Zadania z gwiazdką - seria I, szkice rozwiązań

1. Rozstrzygnij, czy język

$$L = \{w \in \{a, b, c\}^* \mid w = uv \text{ oraz } \#_a(u) + \#_b(u) = \#_b(v) + \#_c(v)\}$$

jest regularny.

Szkic rozwiązania Język L nie jest regularny, wykażemy, że relacja $\sim_L$ ma nieskończenie wiele klas abstrakcji. Niech $w_n = a^n$, pokażemy, że $w_i \not\sim_L w_j$ dla $i \neq j$. Rozważmy słowo $u = bc^i$. Nietrudno sprawdzić, że $w_i u = a^i bc^i \notin L$, ale $w_j u = a^j bc^i \in L$, co kończy dowód. (W realnym rozwiązaniu oczekiwałbym sprawdzenia tego należenia). $\square$

2. Niech $\mathcal{F}$ to najmniejsza klasa zawierająca wszystkie języki skończone nad wszystkimi skończonymi alfabetami, która jest zamknięta na:

- skończoną sumę, dopełnienie i konkatenację;
- skończoną sumę, dopełnienie, konkatenację i rzutowanie.

W obu przypadkach rozstrzygnij, czy klasa $\mathcal{F}$ jest równa językom regularnym.
Uwaga: Przez rzutowanie języka $L \subseteq \Sigma^*$ na pewien podalfabet $\Gamma \subseteq \Sigma$ rozumiemy zbiór słów, które powstały ze słów z języka L poprzez usunięcie wszystkich liter z $\Sigma \setminus \Gamma$.
Uwaga 2: Przez zamknięcie na dopełnienie rozumiemy fakt, że jeśli $L \subseteq \Sigma^*$ należy do $\mathcal{F}$, to również $\Sigma^* \setminus L$ należy do $\mathcal{F}$. W szczególności dla ustalonego L można stosować tę regułę dla różnych alfabetów Σ takich, że $L \subseteq \Sigma^*$.

Szkic rozwiązania Na początek zauważmy, że klasa języków regularnych zawiera wszystkie języki skończone oraz jest zamknięta na sumę, dopełnienie i konkatenację. Łatwo również pokazać, że jest ona zamknięta na rzutowanie. W tym celu rozważmy język regularny L i jego automat A. Aby otrzymać automat A' dla języka będącego rzutem L na podalfabet Γ modyfikujemy tranzycje A zamieniając etykiety z podalfabetu $\Sigma \setminus \Gamma$ na ε. To dowodzi, że zarówno klasa z punktu 1) jak i z punktu 2) jest zawarta w klasie języków regularnych.

Punkt 1. Pokażemy teraz, że klasa z punktu 1) jest istotnie mniejsza niż języki regularne. Niech klasa $\mathcal{F}'$ to języki z $\mathcal{F}$, które zawierają się w $\{a\}^*$. Każdy taki język powstaje za pomocą skończonej liczby operacji: suma, dopełnienie, konkatenacja z języków skończonych. Co więcej wszystkie użyte języki są również z $\mathcal{F}'$. Pokażemy, że $\mathcal{F}'$ zawiera tylko języki skończone oraz koskończone (takie, że ich dopełnienia są skończone). Zrobimy to przez indukcję po budowie ich wyprowadzenia. Załóżmy, że $K, L \in \mathcal{F}'$, czyli są skończone lub koskończone. Wówczas:

- Język $K \cup L$ jest również skończony lub koskończony. Istotnie, jeśli oba K i L są skończone, to $K \cup L$ jest skończony. Jeśli natomiast jeden z K lub L jest koskończony, to $K \cup L$ zawiera go i jest też koskończony.

- Język $\bar{L}$ jest również skończony lub koskończony. Istotnie, jeśli L jest skończony, to $\bar{L}$ jest koskończony, a jeśli L jest koskończony, to $\bar{L}$ jest skończony.
- Język KL jest również skończony lub koskończony. Istotnie, jeśli oba K i L są skończone, to KL jest skończony. Niech co najmniej jeden z K, L będzie koskończony, powiedzmy K. A więc K zawiera wszystkie słowa dłuższe niż n, dla pewnego $n \in \mathbb{N}$. Jeśli drugi z nich, w tym wypadku L, jest pusty, to KL jest pusty. Jeśli natomiast L jest niepusty, to zawiera pewne słowo a^k. Zatem KL zawiera wszystkie słowa dłuższe niż $k+n$, czyli jest koskończony.

