Usage and Refactoring Studies Of Python Regular Expressions

Carl Chapman

Iowa State University

carlallenchapman@gmail.com

13 April, 2016

Overview

Basic Regex References Are Missing

feature usage statistics

What features are more important when...

building an analysis tool?

developing test regexes?

creating a toy language for a research project?

feature set summaries for variants and tools

What features does each language and tool support...

so I can port my regex code?

so I can choose the best fitting language?

so I can choose the best analysis tool?

Why Python?

1. back of envelope feature set comparision, seems like a good balance

Project Selection

1. picture of the 32 starting points
2. impl. issues limited size of dataset obtained

Utilization Defined

1. utl. image

re module insights

1. two tables

PCRE Parsing Patterns

1. that image

Ranked features: Languages

1. ranked features, eight languages combined, red circle around n guys, if possible

Ranked features: Analysis Tools

1. ranked features, four tools combined, red circle around n guys, if possible

What Are Regexes Used For?

1. want to know to support use cases

How to Categorize Regex Usages

1. ~~thorough inspection of 55K utilizations~~
2. ~~unguided manual categorization of 13.5k regexes, without objective basis~~
3. ~~cluster by syntactic similarity like Jaccard or longest substring~~
4. ~~formal analytical subsumption, using brics (30% or less)~~
5. Chosen technique: cluster by behavioral similarity using Rex (61%)

Measuring Behavioral Similarity

1. that m100A stuff

MCL example

1. very small example, use pdfs

Six Categories Of Clusters

Table: Cluster categories and sizes, ordered by number of projects containing at least one pattern in the category.

Category	Clusters	Patterns	Projects	% Projects
Multi Matches	21	237	295	40%
Specific Char	17	103	184	25%
Anchored Patterns	20	85	141	19%
Two or More Chars	16	40	120	16%
Content of Parens	10	46	111	15%
Code Search	15	27	92	13%

Survey Goals

- confirm or deny parts 1 and 2:...
- investigate some topics: usage freq., pain points, testing, html parsing, ephem vs pers. comparision

Confirming PT1

- idk

Confirming PT2

- idk

Regex Testing

- image of regex101

Parsing HTML

- idk, maybe some regexes that parse html

Usage Frequency

- idk

Ephemeral vs Persistent Users

Pain Points

hard to compose (11)

...very difficult to write them since I've never read up on them.

...trickiness to getting the expression right

hard to read (7)

long ones can be hard to read

Readability. Edge cases.

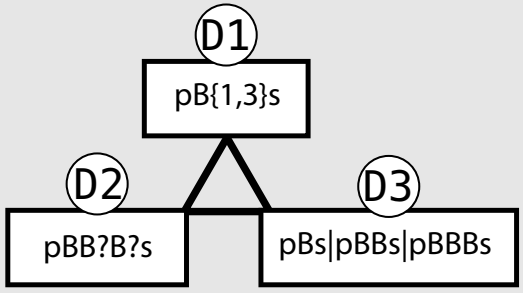
It is terrible to read (especially later after initial development)

inconsistency across implementations (3)

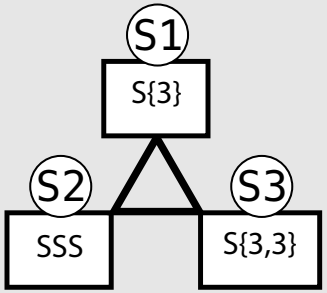
Differences in implementation across languages

Some regexes work differently (or don't work) in some languages.

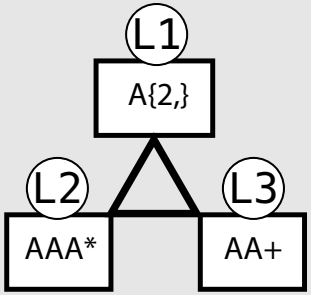
Regex Equivalence Classes



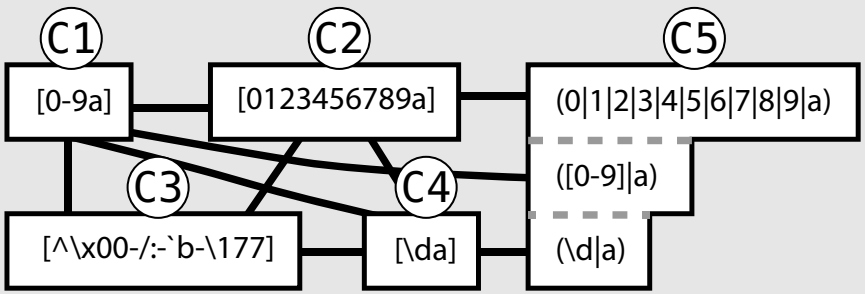
DBB GROUP
using the abstract `pB{1,3}s' where B is any pattern,
p and s are any (possibly empty) prefix, suffix



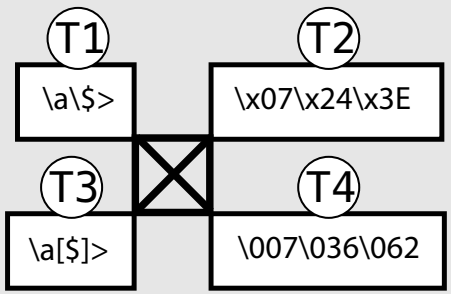
SNG GROUP
using the abstract `S{3}'
where S is any pattern



