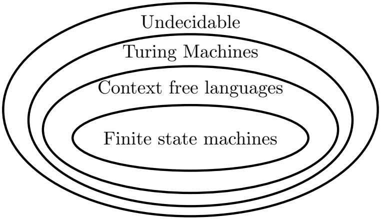
Μια ιεραρχία από κλάσεις γλωσσών



- Γλώσσα είναι ένα σύνολο κωδικοποιήσεων στιγμιοτύπων προβλημάτων (πρόβλημα απόφασης $\leftrightarrow x \in L$).
- Κάθε κλάση γλωσσών χαρακτηρίζεται από τις δυνατότητες του υπολογιστικού μοντέλου που αποδέχεται τα σύνολα της κλάσης.
- Οι δυνατότητες του υπολογιστικού μοντέλου αυξάνουν όσο διευρύνονται οι κλάσεις των γλωσσών.

Ιδιότητες κανονικών συνόλων

Υπάρχουν διάφορες ερωτήσεις σχετικά με κανονικά σύνολα:

1. Δίνεται μια γλώσσα L. Είναι η L κανονικό σύνολο;
2. Αν L είναι κανονική γλώσσα ισχύει $x \in L$;
3. Δίνονται δύο κανονικά σύνολα A και B. Είναι $A = B$;
4. Έστω L κανονική γλώσσα, είναι η L άπειρη·
5. Αν A, B είναι κανονικές γλώσσες ισχύει $A \subseteq B$;
6. $\cdots$

Προσπαθήστε να δώσετε διαισθητικές απαντήσεις ή να χρησιμοποιήσετε τις κλειστότητες των κανονικών συνόλων.

Υποδείξεις για τα προηγούμενα ερωτήματα

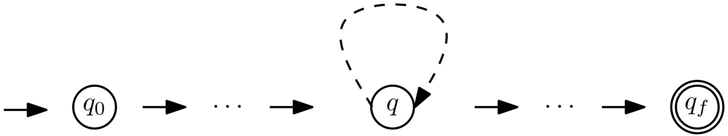
1. Αν μπορείτε δώστε ένα FSM για την L.
2. Γράψτε ένα πρόγραμμα!
3. Ισχύει $A \setminus B = \emptyset$; (σημειώστε ότι $A \setminus B = A \cap \overline{B}$)
4. Υπάρχει κύκλος στο FSM που αναγνωρίζει την L; Τι άλλο πρέπει να υπάρχει;
5. Ισχύει $\overline{A} \cup B = \Sigma^*$;

Μπορούμε να απαντήσουμε σχεδόν όλα τα ενδιαφέροντα ερωτήματα για τα κανονικά σύνολα.

Στη συνέχεια θα παρουσιάσουμε μια τεχνική που μας βοηθάει στο να δείχνουμε ότι μια γλώσσα δεν είναι κανονική.

Ιδιότητα κανονικού συνόλου

Αν μια γλώσσα είναι κανονική τότε γίνεται δεκτή από κάποιο DFA $M = \{Q, \Sigma, \delta, q_0, F\}$ μέ συγκεκριμένο αριθμό καταστάσεων: $|Q| = n$.



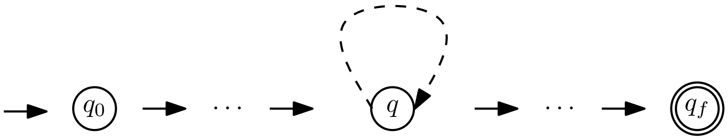
Κάθε στοιχείο $z \in L$ που γίνεται δεκτό από το M με $|z| \geq n$ υποχρεώνει το M να επαναλάβει κάποια κατάσταση.

Αρχή του περιστερώνα

Υπάρχουν $|z|$ περιστέρια που πρέπει να μπούν σε n φωλιές. Αφού $|z| \geq n$ τουλάχιστο μια φωλιά θα έχει πάνω από ένα περιστέρι.