A zatem istotnie $\mathcal{F}'$ zawiera tylko języki skończone lub koskończone. Jednak język słów długości parzystej nad $\{a\}$ jest regularny, ale nie jest skończony ani koskończony. Czyli $\mathcal{F}'$, a więc również $\mathcal{F}$ nie zawiera wszystkich języków regularnych.

Punkt 2. Teraz pokażemy, że klasa z punktu 2) zawiera wszystkie języki regularne. Wystarczy pokazać, że jest ona zamknięta na operację gwiazdki Kleene'go. Wówczas $\mathcal{F}$ zawiera wszystkie języki skończone i jest zamknięta na sumę, konkatenację i gwiazdkę, czyli zawiera wszystkie języki regularne. Niech $L \subseteq \Sigma^*$ należy do $\mathcal{F}$. Niech $\# \notin \Sigma$. Niech $\Sigma_\# = \Sigma \cup \{\#\}$. Zauważmy, że $\Sigma_\#^*$ należy do $\mathcal{F}$, gdyż jest postaci $\Sigma_\#^* \setminus \emptyset$. Rozważmy język

$$L' = (\Sigma_\#^* \cdot \{\#\} \cdot \bar{L} \cdot \{\#\}\Sigma_\#^*) \cup (\Sigma_\#^* \cdot \{\#\} \cdot \bar{L}) \cup (\bar{L} \cdot \{\#\}\Sigma_\#^*) \cup \bar{L}.$$

Po pierwsze $L' \in \mathcal{F}$ jako suma konkatenacji kilku języków z $\mathcal{F}$. Język L' zawiera wszystkie słowa nad $\Sigma_\#$ takie, że istnieje maksymalny blok liter z Σ (czyli bez $\#$), który tworzy słowo spoza L. Rozważmy język $\bar{L'} \in \mathcal{F}$. On zawiera wszystkie słowa nad $\Sigma_\#$ takie, że każdy maksymalny blok liter z Σ należy do L. A więc rzutowanie tego języka na Σ to dokładnie język L^+. A zatem istotnie $L^* = L^+ \cup \{\varepsilon\} \in \mathcal{F}$, co kończy rozwiązanie. $\square$

3. Dla języka $L \subseteq \Sigma^*$ niech jego *gęstość* to funkcja $g_L : \mathbb{N} \to \mathbb{N}$ taka, że dla każdego $n \in \mathbb{N}$ wartość $g(n)$ to liczba słów długości n należących do L. Rozstrzygnij, czy istnieje język regularny L taki, że jego gęstość g_L jest ponad wielomianowa, ale podwykładnicza, czyli dla każdych $c, k > 1$ zachodzi $f = \Omega(n^k)$, ale $f = o(c^n)$.

Szkic rozwiązania Nie istnieje taki język. Rozważmy język regularny i pewien automat deterministyczny A akceptujący L. Rozpatrzymy dwa przypadki.

Przypadek 1. Istnieje pewien stan q osiągalny ze stanu początkowego q_0, z którego można osiągnąć pewien stan końcowy f taki, że istnieją dwa różne cykle proste z q do q. Niech $q_0 \xrightarrow{u} q$, $q \xrightarrow{v_1} q$, $q \xrightarrow{v_2}$, $q \xrightarrow{w} f$ będą wspomnianymi ścieżkami i cyklami. Załóżmy bez straty ogólności, że $|v_1| \leqslant |v_2|$. Dodatkowo v_1 nie jest prefiksem v_2, gdyż w przeciwnym wypadku cykl $q \xrightarrow{v_2} q$ byłby równy cyklowi $q \xrightarrow{v_1} q$, jako, że A jest deterministyczny. Zauważmy, że $v_1 v_2 \neq v_2 v_1$, gdyż mają inny prefiks długości $|v_1|$. A więc L zawiera przynajmniej 2^n słów długości $|u| + n(|v_1| + |v_2|) + |w|$, gdyż każde słowo postaci $us_1 \cdots s_n w$, gdzie $s_i \in \{v_1 v_2, v_2 v_1\}$ należy do L. Zatem gęstość L jest wykładnicza.

