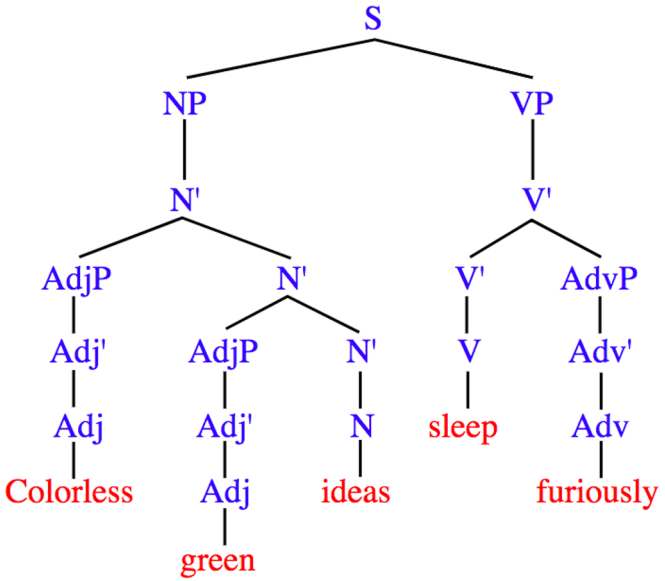
An Intro to Regex in R

Alan Wu

What is regex?

- “Regular expression”
- Sequence of characters that define a search pattern
  - `^\(\d{3}\)\s\d{3}-\d{4}`
  - `(###) ###-####`
- “Regular” refers to a language being definable by certain regular rules



What should I regex?

- Strings

When should I use regex

- Use for
  - Multiple, related conditions or nested if/else
  - Stringr/search functions
    - str_detect(), grepl(): Identify match to a pattern
    - str_extract(): extract match to a pattern
    - str_locate(): locate position of pattern
    - str_replace(), gsub(): replace a pattern
    - str_split(): split a string using a pattern
- Don’t use it when something simpler will suffice

Why should I regex?

- Flexible and structured
  - “Rename visit IDs with either between one and three spaces in their names”
  - “Instances of ‘Pred’, ‘Prad’, ‘Pren’...”
  - “Find all study IDs that start with a 9 and end with a 0”
- Clearer and easier to understand (when done right)
  - Cleaner code
- Easier to change/adapt
  - Reproducibility

I’m sold. How do I regex?

Regex	Meaning	Example	Strings detected
^	Start of a string	^abc	abc, abcdefg, abc123
$	End of a string	abc$	abc, defgabc, 123abc
.	Any character	a.c	abc, acc, a1c
|	Alternation	apple|orange	apple, orange
{...}	Explicit quantifier	ab{2}c	abbc
[...]	Explicit character match	a[bB]c	abc, aBc
(...)	Expression grouping	(abc){2}	abcabc

I’m sold. How do I regex?

Regex	Meaning	Example	Strings detected
\	Escape key for special characters	\$	$32.50, pa$$word
\s	Any white-space	[bB]\s[cd]	B c, b d
\w	Any word	([aeiou])\w+	alien, elephant, igloo
\d	Any digit	(study_id)\d+	study_id5, study_id1
\S, \W, \D	The converse of the previous three	(study_id)\D+	study_idA, study_idf

Looking ahead

- More advanced tactics include:
  - Lookarounds: (?<= …)
    - “Bradley Cooper” vs. “Cooper Hewitt”
  - Conditionals: (? (A)B|C)
    - If A is true, match with B; else match with C
  - Omitting: s\Kt
    - Text matched by left of \K is omitted, so only the first “t” in “streets” would be captured

I’m sold. How do I regex?

Regex	Meaning	Example	Strings detected
[a-z] or [:lower:]	Lowercase letters	[^ABZ[:lower:]]	A, B, C, D, c, d
[:upper:]	Uppercase letters	[ABZ[:upper:]]	A, B
[a-zA-Z] or [:alpha:]	Letters	[a-zA-Z]\w+	Alabama, alabama
[:digit:]	Digits	a[[:digit:]]b	a0b, a1b, a2b
[:alnum:]	Alphanumeric	[[:alnum:]]\w+	A1b2n, 2bdiC3D,
[:punct:]	punctuation	end[[:punct:]]	end., end,, end!

Let’s look at some code