Pattern Matching

- exact pattern matching
- Knuth-Morris-Pratt
- RE pattern matching
- grep

References:
Algorithms in C (2nd edition), Chapter 19
http://www.cs.princeton.edu/introalgsds/63long
http://www.cs.princeton.edu/introalgsds/72regular
exact pattern matching

- Knuth-Morris-Pratt
- RE pattern matching
- grep
Exact pattern matching

**Problem:**
Find first match of a pattern of length \( M \) in a text stream of length \( N \).

- **Pattern:** `needle` \( M = 6 \)
- **Text:** `in a haystack a needle in a`
  \( N = 21 \)
  typically \( N \gg M \)

**Applications.**
- parsers.
- spam filters.
- digital libraries.
- screen scrapers.
- word processors.
- web search engines.
- natural language processing.
- computational molecular biology.
- feature detection in digitized images.
  
  ...
**Brute-force exact pattern match**

Check for pattern starting at each text position.

```java
public static int search(String pattern, String text) {
    int M = pattern.length();
    int N = text.length();

    for (int i = 0; i < N - M; i++) {
        int j;
        for (j = 0; j < M; j++)
            if (text.charAt(i+j) != pattern.charAt(j))
                break;
        if (j == M) return i;  // pattern start index in text
    }
    return -1;  // not found
}
```
Brute-force exact pattern match: worst case

Brute-force algorithm can be slow if text and pattern are repetitive

but this situation is rare in typical applications

Hence, the `indexOf()` method in Java's `String` class uses brute-force
Exact pattern matching in Java

Exact pattern matching is implemented in Java’s `String` class:

`s.indexOf(t, i)`: index of first occurrence of pattern `t` in string `s`, starting at offset `i`.

**Ex: Screen scraping.** Exact match to extract info from website

```java
public class StockQuote {
    public static void main(String[] args) {
        String name = "http://finance.yahoo.com/q?s=";
        In in = new In(name + args[0]);
        String input = in.readAll();
        int start = input.indexOf("Last Trade:", 0);
        int from = input.indexOf("<b>", start);
        int to = input.indexOf("</b>", from);
        String price = input.substring(from + 3, to);
        System.out.println(price);
    }
}
```

```html
...<tr><td class= "yfnc_tablehead1" width= "48%">Last Trade: </td><td class= "yfnc_tabledata1"><big><b>688.04</b></big></td></tr><td class= "yfnc_tablehead1" width= "48%">Trade Time: </td><td class= "yfnc_tabledata1"><hr><big><b>688.04</big></hr></td></tr>
```

```
% java StockQuote goog
688.04
% java StockQuote msft
33.75
```
Algorithmic challenges in pattern matching

Brute-force is not good enough for all applications

Theoretical challenge: Linear-time guarantee. ← fundamental algorithmic problem

Practical challenge: Avoid backup in text stream. ← often no room or time to save text

Now is the time for all people to come to the aid of their party. Now is the time for all
good people to come to the aid of their party. Now is the time for many good people to
come to the aid of their party. Now is the time for all good people to come to the aid of
their party. Now is the time for a lot of good people to come to the aid of their party.
Now is the time for all of the good people to come to the aid of their party. Now is the
time for all good people to come to the aid of their party. Now is the time for each good
person to come to the aid of their party. Now is the time for all good people to come to
the aid of their party. Now is the time for all good Republicans to come to the aid of
their party. Now is the time for all good people to come to the aid of their party. Now is
the time for all good people to come to the aid of their party. Now is the time for all good
Democrats to come to the aid of their party. Now is the time for all people to come to
the aid of their party. Now is the time for all good people to come to the aid of their
party. Now is the time for all good people to come to the aid of their party. Now is the
time for all good people to come to the aid of their party. Now is the time for all good
people to come to the aid of their party. Now is the time for all good people to come to
the aid of their party. Now is the time for all good people to come to the aid of their
party. Now is the time for all good people to come to the aid of their party. Now is the
time for all good people to come to the aid of their party. Now is the time for all good
people to come to the aid of their party. Now is the time for all good people to come to
the aid of their party. Now is the time for all good Democrats to come to the aid of their
party. Now is the time for all good people to come to the aid of their party.
- exact pattern matching
- Knuth-Morris-Pratt
- RE pattern matching
- grep
Knuth-Morris-Pratt (KMP) exact pattern-matching algorithm

Classic algorithm that meets both challenges
• linear-time guarantee
• no backup in text stream

Basic plan (for binary alphabet)
• build DFA from pattern
• simulate DFA with text as input

