Regularni izrazi i konačni automati

Miloš Stanojević

Matematička gimnazija, Nedelja informatike

2. april 2015.

Uvod: mnoštvo primena

- Funkcija `grep` u komandnoj liniji (Unix-based sistemi);
- Word-search u text editorima;
- Leksička analiza u kompajleru i mnoge druge!

Alfabeti

- **Alfabet** je konačan skup Σ, čije elemente zovemo **simboli**;
- Svaki skup može biti alfabet, dokle god je konačan;
- Primeri:
  - $\Sigma_1 = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$
  - $\Sigma_2 = \{a, b, c, \ldots, x, y, z\}$
  - slova azbuke
- Ne-primer:
  - $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$

Stringovi

- **String dužine** n nad alfabetom Σ, je uređena n-torka elemenata iz Σ, napisana bez znakova interpunkcije;
- Primer: Neki stringovi nad $\Sigma = \{a, b, c\}$ su: a, ab, aac, $bbac\ldots$
- Σ^* je skup svih stringova nad skupom Σ;
- ε je jedinstveni **prazan string** za svako Σ.

Konkatenacija stringova

- **Konkatenacija** (ili spajanje) dva stringa $u, v \in \Sigma^*$ je string uv, koji dobijamo kada spojimo kraj prvog sa početkom drugog stringa;
- Primer: $u = ab$, $v = ra$, $w = cad$, tada je $vu = raab$, $uu = abab$ i $wv = cadra$;
- Ovo generalizujemo u spajanje dva ili više stringova (npr. $uvwuv = abracadabra$).

Regularni izrazi nad alfabetom Σ

- svaki simbol $a \in \Sigma$ je regularni izraz;
- ε je regularni izraz;
- $\varnothing$ je regularni izraz;
- ukoliko su r i s regularni izrazi, tada je i $(r|s)$ regularni izraz;
- ukoliko su r i s regularni izrazi, tada je i rs regularni izraz;
- ukoliko je r regularni izrazi, tada je i $(r)^*$ regularni izraz;

Svaki regularni izraz gradi se induktivno, *konačnom* primenom ovih pravila. (**NB** pretpostavljamo da ε, $\varnothing$, (,), | i * *nisu* simboli u Σ.)

Prednost operacija: dogovor

- $-^*$ ima veći prioritet od $-\ -$;
- $-\ -$ ima veći prioritet od $-|-$;
- Primer: $r|st^*$ znači $(r|s(t)^*)$, a ne $(r|s)(t)^*$ ili $((r|st))^*$

Uparivanje stringova i regularnih izraza

- u odgovara $a \in \Sigma \iff u = a$;
- u odgovara $\varepsilon \iff u = \varepsilon$;
- ni jedan string ne odgovara $\varnothing$;
- u odgovara $r|s \iff u$ odgovara r ili u odgovara s;
- u odgovara $rs \iff$ se može izraziti kao konkatenacija dva stringa $u = vw$, gde v odgovara r i w odgovara s;
- u odgovara $r^* \iff$ je $u = \varepsilon$ ili se u može izraziti kao konkatenacija dva ili više stringova, od kojih svaki odgovara r.

Primeri uparivanja, $\Sigma = \{0, 1\}$

- $0|1$ odgovaraju svi simboli u Σ;
- $1(0|1)^*$ odgovaraju svi stringovi u Σ^* koji počinju sa '1';
- $((0|1)(0|1))^*$ odgovaraju svi stringovi parne dužine u Σ^*;
- $(0|1)^*(0|1)^*$ odgovaraju svi stringovi u Σ^*;
- $(\varepsilon|0)(\varepsilon|1)|11$ odgovaraju samo stringovi $\varepsilon, 0, 1, 01$ i 11;
- $\varnothing 1|0$ odgovara samo 0.

