LEX10: Regexp to DFA

Lexical Analysis

CMPT 379: Compilers

Instructor: Anoop Sarkar

anoopsarkar.github.io/compilers-class

Regexp with Distinct Symbols

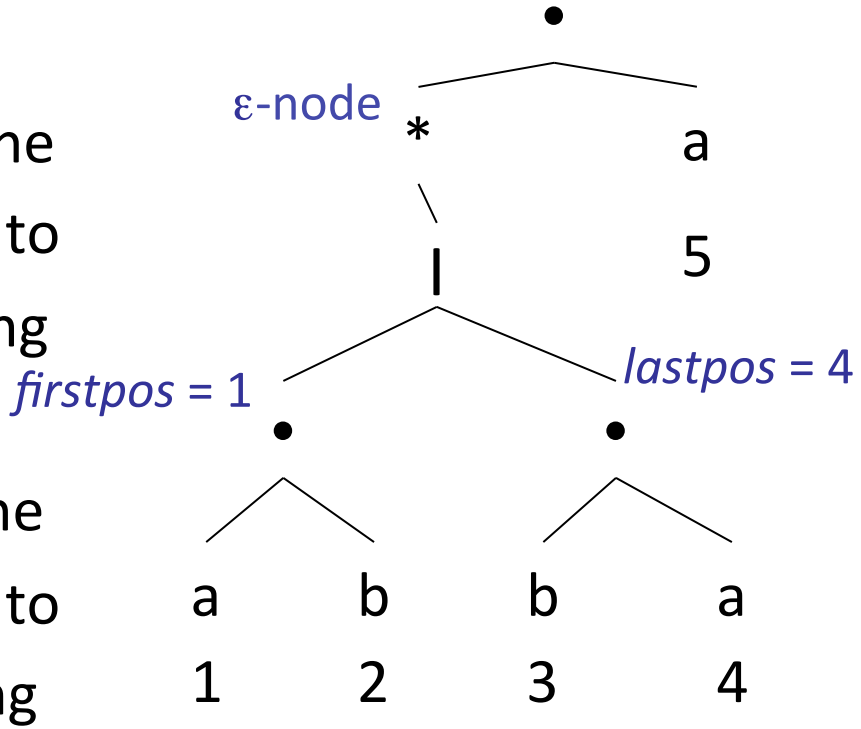
- Associate with each occurrence of a symbol in a regular expression a position
- For example: ((ab)|(ba))*a
  - There are 5 positions:

1:a
2:b
3:b
4:a
5:a

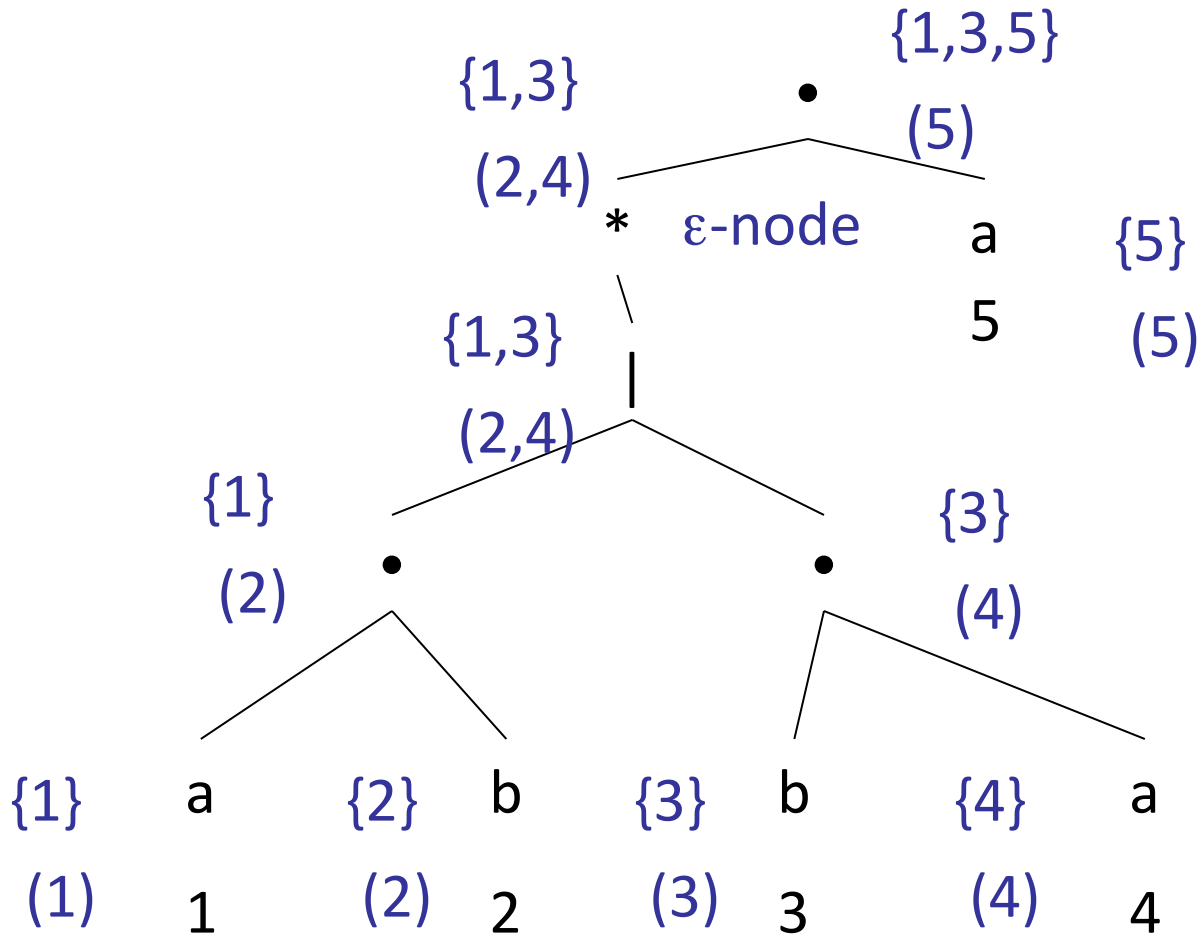
This algorithm was first used by Al Aho in egrep, and used in awk, lex, flex

