NAME

extract − SWISH++ text extractor

SYNOPSIS

extract [*options*] *directory... file...*

DESCRIPTION

extract is the SWISH++ text extractor, a utility to extract what text there is from a (mostly) binary file (similar to the **strings**(1) command) prior to indexing. Original files are untouched.

Text is extracted from the specified files and files in the specified directories; text from files in subdirectories of specified directories is also extracted by default (unless the **−r**, **−−no-recurse**, **−f**, or **−−filter** option or the **RecurseSubdirs** or **ExtractFilter** variable is given).

Ordinarily, text is extracted from files either only if their filename matches one of the patterns in the set specified with either the **−e** or **−−pattern** option or the **IncludeFile** variable (unless standard input is used; see next paragraph) or is not among the set specified with either the **−E** or **−−no-pattern** option or the **ExcludeFile** variable.

If there is a single filename of '`-`', the list of directories and files to extract is instead taken from standard input (one per line). In this case, filename patterns of files to extract need not be specified explicitly: all files, regardless of whether they match a pattern (unless they are among the set not to extract specified with either the **−E** or **−−no-pattern** option or the **ExcludeFile** variable), are extracted, i.e., **extract** assumes you know what you're doing when specifying filenames in this manner.

Ordinarily, the text extracted from a file is written to another file in the same directory having the same filename but with the "`.txt`" extension appended by default, e.g., "`foo.doc`" becomes "`foo.doc.txt`" after extraction. (See also the **−x** or **−−extension** option or the **ExtractExtension** variable.) However, extraction is not performed if the extracted text file exists.

If either the **−f** or **−−filter** option or the **ExtractFilter** variable is given, then only a single file specified on the command line is extracted to standard output. In this case, filename patterns are not used and the existence of an extracted text file is irrelevant.

Filters

Via the **FilterFile** configuration file variable, files having particular patterns can be filtered prior to extraction. (See the examples in **swish++.conf**(4).)

Character Mapping and Word Determination

extract performs the same character mapping, character entity conversions, and word determination heuristics used by **index**(1) but also additionally:

1. Considers all PostScript Level 2 operators that are not also English words to be stop words. Such words in a file usually indicate an encapsulated PostScript (EPS) file and such should not be indexed.
2. Looks specifically for encapsulated PostScript (EPS) data between everything between one of `%%BeginSetup`, `%%BoundingBox`, `%%Creator`, `%%EndComments`, or `%%Title` and `%%Trailer` and discards it.
3. Discards strings of ASCII hex data `Word_Hex_Min_Size` characters or longer, e.g., "`7F454C46`." (Default is 5.)

Motivation

extract was developed to be able to index non-text files in proprietary formats such as Microsoft Office documents. There are a couple of reasons why the functionality of **extract** isn't simply built into **index**(1):

1. Users who do not need to index such documents shouldn't have to pay the performance penalty for doing the extra checks for PostScript and hex data.
2. While **index**(1) can uncompress files on the fly using filters also, uncompressing them every time indexing is performed is excessive. Text extraction, on the other hand, is done only once per file; if the file is updated, the text-extracted version should be deleted and recreated.

OPTIONS

Options begin with either a '`-`' for short options or a "`--`" for long options. Either a '`-`' or "`--`" by itself explicitly ends the options; however, the difference is that '`-`' is returned as the first non-option whereas "`--`" is skipped entirely. Long option names may be abbreviated so long as the abbreviation is unambiguous.

For a short option that takes an argument, the argument is either taken to be the remaining characters of the same option, if any, or, if not, is taken from the next option unless said option begins with a '`-`'.

Short options that take no arguments can be grouped (but the last option in the group can take an argument), e.g., `-lrv4` is equivalent to `-l -r -v4`.

For a long option that takes an argument, the argument is either taken to be the characters after a '`=`', if any, or, if not, is taken from the next option unless said option begins with a '`-`'.

–?
––help Print the usage ("help") message and exit.

–c*c*
––config-file=*c* The name of the configuration file, *c*, to use. (Default is `swish++.conf` in the current directory.) A configuration file is not required: if none is specified and the default does not exist, none is used; however, if one is specified and it does not exist, then this is an error.