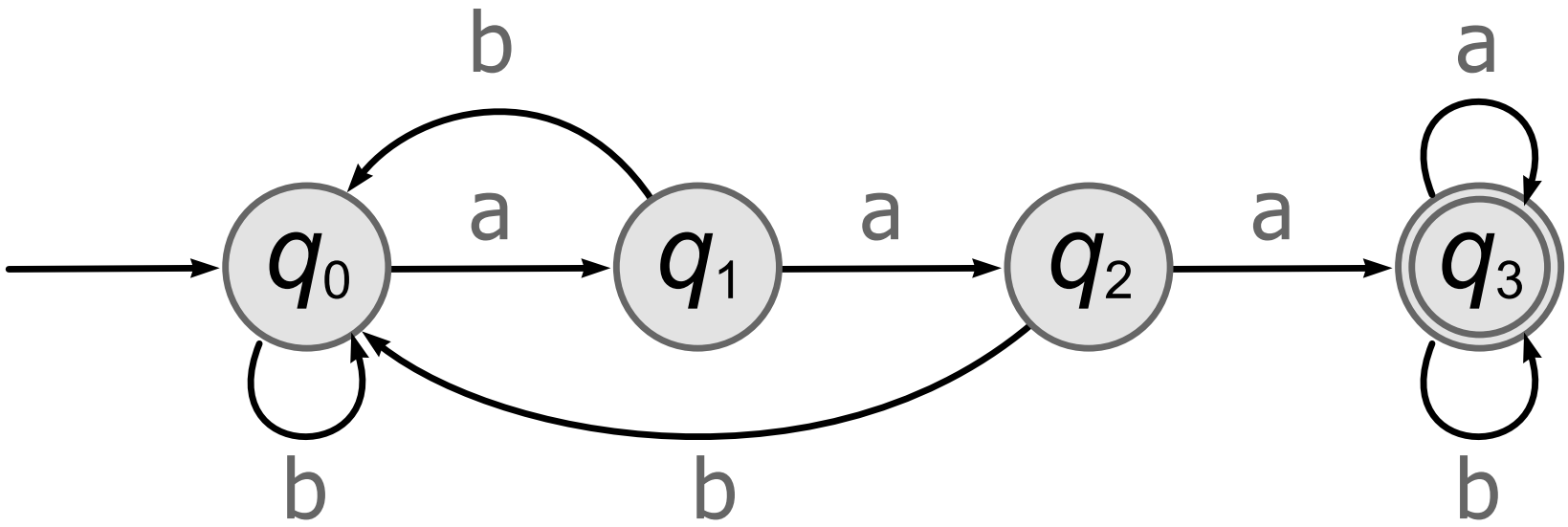
Jezici i regularni izrazi

- (Formalni) **jezik** L nad alfabetom Σ je neki podskup skupa Σ^*;
- **Jezik koji određuje regularni izraz** r nad Σ definiše se kao:

$$L(r) \stackrel{\text{def}}{=} \{u \in \Sigma^* \mid u \text{ odgovara } r\}$$

- Dva regularna izraza r i s (nad istim alfabetom Σ) su **ekvivalentni** akko su $L(r)$ i $L(s)$ jednaki skupovi (odnosno imaju tačno sve iste elemente).

Primer konačnog automata



- Stanja: q_0, q_1, q_2, q_3;
- Ulazni simboli: a, b;
- Prelazi: pogledati strelice na slici;
- Početno stanje: q_0;
- Stanje prihvatanja: q_3.

Jezici i konačni automati

$L(M)$, **jezik koji prihvata konačni automat** M sastoji se od svih stringova u nad njegovim alfabetom ulaznih simbola koji zadovoljavaju $q_0 \xrightarrow{\mathrm{u}}_* q$, gde je q_0 početno stanje i q neko stanje prihvatanja. Ovde

$$q_0 \xrightarrow{u}_* q$$

znači da za neko $u = a_1 a_2 \dots a_n$ postoje neka (ne obavezno različita) stanja $q_1, q_2, \dots, q_n = q$, između kojih su prelazi oblika

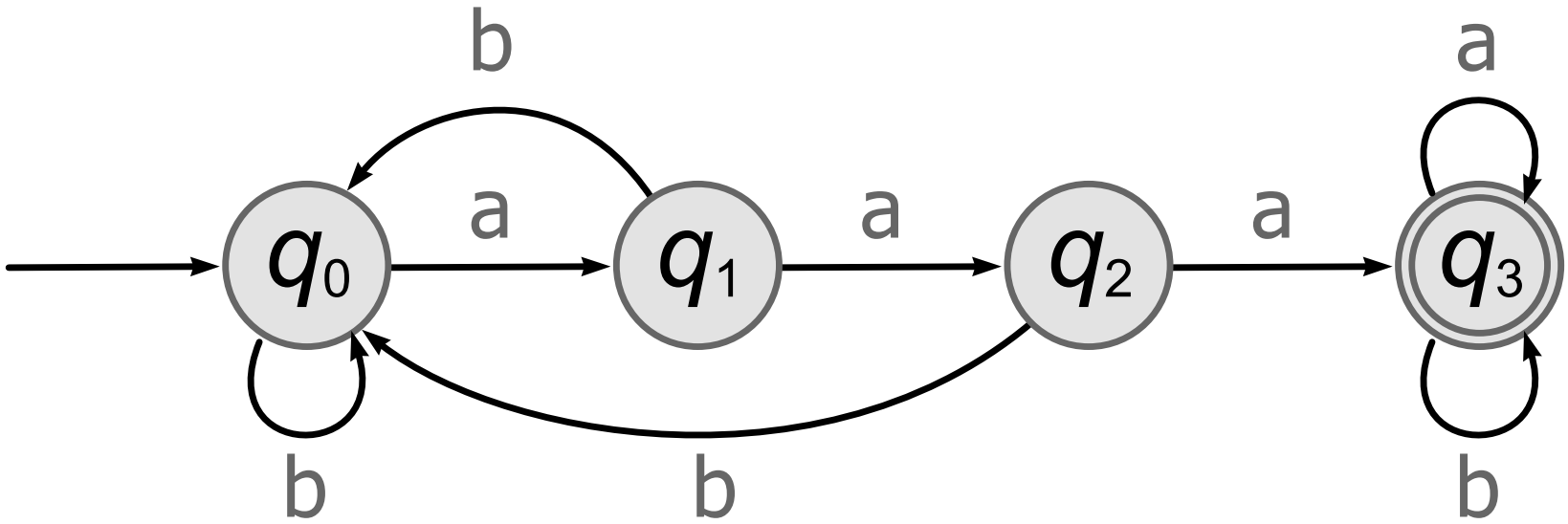
$$q_0 \xrightarrow{a_1} q_1 \xrightarrow{a_2} q_2 \xrightarrow{a_3} \dots \xrightarrow{a_3} q_n = q.$$

NB

slučaj $n = 0$: $q \xrightarrow{\varepsilon}_* q'$ akko $q = q'$

slučaj $n = 1$: $q \xrightarrow{a}_* q'$ akko $q \xrightarrow{a} q'$.

Primeri



- $q_0 \xrightarrow{aaab}_* q_3 \implies aaab \in L(M)$
- $(q_0 \xrightarrow{abaa}_* q \iff q = q_2) \implies abaa \notin L(M)$
- $(q_2 \xrightarrow{baaa}_* q \iff q = q_3) \implies$ nemamo zaključak o $L(M)$
- regularni izraz?

