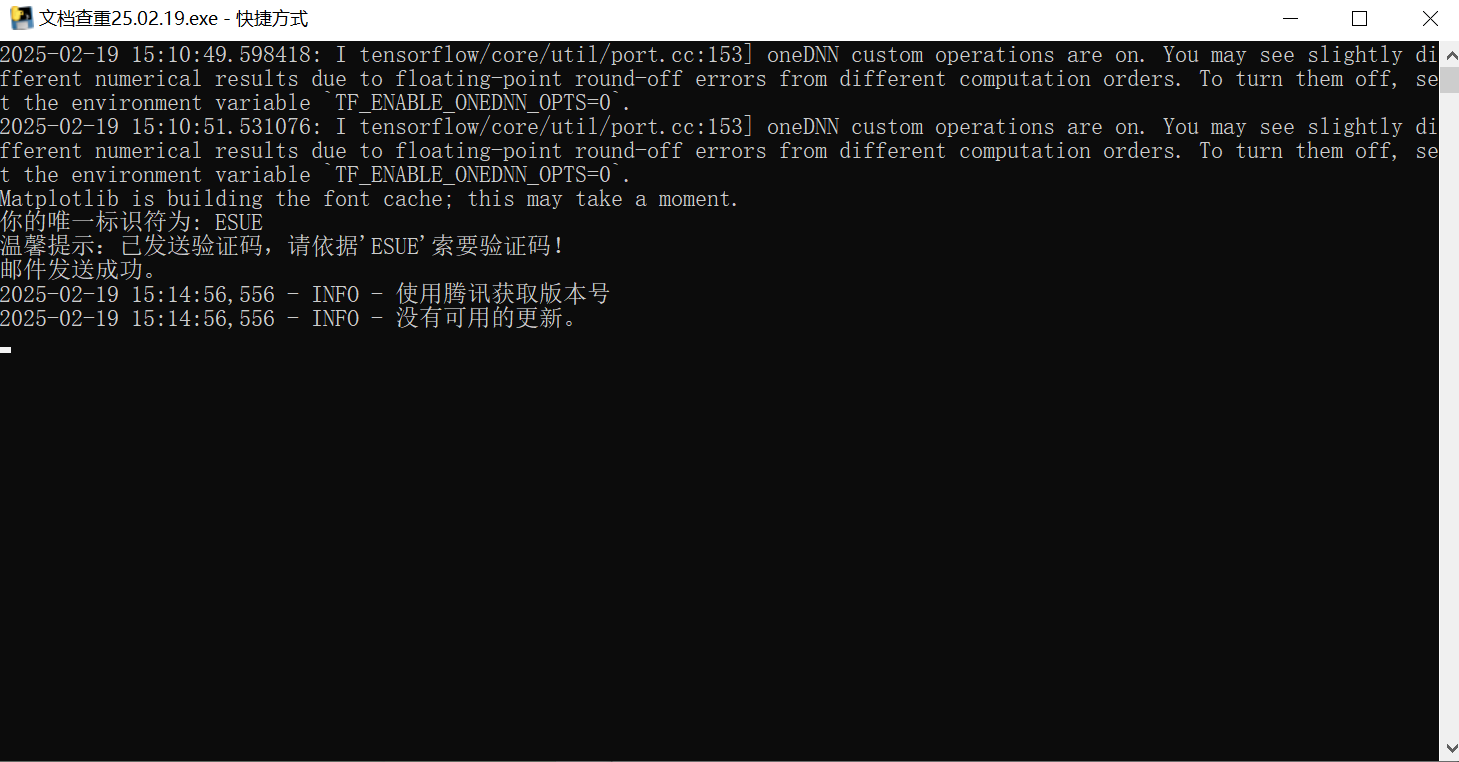
双击“文档查重25.02.19.exe”，出现控制台界面（请耐心等待......）。



