

**NAME**

bogoutil - Dumps, loads, and maintains bogofilter database files

**SYNOPSIS**

**bogoutil** {-h | -V}

**bogoutil** [options] {-d *file* | -H *file* | -l *file* | -m *file* | -w *file* | -p *file*}

**bogoutil** {-r *file* | -R *file*}

**bogoutil** {--db-print-leafpage-count *file* | --db-print-pagesize *file* | --db-verify *file* |  
--db-checkpoint *directory* [flag...] | --db-list-logfiles *directory* | --db-prune *directory* |  
--db-recover *directory* | --db-recover-harder *directory* |  
--db-remove-environment *directory*}

where **options** is

**bogoutil** [-v] [-n] [-C] [-D] [-a *age*] [-c *count*] [-s *min,max*] [-y *date*] [-I *file*] [-O *file*] [-x *flags*]  
[--config-file *file*]

**DESCRIPTION**

Bogoutil is part of the bogofilter Bayesian spam filter package.

It is used to dump and load bogofilter's Berkeley DB databases to and from text files, perform database maintenance functions, and to display the values for specific words.

**OPTIONS**

The **-d** *file* option tells bogoutil to print the contents of the database file to **stdout**.

The **-H** *file* option tells bogoutil to print a histogram of the database file to **stdout**. The output is similar to bogofilter -vv. Finally, hapaxes (tokens which were only seen once) and pure tokens (tokens which were encountered only in ham or only in spam) are counted.

The **-l** *file* option tells bogoutil to load the data from **stdin** into the database file. If the database file exists, **stdin** data is merged into the database file, with counts added up.

The **-m** option tells bogoutil to perform maintenance functions on the specified database, i.e. discard tokens that are older than desired, have counts that are too small, or sizes (lengths) that are too long or too short.

The **-w** *file* option tells bogoutil to display token information from the database file. The option takes an argument, which is either the name of the wordlist (usually wordlist.db) or the name of the directory containing it. Tokens can be listed on the command line or piped to bogoutil. When there are extra arguments on the command line, bogoutil will use them as the tokens to lookup. If there are no extra arguments, bogoutil will read tokens from **stdin**.

The **-p** *file* option tells bogoutil to display the database information for one or more tokens. The display includes a probability column with the token's spam score (computed using bogofilter's default values). Option **-p** takes the same arguments as option **-w**.

The **-r** *file* option tells bogoutil to recalculate the ROBX value and print it as a six-digit fraction.

The **-R** *file* option does the same as **-r**, but saves the result in the training database without printing it.

The **-I** *file* option tells bogoutil to read its input from *file* rather than stdin.

The **-O** *file* option tells bogoutil to write its output to *file* rather than stdout.

The **-v** option produces verbose output on **stderr**. This option is primarily useful for debugging.

The **-C** inhibits reading configuration files and lets bogoutil go with the defaults.

The **--config-file** *file* option tells bogoutil to read *file* instead of the standard configuration file.

The **-D** redirects debug output to stdout (it usually goes to stderr).

The **-x** *flags* option sets debugging flags.

Option **-n** stands for "replace non-ascii characters". It will replace characters with the high bit (0x80) by question marks. This can be useful if a word list has lots of unreadable tokens, for example from Asian spam. The "bad" characters will be converted to question marks and matching tokens will be combined when used with **-m** or **-l**, but not with **-d**.

Option **-a** **age** indicates an acceptable token age, with older ones being discarded. The age can be a date (in form YYYYMMDD) or a day count, i.e. discard tokens older than **age** days.

Option **-c** **value** indicates that tokens with counts less than or equal to **value** are to be discarded.

Option **-s** **min,max** is used to discard tokens based on their size, i.e. length. All tokens shorter than **min** or longer than **max** will be discarded.

Option **-y date** is specifies the date to give to tokens that don't have dates. The format is YYYYMMDD.

The **-h** option prints the help message and exits.

The **-V** option prints the version number and exits.

## ENVIRONMENT MAINTENANCE

The **--db-checkpoint** *dir* option causes bogoutil to flush the buffer caches and checkpoint the database environment.

The **--db-list-logfiles** *dir* option causes bogoutil to list the log files in the environment. Zero or more keywords can be added or combined (separated by whitespace) to modify the behavior of this mode. The default behavior is to list only inactive log files with relative paths. You can add **all** to list all log files (inactive and active). You can add **absolute** to switch the listing to absolute paths.

The **--db-prune** *dir* option causes bogoutil to checkpoint the database environment and remove inactive log files.

The **--db-recover** *dir* option runs a regular database recovery in the specified database directory. If that fails, it will retry with a (usually slower) catastrophic database recovery. If that fails, too, your database cannot be repaired and must be rebuilt from scratch. This is only supported when compiled with Berkeley DB support with transactions enabled. Trying recovery with QDBM or SQLite3 support will result in an error.

The **--db-recover-harder** *dir* option runs a catastrophic data base recovery in the specified database directory. If that fails, your database cannot be repaired and must be rebuilt from scratch. This is only supported when compiled with Berkeley DB support with transactions enabled. Trying recovery with QDBM or SQLite3 support will result in an error.

The **--db-remove-environment** *directory* option has no short option equivalent. It runs recovery in the given directory and then removes the database environment. Use this *before* upgrading to a new Berkeley DB version if the new version to be installed requires a log file format update.

The **--db-print-leafpage-count** *file* option prints the number of leaf pages in the database file *file* as a decimal number, or UNKNOWN if the database does not support querying this figure.

The **--db-print-pagesize** *file* option prints the size of a database page in *file* as a decimal number, or UNKNOWN for databases with variable page size or databases that do not allow a query of the database page size.

The **--db-verify** *file* option requests that bogofilter verifies the database file. It prints only errors, unless in verbose mode.

## DATA FORMAT

Bogoutil reads and writes text files where each nonblank line consists of a word, any amount of horizontal whitespace, a numeric word count, more whitespace, and (optionally) a date in form YYYYMMDD. Blank lines are skipped.

## RETURN VALUES

0 for successful operation. 1 for most errors. 3 for I/O or other errors. Error 3 usually means that something is seriously wrong with the database files.

## AUTHOR

Gyepi Sam <gyepi@praxis-sw.com>.

Matthias Andree <matthias.andree@gmx.de>.

David Relson <relson@osagesoftware.com>.

For updates, see **the bogofilter project page**[1].

## SEE ALSO

bogofilter(1), bogolexer(1), bogotune(1), bogoupgrade(1)

## NOTES

1. the bogofilter project page  
<http://bogofilter.sourceforge.net/>