Separating Words Problem

Csányi Dávid

Témavezető: Pálvölgyi Dömötör

2022. december

1. A separating words problem

1986-ban Goralcik és Koubek[3] felvetette azt a kérdést, hogy egy egyszerű számítási eszközzel milyen nehéz megoldani a legkönnyebb feladatot, két szó megkülönböztetését. Ezt a kérdést fogom általánosabban(több szó esetén) vizsgálni a [2] cikket követve. A kettő szóra vonatkozó állítások bizonyításai a [2] alapján történnek, az ennél több szóra vonatkozó állítások saját eredmények.

1.1. Definíció. Jelölje $\mathrm{sep}_l(a_1, \dots a_l)$ azt a legkisebb k számot, amire létezik egy k állapotú véges automata, ami az $a_1, a_2, \dots a_l$ szavak esetén különböző végállapotokba kerül.

1.2. Definíció. $S_l(n) = \max\{\mathrm{sep}_l(a_1, \dots a_l) : a_1, \dots a_l$ a Σ abc feletti maximum n hosszú páronként különböző szavak.$\}$

A separting words problem célja az $S_l(n)$ függvényre alsó és felső korlátokat találni. Az eddigi legjobb ismert felső korlát $S_2(n) = \tilde{\mathcal{O}}(n^{\frac{1}{3}})$, amelyet Chase bizonyított be 2020-ban[1]. A legjobb ismert alsó korlát $S_2(n) = \Omega(\log(n))$.

1.3. Állítás. $\mathrm{sep}_l(a_1, \dots a_l)$ esetén nem számít a szavak sorrendje.

1.4. Állítás. $sep_3(a, b, c) \leq \min H \cdot \max H$, ahol $H = \{\mathrm{sep}_2(a, b), \mathrm{sep}_2(b, c), \mathrm{sep}_2(c, a)\}$

Bizonyítás. Azt fogjuk belátni, hogy $\mathrm{sep}_3(a, b, c) \leq \mathrm{sep}_2(a, b) \cdot \max\{\mathrm{sep}_2(b, c), \mathrm{sep}_2(c, a)\}$ igaz. Ebből következik az állítás, ugyanis $\mathrm{sep}_2(a, b)$ helyett választhatjuk $\min H$-t és ekkor a szorzat második tagja $\max H$ lesz.

Vegyük azt az $A_{a,b}$ automatát, amely az a és b szavakat különbözteti meg $\mathrm{sep}_2(a, b)$ állapottal. Ha ez az automata a c szóra különböző végállapotba kerül, mint a-ra és b-re akkor készen vagyunk. Tegyük fel, hogy a c és a szóra ugyanabba a végállapotba kerül. Ekkor vegyük az a-t és c-t $\mathrm{sep}_2(a, c)$ állapotban megkülönböztető $A_{a,c}$ automatát. Az $A_{a,b}$ és $A_{a,c}$ direkt szorzata egy $\mathrm{sep}_2(a, b) \cdot \mathrm{sep}_2(a, c)$ állapotú automata, amely mind a három szót különböző végállapotba viszi. Ha a c-t és a b-t viszi egy végállapotba az $A_{a,b}$, akkor hasonlóan járhatunk el, csak most az $A_{b,c}$-vel vesszük a direkt szorzatot. Ezekből következik az állítás. □

2. Különböző hosszú szavak

Ebben a részben meggondoljuk kettő és három szó esetében, hogy ha a szavak hossza eltérő, akkor a megkülönböztetésük nem nehéz és emiatt a továbbiakban feltehetjük hogy a feladatban minden szó hossza n.

2.1. Lemma. [4] Minden $n \geq 2$ természetes számra létezik egy $p \leq 4,4 \log(n)$ prím, amelyre igaz, hogy $p \nmid n$.

Bizonyítás. (Vázlat) Ha az állítás nem igaz, akkor minden $p \leq 4,4 \log(n)$ prím osztja n-et. Ekkor $(\prod_{p \leq 4,4 \log(n)} p) | n$. Definiáljuk a $\theta(x) = \sum_{p \leq x} \log p$ függvényt, amiről megmutatható, hogy

$\theta(x) \geq 0,23x$. Ekkor $\theta(4,4\log(n)) \geq 1,012\log(n)$, amiből következik, hogy egy $q \geq n^{1,012}$ szám osztja n-et, ami ellentmondás. □

2.2. Következmény. Ha $0 \leq i,j \leq n, n \geq 2$ és $i \neq j$, akkor létezik egy p prím, amelyre $p \leq 4,4 \cdot \log(n)$ és $i \not\equiv j \pmod{p}$.

