Regular Expressions

CS 2800: Discrete Structures, Fall 2014

Sid Chaudhuri
Regular expressions ("regex"-es) are defined *inductively*

(start with simple base expressions, construct more complicated ones recursively)
Empty set

$L(\emptyset) = \emptyset$
Empty string

\[ L(\varepsilon) = \{ \varepsilon \} \]
Literal character

$L(x) = \{x\}$

e.g. $L(1) = \{1\}$, $L(2) = \{2\}$, $L(a) = \{a\}$
Concatenation

$L(AB) = \{ab \mid a \in L(A), b \in L(B)\}$

e.g. $L(12) = \{12\}$, $L(aabb) = \{aabb\}$
Concatenation

$L(AB) = \{ab \mid a \in L(A), b \in L(B)\}$

e.g. $L(12) = \{12\}$, $L(aabb) = \{aabb\}$, $L(a\varepsilon) = ?$
Concatenation

$L(AB) = \{ab \mid a \in L(A), b \in L(B)\}$

e.g. $L(12) = \{12\}$, $L(aabb) = \{aabb\}$,
$L(a\varepsilon) = \{a\}$
Concatenation

\[ L(AB) = \{ ab \mid a \in L(A), b \in L(B) \} \]

e.g. \( L(12) = \{12\} \), \( L(aabb) = \{aabb\} \),
\[ L(a\varepsilon) = \{a\} \), \( L(a\varnothing) = ? \)
Concatenation

$L(AB) = \{ab \mid a \in L(A), b \in L(B)\}$

e.g. $L(12) = \{12\}$, $L(aabb) = \{aabb\}$, $L(a\varepsilon) = \{a\}$, $L(a\emptyset) = \emptyset$
Alternation

$L(A|B) = L(A) \cup L(B)$

e.g. $L(1|2) = \{1, 2\}$, $L(aa|bb) = \{aa, bb\}$
Alternation

\[ L(A | B) = L(A) \cup L(B) \]

e.g. \( L(1|2) = \{1, 2\} \), \( L(aa|bb) = \{aa, bb\} \), \( L(a|\epsilon) = ? \)
Alternation

$L(A|B) = L(A) \cup L(B)$

e.g. $L(1|2) = \{1, 2\}$, $L(aa|bb) = \{aa, bb\}$, $L(a|\varepsilon) = \{a, \varepsilon\}$
Alternation

\[ L(A|B) = L(A) \cup L(B) \]

e.g. \( L(1|2) = \{1, 2\} \), \( L(aa|bb) = \{aa, bb\} \),
\( L(a|\epsilon) = \{a, \epsilon\} \), \( L(a|\emptyset) = ? \)
Alternation

\[ L(A|B) = L(A) \cup L(B) \]

e.g. \( L(1|2) = \{1, 2\} \), \( L(aa|bb) = \{aa, bb\} \),
\( L(a|\varepsilon) = \{a, \varepsilon\} \), \( L(a|\emptyset) = \{a\} \)
Kleene star

\[ L(A^*) = \{ \varepsilon \} \cup \{ x_1 x_2 \ldots x_n \mid n \in \mathbb{N}, x_i \in L(A) \} \]

e.g. \[ L(a^*) = \{ \varepsilon, a, aa, aaa, aaaa, \ldots \} \]
Kleene star

$$L(A^*) = \{ \varepsilon \} \cup \{ x_1 x_2 \ldots x_n \mid n \in \mathbb{N}, x_i \in L(A) \}$$

e.g. $L(a^*) = \{ \varepsilon, a, aa, aaa, aaaa, \ldots \}$, $L((ab)^*) = ?$
Kleene star

\[ L(A^*) = \{ \varepsilon \} \cup \{ x_1 x_2 \ldots x_n \mid n \in \mathbb{N}, x_i \in L(A) \} \]

e.g. \(L(a^*) = \{ \varepsilon, a, aa, aaa, aaaa, \ldots \}\),
\(L((ab)^*) = \{ \varepsilon, ab, abab, ababab, \ldots \}\)
Kleene star

\[ L(A^*) = \{ \varepsilon \} \cup \{ x_1 x_2 \ldots x_n \mid n \in \mathbb{N}, x_i \in L(A) \} \]

e.g. \( L(a^*) = \{ \varepsilon, a, aa, aaa, aaaa, \ldots \} \),
\( L((ab)^*) = \{ \varepsilon, ab, abab, ababab, \ldots \} \),
\( L((a|b)^*) = ? \)
Kleene star

\[ L(A^*) = \{ \varepsilon \} \cup \{ x_1 x_2 \ldots x_n \mid n \in \mathbb{N}, x_i \in L(A) \} \]

