# 作业相似度比较系统 V2.0

# 用户手册

**方宏**

（1010402827@qq.com）

## 1 系统介绍

随着信息技术在教育教学领域的广泛应用，越来越多的学生作业及考试采用电子档的形式进行，电子档作业在具有易存储、传输、共享、查阅展示方便等优点的同时，还具有易复制、易修改的特点，使得学生之间互相拷贝作业变得容易，为了能够准确地检测出学生作业之间相似的部分，同时减轻教师的工作量，开发了作业相似度比较系统。

作业相似度比较系统主要检查、比较学生提交的电子档作业之间的相似度，能对程序语言（如 java、c、c#等）类作业、中文文档类作业（如实验报告等）之间的相似度进行比较，进而发现学生之间互相抄袭的行为。

程序类作业的相似度比较基于两个开放系统：一个是基于网络服务的 moss 系统（斯坦福大学开发的支持十几种编程语言代码相似度比较的系统），另一个是本地执行的 sim 系统（支持 java、c 语言）。本系统在它们基础上进行了二次开发和封装，针对 moss 系统，开发出了客户端存取模块，实现了作业提交、结果获取和解析、结果排序等功能；针对 sim，则将其集成到系统中，在 moss 因网络故障等原因不可用时，作为替代产品使用。

中文文本文档作业相似度的比较则基于 shinglecloud 算法（一种基于文本指纹的、语言无关的相似度快速计算方法），文档主要处理过程如下：（1）读取不同格式（txt、doc、pdf 等）的文档，并将其转换成能统一处理的文本；（2）对文本进行预处理：如去掉标点符号、停用词等；（3）使用 shinglecloud 算法计算文本之间的相似度；（4）根据相似度排序，输出比较结果。该算法使用 N-GRAM，是语言无关的，即使中文中有英文，仍能正确处理。

系统的主要特性如下：

（1） 支持多种程序语言（几乎涵盖了实际在用的各种编程语言）、多种格式文档（txt、doc、docx、pdf 等主流文档格式）、多种文字语言（中文、中英混合、英文等）作业相似度的比较。

（2） 比较准确度高，一致性好。经过数千份作业的实际测试，结合人工比对，证明系统给出的结果令人满意，重复比对结果均一致。

（3） 性能良好，支持大量作业相似度的快速比较。

该系统在实际作业相似度检查中显示出优良的性能：比较速度快，较之手工比较，速度提高了几十上百倍，节省了教师大量的时间；比较客观，准确度高，由于是程序自动比较，避免了人工比较中可能出现的主观不确定性。

系统提供图形界面执行方式，程序执行后的初始界面如图 1 所示，基本使用过程是：先点击“选择作业”按钮，选择作业所在的路径；然后选择作业类型，确定相似度门限值，如果是程序作业，还需要进一步确定检查工具类型和相应的

程序语言；接着点击“执行比较”按钮，执行作业相似度比较；比较结束后，点击“查看结果”按钮，查看作业比较结果。


图 1　作业相似度比较系统主界面

## 2 比较程序作业

比较程序作业的步骤如下：**(注意：使用 moss 比较，应将作业拷贝到本系统安装路径的 testdata 子目录下，作业组织及命名可参考 testdata 下的示例)**