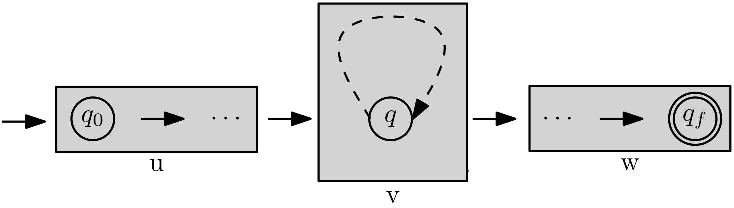
Λήμμα άντλησης για κανονικές γλώσσες

Αν $z \in L$ και η L είναι κανονική, τότε το z μπορεί να μεγαλώσει με την «άντληση» συγκεκριμένων υποσυμβολοσειρών του και να συνεχίσει να ανήκει στην L.



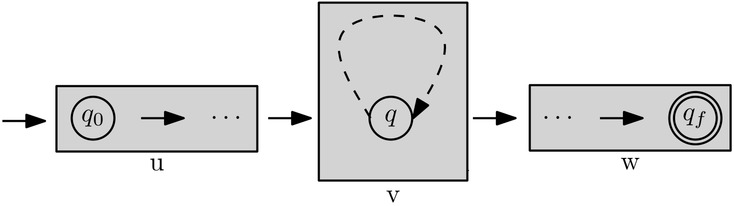
Λήμμα άντλησης για κανονικές γλώσσες

Αν $z \in L$ και η L είναι κανονική, τότε το z μπορεί να μεγαλώσει με την «άντληση» συγκεκριμένων υποσυμβολοσειρών του και να συνεχίσει να ανήκει στην L.



Λήμμα άντλησης για κανονικές γλώσσες

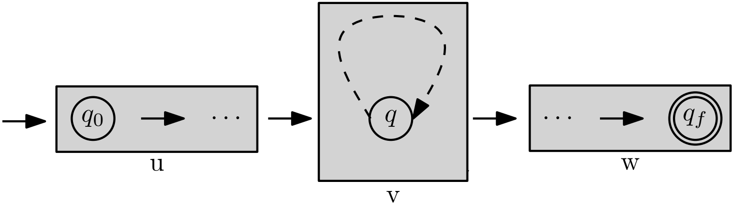
Αν $z \in L$ και η L είναι κανονική, τότε το z μπορεί να μεγαλώσει με την «άντληση» συγκεκριμένων υποσυμβολοσειρών του και να συνεχίσει να ανήκει στην L.



Γίνονται αποδεκτά τα $uw, uvw, uvw, \ldots, uv^iw, \ldots$ για κάθε $i \in \mathbb{N}$.

- Για κάθε κανονική L υπάρχει ένα DFA με n καταστάσεις που την αποδέχεται.
- Κάθε $z \in L$ μπορεί να γραφτεί ως uv^iw με $|uv| \leq n$ και $|v| \geq 1$. Το uv^iw ανήκει στην L για κάθε $i \geq 0$.

Λήμμα άντλησης για κανονικές γλώσσες



- Η ιδιότητα της άντλησης είναι αναγκαία για ένα κανονικό σύνολο δεν είναι όμως ικανή συνθήκη.
- Ισχύει $|uv| \leq n$ και $|v| \geq 1$, δηλαδή κατα την αναγνώριση του z επαναλαμβάνεται μια κατάσταση πριν καταναλωθούν n σύμβολα.
- Μια κατάσταση θα επαναληφθεί κατά την κατανάλωση του uv. Κατά την κατανάλωση του w μπορεί να επαναλαμβάνονται κι άλλες καταστάσεις.

Λήμμα άντλησης για κανονικές γλώσσες

Τυπική διατύπωση

Pumping Lemma

Αν η L είναι κανονική γλώσσα,
$\exists n$ ίσο με τον αριθμό των καταστάσεων ενός DFA για την L,
τέτοιο ώστε $\forall z \in L$, με $|z| \geq n$,
$\exists u, v, w$ τέτοια ώστε $z = uvw$, $|uv| \leq n$, $|v| \geq 1$,
έτσι ώστε $\forall i \geq 0$ το $uv^i w$ ανήκει επίσης στην L.