Przypadek 2. Wszystkie stany osiągalne ze stanów początkowych i osiągające stan końcowy należą do co najwyżej jednego cyklu prostego. Pokażemy, że w tym przypadku gęstość L jest wielomianowa. Rozważmy dowolny bieg akceptujący automatu A i dowolny stan q. Jeśli jest on odwiedzany wiele razy na biegu, to znaczy, że pomiędzy odwiedzeniami bieg pętli się na jedynym cyklu prostym przechodzącym przez q. A zatem bieg można zdekomponować na cykle proste oraz ścieżkę prostą (pozostającą po usunięciu cyklów prostych). Taki bieg wygląda więc następująco:

$$q_0 \xrightarrow{w_1} q_1 \xrightarrow{s_1^*} q_1 \xrightarrow{w_2} q_2 \xrightarrow{s_2^*} q_2 \cdots q_{k-1} \xrightarrow{s_{k-1}^*} q_{k-1} \xrightarrow{w_k} q_k,$$

gdzie $w_1 \cdots w_k$ składają się na etykietowanie ścieżki prostej, a s_i to etykietowania cykli. Zauważmy, że automacie jest skończenie wiele ścieżek prostych, więc wystarczy pokazać, że dla każdej z nich gęstość odpowiadającego jej języka jest wielomianowa. Pytamy więc ile może być słów postaci $w_1(s_1)^* w_2(s_2) * \cdots (s_{k-1})^* w_k$ długości n. Nietrudno zauważyć, że jest to nie większe niż liczba wyborów $2k-2$ punktów podziału słowa długości n (granic kawałków), a więc nie większe niż $\binom{n+1}{2k-2} = O(n^{2k-2})$. Jest to zresztą bardzo grube oszacowanie. A więc istotnie gęstość L jest wielomianowa. □

4. Niedeterministyczny automat jednolicznikowy A nad alfabetem Σ składa się ze zbioru stanów Q, zbioru stanów początkowych $I \subseteq Q$, końcowych $F \subseteq Q$ oraz zbioru tranzycji $\delta \subseteq Q \times \Sigma \times \{-1, 0, 1\} \times Q$. Konfiguracja automatu A to para ze zbioru $Q \times \mathbb{N}$, czyli innymi słowy automat może przyjmować tylko nieujemne wartości licznika. Bieg automatu po słowie $w = a_1 \cdots a_n$ to ciąg konfiguracji $(q_0, c_0), \ldots, (q_n, c_n)$ takich, że dla każdego $i \in \{1, \ldots, n\}$ mamy $(q_{i-1}, a_i, c_i - c_{i-1}, q_i) \in \delta$ oraz oczywiście $c_i \geqslant 0$. Bieg jest akceptujacy gdy $q_0 \in I$, $c_0 = 0$ oraz $q_n \in F$. Słowo jest akceptowane jeśli istnieje bieg akceptujący po nim, a język automatu $L(A)$ to zbiór wszystkich słów akceptowanych. Zaprojektuj algorytm, który dla danego niedeterministycznego automatu jednolicznikowego A odpowiada na pytanie, czy $L(A) = \Sigma^*$.

Szkic rozwiązania Porządek $(X, \leqslant)$ nazwiemy WQO (od ang. well-quasi-order) jeśli dla dowolnego nieskończonego ciągu $x_1, x_2, \ldots$ elementów X istnieją indeksy $i < j$ takie, że $x_i \leqslant x_j$. Dla $x \in \mathbb{N}^d$ niech $x[i]$ oznacza wartość na i-tej współrzędnej wektora x. Przez $\preceq$ oznaczmy porządek na $\mathbb{N}^d$, gdzie $x \preceq y$ o ile $x[i] \leqslant y[i]$ dla każdej współrzędnej $i \in \{1, \ldots, d\}$.

Lemat 1 (Lemat Dicksona)**.** *Porządek $(\mathbb{N}^d, \preceq)$ jest WQO.*

Dowód. Wykażemy to przez indukcję po d. Dla $d = 1$ jest to oczywiste. Załóżmy, że teza lematu jest prawdziwa dla d, pokażemy ją dla $d + 1$. Niech $x_1, x_2, \ldots$ to nieskończony ciąg elementów $\mathbb{N}^{d+1}$. Rozważmy współrzędną $d + 1$-szą tych wektorów. Jeśli jest ona ograniczona, to możemy znaleźć podciąg stały, jeśli natomiast jest ona nieograniczona, to możemy znaleźć podciąg rosnący. W obu wypadkach znajdujemy pewien podciąg $x_{i_1}, x_{i_2}, \ldots$ taki, że d+1-sza współrzędna jest niemalejąca. Z tezy dla d otrzymujemy, że istnieją pewne dwa elementy tego podciągu takie, że $j < k$ oraz $x_{i_j} \preceq x_{i_k}$ po ograniczeniu do pierwszych d współrzędnych. Jednak $x_{i_j}[d + 1] \leqslant x_{i_k}[d + 1]$, więc mamy $x_{i_j} \preceq x_{i_k}$, co kończy dowód lematu. □

Uwaga: Analogicznie dowodzi się, że porządek $((\mathbb{N} \cup \{-1\})^d, \preceq)$ jest WQO, będziemy się na to również powoływali jako na lemat Dicskona.