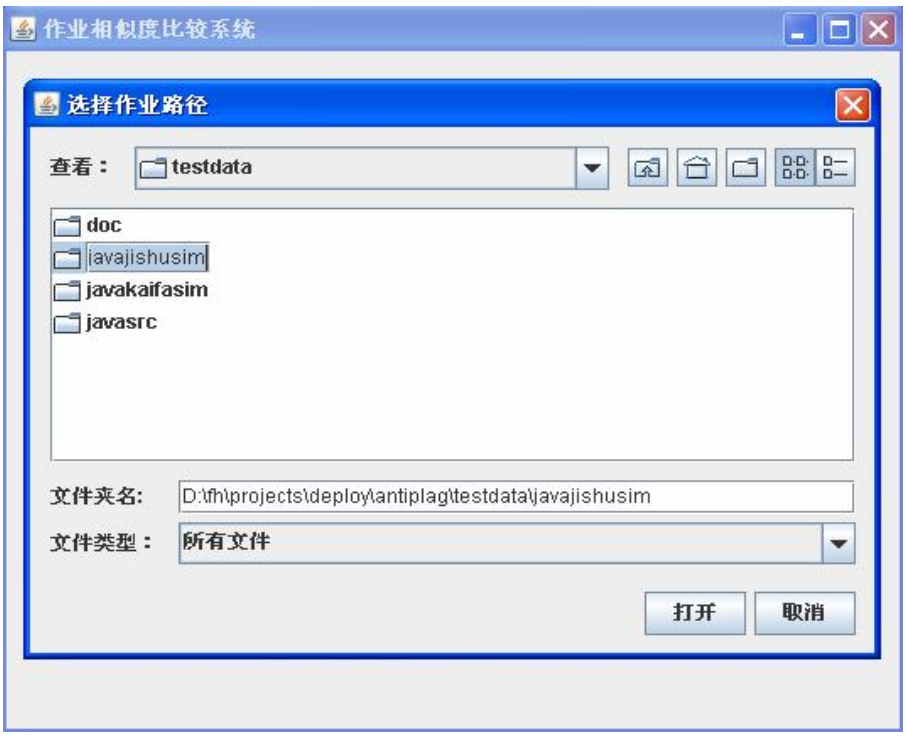


图 2　选择作业路径对话框

（1） 在主界面点击“选择作业”按钮，弹出图 2 所示的对话框，选择程序作业所在的目录路径，点击“打开”按钮，得到如图 3 所示的结果。


图 3　选择作业路径后的主界面

（2）设置比较参数。需要设置作业类型（程序作业、文本作业）、相似度限值（在此限值之上的相似文档才会在比较结果中出现）、相似度检测工具（moss、sim）、程序语言种类。默认作业类型是“程序作业”，无需进一步确定；相似度限值默认是 30，也可以根据需要设置成 0-100 之间的任何值；检测工具默认是 moss；选择作业相应的程序语言。

（3）点击“执行比较”按钮，开始执行比较，等待一段时间后，比较结束弹出图 4 所示的对话框。

（4）点击“查看结果”按钮，弹出如图 5 所示的结果窗口。结果按相似度大小，由大到小输出。


图 4　执行结束对话框

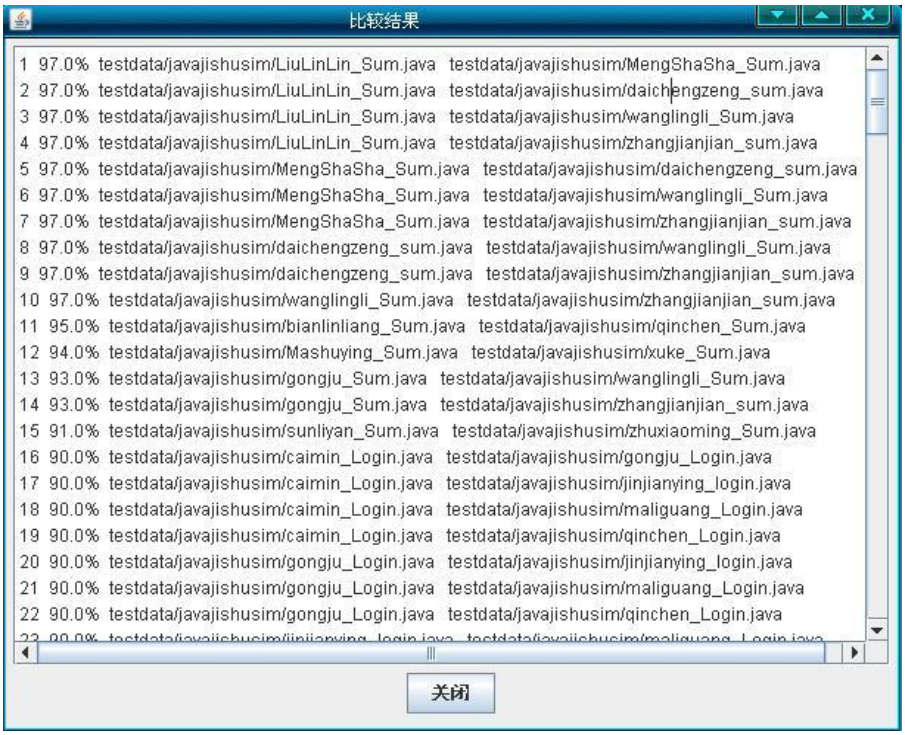


图 5　基于 moss 的相似度比较结果

（5） 如果 点击“执行比较”按钮时，弹出图 6 所示对话框，说明 moss 系统因为网络或主机原因不能访问，此时可以选择 sim 作为相似度检测工具，该工具在本地执行，无需存取网络。Sim 执行比较的结果如图 7 所示。需要注意的是由于算法不同，对于同一批作业，moss 与 sim 比较的结果不尽相同，这不仅体现在具体的相似度数值上，也体现在按大小排序的输出序列结果上。Sim 目前只支持 java 和 c 语言。