2.3. Definíció. Az a szó hosszát a továbbiakban $|a|$ jelöli.

2.4. Állítás. [2] Ha $|a|, |b| \leq n$ és $|a| \neq |b|$, akkor $\text{sep}_2(a,b) = \mathcal{O}(\log(n))$.

Bizonyítás. A 2.2 következmény alapján létezik egy $p = \mathcal{O}(\log(n))$ prím, amire $|a| \not\equiv |b| \pmod{p}$. Egy p hosszú körből álló automata segítségével megkülönböztethető a két szó. □

2.5. Állítás. Ha $|a|, |b|, |c| \leq n$ és páronként különbözőek, akkor $\text{sep}_3(a,b,c) = \mathcal{O}(\log(n))$.

Bizonyítás. $|(|a| - |b|) \cdot (|b| - |c|) \cdot (|c| - |a|)| \leq n^3$, ezért a 2.1 lemma miatt létezik egy $p \leq 4,4 \cdot \log(n^3) = 13,2 \cdot \log(n)$ prím, amely nem osztja ezt a háromtagú szorzatot. Emiatt p nem osztja a szorzat egyik tagját sem, ami azt jelenti, hogy $|a|$, $|b|$ és $|d|$ különböző maradékokat adnak p-vel osztva. Ezért egy p hosszú kört tartalmazó automata más végállapotokba viszi őket. □

2.6. Állítás. Ha a, b és c közül valamelyik szó hossza nem n, akkor $\text{sep}_3(a,b,c) \leq S_2(n) \cdot \mathcal{O}(\log(n))$.

Bizonyítás. A sep_2 és sep_3 közötti egyenlőtlenség(1.4) és a sep_2 különböző hosszú szavak esetére vonatkozó 2.4 állításból következik. □

3. Az abc mérete

A szavak egy Σ véges abc betűiből állnak. Meg fogjuk vizsgálni, hogy ezen abc mérete($|\Sigma| = \kappa$) és a feladat nehézsége között milyen kapcsolat áll fenn.

3.1. Definíció. Jelölje $\text{sep}_l^\kappa(a_1, \ldots a_l)$ a $\text{sep}_l(a_1, \ldots a_l)$ értéket abban az esetben, amikor a szavak egy κ méretű abc felettiek és $S_l^\kappa(n)$ ezek maximumát.

3.2. Állítás. [2] $S_2^\kappa(n) = S_2^2(n)$, ha $\kappa \geq 2$.

Bizonyítás. Az $S_2^2 \leq S_2^\kappa$ irány világos, ugyanis egy kettő méretű abc feletti szó tekinthető egy κ méretű abc feletti szónak. A másik irány belátásához legyen $a \neq b \in \Sigma^n$, ahol $|\Sigma| = \kappa$. Válasszunk egy $1 \leq i \leq n$ indexet, amelyre $a_i \neq b_i$. Definiáljuk a $\Phi : \Sigma \mapsto \{0,1\}$ leképezést az alábbi módon:

$$\Phi(x) = \begin{cases} 1 & \text{ha } x = a_i \\ 0 & \text{ha } x \neq a_i \end{cases}$$

Ezen függvényt az a és b minden betűjére alkalmazva megkapjuk az a' és b' bináris szavakat, amelyek az i-edik bitben különböznek. Ezek megkülönböztethetők egy A automatával, amelynek $k \leq S_2^2(n)$ állapota van. Az automata átmenetfüggvényiben az 1-eket a_i-re, a 0-kat $\Sigma - \{a_i\}$-re cserélve a kapott automata megkülönbözteti a-t és b-t k állapotban. Így $sep_2^\kappa(a,b) \leq S_2^2(n)$ amiből az állítás következik. □

3.3. Állítás. $S_3^\kappa(n) = S_3^2(n)$, ha $\kappa \geq 2$.

