Eötvös Loránd Tudományegyetem

Természettudományi Kar

Csányi Dávid

Szavak megkülönböztetése automatával

Szakdolgozat

Alkalmazott Matematikus Msc

Számítástudomány Szakirány

Témavezető:

Pálvölgyi Dömötör

Számítógéptudományi Tanszék

Budapest, 2023

Tartalomjegyzék

Köszönetnyilvánítás

Szeretnék köszönönetet mondani témavezetőmnek, Pálvölgyi Dömötörnek, aki sokat segített abban, hogy ez a szakdolgazat elkészülhessen. Köszönök szüleimnek, tanáraimnak, rokonaimnak és barátaimnak minden segítséget és támogatást.

Absztrakt

A szakdolgozat célja a separating words problem bemutatása, az eddig megjelent eredmények feldolgozása és azok általánosítása több szóra.

A véges automaták és felhasznált számelméleti tételek bemutatása után ismertetem a problémát és néhány egyszerűbb következtetést a szakirodalomból. Ezek közé tartozik, hogy a feladat megoldása nem függ az abc méretétől, különböző esetek, amelyekben két szó megkülönböztetése egyszerű, és az $\Omega(\log(n))$-es alsó becslés, amely az eddig ismert legjobb. Megfogalmazom a feladat általánosítását több szóra, amelyet tudomásom szerint eddig még nem vizsgáltak korábban. Az eredmények közül számosat általánosítottam több szóra, amelyek bizonyításai szintén ebben a fejezeteben olvashatók.

A harmadik részben bemutatok kettő, az $\tilde{\mathcal{O}}(\sqrt{n})$-es felső korlátra korábban megjelent érvelést. Bebizonyítom az általam elért fő eredményt, hogy az $\tilde{\mathcal{O}}(\sqrt{n})$-es felső becslés több szó megkülönböztetése esetén is igaz, ha azok száma konstans. Ezután leírom a két szó megkülönböztetésére eddig ismert legjobb, $\tilde{\mathcal{O}}(n^{1/3})$-os felső korlát bizonyításának lépéseit az eredeti cikk alapján.

A separating words problem többféle változatát is megfogalmazták és vizsgálták már, amelyek közül összegyűjtök néhányat. Ebben a részben a feladatok megfogalmazásán kívül néhány érdekes eredményt említek meg.

Az utolsó fejezetben ismertetem egy általam írt program alapján, hogy kis n-ekre legrosszabb esetben hány állapot kell kettő és három szó szétválasztásához.

1. A háttér

1.1. Véges automaták

Ebben a fejezetben bevezetem a véges automatákat és néhány ide kapcsolódó definíciót. Ezután megválaszolok pár kérdést, amelyek az automaták izomorfiájával és számával kapcsolatosak.

1.1. Definíció. Az $A = (Q, \Sigma, \delta, q_0, F)$ ötöst determinisztikus **véges automatá**nak (**DFA**) nevezzük, ha Q egy véges halmaz (állapothalmaz), Σ véges abc, $\delta : Q \times \Sigma \mapsto Q$ átmenetfüggvény, $q_0 \in Q$ kezdőállapot és $F \subseteq Q$ a végállapotok.

A véges automata a futtatása során beolvas egy $a \in \Sigma^n$ szót és a Q állapotai között lépked. Kezdetben a q_0 kezdőállapotban van. A szó következő $\sigma \in \Sigma$ betűjét beolvasva az adott betű és állapot (q) alapján átkerül egy új állapotba, amelyet a $\delta(q, \sigma)$ átmenetfüggvény határoz meg.

1.2. Megjegyzés. Egy automatát elérhetőnek nevezünk, ha minden $q \in Q$ állapothoz létezik egy $w \in \Sigma^*$ szó, hogy az automata a w szót olvasva a q állapotba kerül. A nem elérhető állapotok Q-ból való törlésével minden automata elérhetővé tehető, és a továbbiakban csak ilyenekkel fogunk foglalkozni.

1.3. Definíció. Az A véges automata **átmenet gráf**ja az a G irányított gráf, amelynek csúcsai az automata állapotai, és élei az átmenetfüggvény által meghatározott $\{e_{q,\sigma} = (q, \delta(q, \sigma)) : q \in Q, \sigma \in \Sigma\}$ halmaz elemei.

1.4. Definíció. Az $A_1 = (Q_1, \Sigma, \delta_1, q_{01}, F_1)$ és $A_2 = (Q_2, \Sigma, \delta_2, q_{02}, F_2)$ véges automatákat **izomorf**nak nevezünk, ha $|Q_1| = |Q_2|$ és létezik egy $b : Q_1 \mapsto Q_2$ bijekció, amelyre $\forall q \in Q_1, \forall \sigma \in \Sigma$ esetén $b(\delta_1(q, \sigma)) = \delta_2(b(q), \sigma), b(q_{01}) = b(q_{02})$ és $b(F_1) = b(F_2)$.

Tehát ha két automata izomorf, az azt jelenti, hogy ugyanúgy néznek ki és működnek, csak az állapotaik máshogyan vannak elnevezve.

1.5. Kérdés. Hány k állapotú, egymással nem izomorf véges automata van a $\Sigma = \{0, 1\}$ abc felett, ha a végállapotokat nem vesszük figyelembe?

1.6. Definíció. Jelölje $\boldsymbol{\beta_k}$ azon véges és elérhető automaták számát, melyekre $\Sigma = \{0, 1\}, Q = \{0, 1, \ldots k-1\}, q_0 = 0, F = \emptyset$. Jelölje $\boldsymbol{\alpha_k}$ ezen automaták közül az egymással nem izomorfak számát.

A δ átmenetfüggvény leírható egy $2 \times k$-s T táblázattal, ahol $T[\sigma, q] = \delta(q, \sigma)$. A táblázat minden helyére a Q halmaz valamelyik elemét kell írnunk, így k^{2k}-féle eredményt kaphatunk. Előfordulhat, hogy az így kapott δ-hoz tartozó átmenet gráfban nem minden csúcs érhető el q_0-ból(azaz az automata nem elérhető), ekkor a nem elérhető állapotokat elhagyva egy k-nál kevesebb állapotú automatát kapunk. Ilyen módon minden maximum k állapotú automata előáll, tehát k^{2k} egy felső becslés nemcsak a pontosan k, hanem a maximum k állapotú automaták számára is. Azaz $\sum_{i=1}^{k} \beta_k \leq k^{2k}$.

1.7. Állítás. $\beta_1 = 1$ és $\beta_k = k^{2k} - \sum_{i=1}^{k-1} \binom{k-1}{i-1} k^{2k-2i} \beta_i$, ha $k > 1$.

Bizonyítás. A T táblázat 2^{2k} féle kitöltése megadja a lehetséges automatákat, ezek számából a nem elérhetőek számát kell kivonnunk. Jelölje az i változó azt, hogy a kezdőállapotból hány állapot érhető el. Az i lehetséges értékei egy nem elérhető automata esetén $1, 2, \dots k-1$. Adott i-re a kezdőállapoton kívüli másik $i-1$ elérhető állapot címkéjét $\binom{k-1}{i-1}$ féleképpen választhatjuk ki. Ezen állapotokon az elérhető automaták száma β_i. A többi állapotra a δ tetszőlegesen definiálható, ezek lehetséges száma k^{2k-2i}. Ezen értékeket összeszorozva megkapjuk az i darab állapotot elérő automaták számát, amely $\binom{k-1}{i-1}k^{2k-2i}\beta_i$. Ezeket kivonva a 2^{2k}-ból adódik a β_k értéke. □

1.8. Definíció. Egy $A = (Q, \Sigma, \delta, q_0, F)$ automata és $b : Q \mapsto Q'$ bijekció esetén bevezethetjük a $\mathbf{b(A)} = (Q', \Sigma, \delta', b(q_0), b(F))$ jelölést, ahol $\delta'(q', \sigma) = b(\delta(b^{-1}(q'), \sigma))$.