图 6　网络有问题时的对话框

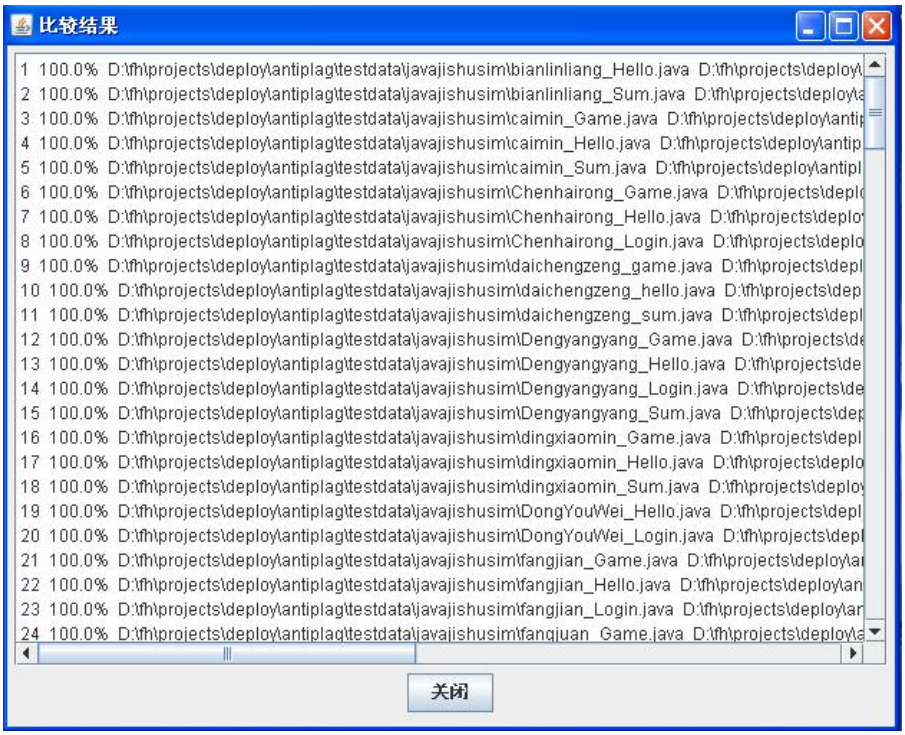


图 7　sim 比较的结果

## 3 比较文本作业相似度

文本作业相似度比较的步骤如下：

（1） 点击“选择作业”按钮，在“选择作业”对话框里，选择、确定文本文档作业所在的路径（这与程序作业路径的选择类似）。

（2） 设置参数。点选作业类型为“文本作业”，设置相似度限值。如图 8

所示。


图 8　文本作业类型

（3） 点击“执行比较”按钮，对文档作业进行比较。比较结束后会弹出相应对话框，如图 9 所示。


图 9　文本作业比较执行完毕

（4） 点击“查看结果”按钮，可以得到如图 10 所示的比较结果。注意结果是按相似度从大到小输出，且只显示相似度在门限值以上的结果。

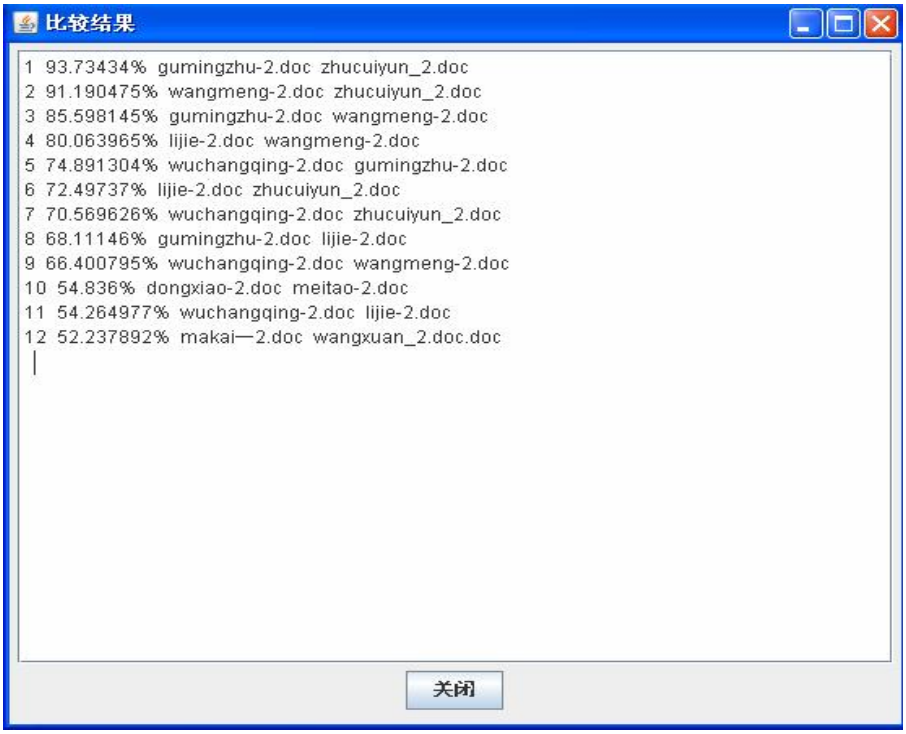


图 10　文本作业比较结果