No backup in a DFA
Linear-time because each step is just a state change
Knuth-Morris-Pratt DFA example

One state for each pattern character

- Match input character: move from $i$ to $i+1$
- Mismatch: move to previous state

DFA for pattern $a\ a\ b\ a\ a\ a$

How to construct? Stay tuned
Knuth-Morris-Pratt DFA simulation

0 a a a b a a b a a a b

1 a a a b a b a a a b

2 a a a b a a b a a a b

3 a a a b a a b a a a b
Knuth-Morris-Pratt DFA simulation

4  a a a b a a b a a a b

5  a a a b a a b a a a b

3  a a a b a a b a a a b

4  a a a b a a b a a a b

5  a a a b a a b a a a b

accept!
Knuth-Morris-Pratt DFA simulation

When in state $i$:
- have found match in $i$ previous input chars
- that is the longest such match

Ex. End in state 4 iff text ends in $aaba$.
Ex. End in state 2 iff text ends in $aa$ (but not $aabaa$ or $aabaaa$).
KMP implementation

DFA representation: a single state-indexed array `next[]`
- Upon character match in state $j$, go forward to state $j+1$.
- Upon character mismatch in state $j$, go back to state `next[j].`
KMP implementation

Two key differences from brute-force implementation:
• Text pointer $i$ never decrements
• Need to precompute `next[]` table (DFA) from pattern.

```java
int j = 0;
for (int i = 0; i < N; i++) {
    if (t.charAt(i) == p.charAt(j)) j++;  // match
    else j = next[j];  // mismatch
    if (j == M) return i - M + 1;  // found
} return -1;  // not found
```

Simulation of KMP DFA
Knuth-Morris-Pratt: Iterative DFA construction

DFA for first i states contains the information needed to build state i+1

Ex: given DFA for pattern aabaaa.
   how to compute DFA for pattern aabaaab?

Key idea
• on mismatch at 7th char, need to simulate 6-char backup
• previous 6 chars are known (abaaaaa in example)
• 6-state DFA (known) determines next state!

Keep track of DFA state for start at 2nd char of pattern
• compare char at that position with next pattern char
• match/mismatch provides all needed info
KMP iterative DFA construction: two cases

Let $x$ be the next state in the simulation and $j$ the next state to build.

If $p[x]$ and $p[j]$ match, copy and increment

```plaintext
next[j] = next[X];
x = X+1
```

DFA for $a a b a a a b$

If $p[x]$ and $p[j]$ mismatch, do the opposite

```plaintext
next[j] = X+1;
x = next[X];
```

DFA for $a a b a a a a$
Knuth-Morris-Pratt DFA construction

DFA

<table>
<thead>
<tr>
<th>State</th>
<th>Input</th>
<th>Next State</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>a</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>b</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>a</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>b</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>a</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>a</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>b</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>b</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>a</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>a</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>b</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>b</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>a</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>a</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>b</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>b</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>a</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>a</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>b</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>b</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>a</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>a</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>b</td>
<td>5</td>
</tr>
</tbody>
</table>

X: current state in simulation
compare p[j] with p[X]

match: copy and increment
next[j] = next[X];
X = X + 1;
mismatch: do the opposite
next[j] = X + 1;
X = next[X];
Knuth-Morris-Pratt DFA construction examples

<table>
<thead>
<tr>
<th>ex: $a\ a\ b\ a\ a\ a\ b$</th>
<th>ex: $a\ b\ b\ a\ b\ b\ b$</th>
</tr>
</thead>
</table>
| \begin{align*}
  &0 \\
  &a \\
  &0 \\
  &0 1 \\
  &a a \\
  &0 0 \\
  &\uparrow \uparrow \\
  &X \\
  &j \\
  &0 1 2 \\
  &a a b \\
  &0 0 2 \\
  &\uparrow \\
  &0 1 2 3 \\
  &a a b a \\
  &0 0 2 0 \\
  &\uparrow \uparrow \\
  &0 1 2 3 4 \\
  &a a b a a \\
  &0 0 2 0 0 \\
  &\uparrow \\
  &0 1 2 3 4 5 \\
  &a a b a a a \\
  &0 0 2 0 0 3 \\
  &\uparrow \uparrow \\
  &0 1 2 3 4 5 6 \\
  &a a b a a a b \\
  &0 0 2 0 0 3 2 \\
  &\uparrow \uparrow \\
\end{align*} | \begin{align*}
  &0 \\
  &a \\
  &0 \\
  &0 1 \\
  &a b \\
  &0 1 \\
  &\uparrow \uparrow \\
  &X \\
  &j \\
  &0 1 2 \\
  &a b b \\
  &0 1 1 \\
  &\uparrow \\
  &0 1 2 3 \\
  &a b b a \\
  &0 1 1 0 \\
  &\uparrow \\
  &0 1 2 3 4 \\
  &a b b a b \\
  &0 1 1 0 1 \\
  &\uparrow \\
  &0 1 2 3 4 5 \\
  &a b b a b b \\
  &0 1 1 0 1 1 \\
  &\uparrow \uparrow \\
  &0 1 2 3 4 5 6 \\
  &a b b a b b b \\
  &0 1 1 0 1 1 4 \\
  &\uparrow \uparrow \\
\end{align*} |