Bizonyítás. Legyenek $a, b, c \in \Sigma^n$ páronként különböző n hosszú szavak a κ betűből álló Σ abc felett. Az 3.2 állítás bizonyításához hasonlóan azt kell megmutatnunk, hogy létezik egy $\Phi : \Sigma \mapsto \{0, 1\}$ leképezés, amelyet betűnként alkalmazva a három szóra, a kapott a', b' és c' bináris szavak páronként különbözőek. Ha ezt sikerül elérni, akkor a', b' és c' megkülönböztethető egy maximum $S_3^2(n)$ állapotú automatával. Ebből az átmenetek megfelelő átnevezésével készíthető egy azonos számú állapotból álló automata, ami megkülönbözteti az eredeti a, b és c szavakat. A továbbiakban azt fogom megmutatni, hogy ilyen Φ függvény létezik, vagy a három szó megkülönböztetése egy speciális esetre korlátozódik.

1. Eset: Létezik $1 \leq i \leq n$ index, amelyre az a_i, b_i, c_i betűk közül kettő megegyezik és a harmadik különböző. A szavak sorrendjének felcserélésével elérhető, hogy $a_i \neq b_i = c_i$. Tudjuk, hogy létezik egy $1 \leq j \leq n$, $j \neq i$ index, amelyre $b_j \neq c_j$. A Φ függvényt úgy készítjük el, hogy $\Phi(a_i) = 0$, $\Phi(b_i) = \Phi(c_i) = 1$ legyen. Továbbá $\Phi(b_j)$ és $\Phi(c_j)$ valamilyen sorrendben a 0 és 1 legyen, ez könnyen elérhető. A többi helyen a Φ tetszőlegesen megválasztható 0-nak vagy 1-nek.

2. Eset: $\forall 1 \leq i \leq n$ indexre az a_i, b_i, c_i betűk páronként különböznek vagy mindhárom azonos. Legyen I azon indexek halmaza, amelyekre a_i, b_i és c_i páronként különböznek. Az I halmaz nem üres, legyen $i \in I$ tetszőleges ilyen index. Ha létezik olyan $j \in I, j \neq i$ index amire $b_j = z \notin \{b_i, c_i\}$ vagy $c_j = z \notin \{b_i, c_i\}$ akkor létezik megfelelő Φ leképezés (elképzelhető, hogy $a_i = z$ igaz):

$$\Phi(x) = \begin{cases} 0 & \text{ha } x \in \{a_i, z\} \\ 1 & \text{ha } x \notin \{a_i, z\} \end{cases}$$

   Ekkor $\Phi(a_i) = 0$, $\Phi(b_i) = \Phi(c_i) = 1$, $\Phi(b_j) = 1$ és $\Phi(c_j) = 0$ vagy $\Phi(c_j) = 1$ és $\Phi(b_j) = 0$. Ha nem létezik ilyen j index, akkor minden $j \in I$-re $b_j, c_j \in \{b_i, c_i\}$.

   Az előbbi érvelés elmondható az $a_j = z \notin \{a_i, b_i\}$ vagy $b_j = z \notin \{a_i, b_i\}$ vagy $a_j = z \notin \{a_i, c_i\}$ vagy $c_j = z \notin \{a_i, c_i\}$ esetekben is, ekkor vagy hasonlóan tudunk készíteni egy megfelő Φ függvényt, vagy $a_j, b_j \in \{a_i, b_i\}$ és $a_j, c_j \in \{a_i, c_i\}$ adódik. Tehát ha egyik eset sem igaz, akkor $a_j = a_i, b_j = b_i, c_j = c_i$ igaz minden $j \in I$-re. Ekkor az (a, b, c) szóhármas nagyon speciálisan néz ki: léteznek az $x, y, z \in \Sigma$ páronként különböző betűk, hogy minden $1 \leq i \leq n$ indexre $a_i = x, b_i = y, c_i = z$ vagy $a_i = b_i = c_i$. A következő állításban be fogom látni, hogy ebben az esetben $sep_3^\kappa(a, b, c) = \mathcal{O}(\log(n))$. Az $S_3(n) \geq S_2(n) = \Omega(\log(n))$ ismert alsó korlát, amiből következik hogy ebben az esetben is $sep_3^\kappa(a, b, c) \leq S_3^2(n)$.

□

3.4. Definíció. Az (a, b, c) szóhármast unalmasnak nevezzük, ha léteznek az $x, y, z \in \Sigma$ páronként különböző betűk, hogy minden $1 \leq i \leq n$ indexre $a_i = x, b_i = y, c_i = z$ vagy $a_i = b_i = c_i$. Hívjuk x, y, z-t az a, b, c eltérési értékeinek, és azokat az i indexeket ahol $a_i = x, b_i = y, c_i = z$ eltérési helyeknek.

