JURAFSKY, Daniel and James H. MARTIN. *Speech and Language Processing.* Prentice Hall, 2000.

CHAPTER 2. REGULAR EXPRESSIONS AND AUTOMATA

RE patterns

plaintext	/searchterm/	
options	/[abc]/ /[0123456789]/	
match	/[A-Z][a-z]/ /[0-9]/	
not	/[^A-Z]/ *not an uppercase letter*	
one option	/colou?r/ *"u" once or not at all*	
one char	/beg.n/ *matches begin, began, begun*	
more	/a*/ *matches a, aa, aaaa, aaaaaaa*	/.*/ *matches anything*
either	/cat|dog/ /gupp(y|ies)/	
start/end	/^start and end$/	

RE special characters

\d	any digit
\D	any non-digit
\w	any alpha-numeric
\W	any non-alphanumeric
\s	whitespace (\t tab, \r return, \n newline)
\b	word boundary
*	zero or more of previous
?	zero or one of previous
+	one or more of previous
{*n*}	*n* occurrences of previous
{*n*, *m*}	*n* to *m* occurrences of previous

Substitution:

s/oldthing/newthing/*flags*
s/throw_out(*keep*)/\1/ *backreference to keep*
s/^The (.*) and (.*)/The \2 and \1/ *switch the first and second results*

Definitions:

State is a pointer or position on a running program
Deterministic algorithm contains no choices (no if-statements)
Non-deterministic algorithm contains choices (if-statements)
Epsilon-transition (ε) allows you to step backwards without looking at input
L(*m*) is a formal language defined by *m* where *m* is a set of keyword strings
NFSA algorithm (figure 2.21, p. 44)

State-space search algorithms, isolate a (memory) space to search

- LIFO (last-in, first-out), like checking a stack of plates for food stains
- FIFO (first-in, first-out), like checking people in a line for ID—aka linear search
- *binary search* (or binary tree search), cut the space in half each time and search that
- *jump search* (for a sorted data set), jump by *x* steps and check if *search term* is greater than or less than current state