$x$: current state in simulation
compare $p[j]$ with $p[X]$

**match**: copy and increment
\[
\text{next}[j] = \text{next}[X]; \\
X = X + 1;
\]

**mismatch**: do the opposite
\[
\text{next}[j] = X + 1; \\
X = \text{next}[X];
\]
DFA construction for KMP: Java implementation

Takes time and space proportional to pattern length.

```java
int X = 0;
int[] next = new int[M];
for (int j = 1; j < M; j++)
{
    if (p.charAt(X) == p.charAt(j))
    {
        // match
        next[j] = next[X];
        X = X + 1;
    }
    else
    {
        // mismatch
        next[j] = X + 1;
        X = next[X];
    }
}
```

DFA Construction for KMP (assumes binary alphabet)
Optimized KMP implementation

Ultimate search program for any given pattern:
- one statement comparing each pattern character to next
- match: proceed to next statement
- mismatch: go back as dictated by DFA
- translates to machine language (three instructions per pattern char)

```c
int kmpsearch(char t[])
{
    int i = 0;
    s0: if (t[i++] != 'a') goto s0;
    s1: if (t[i++] != 'a') goto s0;
    s2: if (t[i++] != 'b') goto s2;
    s3: if (t[i++] != 'a') goto s0;
    s4: if (t[i++] != 'a') goto s0;
    s5: if (t[i++] != 'a') goto s3;
    s6: if (t[i++] != 'b') goto s2;
    s7: if (t[i++] != 'b') goto s4;
    return i - 8;
}
```

Lesson: Your computer is a DFA!
KMP summary

**General alphabet**
- more difficult
- easy with \texttt{next[]}[] indexed by mismatch position, character
- KMP paper has ingenious solution that is not difficult to implement
  [ build \texttt{NFA}, then prove that it finishes in $2N$ steps ]

**Bottom line:** linear-time pattern matching is possible (and practical)

**Short history:**
- inspired by esoteric theorem of Cook
  [ linear time 2-way pushdown automata simulation is possible ]
- discovered in 1976 independently by two theoreticians and a hacker
  - Knuth: discovered linear time algorithm
  - Pratt: made running time independent of alphabet
  - Morris: trying to build a text editor.
- theory meets practice
Exact pattern matching: other approaches

Rabin-Karp: make a digital signature of the pattern
• hashing without the table
• linear-time probabilistic guarantee
• plus: extends to 2D patterns
• minus: arithmetic ops much slower than char comparisons

Boyer-Moore: scan from right to left in pattern
• main idea: can skip $M$ text chars when finding one not in the pattern
• needs additional KMP-like heuristic
• plus: possibility of sublinear-time performance ($\sim N/M$)
• used in Unix, emacs

<table>
<thead>
<tr>
<th>pattern</th>
<th>s y z y g y</th>
</tr>
</thead>
<tbody>
<tr>
<td>text</td>
<td>a a a b b a a b a b a a a b b a a a b a a</td>
</tr>
<tr>
<td></td>
<td>s y z y g y</td>
</tr>
</tbody>
</table>
**Exact pattern match cost summary**

### Cost of searching for M-character pattern in N-character text

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Typical</th>
<th>Worst-case</th>
</tr>
</thead>
<tbody>
<tr>
<td>brute-force</td>
<td>$1.1 \times N$ char compares $^\dagger$</td>
<td>$M \times N$ char compares</td>
</tr>
<tr>
<td>Karp-Rabin</td>
<td>$3N$ arithmetic ops</td>
<td>$3N$ arithmetic ops $^\ddagger$</td>
</tr>
<tr>
<td>KMP</td>
<td>$1.1 \times N$ char compares $^\dagger$</td>
<td>$2N$ char compares</td>
</tr>
<tr>
<td>Boyer-Moore</td>
<td>$\sim \frac{N}{M}$ char compares $^\dagger$</td>
<td>$3N$ char compares</td>
</tr>
</tbody>
</table>