## 4 转换文件

无论是程序类作业还是文档类作业，比较时都要求遵照指定的目录文件结构：作业根路径下，放置唯一命名的文件，具体如图 11 所示。

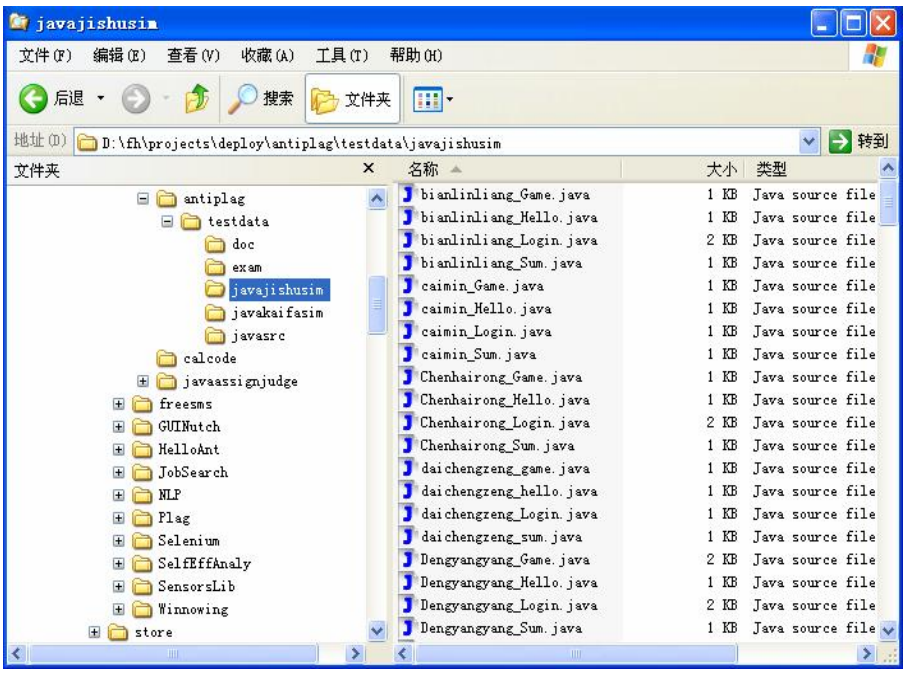


图 11　比较文件的目录结构

但实际教学中，特别是程序类作业，往往以项目方式提交整个作业，里面有多个文件，而实际只需要比较源代码，所以需要一个转换工具，将源代码或文档提取出来并组织成比较需要的目录文件形式。转换文件功能就是为此目的而设计的。具体操作如下：

（1） 在主界面点击“转换文件”按钮，出现如图 12 所示的对话框。

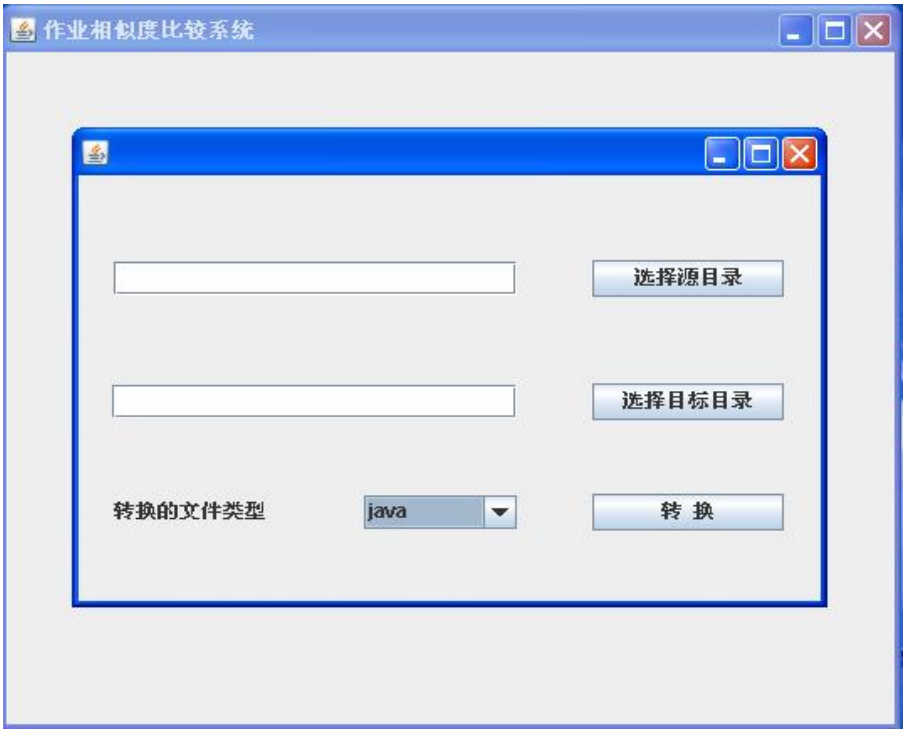


图 12 转换文件对话框

（2）在转换文件对话框中点击“选择源目录”按钮，弹出图 13 所示的源文件选择对话框，选择作业所在的根路径（该路径下是提交的每份作业文件或子目录，也可以是 zip 形式的压缩文件，系统支持自动解压）。路径确定后，点“打开”按钮。