Ισχύει δηλαδή ότι:

$$L \text{ κανονική γλώσσα} \rightarrow \text{ιδιότητα άντλησης}$$

Εμείς θα χρησιμοποιήσουμε το αντιθετοαντίστροφο:

$$\neg(\text{ιδιότητα άντλησης}) \rightarrow \neg(L \text{ κανονική γλώσσα})$$

Άρνηση της ιδιότητας άντλησης

Αν για μια γλώσσα L,
$\forall n$ ίσο με τον αριθμό των καταστάσεων οποιουδήποτε DFA για την L,
$\exists z \in L$ με $|z| \geq n$ τέτοιο ώστε
$\forall u, v, w$ με $z = uvw$ και $|uv| \leq n$, $|w| \geq 1$
$\exists i \geq 0$ τέτοιο ώστε το uv^iw δεν ανήκει στην L,
τότε η L δεν είναι κανονική γλώσσα.

- Για να αποδείξουμε ότι μια L δεν είναι κανονική παίζουμε ένα παιχνίδι με ένα αντίπαλο που συνεχώς θέλει να μας δυσκολεύει.
- Ο αντίπαλος επιλέγει τις οντότητες με τους καθολικούς ποσοδείκτες και εμείς επιλέγουμε τους υπαρξιακούς ποσοδείκτες.
- Ο αντίπαλος «δεν χαρίζεται». Εμείς πρέπει να κάνουμε τέτοιες επιλογές έτσι ώστε ο αντίπαλος να υποχρεωθεί σε ήττα.

Εφαρμογή της άρνησης της ιδιότητας άντλησης

Μια περίπτωση λανθασμένης τακτικής στο παιχνίδι

Το παρακάτω παιχνίδι έχει σκοπό να αποδείξει ότι η $L = \{w | w \text{ είναι παλίνδρομο}\}$ δεν είναι κανονική.

- Ο αντίπαλος ισχυρίζεται ότι η L είναι κανονική: αναγνωρίζεται από ένα DFA με n καταστάσεις.
- Εμείς επιλέγουμε $z = 0^{\lceil n/2 \rceil} 1 0^{\lceil n/2 \rceil}$ (ισχύει $z \in L$ και $|z| \geq n$).
- Ο αντίπαλος θέτει $u = 0^{\lceil n/2 \rceil}$, $v = 1$ και $w = 0^{\lceil n/2 \rceil}$ (ισχύει $|uv| \leq n$ και $|v| \geq 1$).
- Έχουμε χάσει το παιχνίδι γιατί κάθε επιλογή του i κάνει το $uv^i w$ να είναι παλίνδρομο (ανήκει δηλαδή στην L).

Χάσαμε γιατί δεν επιλέξαμε καλά το z! Ο αντίπαλος μας εξανάγκασε να «αντλήσουμε» 1 (το z παρέμεινε παλίνδρομο)

Εφαρμογή της άρνησης της ιδιότητας άντλησης

Σωστή τακτική στο παιχνίδι

Το παρακάτω παιχνίδι δείχνει ότι η $L = \{w | w \text{ είναι παλίνδρομο}\}$ δεν είναι κανονική:

- Ο αντίπαλος ισχυρίζεται ότι η L είναι κανονική: αναγνωρίζεται από ένα DFA με n καταστάσεις.
- Εμείς επιλέγουμε $z = 0^n 1 0^n$ (ισχύει $z \in L$ και $|z| \geq n$).
- Ο αντίπαλος θέτει $0^n 1 0^n = uvw$ έτσι ώστε $|uv| \leq n$ και $|v| \geq 1$).
- Εμείς παρατηρούμε ότι πρέπει να ισχύει $v = 0^k$ με $0 \leq k \leq n$. Επιλέγουμε $i = 2$, τότε $uv^2 w = uvvw = 0^n 0^k 1 0^n = 0^{n+k} 1 0^n$. Έχουμε κερδίσει γιατί το $0^{n+k} 1 0^n$ δεν είναι παλίνδρομο.