第一次开启程序时，是需要进行验证码验证的，并且一次验证，终身有效。请依据控制台生成的唯一标识符“ESUE”向管理员索要验证码，然后将验证码输入。



验证码正确后，在程序目录会生成一个记录验证通过信息的“验证通过记录（勿删）.json”文件，切勿删除或更改它，否则将导致验证失效。





随即进入主程序（运行主程序时，切记不可关闭黑色控制台）。验证成功。

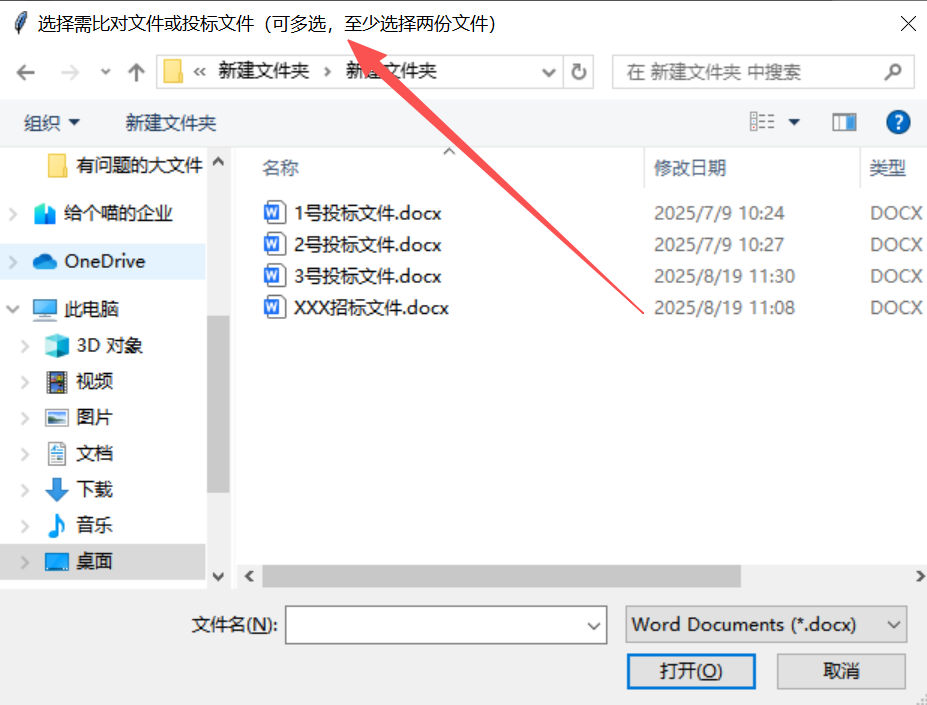
**三、使用**

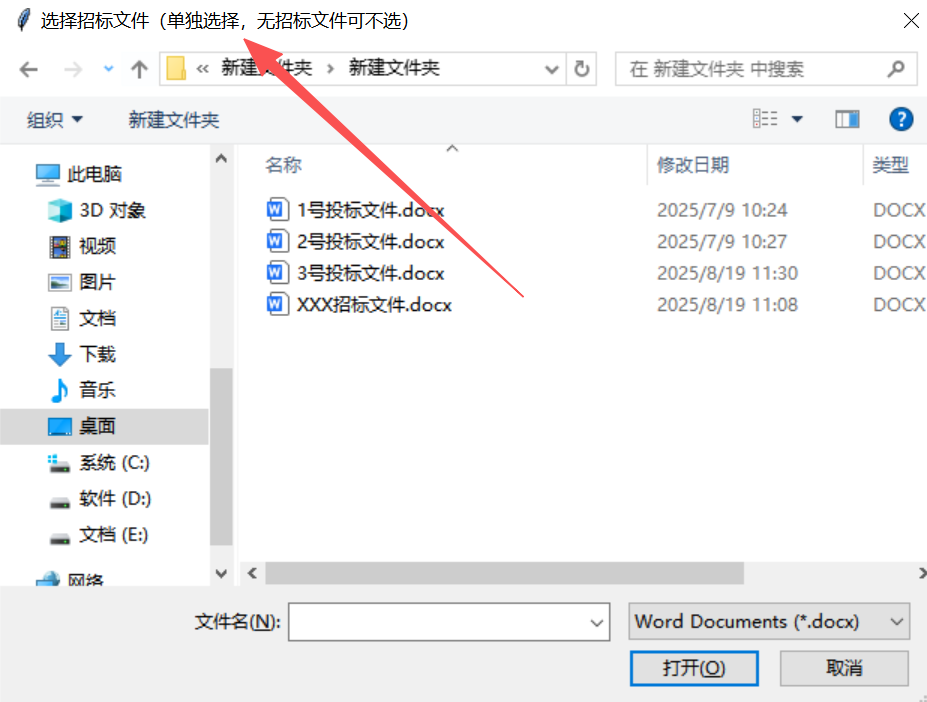


1、显示区

这部分将对导入文件的路径及文件名、输出路径、进度信息进行展示。

1. 添加docx文件





点击“添加docx文件”，将需要进行比对的文件导入，可一次导入多个，也可多次不同路径进行导入。可选择多个投标文件和一个招标文件。**注意这里只能导入docx文件进行比对，不可将doc文件的后缀改为docx，需要将doc文件另存为docx文件才可以**。

3、清空文件列表

可以对已经导入的文件进行清空。

4、个性化设置



海明阈值：hamming图片相似度(在0到64之间)，一般为5的整数，数值越小，要求越高。如果需要完全相同的图片，这里可以设置为0，但是考虑到裁剪、调整大小可能会造成变化，建议这里为5。

相似度阈值：similarity图片相似度(在0到1之间)，一般为0.9的浮点数，数值越大，要求越高。如果需要完全相同的图片，这里可以设置为1，但是考虑到裁剪、调整大小可能会造成变化，建议这里为0.9。