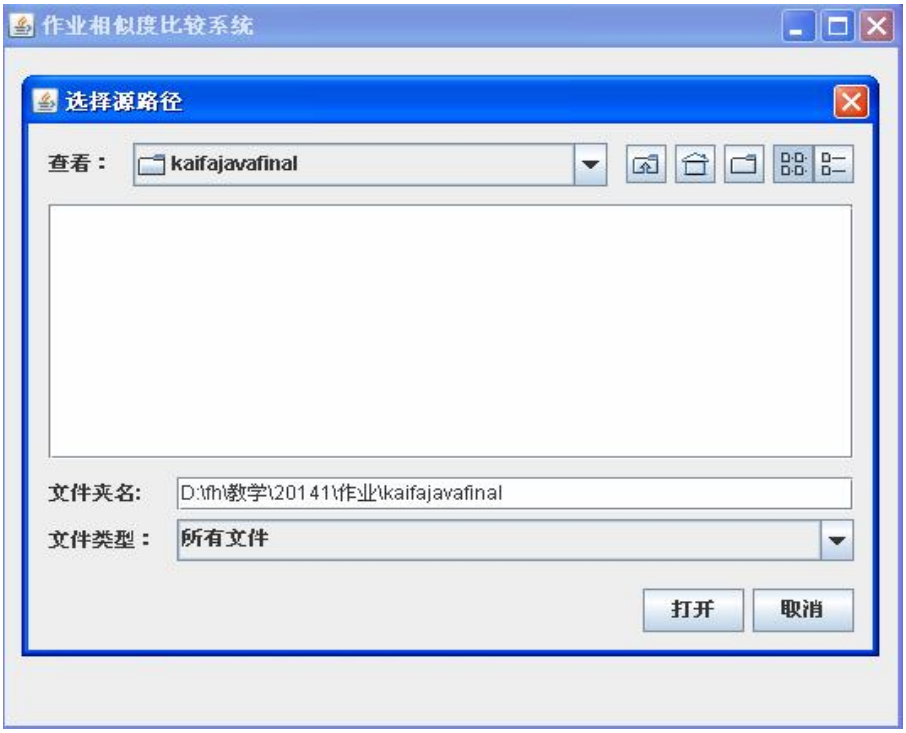


图 13 选择源路径对话框

（3）回到转换文件对话框后，点选“选择目标目录”按钮，出现如图 14 所示的选择目标路径对话框，在此指定被转换抽取的文件所放置的文件目录（ 如果该目录不存在，需要预先创建一个，目标目录应为空目录）。路径确定后，点“打开”