$^\dagger$ assumes appropriate model  
$^\ddagger$ randomized
- exact pattern matching
- Knuth-Morris-Pratt
- RE pattern matching
- grep
Regular-expression pattern matching

**Exact pattern matching:**
Search for occurrences of a single pattern in a text file.

**Regular expression (RE) pattern matching:**
Search for occurrences of one of multiple patterns in a text file.

**Ex. (genomics)**
- Fragile X syndrome is a common cause of mental retardation.
- Human genome contains triplet repeats of \texttt{cgg} or \texttt{agg} bracketed by \texttt{gcg} at the beginning and \texttt{ctg} at the end.
- Number of repeats is variable, and correlated with syndrome.
- Use regular expression to specify pattern: \texttt{gcg(cgg|agg)*ctg}
- Do RE pattern match on person’s genome to detect Fragile X

<table>
<thead>
<tr>
<th>Pattern (RE)</th>
<th>Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>\texttt{gcg(cgg</td>
<td>agg)*ctg}</td>
</tr>
</tbody>
</table>
RE pattern matching: applications

Test if a string matches some pattern.
- Process natural language.
- Scan for virus signatures.
- Search for information using Google.
- Access information in digital libraries.
- Retrieve information from Lexis/Nexis.
- Search-and-replace in a word processors.
- Filter text (spam, NetNanny, Carnivore, malware).
- Validate data-entry fields (dates, email, URL, credit card).
- Search for markers in human genome using PROSITE patterns.

Parse text files.
- Compile a Java program.
- Crawl and index the Web.
- Read in data stored in ad hoc input file format.
- Automatically create Java documentation from Javadoc comments.
Regular expression examples

A regular expression is a notation to specify a set of strings.

<table>
<thead>
<tr>
<th>operation</th>
<th>example RE</th>
<th>in set</th>
<th>not in set</th>
</tr>
</thead>
<tbody>
<tr>
<td>concatenation</td>
<td>aabaab</td>
<td>aabaab</td>
<td>every other string</td>
</tr>
<tr>
<td>wildcard</td>
<td>.u.u.u.</td>
<td>cumulus</td>
<td>succubus</td>
</tr>
<tr>
<td></td>
<td></td>
<td>jugulum</td>
<td>tumultuous</td>
</tr>
<tr>
<td>union</td>
<td>aa</td>
<td>baab</td>
<td>aa</td>
</tr>
<tr>
<td>closure</td>
<td>ab*a</td>
<td>aa</td>
<td>abba</td>
</tr>
<tr>
<td></td>
<td>a (a</td>
<td>b) aab</td>
<td>aaaaab</td>
</tr>
<tr>
<td>parentheses</td>
<td>(ab) *a</td>
<td>a</td>
<td>ababababa</td>
</tr>
</tbody>
</table>
Notation is surprisingly expressive

<table>
<thead>
<tr>
<th>regular expression</th>
<th>in set</th>
<th>not in set</th>
</tr>
</thead>
<tbody>
<tr>
<td>.<em>spb.</em></td>
<td>raspberry</td>
<td>subspace</td>
</tr>
<tr>
<td>contains the trigraph spb</td>
<td>crispbread</td>
<td>subspecies</td>
</tr>
<tr>
<td>a*</td>
<td>(a<em>ba</em>a<em>ba</em>)*</td>
<td>bbb</td>
</tr>
<tr>
<td>number of b’s is a multiple of 3</td>
<td>aaa</td>
<td>bb</td>
</tr>
<tr>
<td></td>
<td>bbbbaabbbbaa</td>
<td>baabbbbaa</td>
</tr>
<tr>
<td>.*0....</td>
<td>1000234</td>
<td>111111111</td>
</tr>
<tr>
<td>fifth to last digit is 0</td>
<td>98701234</td>
<td>403982772</td>
</tr>
<tr>
<td>gcg (cgg</td>
<td>agg)*ctg</td>
<td>gcgctg</td>
</tr>
<tr>
<td>fragile X syndrome indicator</td>
<td>gcgcggctg</td>
<td>gcgcggkgggctg</td>
</tr>
<tr>
<td>gcgcggaggctg</td>
<td>gcgcaggctg</td>
<td>gcgcaggctg</td>
</tr>
</tbody>
</table>

and plays a well-understood role in the theory of computation
## Generalized regular expressions

### Additional operations are often added

- **Ex:** \([a-e]^+\) is shorthand for \((a|b|c|d|e)(a|b|c|d|e)^*\)
- **for convenience only**
- **need to be alert for non-regular additions (Ex: Java /)**