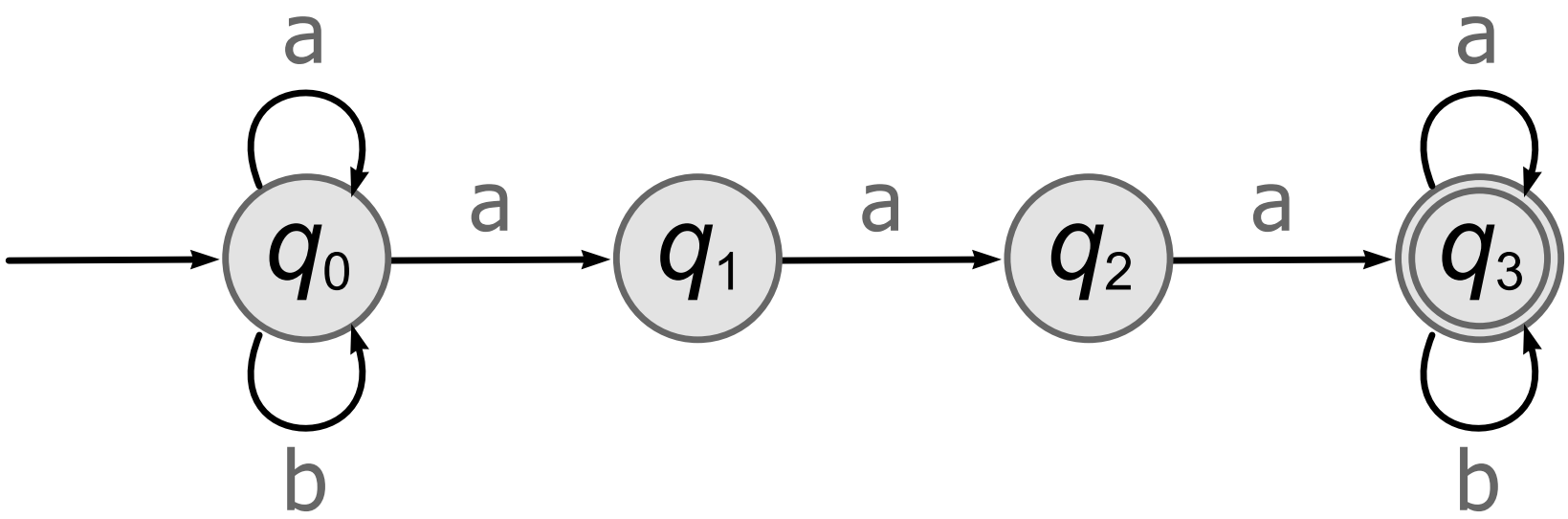
Ne-Deterministički konačni automat

Nedeterministički konačni automat (NFA), M, definisan je sa:

- konačnim skupom **stanja** $Stanja_M$;
- konačnim skupom Σ_M (**alfabet ulaznih simbola**)
- za svako $q \in Stanja_M$ i za svako $a \in \Sigma_M$, podskupom $\Delta_M(q, a) \subseteq Stanja_M$ (skup stanja do kojih se može doći sa jednim **prelazom** pošavši od q)
- elementom $s_M \in Stanja_M$ (**početno stanje**)
- podskupom $Prihvata_M \subseteq Stanja_M$ (**stanja prihvatanja**)

Primer: NFA

Ulazni alfabet: $\{a, b\}$



Jezik koji ovaj automat prihvata isti je kao i jezik automata iz prethodnog primera, naime

$$\{u \in \{a, b\}^* \mid u \text{ sadrži tri uzastopna } a\}$$

Deterministički konačni automat

Deterministički konačni automat (DFA) definiše se kao NFA M sa svojstvom da za svako $q \in Stanja_M$ i $a \in \Sigma_M$, konačni skup $\Delta_M(q, a)$ sarži tačno jedan element, koji zovemo $\delta_M(q, a)$.

Zato su prelazi u M u ovom slučaju definisani **funkcijom sledećeg stanja**, δ_M, koja preslikava svaki (stanje, simbol)-par (q, a) u jedinstveno stanje $\delta_M(q, a)$ do kojeg se od q dolazi prelazom obeleženim sa a:

$$q \xrightarrow{a} q' \quad \text{akko} \quad q' = \delta_M(q, a)$$

Nedeterministički konačni automat sa ε-prelazima

Nederministički konačni automat sa ε-prelazima (NFA$^{\varepsilon}$) definiše se kao NFA M, sa dodatnom binarnom relacijom.