按钮。

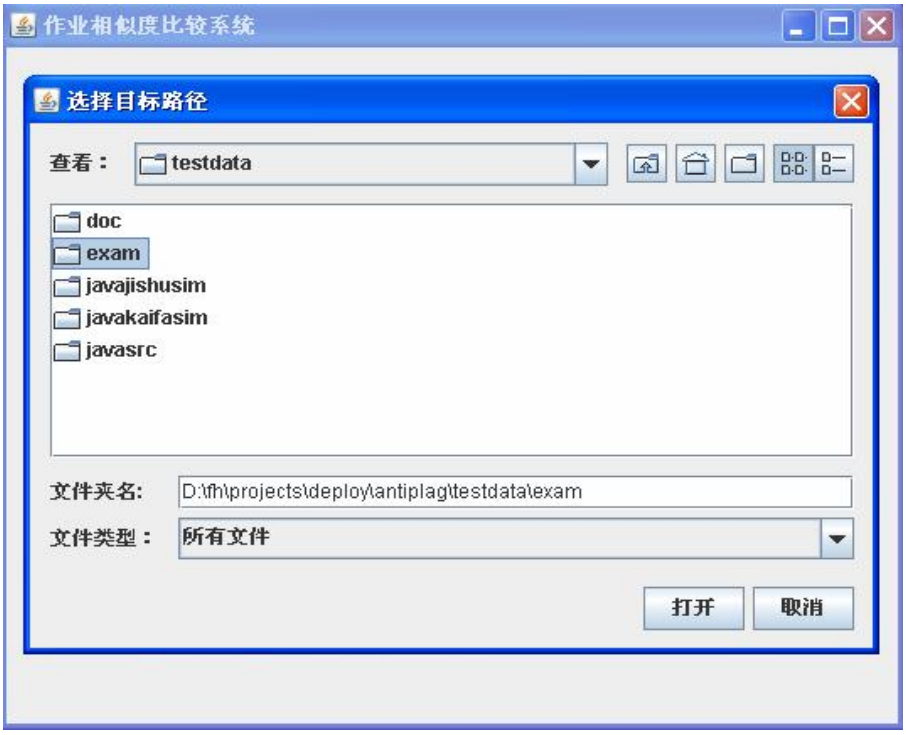


图 14 选择目标路径对话框

（4）确定了源路径、目标路径后，还需要确定提取转换的作业文件类型（java、c、doc 等），结果如图 15 所示。

（5）点击“转换”按钮，完成转换。


图 15 源路径、目标路径确定后的结果

# 5 系统安装配置

作业相似度比较软件的发布形式是一个压缩包，基本上解压即用，但由于系统使用 java 语言开发，程序运行离不开 java 运行时环境，所以需要预先安装 jdk。

## 5．1　JDK 安装配置

1、 下载安装程序

JDK需要1.6以上版本，下载地址如下：

http://www.oracle.com/technetwork/java/javase/downloads/index.html

这里使用的是 jdk-6u22-windows-i586.exe Window 安装版本。

2、 进行安装

双击安装程序进行安装，初始安装界面如图 16 所示，以后按照提示，一路点“下一步”按钮即可。其中安装路径使用默认路径。


图 16　JDK 安装界面

3、配置path和classpath环境变量

对java程序进行编译和运行，主要用到两个java处理程序：javac.exe(用于编译java源程序)、java.exe(用于运行编译好的.class文件)，其所处位置如下，也就是jdk安装目录下的\bin子目录下面，如图17所示：


图 17　JDK 安装 bin 路径下文件

安装完毕jdk后，java.exe和javac.exe还不是系统内部处理命令，打开命令界面（Win+R在弹出对话框中输入cmd然后回车就进入命令界面，如图18、19所示）


图18　运行命令对话框

在命令界面下输入javac命令，发现不是内部命令，即无法进行java源程序编译。


图19 命令窗口

下面要进行path和classpath配置，使得javac命令能在命令界面下运行。依次点击：“我的电脑”－(右键)“属性”（图20）－“高级”－“环境变量”，弹出Windows的“环境变量”配置窗口（图21所示）。


图20 系统属性窗口

在“系统变量”栏下执行三项操作：

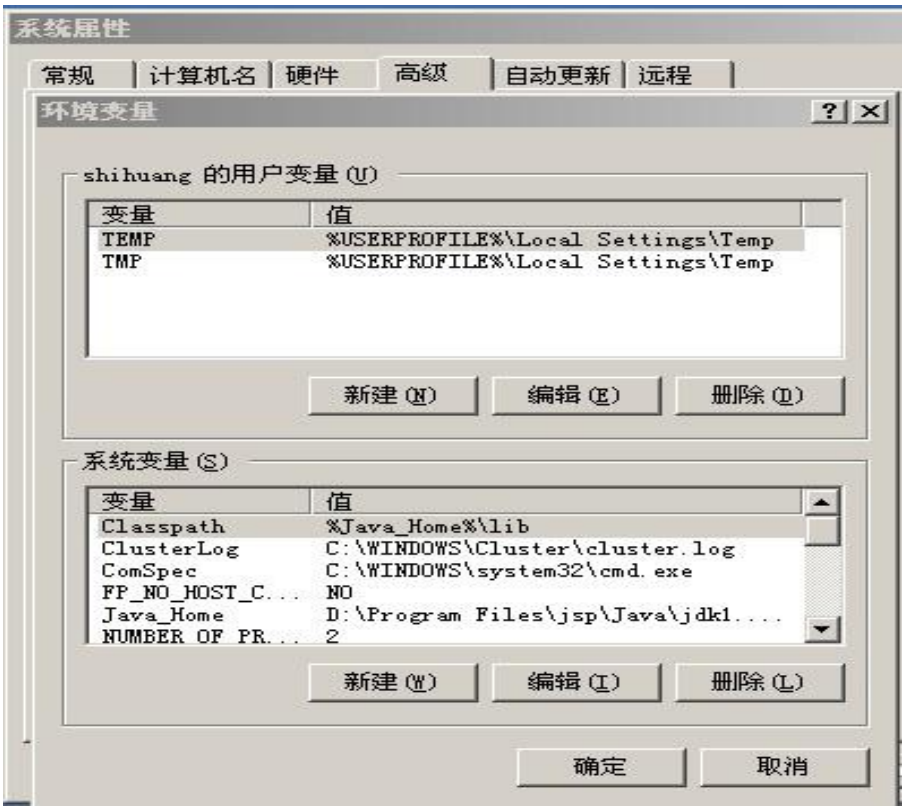


图21 环境变量设置窗口

一是新建“Java_Home”，设置其值为 JDK所在的绝对路径(例如 D:\Program Files\Java\jdk1.6.0_29)。　二是新建“Classpath”(如果已有，则直接编辑)，设置其值为 .;%Java_Home%\lib (若值中原来有内容，用分号与之隔开)。　注意路径前的符号为.;不能漏掉。三是新建“Path”(如果已有，则直接编辑),值：%Java_Home%\bin; (若值中原来有内容，用分号与之隔开)。重新打开一个字符命令界面发现javac命令可用了（图22所示）。