<table>
<thead>
<tr>
<th>operation</th>
<th>example</th>
<th>in set</th>
<th>not in set</th>
</tr>
</thead>
<tbody>
<tr>
<td>one or more</td>
<td>(a(bc) + de)</td>
<td>(abcde)</td>
<td>(ade)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(abcbcede)</td>
<td>(bcde)</td>
</tr>
<tr>
<td>character classes</td>
<td>([A-Za-z][a-z]^*)</td>
<td>(word)</td>
<td>(camelCase)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Capitalized)</td>
<td>(4illegal)</td>
</tr>
<tr>
<td>exactly k</td>
<td>([0-9]{5})-[0-9]{4})</td>
<td>08540-1321</td>
<td>111111111</td>
</tr>
<tr>
<td></td>
<td></td>
<td>19072-5541</td>
<td>166-54-111</td>
</tr>
<tr>
<td>negations</td>
<td>([^aeiou]{6})</td>
<td>(rhythm)</td>
<td>(decade)</td>
</tr>
</tbody>
</table>
Regular expressions in Java

RE pattern matching is implemented in Java’s String class
• basic: `match()` method
• various other methods also available (stay tuned)

Ex: Validity checking. Is input in the set described by the re?

```java
public class Validate {
    public static void main(String[] args) {
        String re = args[0];
        String input = args[1];
        System.out.println(input.matches(re));
    }
}
```

% java Validate "..oo..oo." bloodroot  
true

% java Validate "[$_A-Za-z][$_A-Za-z0-9]*" ident123  
true

% java Validate "[a-z]+@[a-z]+\.(edu|com)" rs@cs.princeton.edu  
true

% java Validate "[0-9]{3}-[0-9]{2}-[0-9]{4}" 166-11-4433  
true

need help solving crosswords?
legal Java identifier
valid email address (simplified)
Social Security number
Regular expressions in other languages

Broadly applicable programmer's tool.
- originated in UNIX in the 1970s
- many languages support extended regular expressions
- built into grep, awk, emacs, Perl, PHP, Python, JavaScript

```bash
grep NEWLINE */*.java
```
print all lines containing NEWLINE which occurs in any file with a .java extension

```bash
egrep '^[qwertyuiop]*[zxcvbnm]*$' dict.txt | egrep '............'
```

**PERL.** Practical Extraction and Report Language.

```bash
perl -p -i -e 's|from|to|g' input.txt
```
replace all occurrences of from with to in the file input.txt

```bash
perl -n -e 'print if /^[A-Za-z][a-z]*$/' dict.txt
```
do for each line
Regular expression caveat

Writing a RE is like writing a program.
- need to understand programming model
- can be easier to write than read
- can be difficult to debug

"Sometimes you have a programming problem and it seems like the best solution is to use regular expressions; now you have two problems."
Can the average web surfer learn to use REs?

Google. Supports * for full word wildcard and | for union.
Can the average TV viewer learn to use REs?

**TiVo. WishList has very limited pattern matching.**

Using * in WishList Searches. To search for similar words in Keyword and Title WishList searches, use the asterisk (*) as a special symbol that replaces the endings of words. For example, the keyword AIRP* would find shows containing “airport,” “airplane,” “airplanes,” as well as the movie “Airplane!” To enter an asterisk, press the SLOW ( playbook) button as you are spelling out your keyword or title.

The asterisk can be helpful when you’re looking for a range of similar words, as in the example above, or if you’re just not sure how something is spelled. Pop quiz: is it “irresistible” or “irresistable?” Use the keyword IRRESIST* and don’t worry about it!

Two things to note about using the asterisk:

- **It can only be used at a word’s end;** it cannot be used to omit letters at the beginning or in the middle of a word. (For example, AIR*NE or *PLANE would not work.)

Reference: page 76, Hughes DirectTV TiVo manual
Can the average programmer learn to use REs?