容差：表格单元格重复率(在0到1之间)，一般为0.9的浮点数，数值越大，要求越高。

最小字符数：文本连续最字符串值(大于等于1)，一般为50的整数，数值越小，谨慎度越高，相应地比对效率越低。

批次大小：每次处理的任务数量(大于等于20)，默认为100的整数，数值越大，内存占用可能越高。计算机性能较好时，可以增加批次大小，提高比对效率。

过滤字符：您需要过滤掉的字符，如果是多个，用“；”分号隔开。忽略某些字符进行文字比对。

过滤所有标点符号：忽略所有标点符号进行文字比对，可以打勾。

启动招标文件表格与图片过滤（严格模式）：会自动剔除重复内容中与招标文件相似的“表格”与“图片”，干系重大，慎用！

1. 状态信息

包括是否导入了招标文件，是否进行某些字符的忽略，是否使用严格模式。

6、输出路径

选择比对结果的输出文档路径。

7、开始比较

信息输入完成后，点击开始比较。

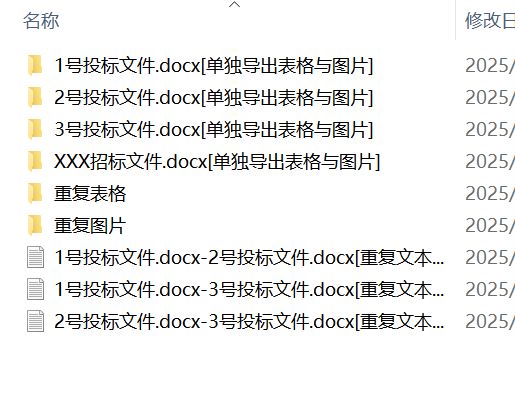
8、进度条

程序进度条。

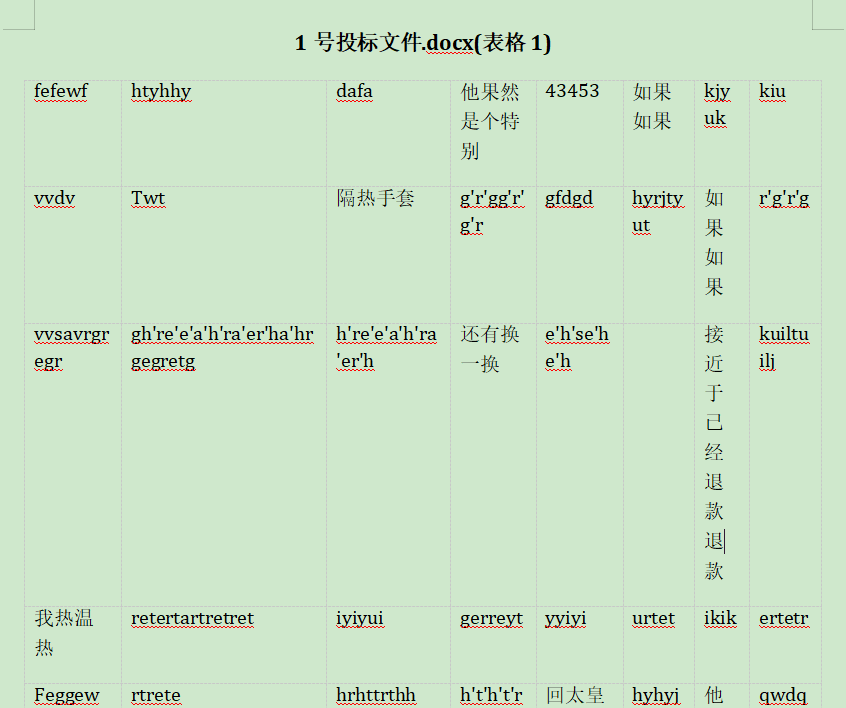