1.9. Állítás. $\alpha_k = \dfrac{\beta_k}{(k-1)!}$

Bizonyítás. A $Q = \{0, 1, \dots k-1\}$ állapothalmazon lévő automatákat vizsgáljuk, amelyek kezdőállapota a 0. Ha egy A és egy B automata izomorf, akkor létezik egy $b : \{0, \dots k-1\} \mapsto \{0, \dots k-1\}$ bijekció az állapotaik között, amelyre igazak a megfelelő tulajdonságok. Igaz az is, hogy $b(0) = b(q_{A0}) = q_{B0} = 0$, emiatt maximum $(k-1)!$ lehet az ilyen bijekciók száma. Ebből következik, hogy minden izomorfiaosztály mérete maximum $(k-1)!$ lehet.

Azt szeretnénk még belátni, hogy minden izomorfiaosztály mérete pontosan $(k-1)!$. Azaz egy tetszőleges $b : \{0, \dots k-1\} \mapsto \{0, \dots k-1\}$ bijekcióra, amely nem az identitás, és amelyre $b(0) = 0$ és tetszőleges A automatára igaz, hogy A és $b(A)$ különbözőek (de izomorfak).

Könnyen láthatjuk, hogy ha A és $b(A)$ megegyeznek, akkor b-nek az identitásnak kell lennie. A kezdőcsúcsból kilépve a 0 élen nevezzük q-nak azt az állapotot, amibe kerülünk. A és $b(A)$ esetén ez ugyanaz a q, mert az automaták megegyeznek. (Esetleg lehetséges, hogy $q = q_0$.) Ekkor az alábbi összefüggést írhatjuk fel: $b(q) = b(\delta(q_0, 0)) = \delta'(q'_0, 0) = q$. Azaz $b(q) = q$. Ezt az érvelést ismételjük egymás után többször a kilépő 0 és 1 élekre is az olyan állapotok esetén, amelyekre $b(q) = q$-t már beláttuk. Az automata elérhetőségét felhasználva azt kapjuk, hogy b az identitás.

Ezekből adódik, hogy minden izomorfiaosztály mérete $(k-1)!$ és $\alpha_k = \dfrac{\beta_k}{(k-1)!}$. □

Az α_k sorozat első néhány eleme az alábbi táblázatban látható és a `https://oeis.org/search?q=1%2C+12%2C+216%2C+5248%2C+160675%2C+5931540&sort=&language=english&go=Search` oldalon további információk találhatók róla.

k	1	2	3	4	5	6
α_k	1	12	216	5248	160675	5931540

1.10. Kérdés. Hogyan tudjuk legenerálni az egymással nem izomorf k állapotú automatákat, melyek száma α_k?

Egy automatából készítsük el az átmenet gráfját, amit nevezzünk G-nek. Írjuk fel a q_0-nak megfelelő csúcsot, ez lesz s, a többi csúcs nevét felejtsük el. Jegyezzük

meg, hogy melyik élhez a Σ abc melyik betűje tartozik. Ekkor egy $(G, s, e \mapsto \sigma(e))$ hármast kapunk, amely egyértelműen meghatározza, hogy az automata hogyan működik bármely inputon. Könnyen látható, hogy izomorf automatákhoz megegyező, nem izomorf automatákhoz különböző ilyen fog tartozni.

Járjuk be a G-t az $s = 0$ csúcsból indulva szélességi kereséssel úgy, hogy minden csúcsból először azon az e élen lépünk ki, amelyre $\sigma(e) = 0$, majd a másikon. Ha egy új csúcsba érünk, akkor azt nevezzük el a legkisebb olyan egész számmal, amelyet még nem használtunk fel. Ez a számozás az adott hármashoz egyértelműen hozzárendel egy automatát, amely a reprezentánsa lesz. Így olyan automatákat kapunk, amelyek az egyes izomorfiaosztályokat reprezentálják.

A szélességi keresés annak felel meg, hogy a T táblázatot balról jobbra haladva töltjük ki, minden oszlopban először a felső majd az alsó sorhoz tartozó értéket írva be. Ha az eddig beért értékek a $\{0, 1 \dots \max\}$ halmaz elemei, akkor az aktuális mezőbe a $\{0, 1 \dots \max +1\}$ értékek közül választhatunk. A $\max +1$-et akkor írjuk, ha a szélességi bejárás egy új csúcsot lát meg, ellenkező esetben egy korábbi csúcsba lépett vissza. Ha a kitöltés során a j-edik oszlophoz érünk és a $\max < j$, akkor a 0 kezdőállapotból nem tudunk elérni a j-be, ezért egy nem elérhető automatát fogunk kapni,

Összefoglalva az eddigieket azt mondhatjuk, hogy az egymással nem izomorf, k állapotú automatákat az alábbi módon kaphatjuk meg:

1. Egy $2 \times k$-s T táblázatot kell kitöltenünk, melyek sorindexei $0, 1$, oszlopindexei $0, 1 \dots k-1$.
2. Minden helyre a $\{0, 1 \dots k-1\}$ értékek valamelyikét írhatjuk.
3. A kitöltésnél az oszlopok sorban következnek balról jobbra, először mindig a felső majd az alsó pozíciójuk.
4. A soron következő mezőbe nem írhatunk $\max +1$-nél nagyobb számot.
5. A $0, 1 \dots j-1$ oszlopok valamelyikében szerepelni kell j-nek, minden $j \in \{1, \dots k-1\}$-re.

1.11. Eljárás (Véges automaták generálása)**.** Ezen észrevételek segítségével adhatunk egy hatékony algoritmust, amely végigjárja az egyes izomorfiaosztályok reprezentánsait. A T táblázatot most egy $1 \times 2k$-s A tömbnek fogjuk képzelni, ahol $T[\sigma, q] = A[2q + \sigma]$. Az algoritmus az alábbi két fő lépésből áll:

1. $a_1, \dots a_{k-1}$ azok a helyek ahol a max értéke növekszik, az a_i helyen i-re. Ezekre teljesülnek az alábbi egyenlőtlenségek:

   $0 \le a_1 \le 1$

   $a_{i-1} < a_i \le 2i - 1$

   Az a_i növekedési helyek meghatározása után az $A[a_i] = i$ értékeket beállítjuk.

2. Ha $0 \le a \le 2k-1$ és $a \notin \{a_1, \dots a_{k-1}\}$, akkor valamely $i \in \{1, 2, \dots k\}$-re $a_{i-1} < a < a_i$. (Itt $a_0 = -1$ és $a_k = 2k$.)

   Ekkor $A[a] \in \{0, \dots i-1\}$ választható.

A lehetséges növekedési helyek sorozatai sorbaállíthatóak, és egy elemből a következő könnyen meghatározható. Hasonlóan a növekedési helyek közötti mezők lehetséges kitöltésein is egymás után tudunk lépkedni. Így lényegében a külső ciklus a lehetséges növekedési helyeken iterál és a belső ciklus az adott növekedési helyek közötti mezők lehetséges kitöltésein.

Ezen algoritmus segítségével legenerálható az összes egymással nem izomorf k állapotú, elérhető véges automata. Ezt később arra fogom használni, hogy ezen automaták kipróbálásával meghatározom a separating words problem értékeit kis n-ek esetén. Az elért eredményeket az 5. fejezetben mutatom be.

1.2. Felhasznált számelméleti állítások

A separating words problemre adott korlátokhoz sokszor számelméleti tételeket fogunk használni. Ebben a fejezetben ezeket az állításokat mondom ki bizonyítások nélkül.

Egy számot nem osztó prímről

1.12. Lemma (Shallit, Breitbart [10])**.** 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 $\theta(x) \geq 0,23x$ ([9] Rosser, Schoenfeld, 10. Tétel). 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. □

1.13. 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 $i \not\equiv j \pmod{p}$.

Prímszámtétel

1.14. Definíció. $\mathbf{f(x) \sim g(x)}$, ha $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1$. Ilyenkor azt mondjuk, hogy az f és a g aszimptotikusan egyenlő.

1.15. Definíció. Jelölje $\boldsymbol{\pi}(\mathbf{x})$ az x-nél nem nagyobb pímek számát.

1.16. Tétel (Prímszámtétel)**.** $\pi(x) \sim \frac{x}{\ln(x)}$.

Az alábbi tétel ekvivalens a prímszámtétellel:

1.17. Tétel. Jelölje p_n az n-edik prímszámot. Ekkor $p_n \sim n\ln(n)$.

A prímszámtétel kettő következményét fel fogjuk használni a későbbiekben:

1.18. Következmény. $\sum_{i=1}^{k} p_k \sim \frac{1}{2}\frac{k^2}{\ln k}$.

1.19. Következmény. Legyen $0 < \alpha < 1$ és $n \in \mathbb{Z}^+$. Az $n^{1-\alpha}$ utáni $\frac{2n^{1-\alpha}}{1-\alpha}$-adik prím nem nagyobb $c(\alpha)n^{1-\alpha}\log(n)$-nél, ahol $c(\alpha)$ az α-tól függő konstans.