Perl RE for Valid RFC822 Email Addresses

```
(?:(?:\r\n)?[ \t]+)*
(?::\r\n)?[\t]+)?
\[[ ^()<>@,;:\".\[\]\000-\031]+(?:\?::(?:\r\n)?[ \t]+)
\[[ ^()<>@,;:\".\[\]\000-\031]+(?:\?::(?:\r\n)?[ \t]+)
\[[ ^()<>@,;:\".\[\]\000-\031]+(?:\?::(?:\r\n)?[ \t]+)
\[[ ^()<>@,;:\".\[\]\000-\031]+(?:\?::(?:\r\n)?[ \t]+)
\[[ ^()<>@,;:\".\[\]\000-\031]+(?:\?::(?:\r\n)?[ \t]+)
\[[ ^()<>@,;:\".\[\]\000-\031]+(?:\?::(?:\r\n)?[ \t]+)
\[[ ^()<>@,;:\".\[\]\000-\031]+(?:\?::(?:\r\n)?[ \t]+)
\[[ ^()<>@,;:\".\[\]\000-\031]+(?:\?::(?:\r\n)?[ \t]+)
\[[ ^()<>@,;:\".\[\]\000-\031]+(?:\?::(?:\r\n)?[ \t]+)
\[[ ^()<>@,;:\".\[\]\000-\031]+(?:\?::(?:\r\n)?[ \t]+)
\[[ ^()<>@,;:\".\[\]\000-\031]+(?:\?::(?:\r\n)?[ \t]+)
\[[ ^()<>@,;:\".\[\]\000-\031]+(?:\?::(?:\r\n)?[ \t]+)
\[[ ^()<>@,;:\".\[\]\000-\031]+(?:\?::(?:\r\n)?[ \t]+)
\[[ ^()<>@,;:\".\[\]\000-\031]+(?:\?::(?:\r\n)?[ \t]+)
\[[ ^()<>@,;:\".\[\]\000-\031]+(?:\?::(?:\r\n)?[ \t]+)
\[[ ^()<>@,;:\".\[\]\000-\031]+(?:\?::(?:\r\n)?[ \t]+)
\[[ ^()<>@,;:\".\[\]\000-\031]+(?:\?::(?:\r\n)?[ \t]+)
\[[ ^()<>@,;:\".\[\]\000-\031]+(?:\?::(?:\r\n)?[ \t]+)
\[[ ^()<>@,;:\".\[\]\000-\031]+(?:\?::(?:\r\n)?[ \t]+)
\[[ ^()<>@,;:\".\[\]\000-\031]+(?:\?::(?:\r\n)?[ \t]+)
\[[ ^()<>@,;:\".\[\]\000-\031]+(?:\?::(?:\r\n)?[ \t]+)
\[[ ^()<>@,;:\".\[\]\000-\031]+(?:\?::(?:\r\n)?[ \t]+)
```

“Implementing validation with regular expressions somewhat pushes the limits of what it is sensible to do with regular expressions, although Perl copes well.”

37 more lines
- exact pattern matching
- Knuth-Morris-Pratt
- RE pattern matching
- grep
GREP implementation: basic plan

Overview is the same as for KMP!

- linear-time guarantee
- no backup in text stream

Basic plan for GREP

- build DFA from RE
- simulate DFA with text as input

No backup in a DFA
Linear-time because each step is just a state change

Pattern in text

Pattern NOT in text

DFA for pattern

gcg (cgg|agg)*ctg

Text

actgtgcaggagggcgcgcggaggagggctggcgag
Deterministic finite-state automata

DFA review.

```java
int pc = 0;
while (!tape.isEmpty())
{
    boolean bit = tape.read();
    if (pc == 0) { if (!bit) pc = 0; else pc = 1; }
    else if (pc == 1) { if (!bit) pc = 1; else pc = 2; }
    else if (pc == 2) { if (!bit) pc = 2; else pc = 0; }
}
if (pc == 0) System.out.println("accepted");
else System.out.println("rejected");
```
Duality

RE. Concise way to describe a set of strings.
DFA. Machine to recognize whether a given string is in a given set.

Kleene’s theorem.
• for any DFA, there exists a RE that describes the same set of strings
• for any RE, there exists a DFA that recognizes the same set of strings

Ex: set of strings whose number of 1’s is a multiple of 3

Good news: The basic plan works
(build DFA from RE and run with text as input)
Bad news : The DFA can be exponentially large (can’t afford to build it).
Consequence: We need a smaller abstract machine.
Nondeterministic finite-state automata

**NFA.**
- may have 0, 1, or more transitions for each input symbol
- may have $\varepsilon$-transitions (move to another state without reading input)
- accept if *any* sequence of transitions leads to accept state

**Ex: set of strings that do not contain 110**

```
in set: 111, 00011, 101001011
not in set: 110, 00011011, 00110
```

*Convention:* unlabelled arrows are $\varepsilon$-transitions

**Implication of proof of Kleene’s theorem:** RE $\rightarrow$ NFA $\rightarrow$ DFA

**Basic plan for GREP (revised)**
- build **NFA** from RE
- simulate **NFA** with text as input
- give up on linear-time guarantee
Simulating an NFA

**How to simulate an NFA?** Maintain set of all possible states that NFA could be in after reading in the first $i$ symbols.

