

# Regex basics

Chris Potts, Ling 278: Programming for Linguists, Fall 2009

Sep 30

## Primitives

$\alpha$  and  $\beta$  are sets of strings drawn from our alphabet:

regular expression	corresponding set of strings	notes
a	{a}	and so on for all letters and digits
\*	{*}	and so on for all the control characters
( $\alpha \beta$ )	{ $\alpha$ } $\cup$ { $\beta$ }	
$\alpha\beta$	{ $\alpha$ }{ $\beta$ }	i.e., {conc( $x, y$ )   $x \in \alpha$ & $y \in \beta$ }
$\alpha^*$	0 or more concatenations of elements of $\alpha$	

## Some supported abbreviations

.	abbreviates	disjunction of every character in the alphabet
[a-z]	abbreviates	(a b c d e f g h i j k l m n o p q r s t u v w x y z)
[^bs0]	abbreviates	everything in the vocabulary <i>except</i> b, s, and 0
a+	abbreviates	a(a)*
a{4}	abbreviates	aaaa
a{2,4}	abbreviates	(aa aaa aaaa)
a{3,}	abbreviates	aaaa*
a?	abbreviates	( a)
\s	abbreviates	space
\d	abbreviates	[0-9]
\w	abbreviates	[0-9a-zA-Z]

## Very important special characters

^	start of line (except when right after the left square bracket)
\$	end of line
\n	newline
\r	carriage return (often hard to distinguish from newline)
\b	word boundary

Capitalizing special characters generally negates them. For example, \S matches non-space, and \D matches non-digits.