Συμβολοσειρές και Γλώσσες

Δημήτρης Φωτάκης

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων
Πανεπιστήμιο Αιγαίου

Αλφάβητα και Συμβολοσειρές

- **Αλφάβητο**: Πεπερασμένο μη-κενό σύνολο Σ.
  Μέλη ονομάζονται **σύμβολα** ή **γράμματα**.
  Π.χ. $\Sigma_2 = \{0,1\}$, $\Sigma = \{a, b, \ldots, z\}$.
- **Συμβολοσειρά** (string) w: Πεπερασμένη ακολουθία συμβόλων του Σ.
  Μήκος συμβολοσειράς w: #συμβόλων στη w.
  Κενή συμβολοσειρά e ή ε: (μοναδική) συμβολοσειρά μήκους 0.
- Σύνολο συμβολοσειρών μήκους k: Σ^k.
  Π.χ. $\{0,1\}^2 = \{00, 01, 10, 11\}$.
- Σύνολο συμβολοσειρών του Σ: Σ^*.
  Π.χ. $\{0,1\}^* = \{\varepsilon, 0, 1, 00, 01, 10, 11, 000, \ldots\}$.

Πράξεις με Συμβολοσειρές

- **Παράθεση** (concatenation) x και y: $x \circ y$ ή xy
  Π.χ. Η παράθεση των $abcd$ και $efgh$ είναι $abcdefgh$.
- x **υποσυμβολοσειρά** του w: $\exists y, z \in \Sigma^*$, $w = yxz$.
  Πρόθεμα αν $y = \varepsilon$. Κατάληξη αν $z = \varepsilon$.
- **Επανάληψη** της w για k φορές: $w^k = \overbrace{w \circ \cdots \circ w}^{k \text{ times}}$.
  Π.χ. $(ab)^3 = ababab$, $(010)^4 = 010\,010\,010\,010$.
- **Αντίστροφη** w^R της w: συμβολοσειρά που προκύπτει διαβάζοντας w από τέλος προς αρχή (αντίστροφα).
  Π.χ. $(abcd)^R = dcba$, $(10010)^R = 01001$.
- Ν.δ.ο. για κάθε $x, y \in \Sigma^*$, $(x \circ y)^R = y^R \circ x^R$.
  Υπόδειξη: μαθηματική επαγωγή. Λύση: βιβλίο, σελ. 74.

Γλώσσες

- **Γλώσσα** στο αλφάβητο Σ: οποιοδήποτε υποσύνολο του Σ^*.
  Δηλαδή, γλώσσα = (πεπερασμένο ή άπειρο) **σύνολο** συμβολοσειρών.
- Μερικές γλώσσες στο $\Sigma = \{0,1,2\}$:
  - $L_1 = \{012, 021, 120, 102, 210, 201\}$.
  - $L_2 = \{\varepsilon, 1, 11, 111, 1111, \ldots, \}$.
  - $L_3 = \{x \in \Sigma^* : \text{ το τελευταίο ψηφίο είναι } 0\}$.
  - $L_4 = \{\} = \emptyset$.
  - $L_5 = \{\varepsilon\}$.

Πράξεις με Γλώσσες

- **΄Ενωση** γλωσσών L_1 και L_2: $L_1 \cup L_2$.
- **Τομή** γλωσσών L_1 και L_2: $L_1 \cap L_2$.
- **Διαφορά** L_1 από L_2: $L_1 - L_2$ ή $L_1 \setminus L_2$
  (συμβολοσειρές που ανήκουν στη L_1 και δεν ανήκουν στη L_2).
- **Συμπλήρωμα** γλώσσας L: $\Sigma^* - L$
  (συμβολοσειρές που δεν ανήκουν στην L).
- **Παράθεση** γλωσσών L_1 και L_2: $L_1 \circ L_2$ ή L_1L_2.
  $L_1 \circ L_2 = \{w \in \Sigma^* : w = x_1 \circ x_2 \text{ για κάποια } x_1 \in L_1 \text{ και } x_2 \in L_2\}$
  Π.χ. αν $L_1 = \{0, 00, 11\}$ και $L_2 = \{\varepsilon, 1\}$, τότε
  $L_1L_2 = \{0, 00, 11, 01, 001, 111\}$.