*One step in simulating an NFA*
NFA Simulation

An NFA trace
NFA representation. Maintain several digraphs, one for each symbol in the alphabet, plus one for $\varepsilon$. 

- $\varepsilon$-graph
- 0-graph
- 1-graph
public class NFA
{
    private int START = 0;       // start state
    private int ACCEPT = 1;      // accept state
    private int N = 2;           // number of states
    private String ALPHABET = "01";  // RE alphabet
    private int EPS = ALPHABET.length(); // symbols in alphabet
    private Digraph[] G;

    public NFA(String re)
    {
        G = new Digraph[EPS + 1];
        for (int i = 0; i <= EPS; i++)
            G[i] = new Digraph();
        build(0, 1, re);
    }

    private void build(int from, int to, String re) { }
    public boolean simulate(Tape tape) { }
}
How to simulate an NFA?

- Maintain a **SET** of all possible states that NFA could be in after reading in the first $i$ symbols.
- Use **Digraph** adjacency and reachability ops to update.

<table>
<thead>
<tr>
<th>pc</th>
<th>next = neighbors of pc in $G[c]$</th>
<th>states reachable from next in $G[\varepsilon]$</th>
<th>updated pc</th>
</tr>
</thead>
<tbody>
<tr>
<td>all states reachable after reading $i$ symbols</td>
<td>possible transitions on reading $(i+1)^{st}$ symbol $c$</td>
<td>possible null transitions before reading next symbol</td>
<td>all states reachable after reading $i+1$ symbols</td>
</tr>
</tbody>
</table>
public boolean simulate(Tape tape) {
    SET<Integer> pc = G[EPS].reachable(START);

    while (!tape.isEmpty()) {
        // Simulate NFA taking input from tape.
        char c = tape.read();
        int i = ALPHABET.indexOf(c);
        SET<Integer> next = G[i].neighbors(pc);
        pc = G[EPS].reachable(next);
    }

    for (int state : pc)
        if (state == ACCEPT) return true;
    return false;
}
Converting from an RE to an NFA: basic transformations

Use generalized NFA with full RE on transitions arrows
• start with one transition having given RE
• remove operators with transformations given below
• goal: standard NFA (all single-character or epsilon-transitions)
Converting from an RE to an NFA example: \( ab^* \mid ab^* \)
private void build(int from, int to, String re) {
    int or = re.indexOf(' |');
    if (re.length() == 0) G[EPSILON].addEdge(from, to);
    else if (re.length() == 1)
    {
        char c = re.charAt(0);
        for (int i = 0; i < EPSILON; i++)
            if (c == ALPHABET.charAt(i) || c == '.')
                G[i].addEdge(from, to);
    }
    else if (or != -1)
    {
        build(from, to, re.substring(0, or));
        build(from, to, re.substring(or + 1));
    }
    else if (re.charAt(1) == '*')
    {
        G[EPSILON].addEdge(from, N);
        build(N, N, re.substring(0, 1));
        build(N++, to, re.substring(2));
    }
    else
    {
        build(from, N, re.substring(0, 1));
        build(N++, to, re.substring(1));
    }
}
Grep running time

**Input.** Text with $N$ characters, RE with $M$ characters.

**Claim.** The number of edges in the NFA is at most $2M$.
- Single character: consumes 1 symbol, creates 1 edge.
- Wildcard character: consumes 1 symbol, creates 2 edges.
- Concatenation: consumes 1 symbols, creates 0 edges.
- Union: consumes 1 symbol, creates 1 edges.
- Closure: consumes one symbol, creates 2 edges.

**NFA simulation.** $O(MN)$ since NFA has $2M$ transitions
- bottleneck: 1 graph reachability per input character
- can be substantially faster in practice if few $\epsilon$-transitions

**NFA construction.** Ours is $O(M^2)$ but not hard to make $O(M)$.

**Surprising bottom line:**
Worst-case cost for grep is the same as for elementary exact match!
Industrial-strength grep implementation

To complete the implementation,
- Deal with parentheses.
- Extend the alphabet.
- Add character classes.
- Add capturing capabilities.
- Deal with meta characters.
- Extend the closure operator.
- Error checking and recovery.
- Greedy vs. reluctant matching.
Regular expressions in Java (revisited)