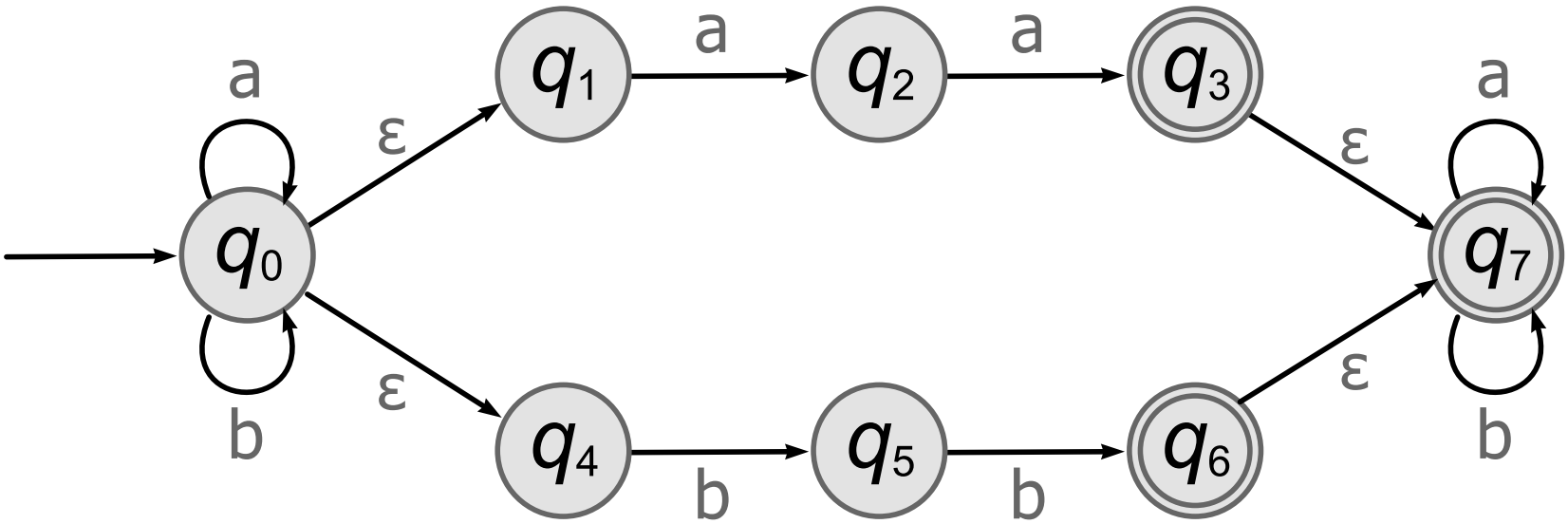
Ta relacija zove se ε-**prelaz**, definisana je na skupu $Stanja_M$ i pišemo

$$q \xrightarrow{\varepsilon} q'$$

da bismo naznačili da je par stanja (q, q') u ε-relaciji.

Primer: NFA$^{\varepsilon}$

Ulazni alfabet: $\{a, b\}$

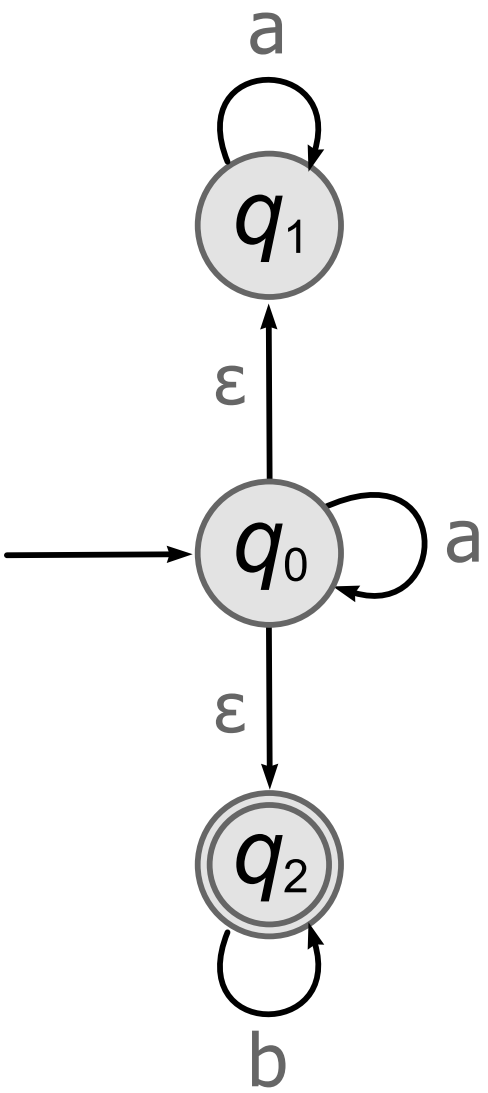


Jezik koji prihvata NFA$^{\varepsilon}$

$L(M)$, **jezik koji prihvata NFA$^{\varepsilon}$** M sastoji se od svih stringova u nad alfabetom Σ_M ulaznih simbola koji zadovoljavaju $q_0 \overset{u}{\Rightarrow} q$ gde je q_0 početno, a q neko stanje prihvtanja. Ovde definišemo operaciju $\cdot \overset{-}{\Rightarrow} \cdot$ na sledeći način:

- $q \overset{\varepsilon}{\Rightarrow} q'$ akko $q = q'$, odnosno postoji sekvenca $q \overset{\varepsilon}{\rightarrow} \ldots q'$ od jednog ili više ε-prelaza u M od q do q'
- $q \overset{a}{\Rightarrow} q'$ (za $a \in \Sigma_M$) akko $q \overset{\varepsilon}{\Rightarrow} \cdot \overset{a}{\rightarrow} \cdot \overset{\varepsilon}{\Rightarrow} q'$
- $q \overset{ab}{\Rightarrow} q'$ (za $a, b \in \Sigma_M$) akko $q \overset{\varepsilon}{\Rightarrow} \cdot \overset{a}{\rightarrow} \cdot \overset{\varepsilon}{\Rightarrow} \cdot \overset{b}{\rightarrow} \cdot \overset{\varepsilon}{\Rightarrow} q'$

Primer podskup konstrukcije



δ_{PM}	a	b
∅	∅	∅
$\{q_0\}$	$\{q_0, q_1, q_2\}$	$\{q_2\}$
$\{q_1\}$	$\{q_1\}$	∅
$\{q_2\}$	∅	$\{q_2\}$
$\{q_0, q_1\}$	$\{q_0, q_1, q_2\}$	$\{q_2\}$
$\{q_0, q_2\}$	$\{q_0, q_1, q_2\}$	$\{q_2\}$
$\{q_1, q_2\}$	$\{q_1\}$	$\{q_2\}$
$\{q_0, q_1, q_2\}$	$\{q_0, q_1, q_2\}$	$\{q_2\}$

Bitna teorema

Teorema. Za svaku NFA$^{\varepsilon}$ M postoji DFA PM sa istim alfabetom, skupom ulaznih simbola i koja prihvata sve iste stringove kao i M, odnosno važi $L(PM) = L(M)$.