Δείξαμε ότι για τη γλώσσα L ισχύει η άρνηση της ιδιότητας άντλησης, συνεπώς η γλώσσα L δεν είναι κανονική.

Εφαρμογή της άρνησης της ιδιότητας άντλησης

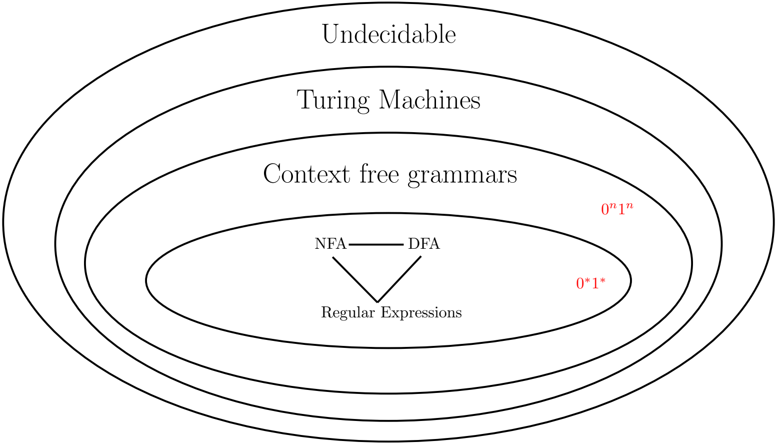
Το παρακάτω παιχνίδι δείχνει ότι η $L = \{w | w = 0^{k^2}\}$ δεν είναι κανονική:

- Ο αντίπαλος ισχυρίζεται ότι ένα DFA n καταστάσεων αποδέχεται την L.
- Εμείς επιλέγουμε $z = 0^{n^2}$ (ισχύει $z \in L$ και $|z| \geq n$).
- Ο αντίπαλος θέτει $0^{n^2} = uvw$ έτσι ώστε $|uv| \leq n$ και $|v| \geq 1$).
- Εμείς παρατηρούμε ότι πρέπει να ισχύει $v = 0^k$ με $0 \leq k \leq n$. Επιλέγουμε $i = 2$, τότε $uv^2w = uvvw = 0^{n^2+k}$. Έχουμε κερδίσει γιατί το 0^{n^2+k} δεν είναι της μορφής 0^{m^2} για οποιοδήποτε m.

  $n^2 < n^2 + k \leq n^2 + n < n^2 + 2n + 1 < (n+1)^2$

Δείξαμε ότι για τη γλώσσα L ισχύει η άρνηση της ιδιότητας άντλησης, συνεπώς η γλώσσα L δεν είναι κανονική.

Μια ματιά στο σύμπαν

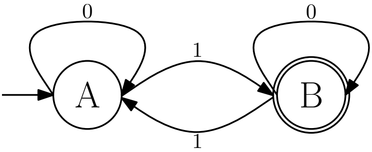


Γραμματικές

Οι γραμματικές είναι συστήματα που παράγουν γλώσσες.

- Οι FSM αποδέχονται τα στοιχεία μιας γλώσσας και απορρίπτουν τα στοιχεία που δεν ανήκουν στη γλώσσα.
- Κάθε FSM έχει μια αντίστοιχη γραμματική.
- Η γραμματική παράγει τα στοιχεία της γλώσσας που το FSM αποδέχεται ενώ δεν μπορεί να παραγάγει τα στοιχεία που το FSM απορρίπτει.
- Άτυπα μια γραμματική είναι ένα σύνολο κανόνων παραγωγής. Ένας κανόνας παραγωγής αντικαθιστά μη τερματικά σύμβολα με κάποια συμβολοσειρά.
- Σκοπός είναι να λάβουμε τα στοιχεία μια γλώσσας με ακολουθίες αντικαταστάσεων όπου εξαλείφονται τα μη τερματικά σύμβολα.