Niech A będzie automatem jednolicznikowym nad Σ o d stanach. Przez $p(i)$ oznaczamy konfigurację: stan p, wartość licznika i. Przez $p(i) \xrightarrow{u} q(j)$ będziemy oznaczać, że z konfiguracji $p(i)$ automat A może dojść po słowie u do konfiguracji $q(j)$. Niech $S_w = \{p(i) \mid q_0(0) \xrightarrow{w} p(i)\}$ będzie zbiorem wszystkich konfiguracji, do których można dojść po słowie w z konfiguracji początkowej $q_0(0)$. Zauważmy, że jeśli $q(i) \xrightarrow{u} r(k)$, to dla $j > i$ również $q(j) \xrightarrow{u} r(k + j - i)$. A więc istotne są jedynie konfiguracje o największej możliwej wartości w danym stanie. Niech $Q = \{q_1, \ldots, q_d\}$. Niech *profil* słowa $w \in \Sigma^*$, $\mathrm{prof}(w) \in (\mathbb{N} \cup \{-1\})^d$, będzie zdefiniowany następująco: $\mathrm{prof}(w)[i]$ to największa możliwa wartość j taka, że $q_i(j) \in S_w$, oraz -1 jeśli dla żadnego $j \in \mathbb{N}$ konfiguracja $q_i(j)$ nie należy do zbioru S_w.

Lemat 2. *Jeśli* $\mathrm{prof}(u) \preceq \mathrm{prof}(v)$, *to wówczas dla dowolnego* $w \in \Sigma^*$ *takiego, że* $uw \in L$ *mamy również* $vw \in L$.

Istotnie, jeśli $uw \in L$, to $q_0(0) \xrightarrow{u} q(i) \xrightarrow{w} f(k)$. Jednak skoro $\mathrm{prof}(u) \preceq \mathrm{prof}(v)$, to istnieje bieg $q_0(0) \xrightarrow{v} q(j)$ dla pewnego $j \geqslant i$. A więc $q_0(0) \xrightarrow{v} q(j) \xrightarrow{w} f(k + j - i)$, czyli $vw \in L$.

Przypuśćmy teraz, że $L \neq \Sigma^*$. Niech w to najkrótsze słowo nie należące do L, niech jego długość to n. Spójrzmy teraz na prefiksy słowa w: $w_0 = \varepsilon, w_1 = w[1..1], w_2 = w[1..2], \ldots, w_n = w[1..n] = w$. Z pewnością nie jest prawdą, że $\mathrm{prof}(w_i) \preceq \mathrm{prof}(w_j)$ dla pewnych $i < j$. Istotnie, przypuśćmy, że tak jest. Mamy wtedy, że $w[1..i] \cdot w[j + 1..n] \in L$, jako, że jest to słowo krótsze niż w. Jednak $\mathrm{prof}(w_i) \preceq \mathrm{prof}(w_j)$, więc z lematu 2 mamy, że również $w = w[1..j] \cdot w[j+1..n] \in L$, sprzeczność.

Jesteśmy już gotowi do zaprezentowania algorytmu. Będziemy szukać potencjalnego słowa $w \notin L$. W tym celu budujemy drzewo słów. Każde słowo jest wierzchołkiem, jego rodzicem w drzewie jest jego prefiks bez ostatniej litery. Czyli korzeniem drzewa jest słowo ε, jego dziećmi słowa długości 1, potem słowa długości 2 itd. W każdym kroku budujemy nowy poziom drzewa, potencjalnie nieskończonego oraz obliczamy profil nowych słów. Jeśli dla pewnego wierzchołka drzewa, odpowiadającemu słowu w mamy $\mathrm{prof}(w) \succeq \mathrm{prof}(w_i)$ dla pewnego prefiksu w_i słowa w, to obcinamy drzewo na tym poziomie i przestajemy przedłużać tę część drzewa, gdyż (jak pokazaliśmy wcześniej) najkrótsze słowo nie należące do L na pewno tu nie leży.