Πράξεις με Γλώσσες

- **Kleene star** γλώσσας L: L^* αποτελείται από συμβολοσειρές που προκύπτουν από **παράθεση** οποιουδήποτε αριθμού συμβολοσειρών της L.
  $L^* = \{w \in \Sigma^* : w = w_1 \circ \cdots \circ w_k \text{ για } k \geq 0 \text{ και } w_1, \ldots, w_k \in L\}$
  - Αν $L = \{1, 00\}$, τότε
    $L^* = \{\varepsilon, 1, 00, 11, 100, 001, 0000, 111, 1100, 10000, 1001, \ldots\}$.
  - Αν $L = \{\varepsilon\}$, τότε $L^* = \{\varepsilon\}$.
  - Αν $L = \{\}$, τότε $L^* = \{\varepsilon\}$ (για κάθε L, $\varepsilon \in L^*$).
- $L^+ = \{w \in \Sigma^* : w = w_1 \circ \cdots \circ w_k \text{ για } k \geq 1 \text{ και } w_1, \ldots, w_k \in L\}$
- Πότε οι L^* και L^+ είναι διαφορετικές;
  Ποιά είναι η διαφορά τους $L^* - L^+$;

Μετρήσιμα Σύνολα

- Πόσες **συμβολοσειρές** έχει ένα (πεπερασμένο) αλφάβητο Σ;
  $|\Sigma^*| = \infty$.
- Πόσες **γλώσσες** ορίζονται στο Σ;
  $|2^{\Sigma^*}| = \infty$.
- Διαισθητικά σαφές ότι $|2^{\Sigma^*}|$ "μεγαλύτερο" του $|\Sigma^*|$.
- **Ισάριθμα** σύνολα A και B: **Αμφιμονοσήμαντη** αντιστοιχία $f : A \mapsto B$.
- **Πεπερασμένο** σύνολο A: ισάριθμο με $\{1, \ldots, |A|\}$.
- **Μετρήσιμο** σύνολο A: είτε πεπερασμένο είτε ισάριθμο του $\mathbb{N}$.
  - Σύνολο ζυγών αριθμών $\{0, 2, 4, 6, \ldots\}$ (αντιστοιχία $f(n) = n/2$).
  - Σύνολο ακεραίων αριθμών $\{0, 1, -1, 2, -2, 3, -3, \ldots\}$.
  - Σύνολο $\mathbb{N} \times \mathbb{N} = \{(0,0), (0,1), (1,0), (0,2), (1,1), (2,1) \ldots\}$.
    Αντιστοιχία $f(i,j) = \frac{1}{2}[(i+j)^2 + 3i + j]$.
  - Σύνολο **ρητών** αριθμών;

Μετρήσιμα Σύνολα

- Το Σ^* είναι **μετρήσιμο**.
- Αλφάβητο Σ πεπερασμένο (εξ΄ ορισμού). ΄Εστω διάταξη συμβόλων Σ.
  Λεξικογραφική σειρά των συμβολοσειρών του Σ^*:
  - Πρώτα η συμβολοσειρά ε μήκους 0,
  - μετά συμβολοσειρές μήκους 1 (ταξινομημένες),
  - μετά συμβολοσειρές μήκους 2 (ταξινομημένες), κοκ.
- Π.χ. αν $\Sigma = \{0, 1\}$,
  $\Sigma^* = \{\varepsilon, 0, 1, 00, 01, 10, 11, 000, 001, 010, 011, 100, 101, \ldots\}$.
- Στη λεξικογραφική σειρά, **αριθμούμε** τις συμβολοσειρές του Σ^* με τους φυσικούς αριθμούς.

