Java Application: FindDupFiles

Written by: Keith Fenske, http://www.psc-consulting.ca/fenske/
First version: Wednesday, 29 October 2008
Document revised: Saturday, 13 February 2010

Description

When collecting a large number of files of any kind, there will be duplicates with the same file under two different names or in more than one place. FindDupFiles is a Java 1.4 application to find duplicate files by searching for files that have the same size and the same MD5 checksum. It won't find files that are merely similar, such as two consecutive photos of the same subject, or two MP3 songs encoded at different times. Possible duplicates are reported to the user, who can then verify that the files are identical, either by inspection or by doing a byte-by-byte comparison with the "comp" command in DOS/Windows or the "cmp" command in UNIX. What to do with files is the user's choice; the program does nothing except report the duplicates. The probability of two files with different contents having the same size and MD5 checksum is extremely small.

To avoid wasting CPU time, MD5 checksums are only calculated if two or more files have the same size. This program took two minutes on an Intel Pentium 4 processor at 2.4 GHz to scan a collection of 16,362 font files of various sizes up to 39.6 MB and using a total of 5.2 GB. Almost half of the files (7,393) had the same size as another file, which forced the MD5 to be computed. Peak memory usage was under 23 MB when run as a graphical application, and 12 MB when run as a console application.

See the DeleteDupFiles application to delete duplicate files when there is a "known good" folder and a folder of unknown files. See the CompareFolders application for comparing two folders to determine if files and subfolders are identical. See the FileChecksum application to generate or test checksums for a single file.

GNU General Public License (GPL)

Installation

You must have the Java run-time environment (JRE) installed on your computer. FindDupFiles was developed with Java 1.4 and should run on later versions. It may also run on earlier versions, but this has not been tested. For Macintosh computers, the version of Java is determined by your version of MacOS. For Windows, Linux, and Solaris, you can download the JRE from Sun Microsystems:

Sun Java
- JRE for end users: http://www.java.com/getjava/
- SDK for programmers: http://developers.sun.com/downloads/
- IDE for programmers: http://www.netbeans.org/

Once Java is installed, you need to put the program files for FindDupFiles into a folder (directory) on your hard drive. The name of the folder and the location are your choice, except it is easier if the name does not include spaces. Assume that files will go into a C:\JAVA folder. Then create the folder and unpack the Java *.class files into this folder (if you received the program as a ZIP file). The files look something like this:

```
FindDupFiles3.class  (21 KB, executable program)
FindDupFiles3.doc  (33 KB, this documentation in Microsoft Word format)
FindDupFiles3.gif  (14 KB, sample program image)
FindDupFiles3.ico  (4 KB, icon for Windows)
FindDupFiles3.jar  (12 KB, archive file with same class files inside)
FindDupFiles3.java  (63 KB, source code)
FindDupFiles3.manifest  (1 KB, main class manifest for archive file)
FindDupFiles3.pdf  (73 KB, this documentation in Adobe Acrobat format)
FindDupFiles3User.class  (1 KB, helper class for main program)
GnuPublicLicense3.txt  (35 KB, legal notice)
RunJavaPrograms.pdf  (88 KB, more notes about running Java)
```

To run the program on Windows, start a DOS command prompt, which is Start button, Programs, Accessories, Command Prompt on Windows 2000/XP. Change to the folder with the program files and run the program with a "java" command:

```
c:
cd  \java
java  FindDupFiles3
```

The program name "FindDupFiles3" must appear exactly as shown; uppercase and lowercase letters are different in Java names. Some systems (Macintosh) will run a main "class" file by clicking on the class file name while viewing a directory in the file browser (Mac Finder). Many systems will run a "jar" file by clicking (or double clicking) on the jar file name (Windows

Explorer). The command line is the only guaranteed way of running a Java program. Should you find this program to be popular, you can create a Start menu item or desktop shortcut on Windows 2000/XP with a target of "java.exe FindDupFiles3" starting in the "c:\java" folder.

One complication may arise when trying to run this program. Java looks for an environment variable called CLASSPATH. If it finds this variable, then that is a list of folders where it looks for *.class files. It won't look anywhere else, not even in the current directory, unless the path contains "." as one of the choices. The symptom is an error message that says:

```
Exception in thread "main" java.lang.NoClassDefFoundError: FindDupFiles3
```

To find out if your system has a CLASSPATH variable defined, type the following command in a DOS window:

```
set  CLASSPATH
```

To temporarily change the CLASSPATH variable to the current directory, use the following command line:

```
java  -cp  .  FindDupFiles3
```

To permanently change the CLASSPATH, you must find where it is being set. This may be in an old AUTOEXEC.* file in the root directory of your system disk (usually the C:\ folder), or it may be in Control Panel, System, Advanced, Environment Variables on Windows 2000/XP.

Removal or Uninstall

To remove this program from your computer, delete the installation files listed above. If the folder that contained the files is now empty, you may also delete the folder ... if you created the folder, of course, not the system. If you created desktop shortcuts or Start menu items, then delete those too. There are no hidden configuration or preference files, and no information is stored in the Windows system registry. You don't need an "uninstall" program.

Graphical Versus Console Application

The Java command line may contain options or file and folder names. If no file or folder names are given on the command line, then this program runs as a graphical or "GUI" application with the usual dialog boxes and windows. See the "-?" option for a help summary:

```
java  FindDupFiles3  -?
```

The command line has more options than are visible in the graphical interface. An option such as -u14 or -u16 is recommended because the default Java font is too small. If file or folder names are given on the command line, then this program runs as a console application without a graphical interface. A generated report is written on standard output, and may be redirected with the ">" or "1>" operators. (Standard error may be redirected with the "2>" operator.) An example command line is:

```
java FindDupFiles3 -s d:\fonts >report.txt
```

The console application will return an exit status equal to the number of duplicate files found. The graphical interface can be very slow when the output text area gets too big, which will happen if thousands of files are reported.

---

file: FindDupFiles3.doc 2010-02-13