Dakle, na osnovu ove teoreme PM je definisana sa:

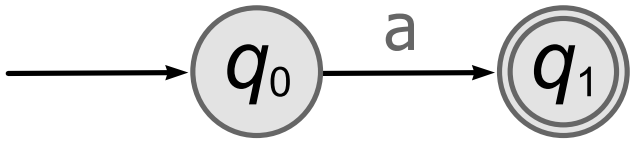
- $Stanja_{PM} \stackrel{\text{def}}{=} \{S \mid S \subseteq Stanja_M\}$
- $\Sigma_{PM} \stackrel{\text{def}}{=} \Sigma_M$
- $S \stackrel{a}{\to} S'$ je u PM akko $S' = \delta_{PM}(S, a)$, gde je $\delta_{PM}(S, a) \stackrel{\text{def}}{=} \{q' \mid (\exists q \in S)(q \stackrel{a}{\Rightarrow} q' \text{ u } M)\}$
- $s_{PM} \stackrel{\text{def}}{=} \{q \mid s_M \stackrel{\varepsilon}{\Rightarrow} q\}$
- $Prihvata_M \stackrel{\text{def}}{=} \{S \in Stanja_{PM} | (\exists q \in S)(q \in Prihvata_M)\}$

Klejnieva teorema

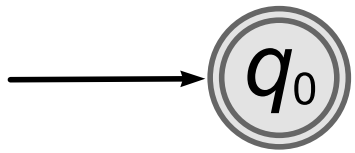
Definicija. Jezik je **regularan** akko ukoliko neki deterministički konačni automat prihvata sve njegove stringove.

Klejnieva teorema
(a) Za svaki regularni izraz r, $L(r)$ je regularni jezik.
(b) Obrnuto, svaki regularni jezik je forme $L(r)$, za neki regularni izraz r.