2. Szavak megkülönböztetése

1986-ban Goralčík és Koubek [6] felvetette azt a kérdést, hogy egy egyszerű számítási eszközzel hogyan lehet megoldani az elképzelhető legkönnyebb feladatot: két szó megkülönböztetését. Ezt a kérdést fogom körüljárni a Remarks on separating words (Demaine, Eisenstat, Shallit, Wilson [5]) cikket követve és általánosabban (több szó esetén) is megvizsgálni. A kettő szóra vonatkozó állítások bizonyításai az [5] cikk alapján történnek, az ennél több szóra vonatkozó állítások saját eredmények.

2.1. Definíció. Jelölje $\mathbf{sep_l(a_1, \dots a_l)}$ 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.

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

A separating words problem célja az $S_2(n)$ függvényre alsó és felső korlátokat találni. Az általánosított feladat esetén $S_l(n)$-t szeretnénk vizsgálni. Az $l = 2$ esetben a legjobb ismert felső korlát $S_2(n) = \tilde{\mathcal{O}}(n^{\frac{1}{3}})$, amelyet Chase bizonyított be 2020-ban [4]. A legjobb ismert alsó korlát $S_2(n) = \Omega(\log(n))$. Tudomásom szerint az $l > 2$ esetben még nem jelent meg eredmény ezen a területen.

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

2.4. 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. □

2.1. Különböző hosszú szavak

Ebben a részben meggondoljuk kettő és három szó esetében, hogy ha a szavak hossza eltérő, akkor a megkülönböztetésük nem nehéz. Emiatt a továbbiakban feltehetjük, hogy a feladatban $l \in \{2, 3\}$ esetén minden szó hossza n.

2.5. Definíció. Az a **szó hosszá**t a továbbiakban $|a|$ jelöli.

2.6. Állítás ([5])**.** Ha $|a|, |b| \leq n$ és $|a| \neq |b|$, akkor $\mathrm{sep}_2(a, b) = \mathcal{O}(\log(n))$.

Bizonyítás. A 1.13 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. □

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

Bizonyítás. $|(|a|-|b|)\cdot(|b|-|c|)\cdot(|c|-|a|)| \leq n^3$, ezért a 1.12 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 $|c|$ különböző maradékokat adnak p-vel osztva. Ezért egy p hosszú kört tartalmazó automata más végállapotokba viszi őket. □

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

Bizonyítás. Ha mindhárom szó különböző hosszú, akkor az előző (2.7) állítás miatt igaz. Ha két szó egyforma hosszú és a harmadik különböző, akkor a sep_2 és sep_3 közötti egyenlőtlenség(2.4) és a sep_2 különböző hosszú szavak esetére vonatkozó 2.6 állítás miatt vagyunk készen. □

2.9. Következmény. Az $S_2(n)$ további vizsgálatánál feltehetjük, hogy $|a| = |b|$ és az $S_3(n)$ esetén, hogy $|a| = |b| = |c|$.

2.2. 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.

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

2.11. Állítás ([5])**.** $S_2^\kappa(n) = S_2^2(n)$, 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 átmenetfüggvényeiben 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. □

Most megmutatom, hogy ez igaz három szóra is. A bizonyítás ugyanazon az ötleten múlik, de kompikáltabb az esetek szétválasztása miatt.

2.12. 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 2.11 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 $S_3(n) \geq S_2(n) = \Omega(\log(n))$ alsó korlát miatt(bizonyítása az alsó korlátok részben olvasható (2.30)) következik hogy ebben az esetben is $sep_3^\kappa(a, b, c) \leq S_3^2(n)$.

□

2.13. Definíció. Az (a, b, c) szóhármast **unalmas**nak 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.

2.14. 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 1.12 lemma miatt létezik egy $p \leq 4, 4 \log(n) = \mathcal{O}(\log(n))$ prím, amely nem osztja $|I|$-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. □

2.15. Definíció. Az $(a_1, a_2, \ldots a_l)$ szósorozatot **monoton**nak nevezzük, ha $a_1[i] \leq a_2[i] \leq \ldots \leq a_l[i]$ igaz minden $1 \leq i \leq n$ indexre, ahol $a_j[i]$ az a_j szó i-edik betűjét jelöli. Ha ezen felül a sorozat tagjai páronként különböznek, akkor szigorúan monotonnak nevezzük.

2.16. Megjegyzés. Ha az $(a_1, \ldots a_l)$ szósorozat szigorúan monoton, akkor az előző bizonyításhoz hasonlóan egy f függvény segítségével megkülönböztethetők az elemei.

A belátott eredmények alapján a továbbiakban feltehetjük, hogy $\Sigma = \{0, 1\}$, ha $l \in \{2, 3\}$.

2.3. Eltérések a szavak elején

2.17. Állítás. Tegyük fel, hogy az $a_1, a_2, \ldots a_l \in \{0, 1, \ldots \kappa - 1\}^n$ szósorozatra igaz, hogy bármely két szó különbözik az első i hely valamelyikén. Ekkor $sep_l^\kappa(a_1, \ldots a_l) \leq 1 + \lfloor \frac{l}{2} \rfloor (i - 1) + l \approx \frac{l}{2}(i + 1)$.

2.18. Megjegyzés. Az állításban szereplő becslés tovább javítható az alábbi módon. Legyen g az az egész szám, amelyre $\kappa^{g-1} < \lfloor \frac{l}{2} \rfloor \leq \kappa^g$. Ekkor $sep_l^\kappa(a_1, \ldots a_l) \leq 1 + \kappa + \kappa^2 + \cdots + \kappa^{g-1} + \lfloor \frac{l}{2} \rfloor (i - g) + l \leq 1 + \lfloor \frac{l}{2} \rfloor (i - 1) + l$.

2.19. Megjegyzés. Ha az $a_1, \ldots a_l \in \{0, 1, \ldots \kappa - 1\}^n$ szósorozatra igaz, hogy bármely két szó különbözik az első i hely valamelyikén, akkor $l \leq \kappa^i$.

2.20. Állítás (2.17 speciális esete $l = 2, \kappa = 2$)**.** Ha az a és b bináris szavak különböznek az i-edik helyen, akkor $\text{sep}(a, b) \leq i + 2$.

Bizonyítás. (2.17 Állítás) 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 $0, 1, \ldots \kappa - 1$ 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 $a_1, a_2, \ldots a_l$ szavakat különböző levélbe viszi.

Futtassuk az automatát az $a_1, \ldots a_l$ 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 külső 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 $\lfloor \frac{l}{2} \rfloor$ 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 $1 + \lfloor \frac{l}{2} \rfloor (i-1)$. Igaz továbbá az is, hogy a gyökér alatti első szinten maximum κ csúcs van, az az alatti(második) szinten maximum κ^2 és így tovább, a $g-1$-edik szinten legfeljebb κ^{g-1} csúcs van. (g az az egész szám, amelyre $\kappa^{g-1} < \lfloor \frac{l}{2} \rfloor \leq \kappa^g$.) Az ez alatti szinteken már lehet $\lfloor \frac{l}{2} \rfloor$ darab csúcs. Így a belső csúcsokra korábban adott becslés $1+\kappa+\kappa^2+\dots\kappa^{g-1}+\lfloor \frac{l}{2} \rfloor (i-g)$-re javítható. Hozzáadva a külső csúcsok számát(l) adódik az állítás. □

2.4. Eltérések a szavak végén

2.21. Állítás. Ha az $a_1, a_2, \dots a_l \in \{0, 2, \dots \kappa - 1\}^n = \Sigma^n$ szósorozatra igaz, hogy bármely két szó különbözik az utolsó i hely valamelyikén, akkor $sep_l^\kappa(a_1, \dots a_l) \leq \sum_{j=0}^{i} \min(\kappa^j, l) \leq 1 + i \cdot l$.