3.5. Állítás. Ha (a, b, c) unalmas, akkor $sep_3^\kappa(a, b, c) = \mathcal{O}(\log(n))$.

Bizonyítás. Legyenek $x, y, z \in \Sigma$ betűk az a, b, c eltérési értékei. Definiáljuk a $w \in \Sigma^n$-re $f(w) = 0 \cdot |w|_x + 1 \cdot |w|_y + 2 \cdot |w|_z$ függvényt, ahol $|w|_x, |w|_y, |w|_z$ jelöli az x, y, z betűk előfordulásinak számait a w szóban. Ha I az a, b, c eltéréseinek helyei és o az f függvény értéke az a, b, c szavak nem eltérési helyeken vett részének, akkor $f(a) = o, f(b) = o + |I|, f(c) = o + 2|I|\cdot$ egymástól eltérő 0 és $2n$ közötti értékek. A 2.1 lemma miatt létezik egy $p \leq 4, 4 \log(2n) = \mathcal{O}(\log(n))$ prím, amely nem osztja $2|I|$-t. Így $f(a), f(b), f(c)$ páronként inkongruensek modulo p, tehát egy p állapotú automatával az $f(w) \mod p$ értéket számolva megkülönböztethető a három szó. □

4. Eltérések a szavak elején

4.1. Állítás. Ha az $a_1, a_2, \dots a_l \in \{0, 2, \dots \kappa - 1\}^n$ szósorozatra igaz, hogy bármely két szó különbözik az első i hely valamelyikén, akkor $sep_l^\kappa(a_1, \dots a_l) \leq 1 + \left\lfloor \frac{l}{2} \right\rfloor (i-1) + l = 1 + \left\lceil \frac{l}{2} \right\rceil + \left\lfloor \frac{l}{2} \right\rfloor i$.

Bizonyítás. Tekintsük azt a gyökeres fenyőt, amelyben a leveleken kívül minden csúcsnak κ gyereke van és minden levél távolsága a gyökértől i. Minden nem levél csúcsból kimenő élekre írjuk a $0, 1, \dots \kappa - 1$ számokat. Minden csúcsra írjuk a gyökértől oda vezető úton lévő élekhez tartozó betűk sorozatát. Ez egy automata állapotgráfja, amely egy szót abba a levélbe visz, amire a szó i hosszú kezdőszelete van írva. A feltétel miatt az $a_1, a_2, \dots a_l$ szavakat különböző levélbe viszi.

Futtassuk az automatát az $a_1, \dots a_l$ szavakra és minden állapotra(csúcsra) számoljuk, hogy hányszor jártunk ott összesen. Azon csúcsokat amelyekben legalább kétszer jártunk nevezzük belső csúcsoknak. A belső csúcsok azon gyerekeit, amelyek nem belső csúcsok nevezzük szélső csúcsoknak. A többi csúcsot lényegtelennek nevezzük és ezeket elhagyhatjuk. Az így kapott fenyőre igaz, hogy az $a_1, \dots a_l$ szavak mindegyike egy különböző levélbe(külső csúcsba) kerül és minden levélbe kerül egy szó. Tehát ezen gráfhoz tartozó automata is megkülönbözteti az adott szavakat.

A belső csúcsok részfenyőjének maximum $\left\lfloor \frac{l}{2} \right\rfloor$ levele van. Egy levél és a gyökér közötti úton maximum $i - 2$ másik csúcs lehet. Ezen utak uniója lefedi az összes belső csúcsot, ezért azok száma maximum $1 + \left\lfloor \frac{l}{2} \right\rfloor (i - 1)$. Hozzáadva a külső csúcsok számát(l) adódik az állítás. □

5. Átlagos eset

5.1. Állítás. [2] Tegyük fel, hogy az a, b szópárt egyenletes eloszlás szerint választjuk a κ méretű abc feletti, n hosszú, különböző szavakból álló párok halmazából. Formálisan $(a, b) \in_R \{(a, b) : a, b \in \{0, 1, \dots \kappa - 1\}^n, a \neq b\}$. Ekkor $\text{sep}_2^\kappa(a, b)$ várható értéke konstans.