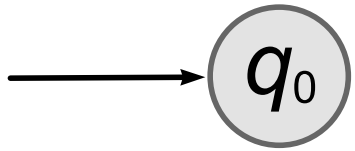
NFAovi atomskih izraza



prihvata samo jedno-simbolni string a

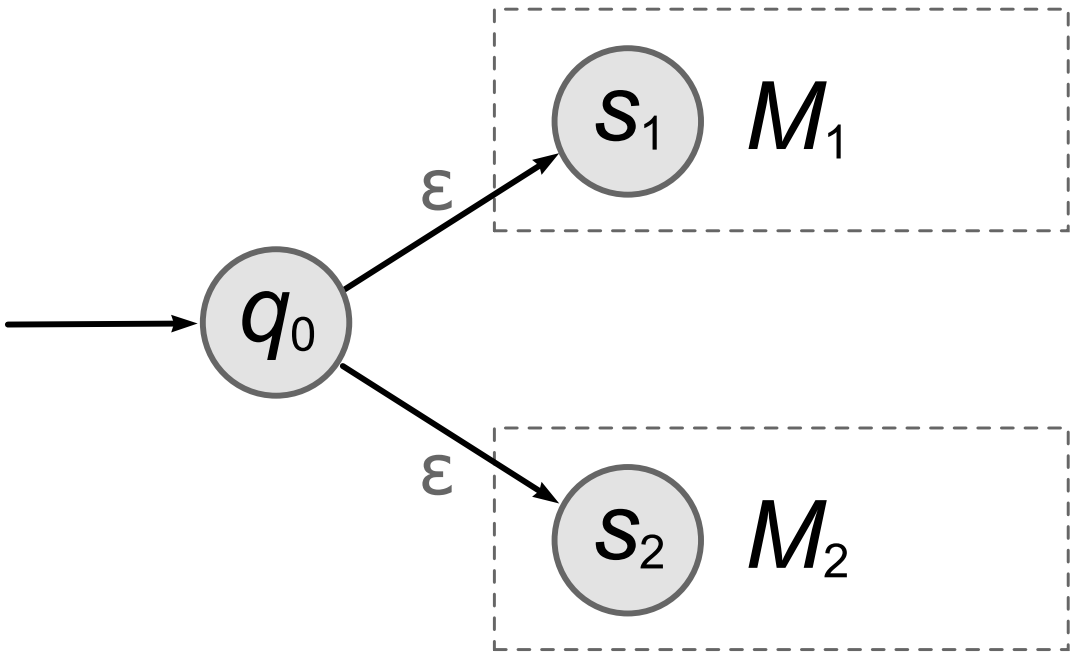


prihvata samo prazan string, ε



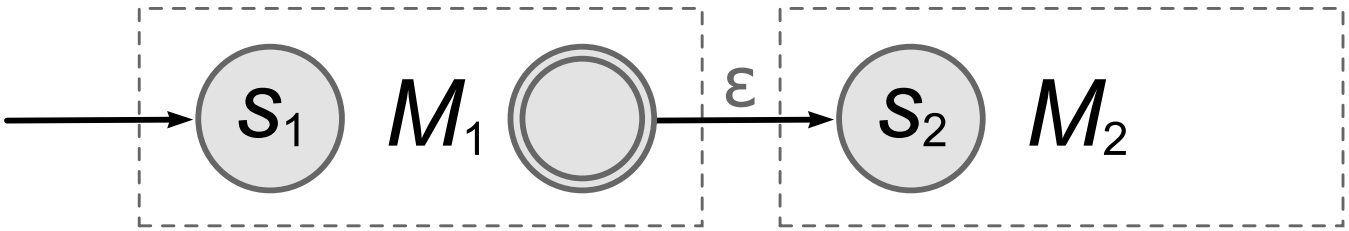
ne prihvata ni jedan string

$Union(M_1, M_2)$



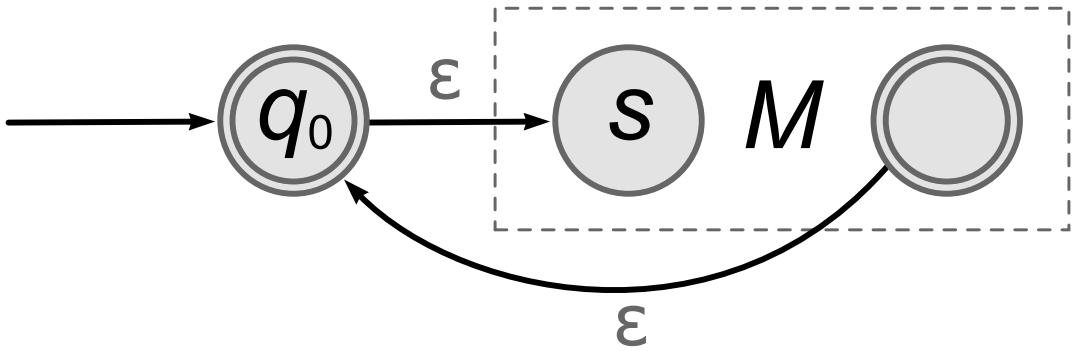
Skup stanja prihvatanja je unija $Prihvata_{M_1}$ i $Prihvata_{M_2}$.

$Concat(M_1, M_2)$



Skup stanja prihvatanja je unija $Prihvata_{M_2}$.

$Star(M)$

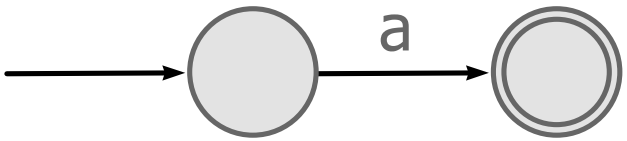


Skup stanja prihvatanja je unija $Prihvata_{M_1}$ i $Prihvata_{M_2}$.

Primer: $(a|b)^*a$

korak: 1

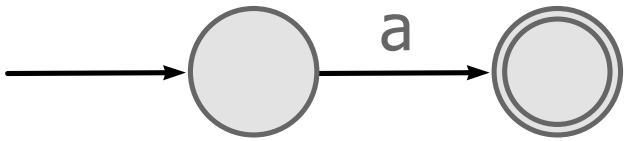
pravimo NFA$^\varepsilon$ za izraz: a

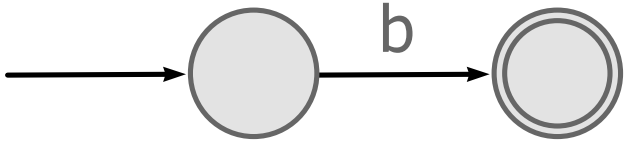


Primer: $(a|b)^*a$

korak: 2

pravimo NFA$^\varepsilon$ za izraz: a i b (odvojeno)