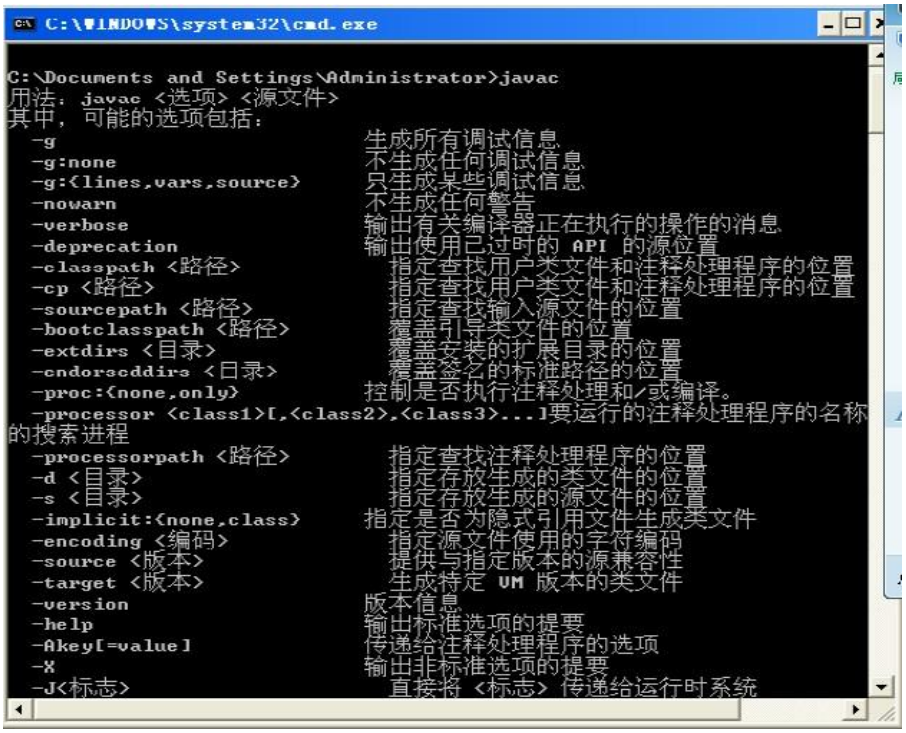


图22 命令窗口执行javac

# 6 常见问题

（1）为何无法执行程序？

系统解压后，可以通过双击解压出的antiplag.jar文件执行程序，如果不行，可以双击run.bat批处理文件来执行程序。如果仍不能执行程序，那么大多数情况下是JDK安装、配置不正确，请参考“系统安装配置”的JDK安装配置部分，正确安装、配置JDK。

（2） 为何使用moss比较时，结果为空？

如果提示“执行失败”，说明执行比较的过程中出现了问题（常见的是网络问题），结果可能为空，可以重新尝试；如果提示“执行完毕，未发现符合限值要求的结果”，说明执行成功，但没有超过相似度限值的结果，此时结果也为空，为了得到结果，可以适当降低相似度限值。也有可能是没有按要求将作业拷贝到testdata目录下。另外使用moss需要存取网络，应保证网络畅通。尽管moss所在的服务器是24小时持续运行的，但偶尔也会停机维护，在此情况下，可以使用sim作为替代。

（3） 通过修改代码的注释、变量名、格式能影响比较结果吗？

简单说：不能。在参考别人代码的基础上，通过一些不影响程序结构的修改，来规避检查，是抄袭者常用的手段，检测算法考虑到了此类问题，针对性的做了处理。

（4） 程序能对文档中的图片进行比较吗？

不能。系统主要对文档中的文本内容进行比较。

（5） 程序能发现从网络抄袭的程序和文档吗？

目前不能。系统只是对提交的作业之间的相似度进行比较，如果作业内容来自网络，又各不相同，系统目前无法检测出此类情况，但考虑在后续版本中添加网络比对功能。

（6） 为何在执行比较的过程中，有时会出现长时间挂起的现象？

经过分析，此类现象可能出现在sim方式下，比较的文件数较大的时候（如500份以上），程序实际仍在执行，但因算法问题，随着比较的文档数增加，耗费的时间将会明显增加。也可能出现在moss方式下，由于网络速度慢或不稳定，造成长时间上传文件（特别是上传文件数较多、较大时）。如果长时间没有响应（如大于10分钟），可以直接点击窗口关闭按钮，关闭程序。

（7） 如何具体查看两份作业之间的相似情况？

系统比较结束后会输出一个按相似度大小排序的作业之间互相比对的结果，可以根据此结果，重点检查相似度高的作业之间的相似度，进而进一步确认是否存在抄袭现象，这需要人工处理。对于moss方式比较的结果，提供了辅助比对的方法，在moss比较结果的最后一行，有一个链接，将此链接输入浏览器，可以得到如图30所示的页面。该页面返回了每个文件之间的相似度及匹配的代码行数，

