CS 3500 - Programming Languages and Translators

Study Guide: Test #1

An *alphabet* is a set of symbols.

A *language* is a set of strings build from an alphabet.

(λ is used to denote the empty string)

There are two problems of importance:

1. *Specification*: How do we accurately describe a language?

2. *Recognition*: Given a string, how to determine whether it belongs to a language?

Regular Expressions

A formalism used to specify languages.
A regular expression is build from symbols from the alphabet and special characters '|' '(' ')' '?' '+' '*'.

Suppose our alphabet is {a,b}. The special characters are used to denote the following operations:

Alternatives:
'|' is used to separates alternatives:
Example: ab|ba : the string ab or the string ba {ab, ba}

Grouping:
'(' and ')' are used to define operator precedence. (as usual)

Quantification:
'*' Indicates zero or many repetitions
Example: (ab)* denotes the language { λ, ab, abab, ababab, ...}

'+' Indicates one or more repetitions
Example: (ab)+ denotes the language { ab, abab, ababab, ... }

'?' Indicates zero or one repetitions
Example: (ab)? denotes the language { λ, ab }

Additionally, we will use '[ ]' to mean "any character between the brackets", sometimes with a range between the brackets.
For example:
[a-z] is a shorthand for (a|b|c|d|e|.....|z)
[0-9] is a shorthand for (0|1|2|3|...|9)

Exercises:

1. Write a regular expression for an integer constant that can be any number of digits, optionally preceded by '+' or '-'

   *(+ | -)? [0-9]**

2. Write a regular expression for an identifier that must start with an uppercase letter followed by any number of digits, letters (upper- or lowercase), and/or underscores

   *[A-Z] ([0-9] | [a-z] | [A-Z] | _)**

3. Write a regular expression for the following simplified URL's defined by the following rules:
   a. Strings can optionally begin with https:// or http://
   b. Strings can then (optionally) be followed with www.
   c. Strings must then be followed by one or more letters, digits, and/or underscores
   d. Strings must then end with .edu or .org or .com

(https:// | http://)? (www.)? ([0-9] | [a-z] | [A-Z] | _)+ (.edu | .org | .com)

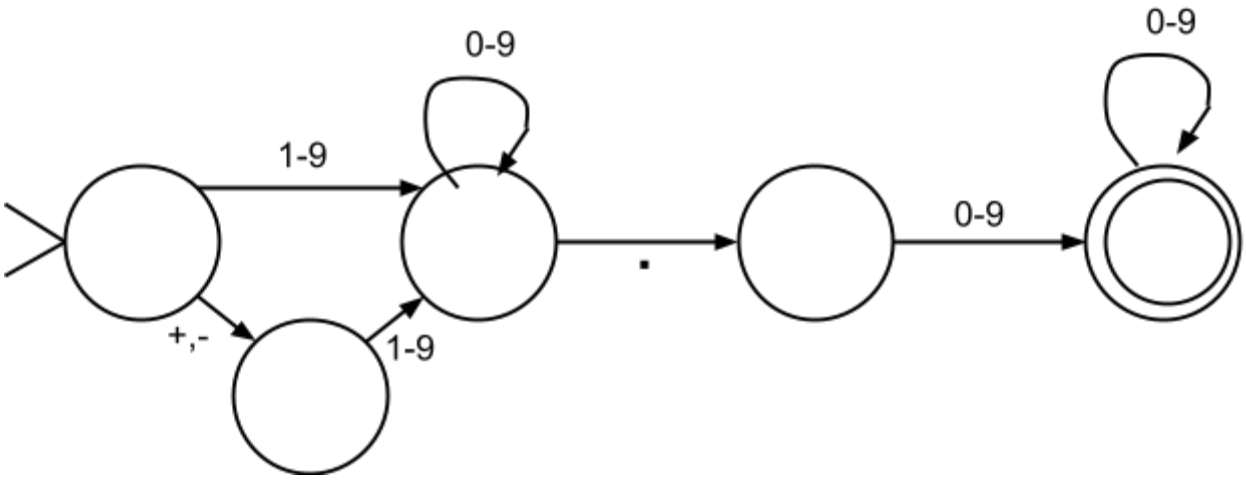
Automata

A formalism used to recognize languages.

Informally, an automata is a set of states, where one state is denoted the “start” state, some states are denoted “accept” states, and the states are connected by directed edges labeled by symbols of the alphabet.

To determine if a string belongs to the language recognized by an automata, beginning at the “start” state, we follow the edges labeled by the symbols in the string in order.
If we end in an “accept” state, the string belongs to the language. Otherwize, the string does not belongs to the language:

Excercises:



Does the preceding automata accepts the following strings:

1. 3.14 *Yes*
2. -.5 *No*
3. 000.3 *No*

4. 6.6.6 *No*

END.