Κατασκευή γραμμικής γραμματικής από FSM



$A \longrightarrow 0A$

$A \longrightarrow 1B$

$B \longrightarrow 0B$

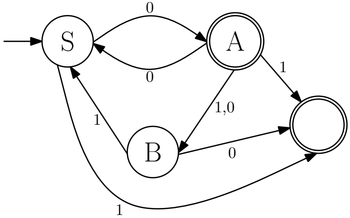
$B \longrightarrow 1A$

$B \longrightarrow \epsilon$

- Παραγωγή του 010: $A \longrightarrow 0A \longrightarrow 01B \longrightarrow 010B \longrightarrow 010$
- Σε ένα κανόνα παραγωγής μιας γραμμικής γραμματικής έχουμε αριστερά και δεξιά του βέλους ένα μόνο μη τερματικό σύμβολο.
- Οι αριστερογραμμικές γραμματικές έχουν κανόνες παραγωγής της μορφής $X \longrightarrow x$ ή $X \longrightarrow xX$, ενώ οι δεξιογραμμικές παραγωγές είναι της μορφής $X \longrightarrow x$ ή $X \longrightarrow Xx$

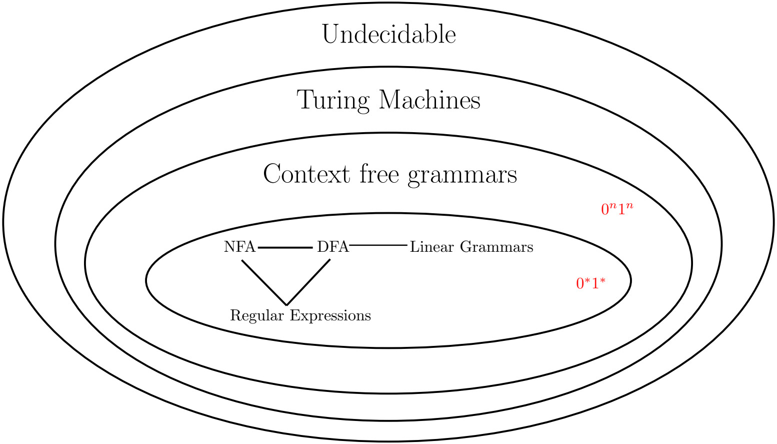
Κατασκευή FSM από γραμμική γραμματική

$$S \longrightarrow 0A$$
$$A \longrightarrow 1B$$
$$A \longrightarrow 0B$$
$$A \longrightarrow 0S$$
$$B \longrightarrow 1S$$
$$A \longrightarrow 1$$
$$B \longrightarrow 0$$
$$S \longrightarrow 1$$
$$A \longrightarrow \epsilon$$



- Στη γενική περίπτωση λαμβάνω ένα NFA.
- Οι γραμματικές είναι (σχεδόν) εξ' ορισμού μη ντετερμινιστική έννοια: πρέπει να μαντέψουν αν θα παραχθεί μη τερματικό ή τερματικό σύμβολο.

Μια ματιά στο σύμπαν



Γραμματικές χωρίς συμφραζόμενα

Context Free Grammars (CFG)

Οι γραμματικές χωρίς συμφραζόμενα έχουν κανόνες παραγωγής της μορφής $X \longrightarrow$ anything. Έχουν δηλαδή αριστερά ένα μη τερματικό σύμβολο ενώ δεξιά δεν υπάρχει κάποιος περιορισμός.

Θα χρησιμοποιήσουμε τη συντομογραφία $A \longrightarrow x \mid y$ που σημαίνει ότι το A παράγει x ή y (x, y γενικά συμβολοσειρές).