RE pattern matching is implemented in Java’s `Pattern` and `Matcher` classes

Ex: Harvesting. Print substrings of input that match `re`

```java
import java.util.regex.Pattern;
import java.util.regex.Matcher;

public class Harvester {
    public static void main(String[] args) {
        String re = args[0];
        In in = new In(args[1]);
        String input = in.readAll();
        Pattern pattern = Pattern.compile(re);
        Matcher matcher = pattern.matcher(input);
        while (matcher.find())
            System.out.println(matcher.group());
    }
}
```

% java Harvester "gcg(cgg|agg)*ctg" chromosomeX.txt
  gcgcggcggcggcggctg
  gcgcgcggcggcggctg
  gcgcggcggcggcggctg

% java Harvester "http://(\w+\.)*(\w+)" http://www.cs.princeton.edu
  http://www.princeton.edu
  http://www.google.com

Ex: Harvesting. Print substrings of input that match `re`

- `compile()` creates a `Pattern` (NFA) from RE
- `matcher()` creates a `Matcher` (NFA simulator) from NFA and text
- `find()` looks for the next match
- `group()` returns the substring most recently found by `find()`

- harvest patterns from DNA
- harvest links from website
Typical application: Parsing a data file

Example. NCBI genome file, ...

```
LOCUS AC146846 128142 bp DNA linear HTG 13-NOV-2003
DEFINITION Ornithorhynchus anatinus clone CLM1-393H9,
ACCESSION AC146846
KEYWORDS HTG; HTGS_PHASE2; HTGS_DRAFT.
SOURCE Ornithorhynchus anatinus (platypus)
ORIGIN
   1 tgtatttcct ttgaccgtgc tgttttttccc cggttttttca gtacggtttt agggagccac
   61 gttgattctgt ttgtttttatg ctgcgcaata gctgctcgat gaaatctctgc atagacagct // a comment
   121 ggcgcaggga gaaatgacca gtttggtgatg acaaaatgta ggaagctgtg ttctttctaaa
   ...
128101 ggaatgcga ccccccacgct aatgacagc ttcttttagat tg
```

```
String regexp = "\[ \]*\[0-9\]+(\[actg \]*)\.*";
Pattern pattern = Pattern.compile(regexp);
In in = new In(filename);
while (!in.isEmpty())
{
    String line = in.readLine();
    Matcher matcher = pattern.matcher(line);
    if (matcher.find())
    {
        String s = matcher.group(1).replaceAll(" ", ":");
        // Do something with s.
    }
}
```
Algorithmic complexity attacks

**Warning.** Typical implementations do not guarantee performance!

SpamAssassin regular expression.

```
java RE "[a-z]+@[a-z]+([a-z.]+\.)+[a-z]+" spammer@x.........................
```

- Takes exponential time.
- Spammer can use a pathological email address to DOS a mail server.
Not-so-regular expressions

Back-references.
- \1 notation matches sub-expression that was matched earlier.
- Supported by typical RE implementations.

```java
Java Harvester "\b(\+\)\1\b" dictionary.txt
beriberi
couscous
```

Some non-regular languages.
- set of strings of the form $ww$ for some string $w$: beriberi.
- set of bitstrings with an equal number of 0s and 1s: 01110100.
- set of Watson-Crick complemented palindromes: atttcggaat.

Remark. Pattern matching with back-references is intractable.
Context

Abstract machines, languages, and nondeterminism.
• basis of the theory of computation
• intensively studied since the 1930s
• basis of programming languages

Compiler. A program that translates a program to machine code.
• KMP string ⇨ DFA.
• grep RE ⇨ NFA.
• javac Java language ⇨ Java byte code.

<table>
<thead>
<tr>
<th>pattern</th>
<th>KMP</th>
<th>grep</th>
<th>Java</th>
</tr>
</thead>
<tbody>
<tr>
<td>string</td>
<td>RE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>unnecessary</td>
<td>check if legal</td>
<td>check if legal</td>
<td></td>
</tr>
<tr>
<td>DFA</td>
<td>NFA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DFA simulator</td>
<td>NFA simulator</td>
<td>JVM</td>
<td></td>
</tr>
</tbody>
</table>
Summary of pattern-matching algorithms

Programmer:
• Implement exact pattern matching by DFA simulation (KMP).
• REs are a powerful pattern matching tool.
• Implement RE pattern matching by NFA simulation (grep).

Theoretician:
• RE is a compact description of a set of strings.
• NFA is an abstract machine equivalent in power to RE.
• DFAs and REs have limitations.

You: Practical application of core CS principles.

Example of essential paradigm in computer science.
• Build intermediate abstractions.
• Pick the right ones!
• Solve important practical problems.