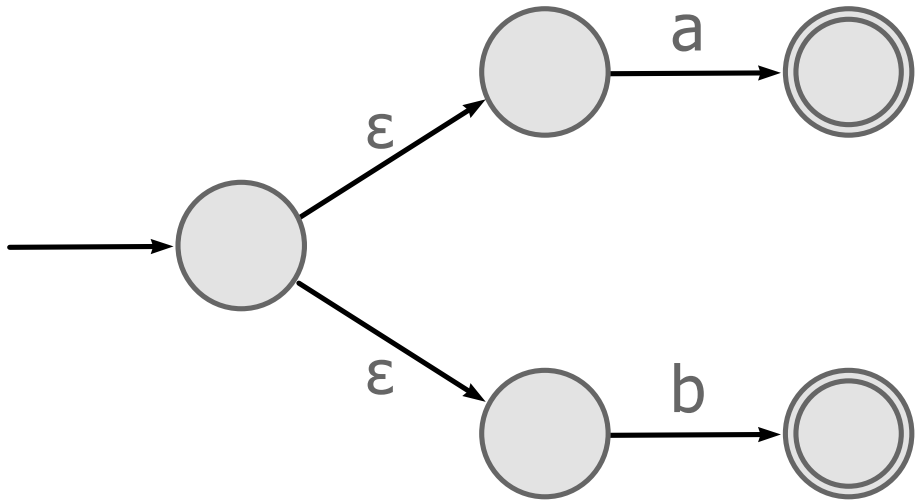




Primer: $(a|b)^*a$

korak: 3

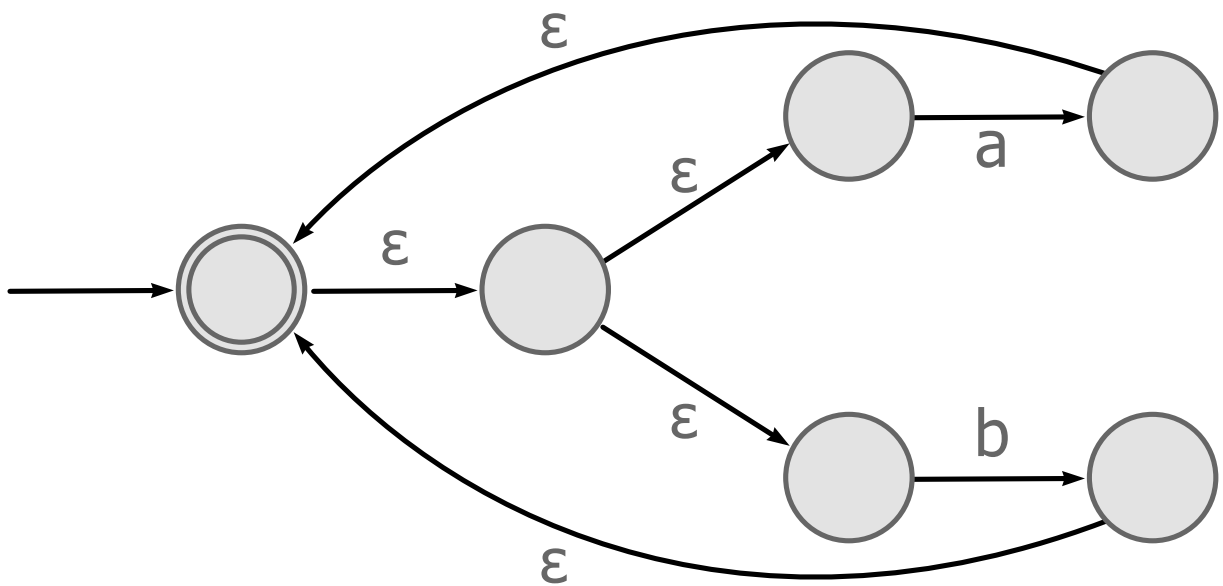
pravimo NFA$^\varepsilon$ za izraz: $a|b$



Primer: $(a|b)^*a$

korak: 4

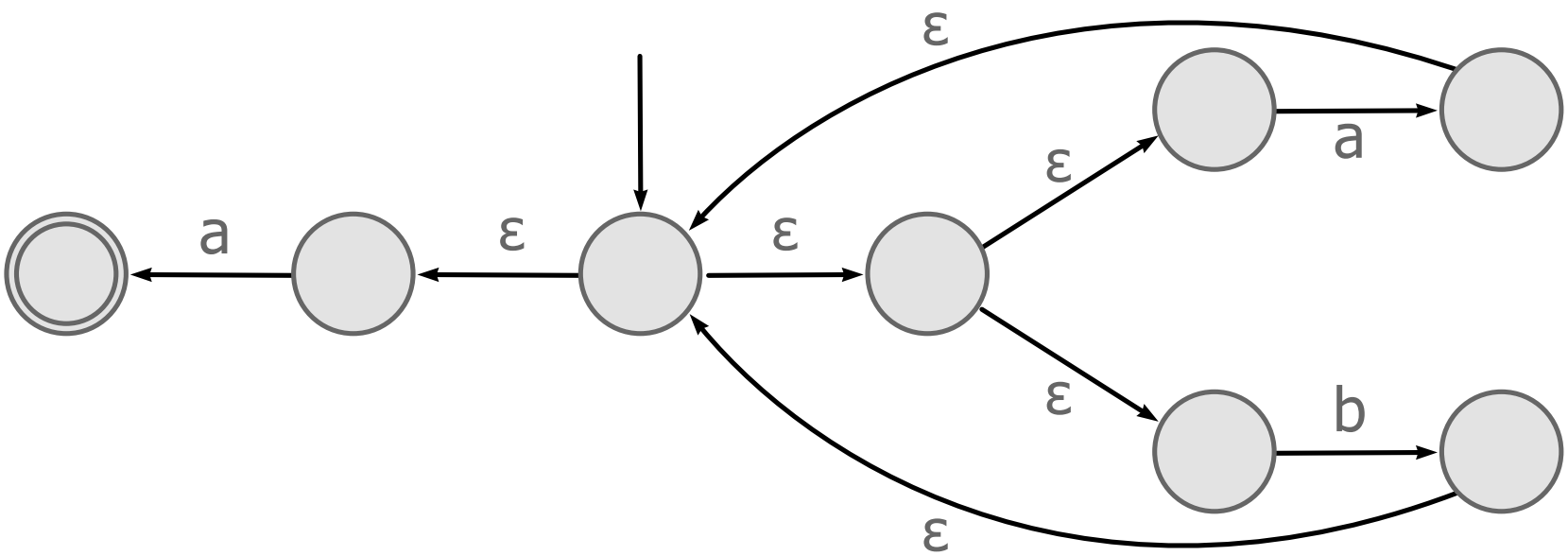
pravimo NFA$^{\varepsilon}$ za izraz: $(a|b)^*$



Primer: $(a|b)^*a$

korak: 5

pravimo NFA$^{\varepsilon}$ za izraz: $(a|b)^*a$



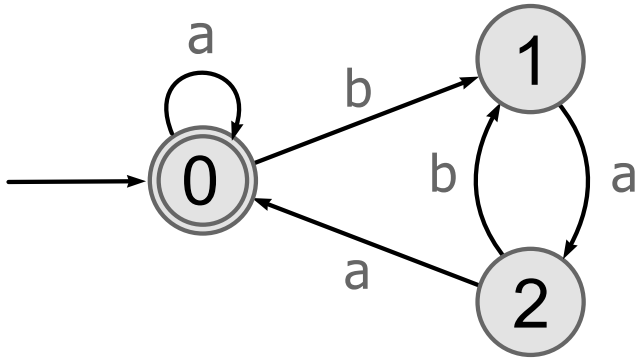
Bitna lema

Lema. Za datu NFA M, za svaki podskup $Q \subseteq Stanja_M$ i svaki par stanja $q, q' \in Stanja_M$, postoji regularni izraz $r^Q_{q,q'}$ koji zadovoljava

$$L(r^Q_{q,q'}) = \{u \in (\Sigma_M)^* \mid q \xrightarrow{u}_* q' \text{ je u } M \text{ sa svim među-stanjima sekvence prelaza u } Q\} \quad (1)$$

Stoga, $L(M) = L(r)$, gde je $r = r_1|\ldots|r_k$ i
$k =$ broj svih stanja prihvatanja,
$r_i = r^Q_{s,q_i}$, gde je $Q = Stanja_M$,
$s =$ početno stanje,
$q_i = i$to stanje prihvatanja.
(U slučaju $k = 0$, uzimamo da je r regularni izraz $\varnothing$.)

