Wykład 3. Star Height.

Języki regularne można definiować za pomocą wyrażeń regularnych skonstruowanych z liter, operacji konkatenacji, sumy i gwiazdki Kleenego. Wysokość gwiazdkowa (starheight) wyrażenia definiujemy jako najwyższy poziom zagniezdzenia gwiazdek w wyrazeniu, np. $Sh((a^*bc)^*) = 2$. Wysokość gwiazdkowa języka to minimalna wysokość gwiazdkowa wśród wyrażeń opisujących ten język. Pytanie o rozstrzygalność problemu, by na podstawie danego DFA $\mathcal{A}$ okreslić $Sh(L(A))$ była postawiona w roku 1963. Roztrzygalność pokazano w 1986, obecnie wiadomo, że problem jest w 2EXPSPACE i jest PSPACE-trudny.

Definiuje się rozszerzone wyrażenia regularne—korzystajace dodatkowo z z operacji dopełnienia (a więc i przekroju). Ponieważ, języki regularne domkniete są na operacje Booleowskie pozostajemy w klasie języków regularnych. Nie wiadomo, czy istnieje język taki, że jego wysokość gwiazdkowa w sensie wyrażen rozszerzonych jest wieksza od 1.

Udowodnimy, ze hierarchia Sh dla (zwykłych) wyrażeń jest nieskończona, a nawet ścisła.

Twierdzenie 1. *Dla dowolnej naturalnej n istnieje język regularny L taki, że $Sh(L) = n$.*

Zdefiniujemy indukcyjnie nastepujący ciąg wyrażeń regularnych oraz słów:

$$\begin{aligned} \alpha_1 &= (ab)^* \\ \alpha_i &= \left(a^{2^{i-1}}\alpha_{i-1}b^{2^{i-1}}\alpha_{i-1}\right)^* \\ w(1,j) &= ab \\ w(i,j) &= a^{2^{i-1}}w(i-1,j)^j b^{2^{i-1}}w(i-1,j)^j \end{aligned}$$

Oczywiście $w(i,j) \in L(\alpha_i)$ oraz $Sh(L(\alpha_i)) \leq i$.
Przez $\#_a(w)$ oznaczamy liczbę liter występujących w słowie w. Niech K_i będzie zbiorem języków L takich, że

1. $\exists t \in \mathbb{Z}\ \forall w \in L\ \#_a(w) = \#_b(w) + t$
2. $\forall j_0 \in \mathbb{N}\ \exists j > j_0\ \exists w \in L\ w(i,j)^j$ jest podsłowem w
3. L ma minimalny Sh spośród jezyków spełniajacych $(1,2)$.

K_i jest niepusta, bo $L(\alpha_i)$ spełnia warunki $(1,2)$. Sh języków z tej klasy to $d_i \leq i$. Rozważmy zatem język $L \in K_i$, dla którego najlepsze (w sensie Sh) wyrażenie ma postać

$$\alpha = \sum \gamma_0^*\gamma_1\gamma_2^*\gamma_3 \dots \gamma_{2n-1}\gamma_{2n}^*$$

gdzie suma jest skończenie indeksowana, a $Sh(\gamma_j) < d_i$.

Lemat 2. *Każde γ_j spełnia warunek 1., żadne γ_j nie spełnia warunku 2..*
Dowód. Zauważmy, że całe wyrażenie $\beta = \gamma_0^* \cdots \gamma_{2n}^*$ spełnia warunek 1.. Dla parzystych γ zachodzi $\#a = \#b$ — w przeciwnym wypadku słowa z inną iloscią „powtórzeń" słowa dopasowujacego sie do takiej γ miałyby różne różnice $\#a - \#b$. Podobnie gdyby do γ_j dla

j nieparzystego dopasowywały sie słowa w', w'' o różnych wartościach tej różnicy, to słowa $w_1w'w_2$, $w_1w'w_2$ dopasowujące się do β również miałyby różne wartości tej różnicy.