Bizonyítás. Annak az eseménynek a valószínűsége, hogy a és b az első $i-1$ helyen megegyezik és az i-edik helyen eltér $\left(\frac{1}{\kappa}\right)^{i-1} \cdot \left(1 - \frac{1}{\kappa}\right)$ és ebben az esetben egy $i + 2$ állapotú automatával megkülönböztethetőek. Ebből az alábbi felső becslés adódik a véletlen szópár megkülönböztetéséhez szükséges állapotok várható értékére:

$$\sum_{i \geq 1} (i + 2) \cdot \left(\frac{1}{\kappa}\right)^{i-1} \cdot \left(1 - \frac{1}{\kappa}\right) = 2 \cdot \sum_{i \geq 1} \left(\frac{1}{\kappa}\right)^{i-1} \cdot \left(1 - \frac{1}{\kappa}\right) + \sum_{i \geq 1} i \cdot \left(\frac{1}{\kappa}\right)^{i-1} \cdot \left(1 - \frac{1}{\kappa}\right)$$

A bal oldali tagban lévő szumma egy $p = 1 - \frac{1}{\kappa}$ paraméterű geometriai eloszláshoz tartozó események teljes rendszeréhez tartozó valószínűségek összege, ezért értéke 1. A jobb oldali tag egy $p = 1 - \frac{1}{\kappa}$ paraméterű geometriai eloszlás várható értéke, ami $\frac{1}{p}$. Ezeket felhasználva adódik az alábbi felső becslés:

$$2 + \frac{1}{1 - \frac{1}{\kappa}} \leq 4$$

□

5.2. Megjegyzés. Az előző bizonyításban a valószínűségeket úgy használtuk, mintha a szavakat választottuk volna egyenletes eloszlás szerint, egymástól függetlenül. Elsőre nem világos, hogy ez miért helyes, ugyanis előfordulhat, hogy ugyanazt a két szót választjuk. Ezt úgy oldhatjuk meg, hogy a véletlen válaszásra a következő módon tekintünk. Választunk egy-egy a és b szót

egyenletes eloszlás szerint, ha megegyeznek, azaz az $1 \leq i \leq n$ indexek egyikén sem térnek el, akkor a szummában valamely $i > n$ taghoz soroljuk őket és ekkor $i+2 > n+2$ állapotot szánunk a megkülönböztetésükre. A valóságban ebben az esetben a két szó nem különböztethető meg és újra kell generálni őket, amíg nem lesznek eltérőek. Így valójában annak a valószínűsége, hogy az egyes $i \leq n$ helyeken eltérnek nagyobb lesz az eredeti eloszlás szerint, mint a bizonyításban írt valószínűség. Ez nem jelent problémát ugyanis ezeket az újragenerálást kívánó eseteket a későbbi tagokban nagyobb állapotszámmal különböztettük meg.

5.3. Állítás. Ha az $a_1, a_2, \ldots a_l$ szósorozatot egyenletes eloszlás szerint véletlen választjuk a $\kappa \geq 2$ méretű abc feletti, n hosszú, páronként különböző szavakból álló l hosszú sorozatok halmazából, akkor $\mathrm{sep}_l^\kappa(a_1, \ldots a_l)$ várható értéke konstans.

6. Tesztek

Az $S_2(n)$ és $S_3(n)$ kis n értékek esetén való kiszámítására készítettem egy programot, amely a következőképpen működik. Az összes n hosszú szópár vagy szóhármas esetén meghatározza a sep_2 vagy sep_3 értéket a lehetséges nem izomorf automaták kipróbálásával és ezek maximumát veszi. A nem izomorf autamaták generálásához egy saját algoritmust használtam. Az alábbi táblázatban láthatók a kiszámolt értékek:

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$\lceil \log_2(n) \rceil$	0	1	2	2	3	3	3	3	4	4	4	4	4	4	4
$S_2(n)$	2	2	2	3	3	3	3	3	3	4	4	4	4	4	4
$S_3(n)$	-	3	3	3	4	4	4	5	5						

Hivatkozások

[1] Zachary Chase. „A new upper bound for separating words”. (2020). URL: `https://arxiv.org/abs/2007.12097`.

[2] Erik D. Demaine, Sarah Eisenstat, Jeffrey Shallit és David A. Wilson. „Remarks on Separating Words”. (2011). URL: `https://arxiv.org/abs/1103.4513v1`.

[3] P. Goralcik és V. Koubek. „On discerning words by automata”. *Lecture Notes Comput. Sci.* 226 (1986), 116–122. old.

[4] J. Shallit és Y. Breitbart. „Automaticity I: Properties of a measure of descriptional complexity.” *J. Comput. System Sci.* 53 (1996), 10–25. old.