Regular Expressions

WHENEVER I LEARN A NEW SKILL I CONCOCT ELABORATE FANTASY SCENARIOS WHERE IT LETS ME SAVE THE DAY.
OH NO! THE KILLER MUST HAVE FOLLOWED HER ON VACATION!
BUT TO FIND THEM WE'D HAVE TO SEARCH THROUGH 200 MB OF EMAILS LOOKING FOR SOMETHING FORMATTED LIKE AN ADDRESS!
IT'S HOPELESS!
EVERYBODY STAND BACK.
I KNOW REGULAR EXPRESSIONS.
PERL!
TAP TAP

By Carlos Guzmán

Regular Expressions

- Flavors/Languages
- Basic Characters
- Quantifiers
- Groups
- Anchors and Boundaries
- Modifiers
- Where to learn
- Resources and sources

Basic characters

- . – Any character except line break
- \d – One digit from 0 to 9
- \w – Alphanumerics and _
- \s – Whitespace
- \D, \W, \S
- \n, \t
- [^A-z] – ASCII range[-] and NOT^
- a|b – Disjunction a OR b

POSIX Bracket expressions

- [:alnum:] - [a-zA-Z0-9]
- [:alpha:] - [a-zA-Z]
- [:punct:] - [!"#$%&'()*+,\-./:;<=>?@[\\\]^_`{|}~]
- [:xdigit:] - [A-Fa-f0-9]
- [:ascii:] - [\x00-\x7F]

Quantifiers

- ? – One or none
- * – Zero or more
- + – One or more
- {n} – Exactly n
- {n, m} – n to m
- “Greedy” vs “Lazy”?

Examples

- Find date JAN-01-1970
- Validate date MMM-DD-YYYY
  - SEP-17-1991 valid; LIT-38-1969 invalid
- Email address
  - Mail::RFC822::Address: regexp-based address validation
- HTML tags
  - <a href=”bit.ly/acmatnyu”>Hello ACM!</a>

Groups

	Positive	Negative
Lookahead	(?=...)	(?!...)
Lookbehind	(?<=...)	(?<!...)

- Backreferences: (\w+)(\d)\1\2 or \k<name>
- Non-capturing: (?: …)
- Named: (?<name> …)
- Subroutines: (\d*\,) (?1)
- Recursion: (?R)
- Conditional: (?(cond)true|false)

Anchors and Boundaries

- ^ – Start of string/line
- $ – End of string/line
- \A to \z – Absolute beginning to absolute end
- \G – Beginning or end of previous match
- \b – Word boundary [^\b] = \B

Examples

- Get only Real Madrid scores :
  - FCB W:5 D:2 L:3 RMA W:9 D:1 L:0 ATM W:7 D:1 L:2
- Password conditions:
  - At least 4 uppercase characters, between 3 and 5 digits, at least one lowercase and a unicorn
- Hex values
  - #ABC012, #DDD

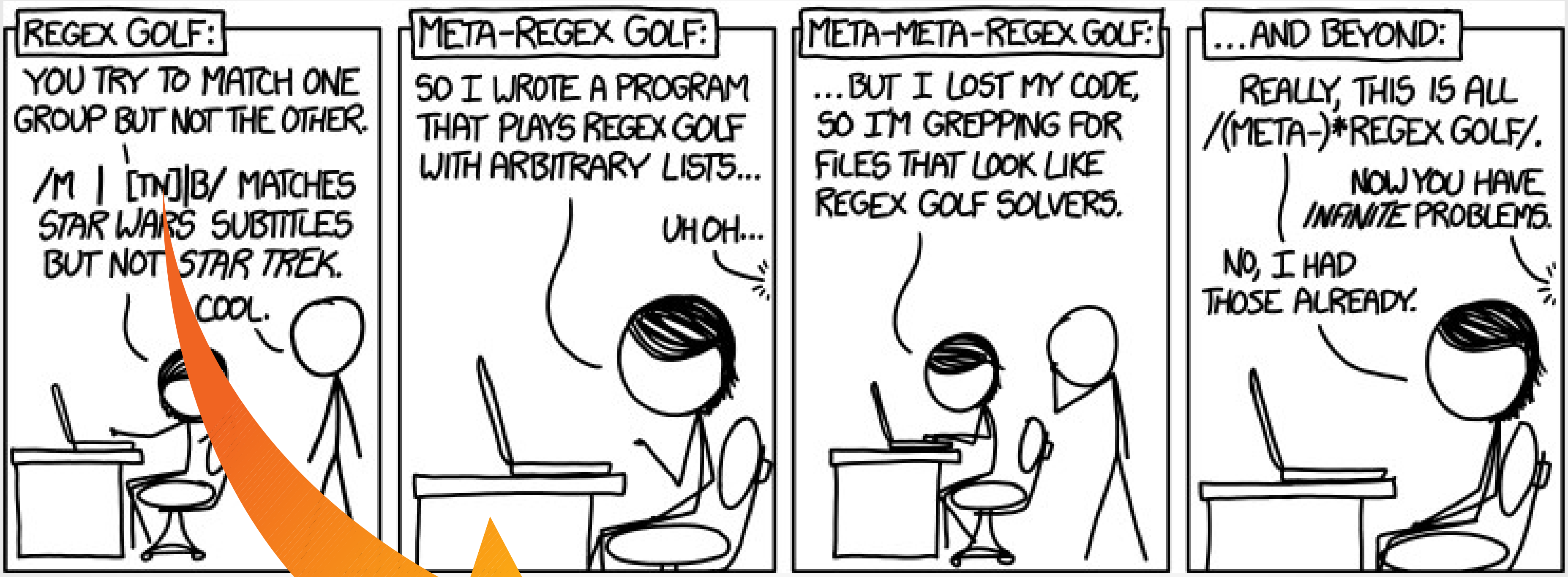
Modifiers (? .)