Regexp to DFA: ((ab)|(ba))*a

- ε-node: if the sub-expression has ε in its language
- firstpos(n): the set of positions in the subtree rooted at n corresponding to the first symbol of at least one string
- lastpos(n): the set of positions in the subtree rooted at n corresponding to the last symbol of at least one string



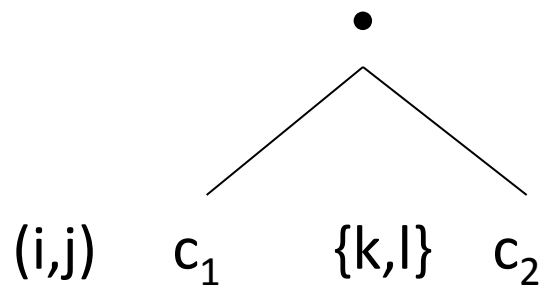
Regexp to DFA: ((ab)|(ba))*a



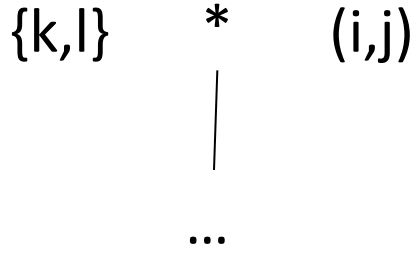
firstpos = {}
lastpos = ()

Regexp to DFA: *followpos*

- *followpos(p):* tells us which positions can follow a position *p*
- There are two rules that use the *firstpos* {} and *lastpos* () information