Μη-Μετρήσιμα Σύνολα

- Το 2^{Σ^*} (σύνολο γλωσσών στο Σ) είναι **μη-μετρήσιμο**.
- Έστω μετρήσιμο: Γλώσσες απαριθμούνται σαν **ακολουθία** $L_1, L_2, L_3, \ldots$.
  Κατασκευάζουμε **γλώσσα** D στο Σ που **διαφέρει** από κάθε L_i, $i = 1, 2, \ldots$
- Έστω $w_1, w_2, \ldots$ λεξικογραφική σειρά συμβολοσειρών του Σ.
  $D = \{w_i : w_i \notin L_i\,,\ i = 1, 2, \ldots\}$, δηλαδή $w_i \in D \Leftrightarrow w_i \notin L_i$.
- Η γλώσσα D διαφέρει (στο w_i) από κάθε L_i, **άτοπο!**
- Άτοπο από αμφιμονοσήμαντη αντιστοιχία ανάμεσα
  στις γλώσσες και τις συμβολοσειρές του Σ.
- Ομοίως, $2^{\mathbb{N}}$ **μη-μετρήσιμο** (Θεώρημα 1.5.2, σελ.54).

Διαγωνιοποίηση

- Έστω $R \subset A \times A$.
  $\forall x \in A$, $R_x = \{y \in A : (x, y) \in R\}$ η αντίστοιχη **γραμμή της** R.
  Έστω **διαγώνιο σύνολο** της R: $D = \{x \in R : (x, x) \notin R\}$

	a	b	c	d	e
a	×		×		×
b			×	×	×
c		×		×	
d	×			×	×
e		×	×		

$D = \{b, c, e\}$
$R_a = \{a, c, e\}$
$R_b = \{c, d, e\}$
$R_c = \{b, d\}$
$R_d = \{a, d, e\}$
$R_e = \{b, c\}$

- Το διαγώνιο σύνολο D είναι **διαφορετικό** από κάθε γραμμή της R.
  Διαφέρει από το R_a στο a, από το R_b στο b, κοκ.
- Ο Georg Cantor εισήγαγε την **αρχή της διαγωνιοποίησης** στα 1891.
- Με διαγωνιοποίηση μπορούμε ν.δ.ο το $\mathbb{R}$ **είναι μη-μετρήσιμο**.

Πεπερασμένη Αναπαράσταση Γλωσσών

- Αναπαράσταση είναι **συμβολοσειρά** κάποιου πεπερασμένου αλφάβητου.
  Δυνατότητα αναπαράστασης για **μετρήσιμα** διαφορετικές γλώσσες.
- Υπάρχουν **μη-μετρήσιμα** διαφορετικές γλώσσες:
  Άρα, υπάρχουν γλώσσες που **δεν έχουν** πεπερασμένη αναπαράσταση.
- Υπάρχει γλώσσα που κανένα πρόγραμμα C δεν την τυπώνει.
- Θα ασχοληθούμε αποκλειστικά με γλώσσες που έχουν πεπερασμένη αναπαράσταση.
- Θα ορίσουμε **μορφές αναπαραστάσεων** και υπολογιστές που τις αναγνωρίζουν.

Υπόθεση του Συνεχούς

- Κάθε **μη-μετρήσιμο** σύνολο **περιέχει** ένα υποσύνολο ισοδύναμο με το $\mathbb{R}$.
- Διαισθητικά, δεν υπάρχει σύνολο μεγαλύτερο του $\mathbb{N}$ και μικρότερο του $\mathbb{R}$.
- Gödel (1937) έδειξε ότι η υπόθεση είναι συμβατή με τα αξιώματα της συνολοθεωρίας (άρα δεν υπάρχει **αντιπαράδειγμα**).
- Cohen (1963) έδειξε ότι η **άρνηση** της υπόθεσης είναι επίσης συμβατή με τα αξιώματα της συνολοθεωρίας (άρα δεν υπάρχει **απόδειξη**).
- Συνεπώς η Υπόθεση του Συνεχούς δεν μπορεί ούτε να αποδειχθεί ούτε να καταρριφθεί.