- g – Global. All matches, not only first
- m – Multiline. ^ and $ match ends of line
- i – Insensitive. [a-zA-Z] ignored
- s – Single line. \n is matched with .
- U – Ungreedy. Lazy by default
- x – Extended. Allows comments with #

Curiosities

- [-~] - All printable characters range
- [^\D2-9]+ - Binary number
- (?i)\b[a-gq-tv-xz]+\b – Left hand only
- ^(?=(?!(.)\1)([^\DO:105-93+30])(?-1)(?<!\d(?<=(?![5-90-3])\d))).[^\WHY?]$ - The meaning of life
- (?:f(?:ive|our)|s(?:even|ix)|t(?:hree|wo)|(?:ni|o)ne|eight) – one to nine

More XKCD


REGEX GOLF:
YOU TRY TO MATCH ONE GROUP BUT NOT THE OTHER.
/M | [TN]|B/ MATCHES STAR WARS SUBTITLES BUT NOT STAR TREK.
COOL.
META-REGEX GOLF:
SO I WROTE A PROGRAM THAT PLAYS REGEX GOLF WITH ARBITRARY LISTS...
UH OH...
META-META-REGEX GOLF:
...BUT I LOST MY CODE, SO I'M GREPPING FOR FILES THAT LOOK LIKE REGEX GOLF SOLVERS.
...AND BEYOND:
REALLY, THIS IS ALL /(META-)*REGEX GOLF/.
NOW YOU HAVE INFINITE PROBLEMS.
NO, I HAD THOSE ALREADY.

/M | [TN]|B/

Extra: Catastrophic Backtracking

- (x+x+)+y
  - What happens when matching 'xxxxxxxxxy'?
  - What happens when matching 'xxxxxxxxxx'?

- Catastrophy

Where to learn

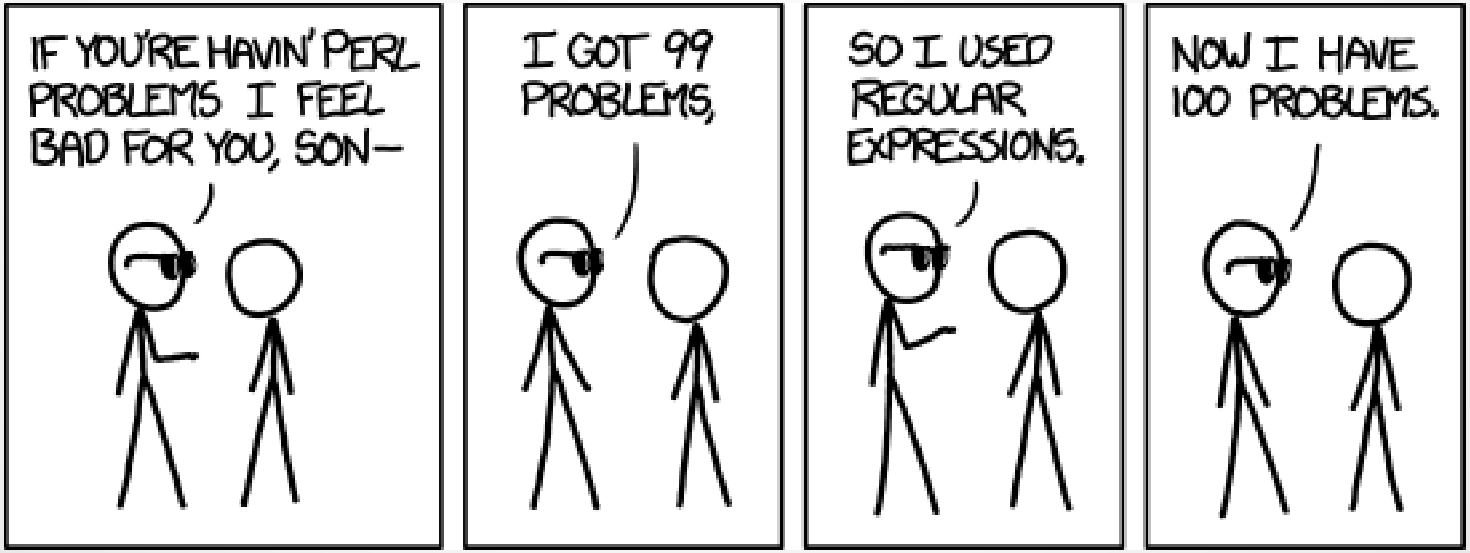
- RegexOne
- Cheatsheet
- Regex Golf
- Regex Tuesday
- Regex Crosswords, specially this one
- Learn Regex The Hard Way

Feedback and questions are appreciated:

carlos.guzman@cs.nyu.edu

Sources

- RexEgg
- RegularExpressions.info


IF YOU'RE HAVIN' PERL PROBLEMS I FEEL BAD FOR YOU, SON—
I GOT 99 PROBLEMS,
SO I USED REGULAR EXPRESSIONS.
NOW I HAVE 100 PROBLEMS.