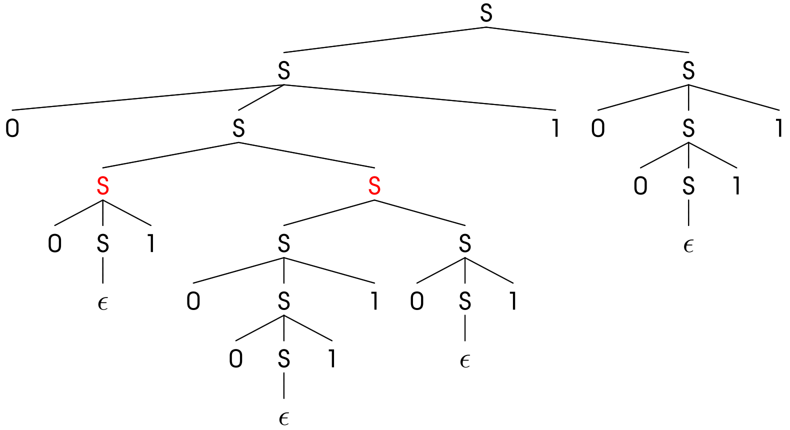
Έστω η γραμματική $S \longrightarrow 0S1 \mid \epsilon$

- Μια παραγωγή μπορεί να είναι:
  $S \longrightarrow 0S1 \longrightarrow 00S11 \longrightarrow 000S111 \longrightarrow 000111$.
- Η γραμματική παράγει στοιχεία της γλώσσας $L = \{0^n1^n,\ n \geq 0\}$.
- Δεν υπάρχει FSM που να αναγνωρίζει την L!

Συντακτικό δέντρο

Γραμματική: $S \longrightarrow 0S1 \mid SS \mid \epsilon$

Παράγει το 00100110110011;



Διφορούμενες γραμματικές

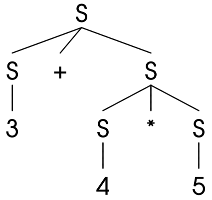
- Αν υπάρχουν δύο ή περισσότερα δέντρα για την παραγωγή της ίδιας συμβολοσειράς τότε η γραμματική ονομάζεται διφορούμενη.
- Η δεξιότερη παραγωγή σχηματίζεται όταν αντικαθιστούμε πάντα το δεξιότερο μη τερματικό σύμβολο: μεταδιατεταγμένη επίσκεψη στο συντακτικό δέντρο.
- Η αριστερότερη παραγωγή σχηματίζεται με την προδιατεταγμένη επίσκεψη στο δέντρο.
- Οι διφορούμενες γραμματικές είναι κακή ιδιότητα για μια γλώσσα: Δεν είναι δυνατό να υπάρξει μεταγλωτιστής για μια διφορούμενη γλώσσα προγραμματισμού.
- $L = \{a^n b^n c^n d^n,\ n > 0\}$ είναι γλώσσα εγγενώς διφορούμενη: όλες οι CFG για την L είναι διφορούμενες.
- Προσοχή στη σημασιολογία!

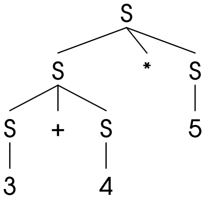
Διφορούμενες γραμματικές

Όταν υπάρχει διαφορά στη σημασιολογία

Γραμματική: $S \longrightarrow S + S \mid S * S \mid 0 \mid 1 \mid 2 \mid \cdots \mid 9$

Υπάρχουν δύο παραγωγές του $3 + 4 * 5$:





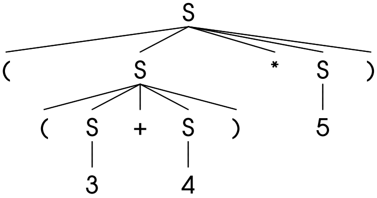
Η γραμματική είναι διφορούμενη, δεν υπάρχει η έννοια της προτεραιότητας κι έτσι χάνουμε στη σημασιολογία.

