LUG Text Processing Talk

Simon Kadesh

February 2025

Motivation

- Powerful
- Portable
- You also want to be a wizard

What's in our toolbox?

- These are everywhere
  - coreutils
  - sed
- These are usually present
  - grep
  - find
  - diff

coreutils

Input

- cat

Input

- cat — Concatenates files and stdin
- head

Input

- cat — Concatenates files and stdin
- head — Outputs the first part of a file
- tail

Input

- cat — Concatenates files and stdin
- head — Outputs the first part of a file
- tail — Outputs the last part of a file
- seq

Input

- cat — Concatenates files and stdin
- head — Outputs the first part of a file
- tail — Outputs the last part of a file
- seq — Generates a sequence of integers

Selecting

- cut

Selecting

- cut — Selects portions of the lines in a file
- tr

Selecting

- cut — Selects portions of the lines in a file
- tr — Map strings of bytes (and also sometimes delete them or get rid of repetition)

Ordering

- sort

Ordering

- sort — Sorts the lines in a file or stdin
- uniq

Ordering

- sort — Sorts the lines in a file or stdin
- uniq — Gets the unique (or duplicated) lines in a file or stdin
- comm

Ordering

- sort — Sorts the lines in a file or stdin
- uniq — Gets the unique (or duplicated) lines in a file or stdin
- comm — Compare the sorted contents of two files
- join

Ordering

- sort — Sorts the lines in a file or stdin
- uniq — Gets the unique (or duplicated) lines in a file or stdin
- comm — Compare the sorted contents of two files
- join — Combine two files by mapping fields together
- shuf

Ordering

- sort — Sorts the lines in a file or stdin
- uniq — Gets the unique (or duplicated) lines in a file or stdin
- comm — Compare the sorted contents of two files
- join — Combine two files by mapping fields together
- shuf — Randomly shuffles the input lines

Output

- tee

Output

- tee — Copy stdin to multiple outputs
- od

Output

- tee — Copy stdin to multiple outputs
- od — Output stdin (or a file) as octal* dump
- wc

Output

- tee — Copy stdin to multiple outputs
- od — Output stdin (or a file) as octal* dump
- wc — Count the number of lines, words, or characters in stdin

Formatting and Converting

- nl

Formatting and Converting

- nl — Generate line numbers for the input
- pr

Formatting and Converting

- nl — Generate line numbers for the input
- pr — Paginate/columnate input
- fold/fmt

Formatting and Converting

- nl — Generate line numbers for the input
- pr — Paginate/columnate input
- fold/fmt — Wrap lines
- expand/unexpand

Formatting and Converting

- nl — Generate line numbers for the input
- pr — Paginate/columnate input
- fold/fmt — Wrap lines
- expand/unexpand — Convert tabs into spaces or convert spaces into the correct form of indentation

More complex tools

sed

- The **s**tream **ed**itor
- Takes input from stdin or a file
- Outputs to stdout or a file
- Takes a "script" as an argument
  - Optional address
  - Command
  - Optional options
  - `s/<regex>/replace/`
  - `d, q, p, n, {}`
  - `y, a, i, c, =, r`

grep

- `g/re/p` — global regex and print
- Finds matching lines within files

find

- Recursively search the directory tree for a matching file
- Match types: empty, executable, name, path, regex
- Actions types: exec, delete, print, prune, quit

diff

- Print the differences between two files
- Output type: normal, brief, ed, rcs, two column