Lightgrep Cheat Sheet

1 Single Characters

c	the character c*
\a	U+0007 (BEL) bell
\e	U+001B (ESC) escape
\f	U+000C (FF) form feed
\n	U+000A (NL) newline
\r	U+000D (CR) carriage return
\t	U+0009 (TAB) horizontal tab
ooo	U+*ooo*, 1–3 octal digits o, $\leq$ 0377
\x*hh*	U+00*hh*, 2 hexadecimal digits h
\x{*hhhhhh*}	U+*hhhhhh*, 1–6 hex digits h
\z*hh*	the byte 0x*hh* (*not* the character!)†
\N{*name*}	the character called *name*
\N{U+*hhhhhh*}	same as \x{*hhhhhh*}
\c	the character c‡

*except U+0000 (NUL) and metacharacters
†Lightgrep extension; not part of PCRE.
‡except any of: adefnprstwDPSW1234567890

2 Named Character Classes

.	any character
\d	[0-9] (= ASCII digits)
\D	[^0-9]
\s	[\t\n\f\r] (= ASCII whitespace)
\S	[^\t\n\f\r]
\w	[0-9A-Za-z_] (= ASCII words)
\W	[^0-9A-Za-z_]
\p{*property*}	any character having *property*
\P{*property*}	any character lacking *property*

3 Character Classes

[*stuff*]	any character in *stuff*
[^*stuff*]	any character **not** in *stuff*

where *stuff* is…

c	a character
a-b	a character range, inclusive
\z*hh*	a byte
\z*hh*-\z*hh*	a byte range, inclusive
[S]	a character class
ST	$S \cup T$ (union)
S&&T	$S \cap T$ (intersection)
S--T	$S - T$ (difference)
S~~T	$S \triangle T$ (symmetric difference, XOR)

8 EnCase GREP Syntax

c	the character c (except metacharacters)
\x*hh*	U+00*hh*, 2 hexadecimal digits h
\w*hhhh*	U+*hhhh*, 4 hexadecimal digits h
\c	the character c
.	any character
#	[0-9] (= ASCII digits)
[a-b]	any character in the range a–b
[S]	any character in S
[^S]	any character not in S
(S)	grouping
S*	repeat S 0 or more times (max 255)
S+	repeat S 1 or more times (max 255)
S?	repeat S 0 or 1 or time
S{n,m}	repeat S n–m times (max 255)
ST	matches S, then matches T
S|T	matches S or T

9 Importing from EnCase into Lightgrep

\w*hhhh*	⟶ \x*hhhh*	S* and S+ are limited to
#	⟶ \d	255 repetitions by EnCase;
S*	⟶ S{0,255}	Lightgrep preserves this in
S+	⟶ S{1,255}	imported patterns.

\w is limited to BMP characters ($\leq$ U+10000) only.

Some people, when confronted with a problem, think "I know, I'll use regular expressions." Now they have two problems.

—JWZ in alt.religion.emacs, 12 August 1997

Notes & Examples

Characters:

```
.*?\x00          (= null-terminated string)
\z50\z4B\z03\z04     (= ZIP signature)
\N{EURO SIGN}, \N{NO-BREAK SPACE}
\x{042F} (= CYRILLIC CAPITAL LETTER YA)
\+12\.5%    (= escaping metacharacters)
```

Grouping: Operators bind tightly. Use (aa)+, not aa+, to match pairs of a's.

Ordered alternation: a|ab matches a twice in aab. Left alternatives preferred to right.

Repetition: Greedy operators match as much as possible. Reluctant operators match as little as possible. a+a matches all of aaaa; a+?a matches the first aa, then the second aa.

.+ will (uselessly) match the **entire** input. Prefer reluctant operators when possible.

Character classes:

```
[abc] = a, b, or c
[^a] = anything but a
[A-Z] = A to Z
[A\-Z]
   = A, Z, or hyphen (!)
[A-Zaeiou] = capitals
   or lowercase vowels
[.+*?\]]
      = ., +, *, ?, or ]
[Q\z00-\z7F]
     = Q or 7-bit bytes
[[abcd][bce]]
      = a, b, c, d, or e
[[abcd]&&[bce]]
                = b or c
[[abcd]--[bce]]
                = a or d
[[abcd]~~[bce]]
              = a, d, or e
[\p{Greek}\d]
      = Greek or digits
[^\p{Greek}7]
  = neither Greek nor 7
[\p{Greek}&&\p{Ll}]
    = lowercase Greek
```

Operators need not be escaped inside character classes.

