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** option or the `RecurseSubdirs` variable is given). Text is extracted from files either only if their filename extension is among the set specified with either the −**e** option or the `IncludeExtension` variable (unless standard input is used; see next paragraph) or is not among the set specified with either the −**E** option or the `ExcludeExtension` 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 extensions of files to extract need not be explicitly specified with the −**e** option; all files, regardless of extension (unless it is among the set specified with the −**E** option), are extracted, i.e., **extract** assumes you know what you're doing when specifying filenames.

In any case, care must be taken not to specify files or subdirectories in directories that are also specified: since directories are recursively extracted by default (unless the −**r** option or the `RecurseSubdirs` variable is given), explicitly specifying a subdirectory or file in a directory that is also specified will result in those files being extracted more than once. (Unlike with indexing, this is harmless: it just wastes time.)

Extracted files have the same filename with the ''`.txt`'' extension appended, e.g., ''`foo.doc`'' becomes ''`foo.doc.txt`'' after extraction. However, extraction is not performed if the extracted text file exists.

Filters

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

Word Determination

extract performs the same character entity conversions and word determination heuristics as **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 penatly 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

−**c***config_file* The name of the configuration file to use. The 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***extension* A filename extension of files to extract text from *without* the ''dot.'' Case is significant. Multiple −**e** options may be specified.

−**E***extension* A filename extension of files *not* to extract text from *without* the ''dot.'' Case is significant. Multiple −**E** options may be specified.

−**l** 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** 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 extensions specified via the −**e** or −**E** options), 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***stop_word_file* The name of a file 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 the built-in set of stop-words to standard output and exit.

−**v***verbosity* The verbosity level, 0-4:

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

CONFIGURATION FILE

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

ExcludeExtension Same as the −**E** option.
FilterExtension (See Filters.)
FollowLinks Same as the −**l** option.
IncludeExtension Same as the −**e** option.
RecurseSubdirs Same as the −**r** option.
StopWordFile Same as the −**s** option.
Verbosity Same as the −**v** option.

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.
30	Unable to read stop-word file.

CAVEATS

Text extraction is not perfect, nor can be.

FILES

`swish++.conf` default configuration file name

SEE ALSO

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

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

AUTHOR

Paul J. Lucas <*pjl@best.com*>