**四、输出内容**

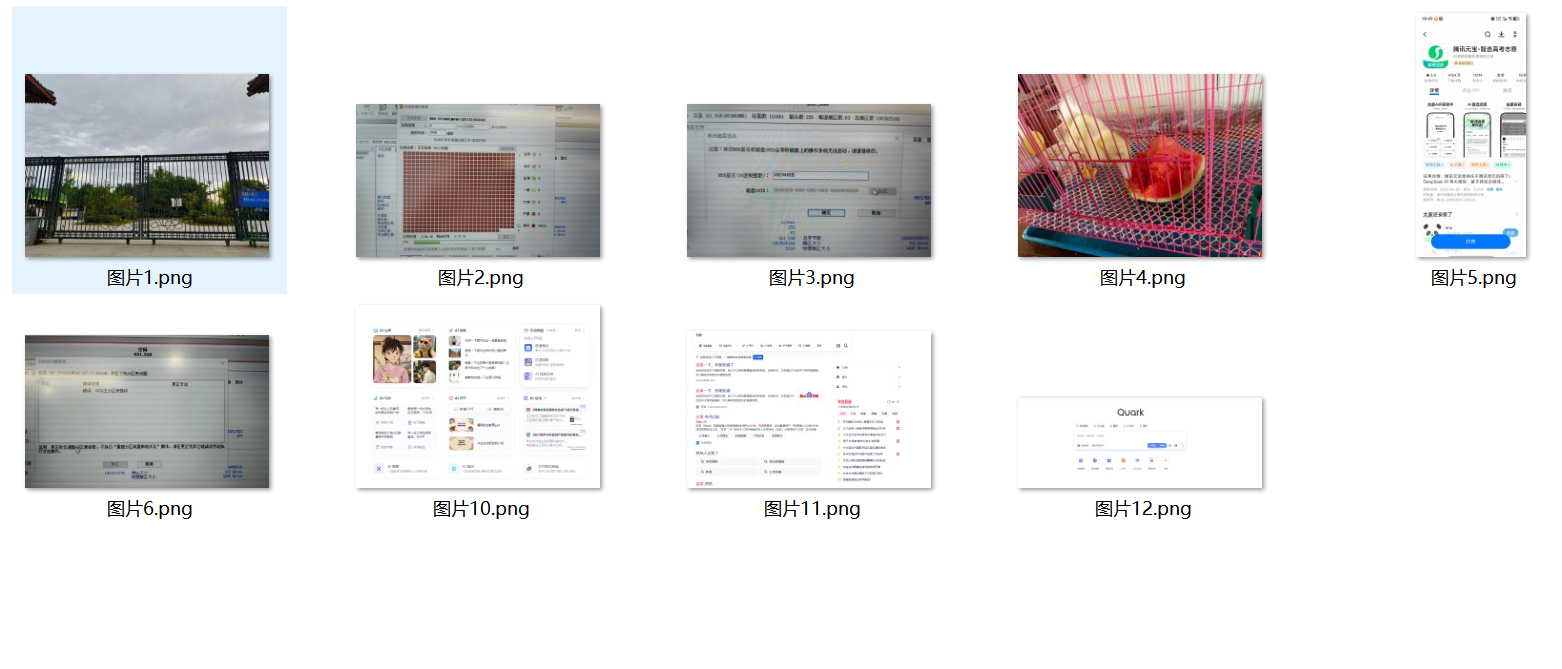


在我选择了“1号投标文件”、“2号投标文件”、“3号投标文件”、“招标文件（找）”进行比对。输出内容如下图



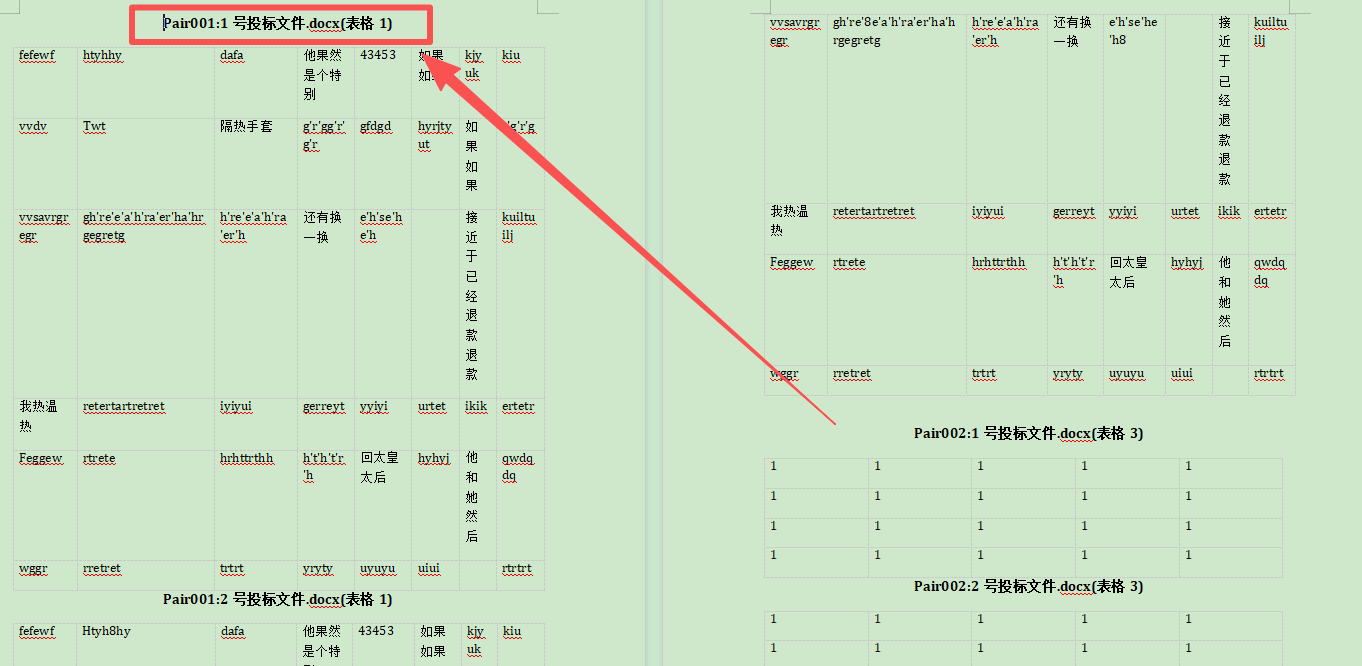
文件夹“1号投标文件.docx-[表格与图片]”（同样“2号投标文件.docx-[表格与图片]”）中保存的是从“1号投标文件.docx”文件导出的所有表格和图片。





