Regular Expressions: Disjunctions

- The string of characters inside the braces specifies a disjunction of characters to match

Pattern	Matches
`[wW]oodchuck`	Woodchuck, woodchuck
`[1234567890]`	Any digit

- Ranges `[A-Z]` the brackets can be used with the dash (-) to specify range any one character in a range.

Pattern	Matches	
`[A-Z]`	An upper case letter	`Drenched Blossoms`
`[a-z]`	A lower case letter	`my beans were impatient`
`[0-9]`	A single digit	`Chapter 1: Down the Rabbit Hole`

Regular Expressions: Negation in Disjunction

Negations [^Ss]

- Carat means negation only when first in [] [^

Pattern	Matches	
[^A-Z] [^A-Z]	Not an upper case letter	Oyfn pripetchik
[^Ss]	Neither 'S' nor 's'	I have no exquisite reason"
[e^] '^'	Either e or ^	Look here
a^b	The pattern a carat b	Look up a^b now

[a b] [W w]

Regular Expressions: More Disjunction

Woodchuck is another name for groundhog!

The pipe | for disjunction

Pattern	Matches
groundhog|woodchuck	woodchuck
yours|mine [yours mine]	yours
a|b|c [a b c]	= [abc]
[gG]roundhog|[Ww]oodchuck	Woodchuck

Regular Expressions: Kleene *, Kleene +

- Kleene *: The set of operators that allows us to say things like "some number of as" are based on the asterisk (*) commonly called the Kleene*.

$aa^* = aa$

- Kleene + means "one or more occurrences of the immediately preceding character or regular expression".
- period (/./), a wildcard expression that matches any single character (except a carriage return).

$a^* = \Lambda, a, aa, aaa \ldots$ $aa\Lambda \ldots$

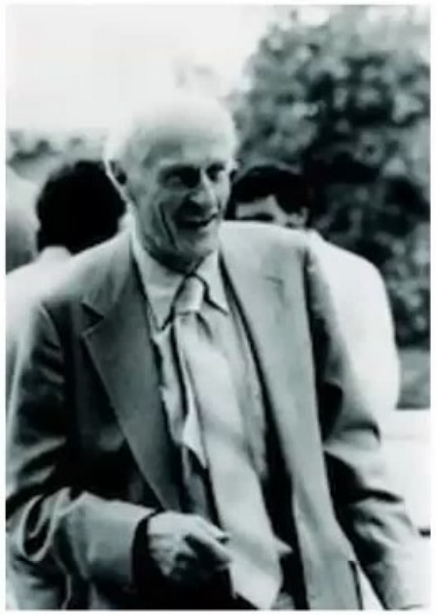
$a^* = a, aa, aaa,$

Regular Expressions: ? *+ .

color | colour

Pattern	Matches	
colou?r	Optional previous char	color colour
oo*h!	0 or more of previous char	oh! ooh! oooh! ooooh!
o+h!	1 or more of previous char	oh! ooh! oooh! ooooh!
baa+		baa baaa baaaa baaaaa
begin	.	begin begun begun beg3n

bas, baas



Stephen C Kleene

Kleene *, Kleene +

Regular Expressions: Anchors ^ $

- Anchors are special characters that anchor regular expressions to particular places in a string.

Line start

Pattern	Matches
^[A-Z]	Palo Alto
^[^A-Za-z]	1 "Hello"
\.$	The end.
.$	The end? The end!

^ [^ => negate

99 99

\b 99 \b

- \b (word boundary) \B (non word boundary)

Regular Expressions:

Disjunction Operator:

- The disjunction operator, also called the pipe symbol |.
- The pattern /cat|dog/ matches either the string cat or the string dog.

fly | flies fly | flies

Precedence:

- To make the disjunction operator apply only to a specific pattern, we need to use the parenthesis operators (and).

fl (y | ies)

Example

Find me all instances of the word "the" in a text.

`the`

The THE

Misses capitalized examples

`[tT]he`

Incorrectly returns `other` or `theology`

`[^a-zA-Z][tT]he[^a-zA-Z]`

Errors

RE the

1 2 3 4 5

The process we just went through was based on fixing two kinds of errors:

1. Matching strings that we should not have matched (there, then, other)

   False positives (Type I errors)

   Increasing precision → minimizing False positive

the

the

2. Not matching things that we should have matched (The)

   False negatives (Type II errors)

   Increasing recall → minimizing False negatives

More Operators

RE	Expansion	Match	First Matches
\d	[0-9]	any digit	Party␣of␣5
\D	[^0-9]	any non-digit	Blue␣moon
\w	[a-zA-Z0-9_]	any alphanumeric/underscore	Daiyu
\W	[^\w]	a non-alphanumeric	!!!!
\s	[␣\r\t\n\f]	whitespace (space, tab)	
\S	[^\s]	Non-whitespace	in␣Concord

More Operators

RE	Match
*	zero or more occurrences of the previous char or expression
+	one or more occurrences of the previous char or expression
?	exactly zero or one occurrence of the previous char or expression
{n}	*n* occurrences of the previous char or expression
{n,m}	from *n* to *m* occurrences of the previous char or expression
{n,}	at least *n* occurrences of the previous char or expression
{,m}	up to *m* occurrences of the previous char or expression

colour

a{3} aaa a{3,} a{,3}

a 2 → 10 (ab){3} ababab

a{2,10}

More Operators

RE	Match	First Patterns Matched
*	an asterisk "*"	"K*A*P*L*A*N"
\.	a period "."	"Dr. Livingston, I presume"
\?	a question mark	"Why don't they come and lend a hand?"
\n	a newline	
\t	a tab	