Bizonyítás. A bizonyításban a végszelet felismerő automata ismert gondolatát fogom általánosítani több szóra.

Legyen q_0 a kezdőállapot és minden $j = 1, \dots i$-re az összes j hosszú szó legyen egy állapot. Ezek együttes száma $\sum_{j=0}^{i} \kappa^j = \frac{\kappa^{i+1}-1}{\kappa-1}$. Úgy képzeljük el, hogy a j hosszú szavakhoz tartozó állapotok a j-edik szinten helyezkednek el, és összesen $i+1$ szint van(beleértve a nulladikat). Egy $w \in \Sigma^j$ csúcsban($0 \leq j \leq i-1$) a $\sigma \in \Sigma$ betűt olvasva kerüljünk át a $w\sigma$ csúcsba, ezen átmenetekhez tartozó éleket húzzuk be a gráfba.

Az i-edik szinten tekintsük azt az l darab csúcsot, amelyek az $a_1, \dots a_l$ szavak i hosszú végszeletei, ezeket jelöljük meg. Ha egy csúcs nincs rajta a q_0-ból valamely jelölt csúcsba vezető úton, akkor töröljük ki. Így minden szinten maximum l darab csúcs marad, és a csúcsok együttes számára teljesül az állításban kimondott $\sum_{j=0}^{i} \min(\kappa^j, l)$ korlát.

Még nem definiáltuk az átmeneteket az i-edik szinten lévő csúcsokból. Továbbá kitöröltünk bizonyos csúcsokat, és a korábban ezekbe vezető élek végpontjait is újra meg kell határoznunk. Ha w egy csúcs és a $\sigma \in \Sigma$ betűre még nincs meghatározva, hogy melyik legyen a következő csúcs, akkor tegyük a következőt. Legyen u a $w\sigma$ szó azon leghosszabb végszelete, amely egy létező csúcs. A w csúcsból a σ-t olvasva lépjünk u-ba. (Megfigyelhetjük, hogy ha a w csúcsból a σ-t olvasva már meghatározott az átmenet, akkor a $w\sigma$ leghosszabb végszelete maga $w\sigma$ és ez a következő csúcs, amibe lépünk.) Könnyen meggondolható, hogy egy a szót olvasva az automata pontosan akkor kerül egy i-edik szinten lévő v csúcsba, ha az a i-hosszú végszelete pontosan v. Ebből következik, hogy az automata elválasztja az $a_1, \dots a_l$ szavakat. □

2.5. Mintaillesztés

Az előző bizonyításban készített automata szorosan kapcsolatban áll a mintaillesztési feladat egyik legismertebb megoldásával, ezért röviden bemutatom a feladatot és ezt az összefüggést.

A mintaillesztés (string matching) feladatban adott egy s és egy t szó a Σ abc felett. A cél, hogy megtaláljuk a t szó előfordulásait részszóként az s szóban. A kérdés általánosítható több szóra, ilyenkor a $t_1, \dots t_l$ szavak mindegyikének előfordulásait keressük az s szóban.

A mintaillesztésre többféle eljárás is ismert, ezek közül az egyik leggyakrabban használt a Knuth-Morris-Pratt algoritmus, amely egy véges automatát használ. Az eltérés a szavak végén című részben belátott 2.21 Állítás segítségével a KMP algoritmus általánosítható. Így egy lineáris idejű módszert kapunk, amely egyszerre több szó előfordulásait találja meg egy szövegben. Ez az Aho-Corasick algoritmus.

2.22. Eljárás (Aho-Corasick algoritmus)**.** [1] A 2.21 állítás bizonyításához hasonlóan készítünk egy automatát, amely a $t_1, \dots t_l$ szavakat ismeri fel végszeletként. Ezután a kapott automatát futtatjuk az s szóra. Ha egy olyan állapotba érünk, amely valamely t_j szóhoz tartozik, akkor megjegyezzük, hogy az adott helyen megtaláltuk a t_j szó egy előfordulását, majd folytatjuk a futtatást. Így az s szó feldolgozása lineáris időben megtehető. Az automata elkészítése a $t_1, \dots t_l$ szavak hosszainak összegében lineáris idő alatt megtehető, további információ található például az alábbi oldalon: [1].

2.6. Átlagos eset

2.23. Állítás ([5])**.** 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 $\mathrm{sep}_2^\kappa(a,b)$ 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 (2.20 állítás). 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:

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

□

2.24. 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 választásra 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.

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

Bizonyítás. Legyen A_i az az esemény, hogy bármelyik két szó az első i hely valamelyikén eltér, azonban ez már nem igaz $(i-1)$-re (azaz van két szó, amely az első $i-1$ helyen azonos). Ha az A_i esemény következett be, akkor egy $1 + \left\lceil \frac{l}{2} \right\rceil + \left\lfloor \frac{l}{2} \right\rfloor i$ állapotú automatával megkülönböztethetőek a szavak (a 2.17 állítás miatt). A keresett várható értékre egy felső becslés a $\sum_{i \geq 1}(1 + \left\lceil \frac{l}{2} \right\rceil + \left\lfloor \frac{l}{2} \right\rfloor i)\mathbb{P}(A_i)$.

Ha az A_i esemény következett be, akkor valamely két szó megegyezik az első $i-1$ helyen a definíció miatt. Ezen valószínűségre felső becslés ha az összes $\binom{l}{2}$ lehetséges szópárra összeadjuk azt a valószínűséget, hogy két szó az első $i-1$ helyen megegyezik.

$$\mathbb{P}(A_i) \leq \binom{l}{2} \left(\frac{1}{\kappa}\right)^{i-1} \left(1 - \frac{1}{\kappa}\right)$$

Ezeket felhasználva az alábbi felső becslés adódik:

$$\begin{aligned} &\sum_{i \geq 1}(1 + \left\lceil \frac{l}{2} \right\rceil + \left\lfloor \frac{l}{2} \right\rfloor i) \cdot \binom{l}{2} \cdot \left(\frac{1}{\kappa}\right)^{i-1} \cdot \left(1 - \frac{1}{\kappa}\right) \\ &= \binom{l}{2} \cdot \left(1 + \left\lceil \frac{l}{2} \right\rceil + \left\lfloor \frac{l}{2} \right\rfloor \frac{1}{1 - \frac{1}{\kappa}}\right) \leq \binom{l}{2} \left(1 + \frac{3}{2}l\right) \end{aligned}$$

□

2.7. Részszavak száma

2.26. Definíció. Jelölje $|a|_w$ a w **szó előfordulásainak számát** részszóként az a szóban.

2.27. Állítás ([5])**.** Ha $a, b \in \{0,1\}^n$ és $|a|_w \neq |b|_w$, akkor $\mathrm{sep}(a,b) = \mathcal{O}(|w| \log(n))$.

Bizonyítás. $|a|_w, |b|_w \leq n$, ezért a 1.13 következmény miatt létezik egy $p = \mathcal{O}(\log(n))$ prím, amelyre $|a|_w \not\equiv |b|_w \mod p$. Egy $|w| \cdot p$ állapotú automatával az $|a|_w \mod p$ értéket számolva megkülönböztethető a két szó. Ez p darab részautomatából áll, amelyek a w végszelet felismerő automatái. Ezekről a 2.21 részben olvashatunk többet. Ha valamelyik részautomata végállapotában vagyunk, akkor átlépünk a következő részautomata megfelelő állapotába (abba az állapotba, amelyikbe a végszelet felismerő automata lép a végállapotából a megfelelő betűt olvasva).

□

2.8. Hamming távolság

Az előző alfejezetben beláttuk, hogy ha a két szó bizonyos értelemben különböző, akkor könnyen meg lehet őket különböztetni. Most azt vizsgáljuk meg, hogy ha nagyon hasonlóak, akkor is szétválaszthatók kevés állapottal.

2.28. Definíció. Az a és b szavak **Hamming-távolság**a, azaz $H(a, b)$ azon $1 \leq i \leq n-1$ indexek száma, ahol a két szó eltér egymástól.

2.29. Állítás ([5])**.** Ha $H(a, b) \leq d$, akkor $\text{sep}(a, b) = \mathcal{O}(d \log(n))$.