e.g. \( L(a^*) = \{ \varepsilon, a, aa, aaa, aaaa, \ldots \} \),
\( L((ab)^*) = \{ \varepsilon, ab, abab, ababab, \ldots \} \),
\( L((a|b)^*) = \{ \varepsilon, a, b, aa, ab, ba, bb, aaa, aab, aba, \ldots \} \)
Kleene star

\[ L(A^*) = \{ \varepsilon \} \cup \{ x_1 x_2 \ldots x_n \mid n \in \mathbb{N}, x_i \in L(A) \} \]

e.g. \( L(a^*) = \{ \varepsilon, a, aa, aaa, aaaa, \ldots \} \),
\( L((ab)^*) = \{ \varepsilon, ab, abab, ababab, \ldots \} \),
\( L((a|b)^*) = \{ \varepsilon, a, b, aa, ab, ba, bb, aaa, aab, aba, \ldots \} \),
\( L(\varepsilon^*) = ? \)
Kleene star

\[ L(A^*) = \{ \varepsilon \} \cup \{ x_1 x_2 \ldots x_n \mid n \in \mathbb{N}, x_i \in L(A) \} \]

e.g. \( L(a^*) = \{ \varepsilon, a, aa, aaa, aaaaa, \ldots \} \),
\( L((ab)^*) = \{ \varepsilon, ab, abab, ababab, \ldots \} \),
\( L((a|b)^*) = \{ \varepsilon, a, b, aa, ab, ba, bb, aaa, aab, aba, \ldots \} \),
\( L(\varepsilon^*) = \{ \varepsilon \} \)
Kleene star

\[ L(A^*) = \{ \varepsilon \} \cup \{ x_1 x_2 \ldots x_n \mid n \in \mathbb{N}, x_i \in L(A) \} \]

e.g. \( L(a^*) = \{ \varepsilon, a, aa, aaa, aaaaa, \ldots \} \),
\( L((ab)^*) = \{ \varepsilon, ab, abab, ababab, \ldots \} \),
\( L((a|b)^*) = \{ \varepsilon, a, b, aa, ab, ba, bb, aaaa, aab, aba, \ldots \} \),
\( L(\varepsilon^*) = \{ \varepsilon \} \), \( L(\emptyset^*) = ? \)
Kleene star

\[ L(A^*) = \{ \varepsilon \} \cup \{ x_1 x_2 \ldots x_n \mid n \in \mathbb{N}, x_i \in L(A) \} \]

e.g. \( L(a^*) = \{ \varepsilon, a, aa, aaa, aaaa, \ldots \} \),
\( L((ab)^*) = \{ \varepsilon, ab, abab, ababab, \ldots \} \),
\( L((a|b)^*) = \{ \varepsilon, a, b, aa, ab, ba, bb, aaa, aab, aba, \ldots \} \),
\( L(\varepsilon^*) = \{ \varepsilon \} \), \( L(\emptyset^*) = \{ \varepsilon \} \)
Playing with regexes

- [http://rubular.com/](http://rubular.com/)
- [http://www.google.com/search?q=online+regex+tester](http://www.google.com/search?q=online+regex+tester)
Kleene's Theorem

- A language is recognized by a regular expression (that is, it is a “regular language”) if and only if it is recognized by a finite automaton
Kleene's Theorem

- A language is recognized by a regular expression (that is, it is a “regular language”) if and only if it is recognized by a finite automaton
  - Regex has FA
    - Relatively simple construction
Kleene's Theorem

- A language is recognized by a regular expression (that is, it is a “regular language”) if and only if it is recognized by a finite automaton
  - Regex has FA
    - Relatively simple construction
  - FA has regex
    - Tricky to prove
Regex $\rightarrow$ FA

• For every regular expression, there is a finite automaton that recognizes the same language
For every regular expression, there is a finite automaton that recognizes the same language

We will construct an $\varepsilon$-NFA
Regex → FA

• For every regular expression, there is a finite automaton that recognizes the same language

• We will construct an $\varepsilon$-NFA
  – ... which can be converted to an NFA
Regex $\rightarrow$ FA

- For every regular expression, there is a finite automaton that recognizes the same language.
- We will construct an $\varepsilon$-NFA
  - ... which can be converted to an NFA
  - ... which can be converted to a DFA
Recap: NFAs with epsilon transitions

- Just like ordinary NFAs, but...
  - Can “instantaneously” change state without reading an input symbol
  - Valid transitions of this type are shown by arcs labeled 'ε'
  - Note that ε does not suddenly become a member of the alphabet. Instead, we assume ε does not belong to any alphabet – it's a special symbol.
Why $\varepsilon$-NFAs?

- Suitable for representing “or” relations
- E.g. $L = \{ a^n \mid n \in \mathbb{N} \text{ is divisible by } 2 \text{ or } 3 \}$

... but they're equivalent to NFAs and DFAs