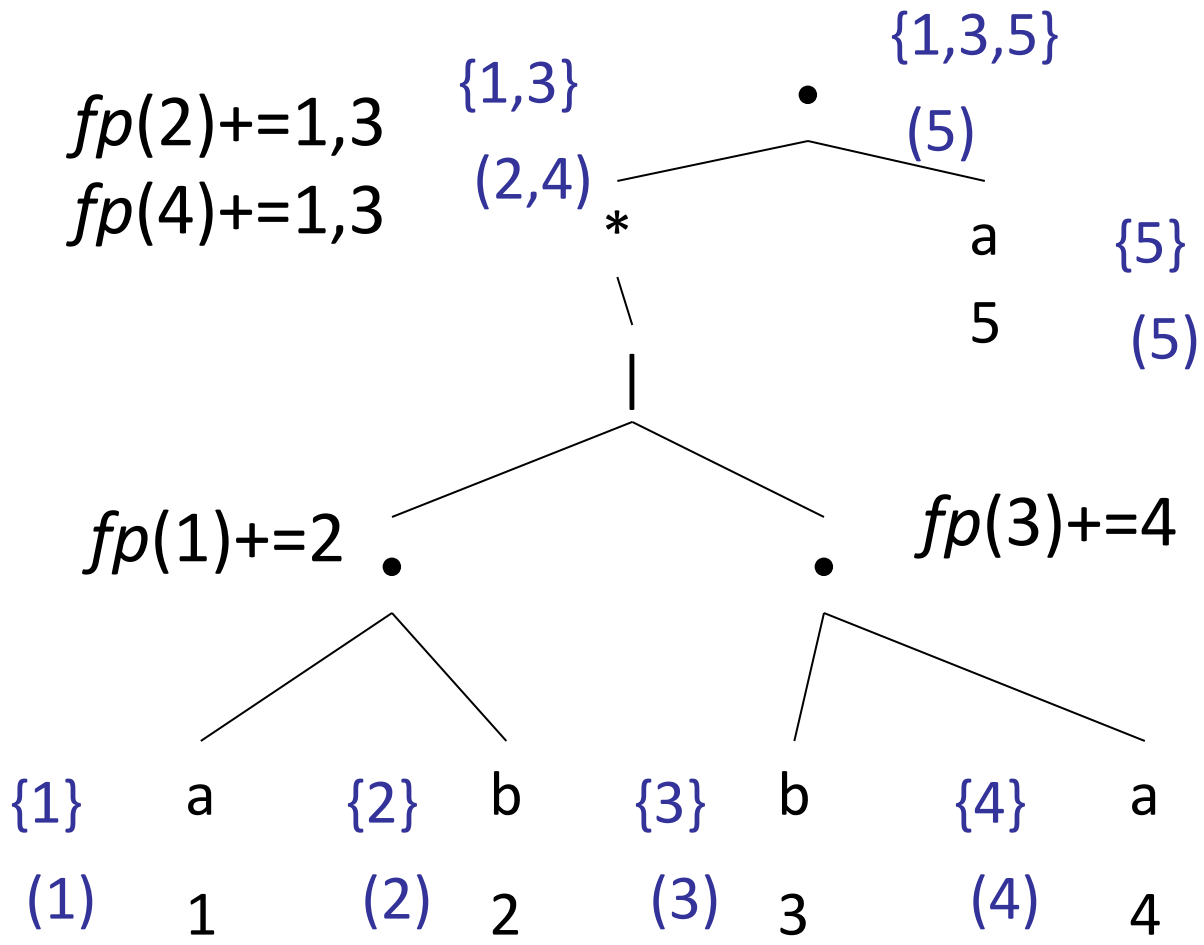


followpos(i)+=k,l
followpos(j)+=k,l



followpos(i)+=k,l
followpos(j)+=k,l

Regexp to DFA: ((ab)|(ba))*a



fp(2)+=1,3
fp(4)+=1,3

fp(2)+=5
fp(4)+=5

fp(1)+=2

fp(3)+=4

root={1,3,5}
fp(1)=2
fp(3)=4
fp(2)=1,3,5
fp(4)=1,3,5

Regexp to DFA: ((ab)|(ba))*a

root={1,3,5}
fp(1)=2
fp(3)=4
fp(2)=1,3,5
fp(4)=1,3,5

1:a
2:b
3:b
4:a
5:a ⟵ *final*

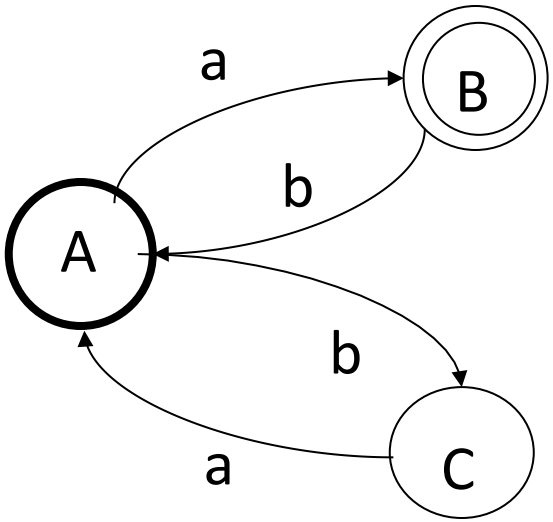
{1,3,5} A

A: *fp*(1),a {2},a B,a
fp(5),a {},a

A: *fp*(3),b {4},b C,b

B: *fp*(2),b {1,3,5},b A,b

C: *fp*(4),a {1,3,5},a A,a



End Marker

- For example: ((ab)|(ba))*

1:a
2:b
3:b
4:a

- Add endmarker ((ab)|(ba))*#

Any state with a transition on # will be marked as final state