Bizonyítás. Jelölje $i_1 < i_2 < \cdots < i_d$ azokat az indexeket, ahol a két szó eltér. Ekkor az $N = (i_2 - i_1) \cdot (i_3 - i_1) \ldots (i_d - i_1) \leq n^{d-1}$ egyenlőtlenség és a 1.12 lemma miatt létezik egy $p = \mathcal{O}(\log(n^{d-1})) = \mathcal{O}(d \log(n))$ prím, amelyre igaz, hogy $p \nmid N$.

Legyen $a_{p,i} = \sum_{j \equiv i \mod p} a_j \mod 2$. Ekkor $a_{p,i_1} \neq b_{p,i_1}$, ugyanis az a és b szavak eltérnek az i_1 helyen és az összes többi $j \equiv i_1 \mod p$ helyen megegyeznek. (Ellenkező esetben $p|(j - i_1)|N$ ellentmondás lenne.) Tehát az a_{p,i_1} és b_{p,i_1} definíciójában szereplő szummák értéke eggyel tér el, ezért mod 2 nem egyenlőek.

Az a_{p,i_1} értékek egy $2p$ állapotú automatával megszámolhatóak modulo 2. Az automata kettő p hosszú körből áll, melyek állapotai 0-tól $p-1$-ig vannak számozva. Ha valamely kör i_1-edik állapotában 1-est olvas az automata, akkor átlép a másik kör $i_1 + 1$-edik állapotába. Különben ugyanazon a körön lép tovább a következő állapotba. A futás végén az a_{p,i_1} érték attól függ, hogy a nulladik vagy az első körön vagyunk. □

2.9. Alsó korlát

Az alábbi állítás az eddig ismert legjobb alsó korlátot mondja ki a separating words problem-re.

2.30. Állítás ([5])**.** $S_2(n) = \Omega(\log(n))$

Bizonyítás. Legyen $a = 0^{k-1}1^{k-1+\text{lcm}(1,\ldots k)}$ és $b = 0^{k-1+\text{lcm}(1,\ldots k)}1^{k-1}$, ahol $\text{lcm}(1, \ldots k)$ az $1, \ldots k$ számok legkisebb közös többszöröse. Ekkor semelyik maximum k állapotú automata nem különbözteti meg a-t és b-t. Ugyanis valamely $x \in \Sigma$ betűre és $q \in Q$ állapotra a $p_i = \delta(q, x^i)$ állapotokat vizsgálva $i = 0, 1, \ldots$ esetén megállapíthatjuk,

hogy azok maximum $k-1$ állapot után ciklizálni fognak k-nál nem nagyobb periódusszámmal. Ebből következik, hogy bármely automata ugyanabba az állapotba fog kerülni, miután beolvasta az a vagy a b szó elején lévő nullákat, majd ugyanabba a végső állapotba fog kerülni, miután beolvasta az egyeseket. Tehát semelyik maximum k állapotú automata nem különböztetheti meg ezen $n = 2(k-1)+\mathrm{lcm}(1,\dots k)$ hosszú szavakat. A prímszámtétel miatt $\mathrm{lcm}(1,\dots k) = e^{k(1+o(1))}$, amiből következik, hogy $S_2(n) = \Omega(\log(n))$. □

3. Felső korlátok

Ebben a fejezetben bemutatok az $S_2(n)$-re adott felső becslések közül néhányat, köztük az eddig ismert legjobb $\tilde{\mathcal{O}}(n^{1/3})$-os korlátot is (Chase [4]). Bebizonyítom továbba a saját eredményemet, hogy az $\tilde{\mathcal{O}}(\sqrt{n})$-es korlát tetszőleges konstans l esetén is igaz $S_l(n)$-re.

3.1. Gyökös felső becslés

Először egy egyszerűbb $\tilde{\mathcal{O}}(\sqrt{n})$-es felső korlátot ismerhetünk meg, melynek bizonyításához a körosztási polinomok egyes tulajdonságait fogjuk használni.

3.1. Definíció. Legyen $a = a_0 \dots a_{n-1} \in \{0,1\}^n$ és $w = w_0 w_1 \dots w_{l-1} \in \{0,1\}^l$. Ekkor $\mathbf{pos_w(a)}$-val jelöljük a w részszó kezdőpozícióinak halmazát az a szóban. Formálisan $\mathrm{pos}_w(a) = \{j : 0 \le j \le n-l, a_j = w_0, \dots a_{j+l-1} = w_{l-1}\}$.

3.2. Definíció. Legyen $m \in \mathbb{Z}^+$ és i egy maradékosztály modulo m. Ekkor jelöljük $\mathbf{pos_w(a)_{m,i}}$-vel a w részszó azon j kezdőpozícióinak halmazát, amelyekre $j \equiv i \mod m$.

Ha az a és b bináris szavak különbözőek, akkor $\mathrm{pos}_1(a) \neq \mathrm{pos}_1(b)$. Következő lépésként azt fogjuk belátni, hogy ha egy p prímre és egy i maradékosztályra igaz, hogy $|\mathrm{pos}_1(a)_{p,i}| \neq |\mathrm{pos}_1(b)_{p,i}|$, akkor az a és b szavakat meg tudjuk különböztetni egy automatával. Ezután meggondoljuk, hogy létezik nem túl nagy ilyen p és hozzá megfelelő i.

3.3. Tétel (Chase [4])**.** $S_2(n) = \mathcal{O}(\sqrt{n} \cdot \log(n)^{1.5})$.

Bizonyítás. Tegyük fel, hogy találtunk egy p prímet, i egy maradékosztályt, melyekre $|\mathrm{pos}_1(a)_{p,i}| \neq |\mathrm{pos}_1(b)_{p,i}|$. (A következő állításban fogjuk megmutatni, hogy tudunk találni ilyen p-t és i-t.) Tudjuk, hogy $|\mathrm{pos}_1(a)_{p,i}|, |\mathrm{pos}_1(b)_{p,i}| \le n$ és a 1.13 következmény miatt létezik egy $q = \mathcal{O}(\log(n))$ prím, amire a két érték osztási maradéka különböző. Készítsünk egy olyan automatát, amely egy a szó esetén a $|\mathrm{pos}_1(a)_{p,i}| \mod q$ értéket számolja. Ezt megtehetjük, ha q darab p hosszú kört használunk és ha egy kör i-edik állapotában 1-est olvasunk, akkor a következő kör $i+1$-edik állapotába lépünk, egyébként ugyanazon a körön a következő állapotba. Ez egy pq állapotú automata, amely a $|\mathrm{pos}_1(a)_{p,i}| \mod q$ értéket számolja és emiatt megkülönbözteti a két szót. A következő számelméleti állítás alapján létezik egy p prím, amely $\mathcal{O}(\sqrt{n\log(n)})$ méretű és hozzá egy i maradékosztály, amelyek megfelelőek (azaz $|\mathrm{pos}_1(a)_{p,i}| \neq |\mathrm{pos}_1(b)_{p,i}|$). Ebből következik, hogy bármelyik kettő n hosszú bináris szóhoz létezik egy $pq = \mathcal{O}(\sqrt{n}\log(n)^{1.5})$ állapotú automata, amely megkülönbözteti őket. □

Lássuk be a bizonyítás teljességéhez hiányzó részt, azaz hogy létezik egy megfelelő p prím és egy i maradékosztály. Ezt az automatákról megfeledkezve, egy A és B egymástól különböző halmazra is megtehetjük, melyek a $\{0, 1, \dots n-1\}$ részhalmazai. Speciális esetként $A = |\mathrm{pos}_1(a)|$ és $B = |\mathrm{pos}_1(b)|$ választással megkapjuk a hiányzó részt.

3.4. Definíció. Legyen $A \subseteq \{0, 1, \dots n-1\}$. Ekkor definiáljuk az $\mathbf{A_{p,i}} = \{j \in A : j \equiv i \mod p\}$ halmazt.

3.5. Állítás (Chase [4])**.** Ha $A \neq B \subseteq \{0, 1 \dots n-1\}$, akkor létezik egy $p = \mathcal{O}(\sqrt{n \log(n)})$ prím és i maradékosztály, amelyekre $|A_{p,i}| \neq |B_{p,i}|$.

