NAME

indexmeister − reads a *LaTeX* or other file and suggests terms for indexing.

SYNOPSIS

indexmeister *filename.tex* [-dfxE]

DESCRIPTION

Indexmeister is designed to speed the process of indexing large documents such as dissertations and nonfiction books. It a file and outputs a list of suggested terms to be included in the index. This list is suitable for semi-automatic indexing using *indexmeister's* sister application *imbrowse.* the first pass in the indexing process.

If either *detex* or *pandoc* is installed then *indexmeister* will be able to read *LaTeX* and plain text files. If *pandoc* is installed then it will also be able to read several additional formats including docx, html, and epub (see the *pandoc* man page for a complete list).

Indexmeister is written in *Python* 2.7 and should run on any Unix-like environment, including Cygwin (although so far it has actually only been tested on Android (w/Termux) and Debian, Ubuntu, and SUSE Linux).

In the default mode *indexmeister* primarily looks for capitalized phrases, which captures all names of people, places, and other proper nouns. It will also include terms that are inside *LaTeX* \emph{ } tags, unless the -E flag is given.

If the -d flag is given it will also suggest words that do not appear in the system dictionary (this requires a properly installed *aspell* or *hunspell* to work). This option is occasionally useful to capture foreign language or scientific terms. If the -f flag is given it will attempt to detect important words using word frequency analysis. This mode is still a 'work in progress' but when it works it can be useful.

Detex often leaves in 'junk' text from the original Tex code in its output, which can make its way into Indexmeister's results. If the file .indexmeister-exclude, containing a list of strings to filer out, is present in the user's home folder *indexmeister* will read it and use it to filter its input stream. This file should contain one string per line. Lines beginning with '#' are comments and will be ignored.

OPTIONS

`-d` Also suggest words that do not appear in the system dictionary (e.g. scientific terms or foreign words)

`-f` Also suggest words based on word frequency analysis (experimental)

`-x` Force the use of *detex* as a back-end even if *pandoc* is installed

`-E` Suppress inclusion of words in \emph{ } tags

DEPENDENCIES

Indexmeister requires that either *pandoc* or *detex* be installed on the path. Detex is available in most LaTeX distributions. If you are writing books in LaTeX you most likely already have it installed. Either *aspell* or *hunspell* is required for the -d mode. Most Linux and Unix distributions ship with one or both.

INDEXING THEORY (THE QUICK VERSION)

For our purposes indexing refers to "the process of creating a list of terms that captures topics of interest within a print item and making a list of those terms with page numbers to aid the reader". In general, these terms fall into two categories: nominal and conceptual.

Nominal Terms are the names of people, places, events, etc. Heuristic indexers like *indexmeister* do a good job of extracting nominal terms. In some genres the majority of the interesting (to the reader) terms are nominal. For example, in a history of Germany the reader might want to know which pages mention "Luther, Martin" or "Wiemar Republic". This is also the case for most textbooks, where the concepts tend to be associated with readily identifiable names. For example in an undergraduate statistics class indexing the occurrences of "Confidence Interval, of a proportion" will also capture the pages that tell you how to calculate that confidence interval.

Conceptual Terms are those which deal with ideas and constructs. They tend to be much harder to pick up with simple heuristic methods. For example the history of Germany above might include several paragraphs discussing general social trends during the industrial revolution. This is probably something well worth indexing. Unfortunately, a program like *indexmeister* unlikely to pick it up. And if it does find it (probably because the words "Social Trends During the Industrial Revolution" occur in a heading, it is likely to make mistakes in the page range. A truly effective conceptual indexer would probably need to rely on machine learning or artificial intelligence. It would also need to be trained on a sufficiently large corpus for each discipline. There is no telling what will happen with the *indexmeister* project in the future. For now, however, you can use *indexmeister* to catch most of the nominal terms, but plan on indexing the conceptual terms manually. Actually, you might find that adding index tags with the main concepts of key paragraphs helps you with your editing process.

A final point to remember is that doesn't understand (and ignores) figures and images. That means that if they get indexed it will be on the basis of their captions. So try to write good descriptive captions for all of your graphics...which is what every style guide tells you to do anyway.

HISTORY

indexmeister and *imbrowse* were developed in-house at Creative Minority Productions and released publicly under the GPL starting with version 0.30.

SEE ALSO

imbrowse(1), pandoc(1), detex(1)

AUTHOR

Kevin A. Straight <longhung@yahoo.com> <www.kevinastraight.com>