Τι θα έκανε σε αυτή την περίπτωση ένας μεταγλωτιστής;

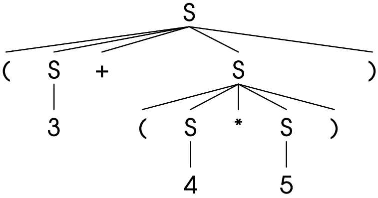
Μη διφορούμενη γραμματική

Γραμματική: $S \longrightarrow (S + S) \mid (S * S) \mid 0 \mid 1 \mid 2 \mid \cdots \mid 9$

Παραγωγή του $((3 + 4) * 5)$:



Παραγωγή του $(3 + (4 * 5))$:



Για κάθε συμβολοσειρά της γλώσσας υπάρχει μοναδικό συντακτικό δέντρο.

Τεχνικές σχεδιασμού γραμματικών

Αναδρομή

Παράδειγμα: η γραμματική $S \longrightarrow 0S1 \mid \epsilon$ για τη γλώσσα $L = \{0^n 1^n,\ n \geq 0\}$

Σημασιολογική ερμηνεία των μη τερματικών συμβόλων

Παράδειγμα: η γραμματική για τη γλώσσα $L = \{w \mid w \text{ έχει ίσο αριθμό 0 και 1}\}$:

$S \longrightarrow 0A \mid 1B \mid \epsilon$	S	:	ίσος αριθμός 0 και 1
$A \longrightarrow 1S \mid 0AA$	A	:	χρωστάω ένα 1
$B \longrightarrow 0S \mid 1BB$	B	:	χρωστάω ένα 0

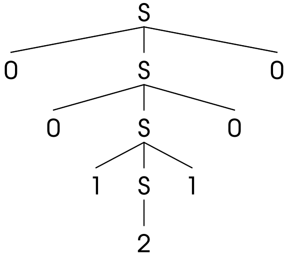
Παράδειγμα

$L = \{w2w^R \mid w \in (0 \cup 1)^*\}$

Γραμματική: $S \longrightarrow 0S0 \mid 1S1 \mid 2$

Παραγωγή 0012100: $S \longrightarrow 0S0 \longrightarrow 00S00 \longrightarrow 001S100 \longrightarrow 0012100$

Συντακτικό δέντρο:



Παράδειγμα

Παλίνδρομες συμβολοσειρές περιττού και άρτιου μήκους.

Περιττό μήκος

Γλώσσα:

$L = \{w2w^R \mid w \in (0 \cup 1)^*\}$

Γραμματική: $S \longrightarrow 0S0 \mid 1S1 \mid 2$

Ντετερμινιστικές παραγωγές.

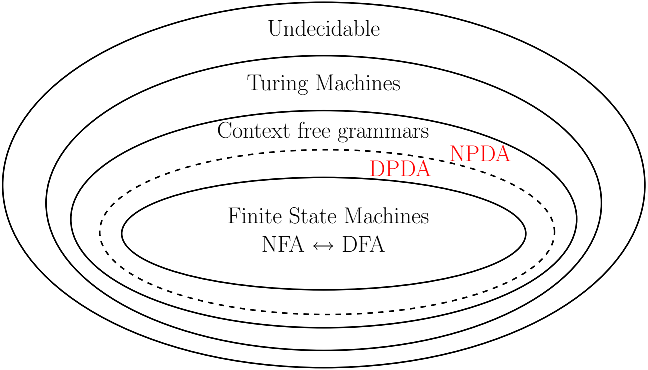
Άρτιο μήκος

Γλώσσα:

$L = \{ww^R \mid w \in (0 \cup 1)^*\}$

Γραμματική: $S \longrightarrow 0S0 \mid 1S1 \mid \epsilon$

Μη ντετερμινιστικές παραγωγές.