3.6. Definíció. Az n-edik **körosztási polinom** az a normált polinom, amelynek gyökei pontosan az n-edik primitív komplex egységgyökök, mindegyik egyszeresen. $\Phi_n(x) = (x - \xi_1) \dots (x - \xi_{\varphi(n)})$, ahol $\xi_1, \dots \xi_{\varphi(n)}$ a primitív n-edik egységgyökök.

3.7. Definíció. Ha $a \in \{0, 1\}^n$ egy bináris szó, akkor jelölje $\mathbf{A(x)} = \sum_{i=0}^{n-1} a_i x^i$ az egyesek generáló függvényét. Ha $A \subseteq \{0, \dots n-1\}$, akkor $\mathbf{A(x)} = \sum_{i=0}^{n-1} \mathbb{1}_A(i) x^i$.

3.8. Lemma (Vyalyi, Gimadeev [13])**.** Tekintsük az $A \neq B \subseteq \{0, \dots n-1\}$ halmazokat és az $m \in \mathbb{Z}^+$ számot. Ha minden $i \in \{0, \dots m-1\}$ maradékosztályra $|A_{m,i}| = |B_{m,i}|$, akkor $\Phi_m(x)$ osztja az $(A(x) - B(x))$ polinomot.

Bizonyítás. Elég megmutatni, hogy $x^m - 1$ osztja az $A(x) - B(x)$ polinomot, ugyanis $\Phi_m(x)$ osztja $x^m - 1$-et.

Rögzített i-re tekintsük $A(x)$ azon monomjait, amelyek x^{i+km} alakúak valamely $k \in \mathbb{Z}^+$-ra, ezek száma $|A_{m,i}|$. Hasonlóan $B(x)$ azon monomjainak száma, amelyek x^{i+km} alakúak $|B_{m,i}|$. A lemma feltétele miatt $|A_{m,i}| = |B_{m,i}|$, ezért az $A(x)$ és $B(x)$ ilyen monomjai párbaállíthatók olyan módon, hogy egy $A(x)$-beli $x^{i+k_1 m}$ alakú tag párja $B(x)$-ben egy $x^{i+k_2 m}$ legyen.

Könnyen ellenőrizhető, hogy $(x^m - 1) | (x^{i+k_1 m} - x^{i+k_2 m})$.

Ezeket felhasználva adódik, hogy $x^m - 1$ osztja az $A(x) - B(x)$ polinomot. □

Bizonyítás. (3.5 Állítás) Tegyük fel, hogy egy adott $k \in \mathbb{Z}^+$-ra minden $p \leq k$ prím és $i \in \{0, \dots p-1\}$ maradékosztály esetén $|A_{p,i}| = |B_{p,i}|$. Az előző lemmából az következik, hogy $\Phi_p(x) | (A(x) - B(x))$ minden $p \leq k$ prímre. Az $A(x) - B(x)$ polinom nem azonosan nulla és a körosztási polinomok relatív prímek, ezért az $A(x) - B(x)$ foka legalább a $\Phi_p(x)$ körosztási polinomok fokainak összege kell legyen.

$$n \geq \deg(A(x) - B(x)) \geq \sum_{p \leq k} \deg(\Phi_p(x)) = \sum_{p \leq k} (p - 1)$$

A Prímszámtétel 1.18 Következménye miatt $\sum_{p \leq k} (p-1) \sim \frac{1}{2} \frac{k^2}{\log k}$, ami miatt $n \geq \frac{1}{2} \frac{k^2}{\log k}(1 + o(1))$. Azt kaptuk, hogy $2n \log(n) \geq 2n \log(k) \geq k^2(1 + o(1))$, azaz $k \leq \mathcal{O}(\sqrt{n \log n})$, amiből következik az állítás. □

3.2. A gyökös becslés általánosítása

Az előző gondolatmenetet általánosítottam, hogy $S_l(n)$-re egy $\tilde{\mathcal{O}}(\sqrt{n})$ felső korlátot kapjak, amelyet a következő tételben mondok ki.

3.9. Tétel. $S_l(n) = \mathcal{O}(l^5 \sqrt{n} \log^{1.5}(n))$.

Először bemutatom a szükséges definíciókat és segédállításokat, majd a tétel bizonyítását.

3.10. Definíció. Legyen $L, m \in \mathbb{Z}^+$. Ha minden $h \neq j \in \{1, 2, \dots L\}$ párhoz tartozik egy $i_{h,j} \in \{0, 1, \dots m-1\}$ maradékosztály modulo m, akkor ezek $i_{1,2}, i_{1,3}, \dots i_{1,l}, i_{2,3}$, $\dots i_{L-1,L}$ gyűjteményét (összesen $\binom{L}{2}$ darab) egy **maradékosztály $\binom{\mathbf{L}}{\mathbf{2}}$-es**nek nevezzük modulo m.

3.11. Lemma. Tekintsük az $A_1, A_2, \dots A_L \subseteq \{0, \dots n-1\}$ páronként különböző halmazokat és az $m \in \mathbb{Z}^+$ számot. Ha minden $i_{1,2}, \dots i_{L-1,L} \in \{0, \dots m-1\}$ maradékosztály $\binom{L}{2}$-esre létezik $h \neq j$, melyekre igaz, hogy $|(A_h)_{m,i_{h,j}}| = |(A_j)_{m,i_{h,j}}|$, akkor $\Phi_m(x)$ osztja az $(A_1)(x), \dots (A_L)(x)$ polinomok közül valamelyik kettő különbségét.

Bizonyítás. Ha létezik egy $h \neq j$, hogy minden i maradékosztályra $|(A_h)_{m,i}| = |(A_j)_{m,i}|$, akkor a 3.8 Lemma miatt $\Phi_m(x)$ osztja az $((A_j)(x) - (A_h)(x))$ polinomot.

Ha nem létezik ilyen, akkor minden lehetséges $h \neq j$-hez van egy $i_{h,j}$, amelyre $|(A_h)_{m,i_{h,j}}| \neq |(A_j)_{m,i_{h,j}}|$. Gyűjtsük össze ezeket egy $i_{1,2}, \dots i_{L-1,L}$ maradékosztály $\binom{L}{2}$-esbe. A lemma feltételének ellentmondó példát találtunk, tehát ez az eset nem lehetséges. □

3.12. Állítás. Ha $A_1, A_2, \dots A_L \subseteq \{0, 1, \dots n-1\}$ és páronként eltérőek, akkor létezik egy $p = \mathcal{O}(L\sqrt{n\log(n)})$ prím és egy $i_{1,2}, \dots i_{L-1,L}$ maradékosztály $\binom{L}{2}$-es, hogy minden $0 \leq h \neq j \leq L-1$-re $|(A_h)_{p,i_{h,j}}| \neq |(A_j)_{p,i_{h,j}}|$.

Bizonyítás. Tegyük fel, hogy egy $k \in \mathbb{Z}^+$-ra minden $p \leq k$ prím és $i_{1,2}, \dots i_{L-1,L} \in \{0, \dots p-1\}$ maradékosztály $\binom{L}{2}$-es esetén valamely $h \neq j$-re $|(A_h)_{p,i_{h,j}}| = |(A_j)_{p,i_{h,j}}|$. Felhasználva az előző lemmát a 3.5 Állításhoz hasonlóan azt kapjuk, hogy:

$$\binom{L}{2} n \geq \deg(\prod_{h \neq j}((A_h)(x) - (A_j)(x))) \geq \sum_{p \leq k} \deg(\Phi_p(x)) = \sum_{p \leq k}(p-1) \sim \frac{1}{2}\frac{k^2}{\log k}$$

Tehát $k \leq \mathcal{O}\left(\sqrt{\binom{L}{2} n \log(n)}\right)$. □

Az $S_l(n)$-re való felső korlátot úgy szeretném belátni, hogy veszem az egyesek előfordulásait ez egyes szavakban, azaz a $\mathrm{pos}_1(a_1), \mathrm{pos}_1(a_2), \dots \mathrm{pos}_1(a_l) \subseteq \{0, 1, \dots n-1\}$ páronként különböző halmazokat. Az előző állítás alapján van egy $p = \mathcal{O}(l\sqrt{n\log(n)})$ prím és egy $i_{1,2}, \dots i_{l-1,l}$ maradékosztály $\binom{l}{2}$-es, melyekre minden $h \neq j$-re $|\mathrm{pos}_1(a_h)_{p,i_{h,j}}| \neq |\mathrm{pos}_1(a_j)_{p,i_{h,j}}|$. Innen készíthető egy automata. Ezzel azonban az a baj, hogy csak akkor működik, ha feltesszük, hogy minden szó hossza pontosan n. Ha lehetnek n-nél rövidebb szavak is, akkor elképzelhető, hogy $a_i \neq a_j$, de $\mathrm{pos}_1(a_i) = \mathrm{pos}_1(a_j)$. Ez pontosan akkor van így, ha az a_j szót az a_i-ből néhány 0 hozzáírásával kapjuk (vagy fordítva).