Primer



Direktnom analizom automata (M) dobijamo:

$r_{i,j}^{\{0\}}$	0	1	2
0			
1	Ø	ε	a
2	aa*	a*b	ε

$r_{i,j}^{\{0,2\}}$	0	1	2
0	a*	a*b	
1			
2			

(nepopunjena polja matrica nam nisu potrebna)

Objašnjenje primera

Pošto počinjemo iz stanja 0, i ono je jedino stanje prihvatanja, jezik svih stringova koje automat prihvata određen je regularnim izrazom $r_{0,0}^{\{0,1,2\}}$ Ukoliko odlučimo da uklonimo stanje 1 iz skupa stanja, imamo

$$L(r_{0,0}^{\{0,1,2\}}) = L(r_{0,0}^{\{0,2\}} | r_{0,1}^{\{0,2\}} (r_{1,1}^{\{0,2\}})^* r_{1,0}^{\{0,2\}}). \tag{2}$$

Direktna analiza pokazuje da je $L(r_{0,0}^{\{0,2\}}) = L(a^*)$ i $L(r_{0,1}^{\{0,2\}}) = L(a^*b)$. Kako bismo izračunali $L(r_{1,1}^{\{0,2\}})$ kao i $L(r_{1,0}^{\{0,2\}})$, biramo da uklonimo satnje 2 iz razmatranja:

$$L(r_{1,1}^{\{0,2\}}) = L(r_{1,1}^{\{0\}} | r_{1,2}^{\{0\}} (r_{2,2}^{\{0\}})^* r_{2,1}^{\{0\}}) \tag{3}$$

$$L(r_{1,0}^{\{0,2\}}) = L(r_{1,0}^{\{0\}} | r_{1,2}^{\{0\}} (r_{2,2}^{\{0\}})^* r_{2,0}^{\{0\}}) \tag{4}$$

Objašnjenje primera, cont'd

Poslednja dva regularna izraza (navedena na prethodnom slajdu) mogu se odrediti sa slike. Stoga, $L(r_{1,1}^{\{0,2\}}) = L(\varepsilon|a(\varepsilon)^*(a^*b))$, što je isto što i $L(\varepsilon|aa^*b)$. Slično, $L(r_{1,0}^{\{0,2\}}) = L(\varnothing|a(\varepsilon)^*(aa^*))$, što je jednako sa $L(aaa^*)$. Zamenom ovih vrednosti u jednačinu (1) dobijamo:

$$L(r_{0,0}^{\{0,1,2\}}) = L(a^* \mid a^*b\ (\varepsilon|aa^*b)^*\ aaa^*) \tag{5}$$

Dakle, $a^*|a^*b(\varepsilon|aa^*b)^*aaa^*$ je regularni izraz koji odgovara automatu iz primera. (Jasno je da bi se ovaj regularni izraz mogao dalje uprostiti, ali nećemo se sada time baviti.)

Dodatak: $Not(M)$

Za datu **DFA** M, $Not(M)$ je DFA sa svojstvima:

- $Stanja_{Not(M)} \stackrel{\text{def}}{=} Stanja_M$
- $\Sigma_{Not(M)} \stackrel{\text{def}}{=} \Sigma_M$
- prelazi u $Not(M)$ = prelazi u M
- početno stanje od $Not(M)$ = početno stanje od M
- $Prihvata_{Not(M)} = \{q \in Stanja_M \mid q \notin Prihvata_M\}$

Tada za svako $u \in L(Not(M))$ važi:

$$L(Not(M)) = \{u \in \Sigma^* \mid u \notin L(M)\}.$$

Dodatak: Ekvivalentni regularni izrazi

Definicija. Za dva regularna izraza r i s kažemo da su ekvivalentni ako $L(r) = L(s)$, tačnije, oni određuju tačno iste skupove stringova prepoznavanjem šablona (matching-om).

Dodatak: Ekvivalentni regularni izrazi, cont'd

Primetimo da su sledeći izrazi ekvivalentni:

$$
\begin{aligned}
& L(r) \subseteq L(s) \wedge L(s) \subseteq L(r) && (6)\\
\iff & (\Sigma^* \setminus L(r)) \cap L(s) = \varnothing = (\Sigma^* \setminus L(s)) \cap L(r) && (7)\\
\iff & L((\sim r)\&s) = \varnothing = L((\sim s)\&r) && (8)\\
\iff & L(M) = \varnothing = L(N) && (9)
\end{aligned}
$$

gde su M i N DFA koji prihvataju skupove stringova koji se uparuju sa regularnim izrazima $(\sim r)\&s$ i $(\sim s)\&r$, respektivno.

Dakle, određivanje da li su dva regularna izraza ekvivalentna, svodi se na proveru da li ovako definisanu automati M i N prihvataju neki string. Ovu proveru ponovljamo konačan broj puta, jer automati imaju konačno mnogo stanja, i jer možemo ukloniti petlje sa dužih puteva.