–e*p***[,***p***...]**
––pattern=*p***[,***p***...]** A filename pattern (or set of patterns separated by commas), *p*, of files to extract text from. Case is significant. Multiple **–e** or **––pattern** options may be specified.

–E*p***[,***p***...]**
––no-pattern=*p***[,***p***...]**
A filename pattern or patterns, *p*, of files *not* to extract text from. Case is significant. Multiple **–E** or **––no-pattern** options may be specified.

–f
––filter Extract a single file to standard output and exit.

–l
––follow-links Follow symbolic links during extraction. The default is not to follow them. (This option is not available under Microsoft Windows since it doesn't support symbolic links.)

–r
––no-recurse Do not recursively extract the files in subdirectories, that is: when a directory is encountered, all the files in that directory are extracted (modulo the filename patterns specified via the **–e**, **––pattern**, **–E**, or **––no-pattern** options or the **IncludeFile** or **ExcludeFile** variables) but subdirectories encountered are ignored and therefore the files contained in them are not extracted. (This option is most useful when specifying the directories and files to extract via standard input.) The default is to extract the files in subdirectories recursively.

–s*f*
––stop-file=*f* The name of a file, *f*, containing the set stop-words to use instead of the built-in set. Whitespace, including blank lines, and characters starting with `#` and continuing to the end of the line (comments) are ignored.

–S
––dump-stop Dump the built-in set of stop-words to standard output and exit.

–v*c*
––verbosity=*v* The verbosity level, *v*, for printing additional information to standard output during indexing. The verbosity levels, 0-4, are:

0 No output is generated (except for errors).
1 Only run statistics (elapsed time, number of files, word count) are printed.
2 Directories are printed as extraction progresses.
3 Directories and files are printed with a word-count for each file.
4 Same as 3 but also prints all files that are not extracted and why.

−V
−−version Print the version number of **SWISH++** and exit.

−x*e*
−−extension=*e* The extension to append to filenames during extraction. (It can be specified with or without the dot; default is `txt`.)

CONFIGURATION FILE

The following variables can be set in a configuration file. Variables and command-line options can be mixed.

ExcludeFile	Same as **−E** or **−−no-pattern**
ExtractExtension	Same as **−x** or **−−extension**
ExtractFilter	Same as **−f** or **−−filter**
FilterAttachment	(See FILTERS in **swish++.conf**(4).)
FilterFile	(See FILTERS in **swish++.conf**(4).)
FollowLinks	Same as **−l** or **−−follow-links**
IncludeFile	Same as **−e** or **−−pattern**
RecurseSubdirs	Same as **−r** or **−−no-recurse**
StopWordFile	Same as **−s** or **−−stop-file**
Verbosity	Same as **−v** or **−−verbosity**

EXAMPLES

Extraction

To extract text from all Microsoft Office files on a web server:

```
cd /home/www/htdocs
extract −v3 −e ’*.doc’ −e ’*.ppt’ −e ’*.xls’ .
```

Filters

(See the examples in **swish++.conf**(4).)

EXIT STATUS

Exits with one of the values given below:

0	Success.
1	Error in configuration file.
2	Error in command-line options.
20	File to extract does not exist.
30	Unable to read stop-word file.

CAVEATS

1. Text extraction is not perfect, nor can be.

2. As with **index**(1), the word-determination heuristics employed are heavily geared for English. Using SWISH++ as-is to extract files in non-English languages is not recommended.

FILES

`swish++.conf` default configuration file name

SEE ALSO

index(1), **search**(1), **strings**(1), **swish++.conf**(4), **glob**(7)

Adobe Systems Incorporated. *PostScript Language Reference Manual, 2nd ed.* Addison-Wesley, Reading, MA. pp. 346-359.

International Standards Organization. ‘‘ISO/IEC 9945-2: Information Technology -- Portable Operating System Interface (POSIX) -- Part 2: Shell and Utilities,’’ 1993.

AUTHOR

Paul J. Lucas <*pauljlucas@mac.com*>