Három szó esetén ez a probléma kiküszöbölhető a 2.7 állítás segítségével. Ilyenkor elég a fent leírt ötletet végigvinni egyforma hosszú szavak esetére. Azonban ha l tetszőleges egész szám lehet akkor nem elég az 1-esek pozícióit vizsgálni és a bizonyítás összetettebb lesz. A következő részben ezt fogom bemutatni.

3.9. Tétel. $S_l(n) = \mathcal{O}(l^5\sqrt{n}\log^{1.5}(n))$.

Bizonyítás. Tekintsük az alábbi $2l$ darab halmazt, amelyek a $\{0, 1, \dots n-1\}$ részhalmazai. (Elképzelhető, hogy vannak közöttük megegyezőek is.)

$$(A_1, A_2, \dots A_{2l}) = (\mathrm{pos}_0(a_1), \mathrm{pos}_0(a_2), \dots \mathrm{pos}_0(a_l), \mathrm{pos}_1(a_1), \mathrm{pos}_1(a_2), \dots \mathrm{pos}_1(a_l))$$

Ezek között a különbözőek számát jelöljük L-el, és alkalmazzuk az előző állítást ezen különböző halmazokra. Azt kapjuk, hogy található egy $p = \mathcal{O}(L\sqrt{n\log(n)}) = \mathcal{O}(l\sqrt{n\log(n)})$ prím és egy megfelelő maradékosztály $\binom{L}{2}$-es.

Ebből készíthetünk egy $i_{1,2}, \dots i_{2l-1,2l}$ maradékosztály $\binom{2l}{2}$-est, ahol minden $1 \le h \ne j \le 2l$-re $|(A_h)_{p,i_{h,j}}| \ne |(A_j)_{p,i_{h,j}}|$ igaz, ha $A_h \ne A_j$. Ezt egyszerűen úgy tehetjük meg, hogy az $i_{h,j}$ az A_h-hoz és A_j-hez tartozó érték lesz a maradékosztály $\binom{L}{2}$-esből ha $A_h \ne A_j$ (így néhány értéket többször fogunk berakni az új $\binom{2l}{2}$-esbe) és különben tetszőleges. Érdemes megfigyelni, hogy itt h és j az $1 \le h, j \le 2l$ intervallumban volt, a segédállításokban $1 \le h, j \le L$, a bizonyítás további részében azonban h-ra és j-re a $1 \le h, j \le l$ korlátok között gondolunk. A továbbiakban a talált $i_{1,2}, \dots i_{2l-1,2l}$ maradékosztály $\binom{2l}{2}$-esnek csak az $i_{1,2}, \dots i_{l-1,l}$ és $i_{l+1,l+2}, \dots i_{2l-1,2l}$ rész maradékosztály $\binom{l}{2}$-eseit fogjuk használni. Így a készített maradékosztály $\binom{2l}{2}$-es többi tagjáról meg is felejtkezhetünk.

Ha az a_h és az a_j szavak különbözőek az azt jelenti, hogy vagy $\mathrm{pos}_0(a_h) \ne \mathrm{pos}_0(a_j)$ vagy pedig $\mathrm{pos}_1(a_h) \ne \mathrm{pos}_1(a_j)$. Tehát bármelyik két szó valamelyik kiválasztott halmazban eltér.

Definiáljuk egy a szóra az alábbi értéket:

$$\begin{aligned} V(a) = |\mathrm{pos}_0(a)_{p,i_{1,2}}| + (n+1)\cdot|\mathrm{pos}_0(a)_{p,i_{1,3}}| + (n+1)^2|\mathrm{pos}_0(a)_{p,i_{1,4}}| + \dots \\ +(n+1)^{\binom{l}{2}-1}|\mathrm{pos}_0(a)_{p,i_{l-1,l}}| + (n+1)^{\binom{l}{2}}|\mathrm{pos}_1(a)_{p,i_{l+1,l+2}}| \\ +(n+1)^{\binom{l}{2}+1}\cdot|\mathrm{pos}_1(a)_{p,i_{l+1,l+3}}| + \dots + (n+1)^{2\cdot\binom{l}{2}-1}|\mathrm{pos}_1(a)_{p,i_{2l-1,2l}}| \end{aligned}$$

Könnyen látható, hogy $V(a_h) \ne V(a_j)$, ha $1 \le h \ne j \le l$, ugyanis a $V(a_h)$ és $V(a_j)$ értékeket $(n+1)$ alapú számrendszerben felírva az $i_{h,j}$-hez tartozó számjegy különböző lesz, ha $\mathrm{pos}_0(a_j) \ne \mathrm{pos}_0(a_h)$ vagy az $i_{l+h,l+j}$-hez tartozó számjegy különböző lesz, ha $\mathrm{pos}_1(a_h) \ne \mathrm{pos}_1(a_j)$.

Tudjuk, hogy $V(a) \le n^{2\binom{l}{2}+1}$, emiatt $\prod_{h\ne j}|V(a_h) - V(a_j)| \le n^{\left(2\binom{l}{2}+1\right)\binom{l}{2}}$. A 1.12 lemma miatt létezik egy q prím, melyre $q = \mathcal{O}\left(\left(2\binom{l}{2}+1\right)\binom{l}{2}\log(n)\right) = \mathcal{O}(l^4\log(n))$ és ami nem osztja a $\prod_{h\ne j}|V(a_h) - V(a_j)|$ szorzatot. Tehát a $V(a_1), \dots V(a_l)$ értékek különböző maradékot adnak modulo q.

Készítsünk egy olyan automatát, amely egy a szó esetén a $V(a) \mod q$ értéket számolja. Ezt megtehetjük, ha q darab p hosszú kört használunk. Ha egy kör $i_{h,j}$-edik állapotában 0-át olvasunk, akkor a (modulo q értve) $(n+1)^x$-el későbbi kör $i_{h,j}+1$-edik állapotába lépünk, ahol $(n+1)^x$ a $V(a)$ képletében a $|\mathrm{pos}_0(a)_{p,i_{h,j}}|$ tag együtthatója. Ha egy kör $i_{l+h,l+j}$-edik állapotában 1-est olvasunk, akkor a (modulo q értve) $(n+1)^x$-el későbbi kör $i_{l+h,l+j}+1$-edik állapotába lépünk, ahol $(n+1)^x$ a

$V(a)$ képletében a $|\mathrm{pos}_1(a)_{p,i_{l+h},l+j}|$ tag együtthatója. Egyébként ugyanazon a körön a következő állapotba lépünk.

Ez egy pq állapotú automata, amely megkülönbözteti a szavakat. Tehát bármelyik n hosszú bináris szó l-eshez létezik egy $pq = \mathcal{O}\left(l^5\sqrt{n}\log(n)^{1.5}\right)$ állapotú automata, amely megkülönbözteti őket. □

3.3. Második gyökös becslés

3.13. Definíció. Egy $a \in \Sigma^n$ szó **p-periodikus**, ha $a_i = a_{i+p}$ igaz minden $0 \leq i \leq n-1-p$ indexre. A legkisebb ilyen p-t nevezzük a **szó periódusá**nak. Egy szót **periodikus**nak nevezünk, ha a periódusa nem nagyobb, mint a hossza fele.

3.14. Lemma (Robson [7])**.** Az $w0$ és $w1$ szavak közül legalább az egyik nem periodikus. (Ahol $w0$ azt jelöli, hogy az w mögé írunk egy 0-t.)