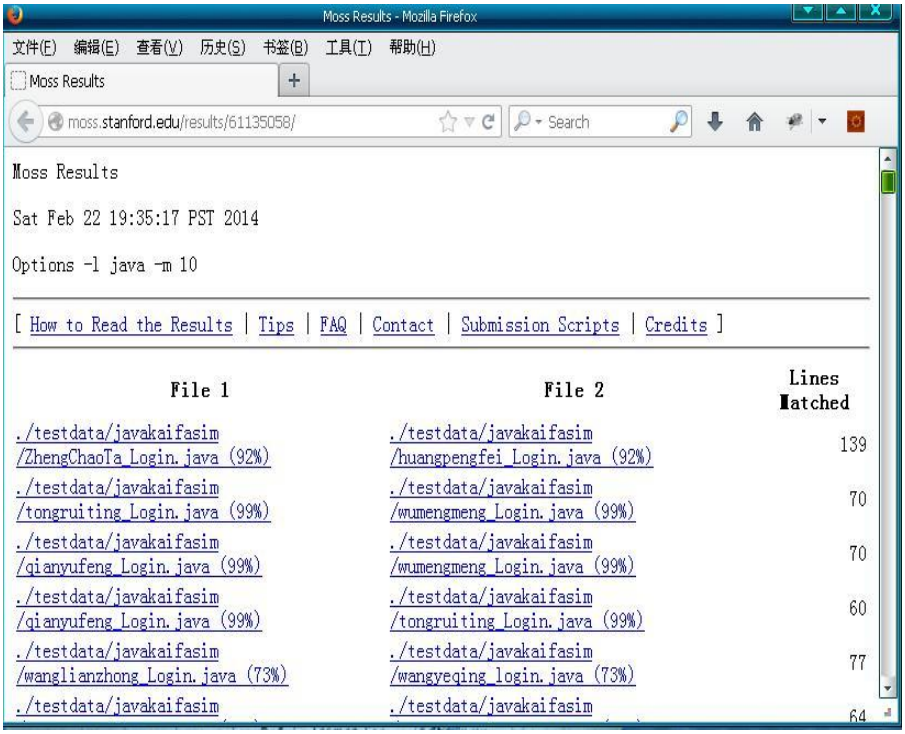


图30 moss比较结果网页

但并没有按序输出。要想看到具体两个文件之间的相似情况，只要点击相应文件名标识的链接，就可以得到如图31所示的结果。在此页面可以看到具体的相似代码的情况（相似代码会被标识出来），有助于分析是否发生了抄袭。

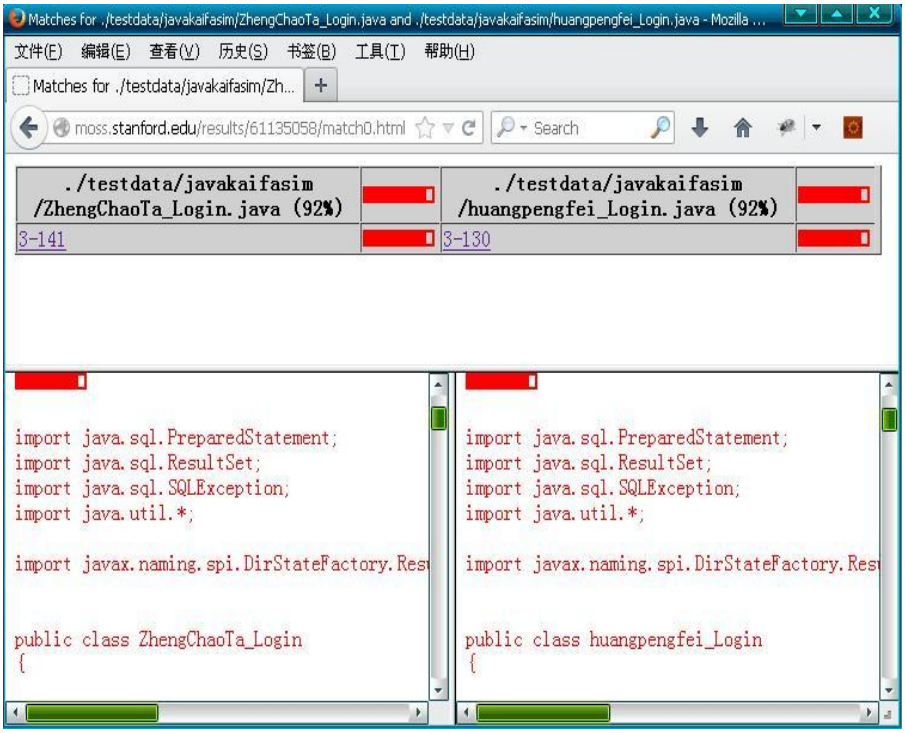


图31 具体文件相似情况比对页面

对于文档作业，查看具体的相似情况，目前只能依靠人工比较查看。但考虑在后续版本中增加与moss类似的文本比对功能。