文件夹“重复表格”为每两个文档比对出的重复表格结果。

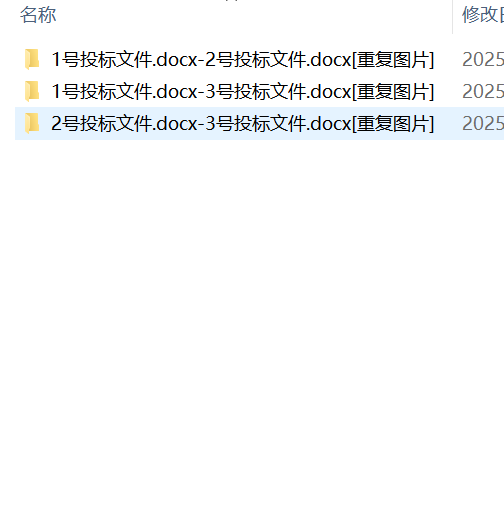


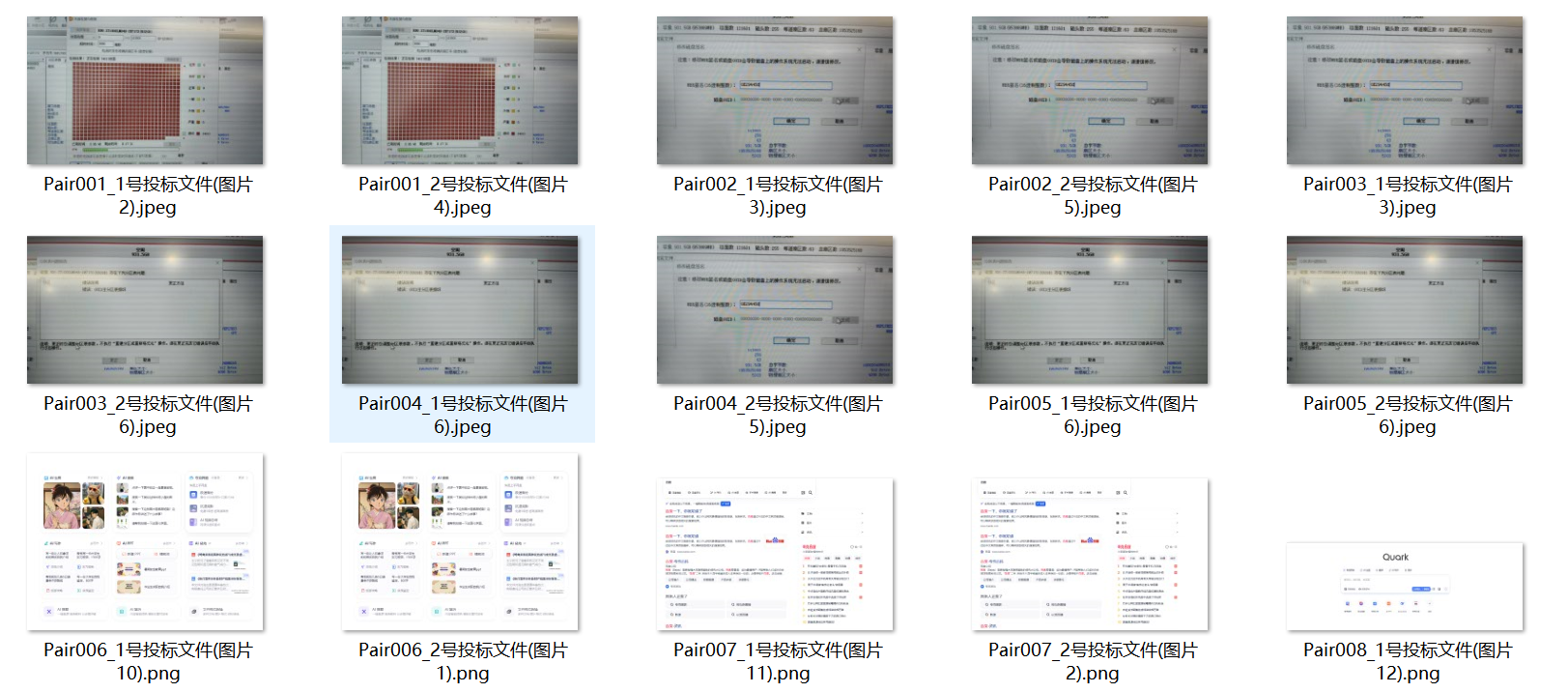


Pair001就表示第一对，1号投标文件.docx(表格1)，就表示这个表格是“1号投标文件.docx”的表格1（表格的编号与“1号投标文件.docx-[表格与图片]”中表格编号对应）

