A logika és a számítástudomány alapjai mérnök informatikus hallgatóknak

10. előadás, reguláris kifejezések

Mihálydeák Tamás, Aszalós László

A tananyag elkészítését az EFOP-3.4.3-16-2016-00021 számú projekt támogatta. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósult meg.

2019. november 4.

Motiváció

Az informatikában nagyon gyakori feladat az elemzés és a generálás.

- A különféle programok/számítógépek közti adatforgalom különféle szabványokat használ (HTML, XML, json, YAML, stb). Ezért a küldő oldalon generálni kell ilyen adatokat, a fogadó oldalon pedig elemezni.
- A forrásprogramból a futtatható kód fordítás során készül el. A fordítás első része a lexikai, szintaktikai és szemantikai elemzés, melyet a kódgenerálás követ.
  (Interpreter esetén az utóbbi a végrehajtásra cserélődik.)

Motiváció - 2

Ilyen feladat mindenki saját gyakorlatában is gyakran előfordul.

Például

- Egy HTML forrásban szeretnénk a nagybetűkkel leírt elemeket (tag) kisbetűsíteni.
- HTML oldalt szeretnénk generálni PHP, Python és egyéb keretrendszerekben.
- Űrlapok esetén
  - A bevitt adat valóban egy email cím?
  - A jelszó tartalmaz nem betű karaktert?
- Mely fájllal kell még dolgoznunk? (Hol szerepel TODO?)

Megjegyzés

Ha valamilyen mintát/szabványt követünk, akkor szabványos megoldásokat használhatunk, és hatékonyabbak lehetünk.

Motiváció - 3

Példa - Lindenmayer rendszer

Formális nyelv

A következőkben jelöljön Σ egy nemüres halmazt! A halmaz elemeit betűknek, a halmazt ábécének nevezzük.

Definíció

Jelölje Σ^+ a Σ betűiből álló $w = a_1 \ldots, a_n$ $(a_i \in \Sigma)$ alakú véges hosszú sorozatok, az úgynevezett Σ feletti véges szavak halmazát. A w hosszán a benne szereplő betűk számát értjük: $|w| = n$.
Az üres szót (melynek nincs egy betűje sem, és a hossza 0) a ε jelöli.
A $\Sigma^\star = \Sigma^+ \cup \{\varepsilon\}$ halmazt a Σ feletti szavak halmazának nevezzük.
A Σ ábécé feletti szavak egy tetszőleges L halmazát a Σ ábécéből alkotott (formális) nyelvnek nevezzük.
Az üres nyelvet, melynek nincs egy szava sem, a $\emptyset$ jelöli.

Formális nyelv műveletei

Definíció

Legyen A és B két Σ feletti nyelv! A következő műveleteket definiáljuk:

- $A \cup B = \{x | x \in A \text{ vagy } x \in B\}$ (unió)
- $A \circ B = \{xy | x \in A \text{ és } y \in B\}$ konkatenáció (összefűzés)
- $A^\star = \{x_1 x_2 \dots x_k | k \geq 0 \text{ és } x_i \in A\}$ Kleene-csillag.

Példa

Legyen $\Sigma = \{a, b, c\}$, $A = \{a, bb, ccc\}$ és $B = \{ac, bc\}$. Ekkor

- $A \cup B = \{a, ac, bb, bc, ccc\}$,
- $A \circ B = \{aac, abc, bbac, bbbc, cccac, cccbc\}$,
- $A^\star = \{\varepsilon, a, bb, ccc, aa, abb, accc, bba, bbbb, bbccc, \dots\}$.

Reguláris kifejezés

Történelem

- Stephen C. Klenne, „reguláris halmazok” (1956)
- Ken Thompson, QED szövegszerkesztő, Just-In-Time fordító (1968)
- Unix: vi, lex, sed, awk, emacs (197?)
- Perl (1986)

Definíció

Azt mondjuk, hogy az R egy reguláris kifejezés, ha

1. Ha $R = \emptyset$
2. Ha $R = \varepsilon$
3. Ha $R = a$, ahol $a \in \Sigma$.
4. Ha $R = (R_1 + R_2)$, ahol R_1 és R_2 reguláris kifejezések
5. Ha $R = (R_1 \cdot R_2)$, ahol R_1 és R_2 reguláris kifejezések
6. Ha $R = R_1^\star$, ahol R_1 reguláris kifejezés

Reguláris kifejezés által felismert nyelv

A reguláris kifejezések és a nyelvek között kapcsolat létesíthető: minden egyes reguláris kifejezésnek megfelel egy nyelv.

Definíció

Az R reguláris nyelv által felismert nyelvet (jelölése L_R) a következő induktív definíció adja meg:

1. Ha $R = \emptyset$, akkor $L_R = \emptyset$.
2. Ha $R = \varepsilon$, akkor $L_R = \{\varepsilon\}$.
3. Ha $R = a$, akkor $L_R = \{a\}$, ahol $a \in \Sigma$.
4. Ha $R = (R_1 + R_2)$, akkor $L_R = L_{R_1} \cup L_{R_2}$.
5. Ha $R = (R_1 \cdot R_2)$, akkor $L_R = L_{R_1} \circ L_{R_2}$.
6. Ha $R = R_1^\star$, akkor $L_R = L_{R_1}^\star$.

Reguláris nyelvek és műveleteik

Definíció

A Σ ábécé feletti L nyelv reguláris, ha létezik egy olyan R reguláris kifejezés, hogy $L_R = L$

Tétel

Ha a Σ ábécé feletti L_1 és L_2 nyelvek regulárisak, akkor az $L_1 \cup L_2$, $L_1 \cap L_2$, $L_1 \backslash L_2$, $\overline{L_1} = \Sigma^\star \backslash L_1$, $L_1 \circ L_2$ és $L_1^\star$ nyelvek is regulárisak.

Reguláris kifejezés a gyakorlatban

Különböző szabványok terjedtek el, melyek további konstrukciókat tartalmaznak:

- $R? = R + \varepsilon$ (0 vagy 1 előfordulás)
- $R^+ = RR^\star$ (1 vagy több előfordulás)
- $R\{n\}$ (pontosan n-szeri előfordulás)
- $R\{n, m\}$ (legalább n-szeri, maximum m-szeri előfordulás)
- $[a_1 \ldots a_n] = a_1 + \cdots + a_n$ (betűhalmaz)
- $[a_1 - a_n]$ (betűintervallum, pl $a - z$)

A Unix a reguláris kifejezések írásakor a $+$ karakter helyett $|$-et használ: $alma|eper$.