

Emacs Regular Expressions

e.g. Download <http://www.stat.wisc.edu/~jgillett/DSCP/emacs/data.txt>:

```
Brown,Joe      123456789  jbrown@wisc.edu      1000
Roukos,Sally  456789123  sroukos@wisc.edu      5000
Chen,Jean     789123456  chen@wisc.edu         24000
Juniper,Jack  345678912  jjuniper@wisc.edu    300000
```

A *regular expression* describes a set of character strings. In an emacs regular expression,

- letters and digits (a-z, A-Z, 0-9) match themselves; e.g. C-x C-f data.txt Enter C-M-s wisc C-g
- . matches any single character (except a newline); C-M-s ., . C-g
- square brackets, [...], enclose a *character set* that matches any one of its characters, except that [^...] matches any one character *not* in the set; a hyphen (-) indicates a range
- ^ matches the beginning of a line and \$ matches the end; e.g. C-M-s [A-Z] C-g, C-M-s ^[A-Z] C-g
- \< matches the beginning of a word and \> matches the end; e.g. C-M-s [13] C-g, C-M-s \<[13] C-g
- repetition quantifiers indicate matching the (smallest) previous expression
 - \{n\} exactly n times; e.g. C-M-s [aeiou]\{2\}
 - \{n,m\} n to m times (shorthand: ? means \{0,1\})
 - \{n,\} n or more times (shorthand: * means \{0, \}, + means \{1, \})

Note: repetition is maximal, except that appending ? to a quantifier makes it minimal. e.g. C-M-s [0-9] [a-z] and then add + and then ?

- \| means “or”; e.g. C-M-s Jo\|Je
- (escaped) parentheses, \(...\), enclose an expression; a backreference \d (where d is a *digit* in 1:9) refers to what the dth enclosed expression matched; e.g. C-M-s \([a-z]\)\1
- To match one of \$^.*+?[\, escape it with \. e.g. \^ matches ^ (and not beginning-of-line).

For more, see https://www.gnu.org/software/emacs/manual/html_node/emacs/Regexps.html.

Examples

Use M-x query-replace-regexp on that file for these operations:

- To replace each vowel with V, replace [aeiou] with V.
- To replace each nonvowel with C, replace [^aeiou] with C.
- To replace four digits with DDDD, replace [0-9]\{4\} with DDDD.
- To replace only four-digit “words” with DDDD, replace \<[0-9]\{4\}\> with DDDD.
- To change the “last,first ID email ...” line format to “first,last,NetID,ID”, replace \([a-zA-Z]+\),\([a-zA-Z]+\) +\([0-9]\{9\}\) \([a-z]+\)@wisc.edu.* with \2,\1,\4,\3.