Pair001就表示第一对，2号投标文件.docx(表格1)，就表示这个表格是“2号投标文件.docx”的表格1（表格的编号与“2号投标文件.docx-[表格与图片]”中表格编号对应）

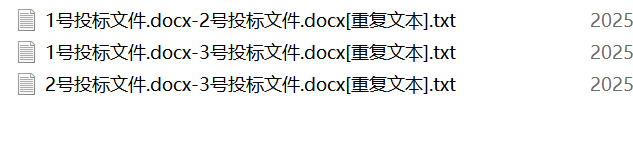
文件夹“重复图片”为每两个文档比对出的重复图片结果。



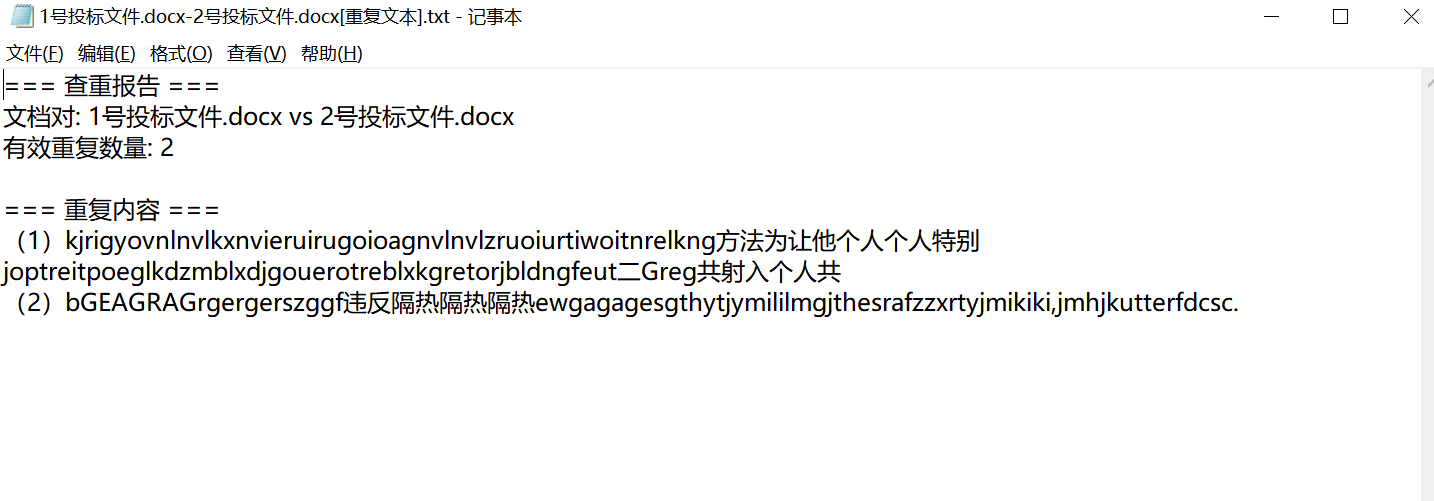


pair001就表示第一对，1号投标文件(图片2)，就表示这个图片是“1号投标文件.docx”的图片2（图片的编号与“1号投标文件.docx-[表格与图片]”中图片编号对应）

pair001就表示第一对，2号投标文件(图片4)，就表示这个图片是“2号投标文件.docx”的图片4（图片的编号与“2号投标文件.docx-[表格与图片]”中图片编号对应）



txt文档“1号投标文件.docx-2号投标文件.docx[重复文本]”中内容为同时存在于“1号投标文件”与“2号投标文件”的文本，但是不存在与招标文件“XXX招标文件”。



重复长度从长到短排列。可依据每一段的内容在你的投标文件中进行查找，然后修改。

**五、更新**

由于本程序包含文字、表格、图片比对，尤其是图片比对采用深度学习模型，所以所占内容较大，如果不能自动更新版本，请联系管理员(QQ：1011489402)。