Jeśli w pewnym momencie znajdziemy słowo, które nie należy do L, to odpowiadamy oczywiście, że $L \neq \Sigma^*$. Jeśli natomiast w pewnym momencie już wszystkie kawałki drzewa zostaną obcięte i nie będzie nic do rozważania, to odpowiadamy, że $L = \Sigma^*$

Należy pokazać dwie rzeczy. Po pierwsze dlaczego ten algorytm zawsze się kończy, a po drugie dlaczego zwraca poprawną odpowiedź. Jeśli $L \neq \Sigma^*$, to istotnie kiedyś znajdziemy słowo, które nie należy do L i zakończymy algorytm z dobrą odpowiedzią, ten przypadek jest prosty. Załóżmy więc, że $L = \Sigma^*$. Zauważmy, że na każdej nieskończonej ścieżce z drzewie: $w_0, w_1, \ldots$, gdzie każde następne słowo powstało przez dodanie litery do poprzedniego mamy ciąg $\mathrm{prof}(w_0), \mathrm{prof}(w_1), \ldots \in (\mathbb{N} \cup \{-1\})^d$. Z lematu Dicksona otrzymujemy, że dla pewnych $i < j$ mamy $\mathrm{prof}(w_i) \preceq \mathrm{prof}(w_j)$, a więc na tej ścieżce drzewo kiedyś zostanie obcięte. Należy teraz pokazać, że jeśli każda ścieżka została kiedyś obcięta, to pewnym poziomie już wszystkie ścieżki zostały obcięte. Rozważmy drzewo po wykonaniu wszystkich obcięć, chcemy pokazać, że jest skończone. Wynika to z lematu Königa, który mówi, że każde drzewo nieskończone o skończonym rozgałęzieniu ma nieskończoną ścieżkę. Nasze drzewo ma skończone rozgałęzienie, tzn. każdy wierzchołek ma skończenie wiele dzieci, dokładnie rzecz biorąc $|\Sigma|$. Lemat Königa można udowodnić następująco: korzeń ma nieskończone poddrzewo, skoro ma skończenie wiele dzieci, to któreś z dzieci ma nieskończone poddrzewo, więc któreś z jego dzieci ma nieskończone poddrzewo itd., w ten sposób skonstruujemy nieskończoną ścieżkę. A zatem skoro w naszym drzewie po obcięciach nie ma nieskończonej ścieżki (bo każda została obcięta), to drzewo jest skończone, w którym momencie wszystko obetniemy. Jeśli wszystko obcięliśmy, to faktycznie nie ma szans na znalezienie najkrótszego słowa spoza L, czyli $L = \Sigma^*$, co kończy dowód. $\square$

5. Pokaż, że jeśli zarówno L jak i $\Sigma^* \setminus L$ są rozpoznawane przez pewne niedeterministyczne automaty jednolicznikowe, to wtedy L jest regularny.

Szkic rozwiązania Będziemy używali podobnych pojęć co w szkicu rozwiązania zadania 4. Niech L będzie rozpoznawany przez automat A do a stanach, a $\bar{L}$ przez automat B o b stanach. Dla słowa $w \in \Sigma^*$ niech $\mathrm{prof}_A(w) \in (\mathbb{N} \cup \{-1\})^a$ będzie jego profilem względem automatu A, natomiast $\mathrm{prof}_B(w) \in (\mathbb{N} \cup \{-1\})^b$ będzie jego profilem względem automatu B. Niech $\mathrm{prof}(w) = (\mathrm{prof}_A(w), \mathrm{prof}_B(w)) \in (\mathbb{N} \cup \{-1\})^{a+b}$. Pokażemy teraz, że L jest regularny. Przypuśćmy przeciwnie, wówczas ma on nieskończenie wiele klas abstrakcji relacji Myhilla-Neroda, niech reprezentanci tych klas to $w_1, w_2, \ldots$. Z lematu Dicksona (w zadaniu 4) otrzymujemy, że istnieją indeksy $i < j$ takie, że $\mathrm{prof}(w_i) \preceq \mathrm{prof}(w_j)$. To oznacza w szczególności, że $\mathrm{prof}_A(w_i) \preceq \mathrm{prof}_A(w_j)$ oraz $\mathrm{prof}_B(w_i) \preceq \mathrm{prof}_B(w_j)$. A zatem, z lematu 2 zaaplikowanego do automatu A wynika, że jeśli $w_i u \in L$, to $w_j u \in L$, natomiast zaaplikowanego do automatu B wynika, że jeśli $w_i u \in \bar{L}$, to $w_j u \in \bar{L}$. A zatem otrzymujemy, że $w_i u \in L \iff w_j u \in L$, czyli $w_i \sim_L w_j$. Sprzeczność z założeniem, że w_i oraz w_j są reprezentantami różnych klas abstrakcji relacji $\sim_L$. $\square$