trish2 User Guide

Robert Bossy

March 15, 2007

Contents

1 What is trish2?

trish2 is a tool for matching strings against a dictionary. It has the following features:

- PATRICIA representation of dictionary, amorticized linear match;
- PATRICIAs can be saved on disk;
- full string match or partial match (fgrep functionality);
- unicode internal representation, handles any encoding supported by the system;
- association of tags to the entries, dictionaries may work as large hashes;
- full control on the output.

2 Installation

2.1 Prerequisites

trish2 can be installed on any linux box provided that it includes:

- coreutils

- GNU make
- GNU CC
- libpopt

2.2 Getting trish2

Get the latest source archive at `http://XXX`.

2.3 Building and Installing

Type `make`, then `make install`.

It will install executables and example documents and decision tree in `/usr/local`. If you want to install it elsewhere, set the `PREFIX` environment variable before building.

3 trish2 usage

```
trish2 [OPTIONS] DICT
```

`DICT` is a dictionary in text format with one entry per line. *trish2* expects strings to be matched in standard output (one per line) andd the result is displayed on standard output.

3.1 General options

`-n, --no-search`
Do not perform search, just read the dictionary.

`-D, --save-as=FILE`
Save the PATRICIA generated from the dictionary to a binary file, we call this file a *compiled* dictionary. `FILE` is the path to the file to save into.

`-P, --optimal-parameters`
Prints command-line options for faster compilation. This option computes optimal parameters for future use of the same dictionary. Include this output to the command line the next time the same dictionary is searched.

3.2 Search options

`-i, --input=FILE`
Read the input from `FILE` instead of standard input.

`-o, --output=FILE`
Write the output into `FILE` instead of standard output

`-x, --search-max-length=LENGTH`
Maximum length of query strings. By default, 1024.

`-e, --search-encoding=ENCODING`

Input and output encoding scheme. To get a list of supported encodings, type `locale -a`.

`-w, --substrings`
Search for substrings. By default, *trish2* searches for the whole line in the dictionary, this option tells to search substrings.

`-W, --words`
Only search for substrings enclosed by non-alphanumeric characters. Ignored if `-w` is not set.

`-C, --capital-insensitive`
Be case insensitive for the first character.

`-c, --case-insensitive`
Be case insensitive for all characters.

`-G, --general-capital-insensitive`
Use General Case Insensitiveness (GCI) on the first character.

`-g, --general-case-insensitive`
Use GCI on all characters.

`-t, --gci-table=PATH`
Set the GCI table.

`-N, --format-no=FORMAT`
Set the output format for queries that do not match. `FORMAT` is a string where \n and \t are respectively interpreted as a newline and a tab character. By default, empty string (will not print anything for unmatched queries).

Additionally it recognizes the special string {`query`} will be replaced by the query.

`-Y, --format-yes=FORMAT`
Set the output format for matched queries. `FORMAT` is similar as the one for `-N` option but it accepts additional {`MNEMONIC`} strings described in the following table:

Mnemonic	**Effect**
`query`	the query string
`line`	line number in input stream
`length`	length of the matched string
`match`	(sub)string matched
`entry`	entry matched
`mismatch`	number of case insensitive and GCI replacements
`tag N`	Nth tag associated to matched entry

Note: the output format will be displayed several times if there are several entries that match (because of case insensitiveness or GCI). If the dictionary is a mult-dictionary, you must enclose the parts to be repeated for each tagset of the same entry between square brackets. Additionally all {`MNEMONIC`} sequences must be inside square brackets.

`-h, --show-headers`
Show tag headers following the `-Y` format.

3.3 Dictionary options

`-d, --compiled-dict`
The dictionary is a compiled one, previously generated by *trish2* with `-D` option.

`-X, --dict-max-length=LENGTH`
Maximal length of a dictionary entry (including tags). 1024, by default.

`-a, --tag-delimiters=DELIMITERS`
Set the delimiters between entry and tags. The first character of `DELIMITERS` splits entry from tag 0, the second character splits tag 0 from tag 1, etc. If the entry or a tag must contain a delimiter, it must be escaped with a \.

`-m, --multi-dict`
Reads a multi-dictionary, where a single entry may have several tagsets. Without this option *trish2* issues a warning for each duplicate key and only stores only the first tagset. With `-m`, all tagsets are stored.

When using a multi-dictionary, the `-Y` format accepts square brackets, and {`MNEMONIC`} *requires* square brackets.

`-E, --dict-encoding=ENCODING`
Dictionary encoding scheme.

`-H, --headers`
The first line in the dictionary sets the headers of the tagset. The header for the entry and the header for the tag 0 are split by the first character of the separator string given by `-a`.

3.4 Verbosity options

`-q, --quiet`
Suppress warnings and information provided on standard output.

`-v, --verbose`
Give more information.

`-V, --locace`
Give even more information.

`-s, --strict`
Stop at the first warning.

3.5 General Case Insensitiveness

General Case Insensitiveness (GCI) allows to match characters that are not identical, in the same way as upper/lower case. For instance, a GCI table

can tell *trish2* that underscores in a query can match spaces or dashes in the dictionary. GCI can also be used to search regardless of some diacritics.

The format of a GCI table is one line per character to replace. The first character is the character to match in the query, the following characters on the same line will match it. For instance, the following GCI table will match underscores with spaces and dashes:

```
_ -
```

This other example will match regardless the diacritics for French language:

```
aàäâ
eéèêë
iïî
oôö
uùüû
cç
```