Lightgrep Search for EnCase®

Fast Search for Forensics

www.lightgrep.com

4 Grouping

(S) makes any pattern S atomic

5 Concatenation & Alternation

ST	matches S, then matches T
S|T	matches S or T, preferring S

6 Repetition

		Repeats S…
Greedy	S*	0 or more times (= S{0,})
	S+	1 or more times (= S{1,})
	S?	0 or 1 time (= S{0,1})
	S{n,}	n or more times
	S{n,m}	n–m times, inclusive
Reluctant	S*?	0 or more times (= S{0,})
	S+?	1 or more times (= S{1,})
	S??	0 or 1 time (= S{0,1})
	S{n,}?	n or more times
	S{n,m}?	n–m times, inclusive

7 Selected Unicode Properties

```
Any                     Assigned
Alphabetic              White_Space
Uppercase               Lowercase
ASCII                   Noncharacter_Code_Point
Name=name               Default_Ignorable_Code_Point
General_Category=category
  L, Letter                   P, Punctuation
  Lu, Uppercase Letter        Pc, Connector Punctuation
  Ll, Lowercase Letter        Pd, Dash Punctuation
  Lt, Titlecase Letter        Ps, Open Punctuation
  Lm, Modifier Letter         Pe, Close Punctuation
  Lo, Other Letter            Pi, Initial Punctuation
  M, Mark                     Pf, Final Punctuation
  Mn, Non-Spacing Mark        Po, Other Punctuation
  Me, Enclosing Mark          Z, Separator
  N, Number                   Zs, Space Separator
  Nd, Decimal Digit Number    Zl, Line Separator
  Nl, Letter Number           Zp, Paragraph Separator
  No, Other Number            C, Other
  S, Symbol                   Cc, Control
  Sm, Math Symbol             Cf, Format
  Sc, Currency Symbol         Cs, Surrogate
  Sk, Modifier Symbol         Co, Private Use
  So, Other Symbol            Cn, Not Assigned
Script=script
  Common Latin Greek Cyrillic Armenian Hebrew Ara-
  bic Syraic Thaana Devanagari Bengali Gurmukhi Gu-
  jarati Oriya Tamil Telugu Kannada Malayalam Sin-
  hala Thai Lao Tibetan Myanmar Georgian Hangul
  Ethiopic Cherokee Ogham Runic Khmer Mongolian
  Hiragana Katakana Bopomofo Han Yi Old_Italic
  Gothic Inherited Tagalog Hanunoo Buhid Tagbanwa
  Limbu Tai_Le Linear_B Ugaritic Shavian Osmanya
  Cypriot Buginese Coptic New_Tai_Lue Glagolitic
  Tifinagh Syloti_Nagri Old_Persian Kharoshthi Ba-
  linese Cuneiform Phoenician Phags_Pa Nko Sudanese
  Lepcha …                    See Unicode Standard for more.
```

Email addresses: `[a-z\d!#$%&'*+/=?^_‘{|}~-][a-z\d!#$%&'*+/=?^_‘{|}~.-]{0,63}@[a-z\d.-]{1,253}\.[a-z\d-]{2,22}`

Hostnames: `([a-z\d]([a-z\d_-]{0,61}[a-z\d])?\.){2,5}[a-z\d][a-z\d-]{1,22}`

N. American phone numbers: `\(?\d{3}[ ).-]{0,2}\d{3}[ .-]?\d{4}\D`

Visa, MasterCard: `\d{4}([ -]?\d{4}){3}`

American Express: `3[47]\d{2}[ -]?\d{6}[ -]?\d{5}`

Diners Club: `3[08]\d{2}[ -]?\d{6}[ -]?\d{4}`

EMF header: `\z01\z00\z00\z00.{36}\z20EMF`

JPEG: `\zFF\zD8\zFF[\zC4\zDB\zE0-\zEF\zFE]` Footer: `\zFF\zD9`

GIF: `GIF8[79]` Footer: `\z00\z3B`

BMP: `BM.{4}\z00\z00\z00\z00.{4}\z28`

PNG: `\z89\z50\z4E\z47` Footer: `\z49\z45\z4E\z44\zAE\z42\z60\z82`

ZIP: `\z50\z4B\z03\z04` Footer: `\z50\z4B\z05\z06`

RAR: `\z52\z61\z72\z21\z1a\z07\z00...[\z00-\z7F]`

Footer: `\z88\zC4\z3D\z7B\z00\z40\z07\z00`

GZIP: `\z1F\z8B\z08`

MS Office 97-03: `\zD0\zCF\z11\zE0\zA1\zB1\z1A\zE1`

LNK: `\z4c\z00\z00\z00\z01\z14\z02\z00`

PDF: `\z25\z50\z44\z46\z2D\z31` Footer: `\z25\z45\z4F\z46`