LWB GROUP
using the abstract `A{2,}'
where A is any pattern



CCC GROUP
using the concrete patern `[0-9a]' and assuming an ASCII charset



LIT GROUP
using the concrete pattern `\a\$>'
and assuming an ASCII charset

Example Equivalences

LIT : `x` ≡ `y`

DBB : another one

CCC : a = b

LWB : another one

SNG : a = b

Table

Imagetest

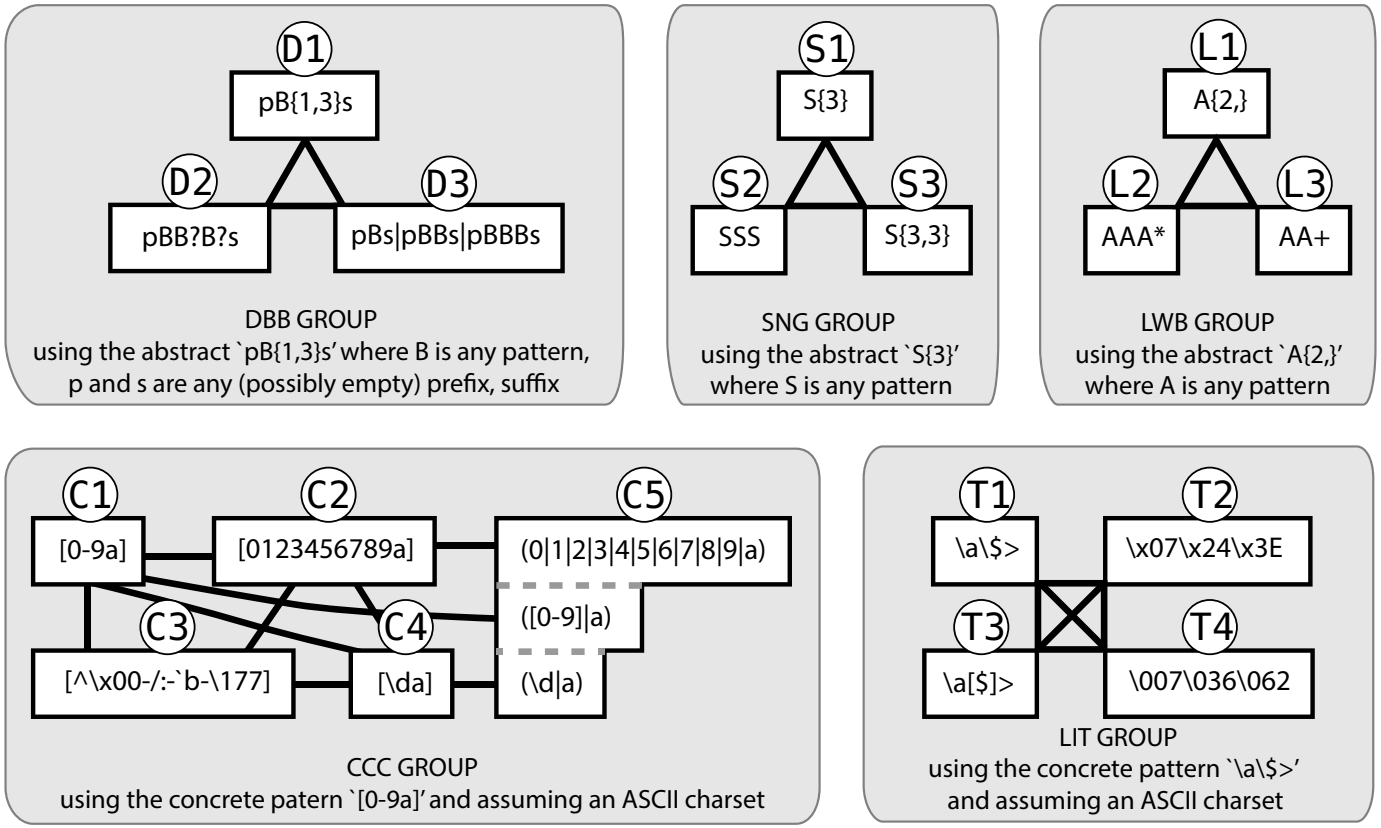


Figure: Equivalence classes with various representations of semantically equivalent representations within each class. DBB = Double-Bounded, SNG = Single Bounded, LWB = Lower Bounded, CCC = Custom Character Class and LIT = Literal

Citation

An example of the `\cite` command to cite within the presentation:

This statement requires citation Smith (2012).

Verbatim

Example (Theorem Slide Code)

```
\begin{frame}
\frametitle{Theorem}
\begin{theorem}[Mass--energy equivalence]
$E = mc^2$
\end{theorem}
\end{frame}
```

regex formatting test

```
ab*c
```

References

John Smith (2012) Title of the publication Journal Name 12(3), 45 – 678.

Questions?