Teraz ponieważ $Sh(\gamma_j) < d_i$, to gdyby γ_j spełniało warunek 2., otrzymalibyśmy sprzeczność z warunkiem 3. należenia do K_i. □

Lemat 3. *W sumie α można wybrać taki składnik β i indeks $0 \leq k \leq n$, że podwyrażenie γ^*_{2k} w β spełnia warunek 2..*

Dowód. Przypuśćmy nie wprost, że nie ma takiego k. Istnieje zatem takie j_0, że dla $j > j_0$ słowo $w(i,j)^j$ nie dopasowuje się do żadnego z $\gamma^*_0, \gamma_1, \gamma^*_2, \ldots$. Rozważmy DFA rozpoznający język β o j_1 stanach. Niech $j_2 > max(j_0, j_1)$, wtedy $w(i,j_2)^{j_2}$ jest podsłowem pewnego słowa z $L(\beta)$. Zatem dla tego słowa na scieżce przejść rozważanego DFA musi istnieć cykl, a wtedy istnieje liczba $0 < m < j_2$ taka, że dla $t \in \mathbb{N}$ i pewnych w_1, w_2 jest $w_1w(i,j_2)^{j_2+mt}w_2 \in L(\beta)$. Możemy wziąć słowo takiej postaci i $j > (2n+1)j_2$ i przyjrzeć się słowu $w_1w(i,j_2)^jw_2$. Nie jest możliwe, aby każda kopia słowa $w(i,j_2)^{j_2}$ zawierała w sobie granicę między γ_l, γ_{l+1} — granic tych jest $2n$, kopii słowa — $2n+1$. Zatem któraś trafia do pewnego γ_i lub γ^*_i, sprzeczność.

Z lematu 2 wynika dodatkowo, że nieparzysta nie może zawierać takiego słowa, więc musi ono znajdować się w którymś γ_{2k}. □

Lemat 4. *W sumie α można wybrać taki składnik β i indeks $0 \leq k \leq n$, że podwyrażenie γ_{2k} w β spełnia warunek $2'$., tzn. 2. z i zastapionym przez $i-1$.*

Dowód. Rozważmy γ_{2k} wybrane w lemacie 3. Wiemy, że γ^*_{2k} spełnia 2., przypuśćmy nie wprost, że γ_{2k} nie spełnia $2'$.. Istnieje zatem j_0 takie, że dla $j > j_0$ $w(i-1,j)^j$ nie jest podsłowem żadnego słowa z $L(\gamma_{2k})$. Istnieje też $j_1 > j_0$ takie, że $w(i,j_1)^{j_1}$ jest podsłowem pewnego $w \in L(\gamma^*_{2k})$. Oznaczmy $w = w_1w_2 \ldots w_m, w_j \in L(\gamma_{2k})$ i zastanówmy się, jak jeden egzemplarz $w(i,j_1) = a^{2^{i-1}} \underbrace{w(i-1,j_1)^{j_1}}_{u} \underbrace{b^{2^{i-1}}w(i-1,j_1)^{j_1}}_{v}$ może być umieszczony w czynnikach w. Niech w_p, w_q oznaczają czynniki, w których zaczynają się odpowiednio słowa u, v, oraz $w_q = xy$ gdzie x jest sufiksem u. Zachodzi $p < q$, ponieważ w przeciwnym przypadku całe $w(i-1,j_1)^{j_1}$ znajduje się w jednym czynniku dopasowującym się do γ_{2k}. Poczyńmy nastepujące spostrzeżenia:

Obserwacja 1: W każdym sufiksie $w(i-1,j_1)^{j_1}$ jest nie mniej liter b niż a.

Obserwacja 2: W każdym prefiksie v jest wiecej liter b niż a.

W całym w_q jest tyle samo a, co b, a x jest sufiksem $w(i-1,j_1)^{j_1}$. W y musi być zatem nie mniej a niż b. Więc albo y zawiera całe v (a dalej jeszcze więcej liter a), co znów daje sprzeczność, albo y jest puste. Ten sam argument stosujemy teraz do w_{q+1}, które zaczyna się tam, gdzie v. □

Dowód Twierdzenia. Pokażemy, że $d_i \geq i$, co razem ze znaną już nierównością przeciwną dowiedzie twierdzenia. Baza indukcji ($i = 1$) jest trywialna.

Rozważmy zatem $L \in K_i$ opisany przez wyrażenie α jak powyżej i wybierzmy γ_{2k} z lematów 3, 4. Wiemy, że $Sh(\gamma_{2k}) < d_i$, a zatem $Sh(L(\gamma_{2k})) < d_i$. Wyrażenie to spełnia warunki 1., 2'., a więc warunki 1., 2. należenia do K_{i-1}, zatem dla $L' \in K_{i-1}$ jest $Sh(L') < d_i$. Oznacza to, że $d_{i-1} \leq d_i - 1$, co przy założeniu indukcyjnym $d_{i-1} \leq i - 1$ daje tezę.