Αυτόματα Στοίβας

Push down automata (PDA)

- Πρόκειται για «αναβάθμιση» των FSM: τα PDA έχουν έξτρα μνήμη με την μορφή στοίβας αλλά περιορίζονται στην πρόσβαση σε αυτή.
- Πρόσβαση στη στοίβα υπάρχει μόνο στην κορυφή της:

  push Τοποθετεί στην κορυφή ένα στοιχείο που δίνεται.

  pop Αφαιρεί από την κορυφή ένα στοιχείο για χρήση.
- Θα τα σχεδιάσουμε σαν τα FSM, όμως οι επικέτες στα βέλη θα είναι της μορφής a, b, c, όπου:

  a είναι το τρέχον σύμβολο που διαβάζει το PDA από την είσοδο,

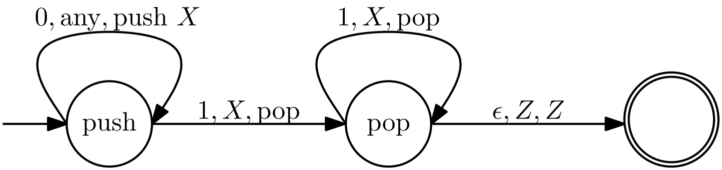
  b είναι το σύμβολο στην κορυφή της στοίβας και

  c συμβολίζει το χειρισμό της στοίβας (όπως push X).
- Τα PDA χρησιμοποιούν επιπλέον των FSM ένα αλφάβητο στοίβας και η σχέση μετάβασης είναι περισσότερο πολύπλοκη.

PDA για $L = \{0^n 1^n,\ n > 0\}$

Διαισθητική περιγραφή

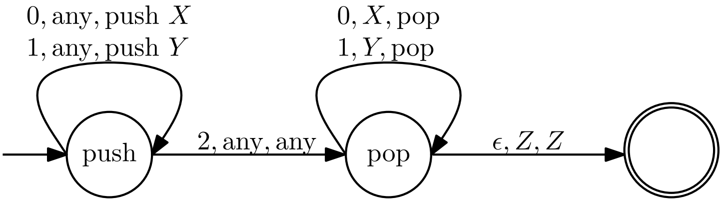
Κάθε φορά που στην είσοδο διαβάζει 0 κάνει push X στη στοίβα. Κάθε φορά που διαβάζει 1 κάνει pop από τη στοίβα. Αν η στοίβα είναι άδεια (Z) και βρίσκεται σε τελική κατάσταση αποδέχεται, αλλιώς απορίπτει.



- Χρειαζόμαστε βέλη για όλους τους συνδιασμούς a, b, c. Βέλη που δεν υπάρχουν υποτίθεται ότι οδηγούν σε καταβόθρες.
- Η ύπαρξη ϵ-κινήσεων δεν κάνει απαραίτητα το PDA μη ντετερμινιστικό.
- Γιατί απορρίπτει τα 0010 και 0001111111;

PDA για $L = \{w2w^R \mid w \in (0 \cup 1)^*\}$

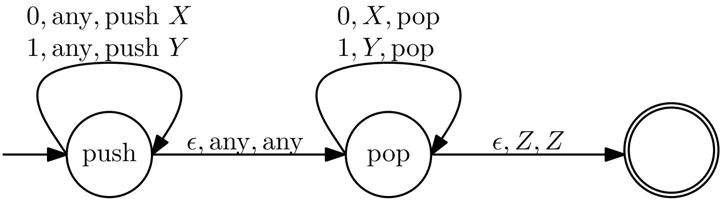
Είναι το σύνολο των παλίνδρομων περιττού μήκους όπου το 2 «μαρκάρει» το μέσο της συμβολοσειράς.



- Μαρκάρει τα 0 με X και τα 1 με Y.
- Όταν διαβάσει 2 ελέγχει αν ταιριάζει η υποσυμβολοσειρά της εισόδου μετά το 2 με αυτή που ήδη διάβασε πριν το 2.
- Είναι ντετερμινιστικό!

PDA για $L = \{ww^R \mid w \in (0 \cup 1)^*\}$

Είναι το σύνολο των παλίνδρομων αρτίου μήκους.



- Η γλώσσα δεν αναγνωρίζεται με ντετερμινιστικό PDA.
- Το PDA «μαντεύει» ότι διάβασε τα μισά σύμβολα της εισόδου και στη συνέχεια επαληθεύει την επιλογή του.
- Γιατί απορρίπτει τα $x \notin L$;