Bizonyítás. Jelölje l a w szó hosszát, azaz $w = w_0 \dots w_{l-1}$. Továbbá legyen p_0 a $w0$ periódusa és p_1 a $w1$ periódusa. Indirekt tegyük fel, hogy $w0$ és $w1$ is periodikus, azaz $p_0 \leq \frac{l+1}{2}$ és $p_1 \leq \frac{l+1}{2}$. Ekkor $w_{l+ap_1 \mod p_0} = 1$ és $w_{l+bp_0 \mod p_1} = 0$ minden $a, b \in \mathbb{Z}$-re. Az $ap_1 + bp_0 = -x \cdot \gcd(p_0, p_1)$ választással, ahol $x \in \mathbb{Z}$ és $0 \leq l - x \cdot \gcd(p_0, p_1) < p_0, p_1$ azt kapjuk, hogy $0 = w_{l - x \cdot \gcd(p_0, p_1)} = 1$. □

3.15. Lemma (Robson [7])**.** Legyen $w \in \{0,1\}^l$, $a \in \{0,1\}^n$ és $l < n$. Ha w periódusa p és $a_i \dots a_{i+l-1} = w = a_j \dots a_{j+l-1}$, akkor $|j - i| \geq p$.

Bizonyítás. Ha $a_i \dots a_{i+l-1} = w = a_j \dots a_{j+l-1}$, akkor a w szó $|j-i|$-periodikus, ezért $|j - i| \geq p$. □

3.16. Következmény. Legyen $w \in \{0,1\}^l$, $a \in \{0,1\}^n$ és $l < n$. Ha w periódusa p, akkor w előfordulásainak száma a-ban maximum $\frac{n}{p}$.

3.17. Lemma (Robson [7])**.** Minden $\alpha < 1$-re ha az $a_{i-l+1} \dots a_i$ részszó nem periodikus és $l \leq n^\alpha$, akkor létezik egy $j \leq c\frac{n\log(n)}{l}$ prím, hogy $a_{k-l+1} \dots a_k \neq a_{i-l+1} \dots a_i$ igaz minden olyan k-ra, amelyre $k \equiv i \mod j$, de $k \neq i$. Itt a c konstans csak α-tól függ.

Bizonyítás. Az alábbi lépéseket fogjuk megtenni: először megbecsüljük hogy a részszó maximum hányszor fordulhat elő az a szóban. Ezután megvizsgáljuk, hogy egy ilyen előfordulás hány prímszámra ronthatja el a lemmában szereplő elvárást.

Az $a_{i-l+1} \dots a_i$ részszó nem periodikus, tehát ha a periódusát p-vel jelöljük akkor $p > \frac{l}{2}$. A 3.16 következmény miatt ezen részszó előfordulásainak száma a-ban maximum $\frac{n}{p} < \frac{2n}{l}$.

Vizsgáljuk meg, hogy a részszó egy ilyen előfordulása milyen j prímekre rontja el a lemma elvárását. Tegyük fel, hogy az $a_{i-l+1} \dots a_i$ részszó vizsgált előfordulása az $a_{k-l+1} \dots a_k$ helyen található, ahol $k \neq i$. Ekkor minden olyan j prímre, amely osztója az $|k - i|$ különbségnek nem teljesül a lemmában megfogalmazott elvárás.

A következő kérdés az, hogy hány darab olyan j prím van, amely osztja a $|k-i|$ különbséget? Azt szeretnénk belátni, hogy nem lehet $(1-\alpha)^{-1}$ darab $\frac{n}{l}$-nél nagyobb prímosztója. Ehhez felhasználjuk a lemma feltételei közül az $l \leq n^{\alpha}$ előírást, ami miatt $\frac{n}{l} \geq \frac{n}{n^{\alpha}} = n^{1-\alpha}$. Az $(1-\alpha)^{-1}$ darab $n^{1-\alpha}$-nál nagyobb prím osztó szorzata már n-nél nagyobb lenne, de $|k-i| \leq n$.

Ott tartunk, hogy a részszónak maximum $\frac{2n}{l}$ darab előfordulása lehet, és minden ilyen előfordulás kevesebb, mint $(1-\alpha)^{-1}$ darab $j > \frac{n}{l}$ prímszámra ronthatja el a lemmában szereplő egyenlőtlenséget.

Így lennie kell egy j-nek az $\frac{n}{l}$-nél nagyobb első $\frac{2n}{l(1-\alpha)}$ prím között, amire igaz a lemma. A prímszámtételből 1.19 következménye miatt igaz a j-re adott felső korlát. □

3.18. Definíció. Az A automata **megtalálja** az a szót, ha annak utolsó betűjének beolvasával kerül először a q^+ elfogadó állapotba.

3.19. Állítás (Robson [7])**.** Ha az $a \neq b$ szavakat szeretnénk megkülönböztetni, akkor elég egy olyan A autamatát készítenünk, amely az $a_0 \dots a_i \neq b_0 \dots b_i$ kezdőszeletek közül az egyiket megtalálja és a másikat beolvasva sosem lép a q^+ elfogadó állapotba.

Bizonyítás. Tegyük fel, hogy az A automata az $a_0 \dots a_i$ szót megtalálja és a $b_0 \dots b_i$ szó esetén nem lép a q^+ elfogadó állapotba. Ha a teljes b szót beolvasva sem lép a q^+ elfogadó állapotba egyszer sem, akkor állítsuk be az átmenetfüggvényt úgy, hogy ha az automata egyszer a q^+ állapotba lép, akkor azután mindig ott marad. Ekkor az A automata megkülönbözteti a két szót.

A másik lehetőség, hogy teljes b szót beolvasva valamely $j > i$ indexre a b_j betű beolvasása után lépünk először a Q állapotba. Ekkor az $a_{i+1} \dots a_{n-1}$ és $b_{j+1} \dots b_{n-1}$ szavak különböző hosszúak, ezért a 2.6 Állítás miatt megkülönböztethetők egy B automatával, amelynek $\mathcal{O}(\log(n))$ állapota van. A q^+ állapotot a B kezdőállapotával azonosítva egy olyan összetett A' automatát kapunk, amely megkülönbözteti a két szót. Ha az A automatának $\mathcal{O}(f(n))$ állapota volt, akkor az A'-nek $\max\{\mathcal{O}(f(n)), \mathcal{O}(\log(n))\}$ állapota van. □

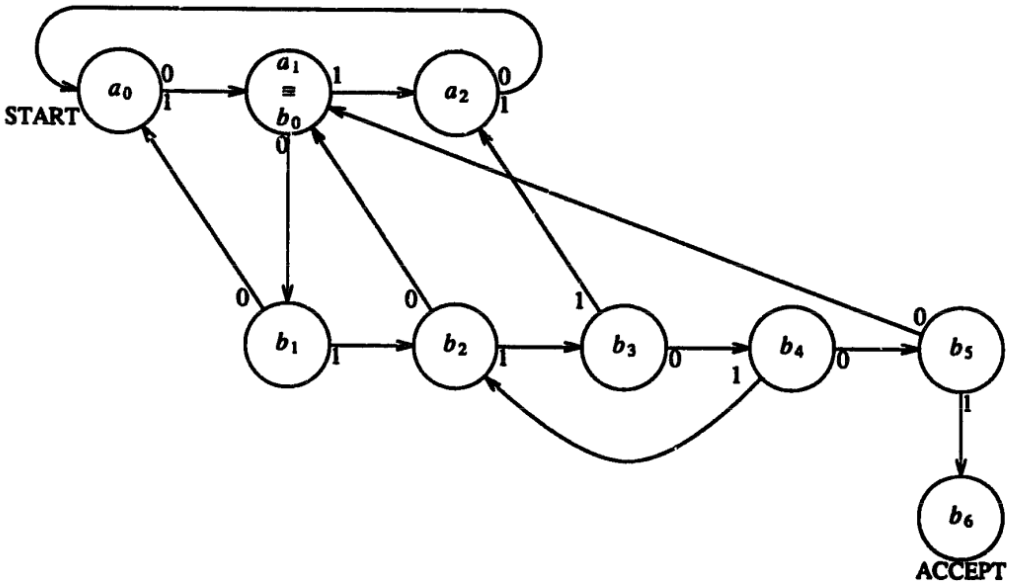
3.20. Tétel (Robson [7])**.** $S_2(n) = \mathcal{O}(\sqrt{n\log(n)})$

Bizonyítás. Legyen $a \neq b \in \{0,1\}^n$. Jelölje i azt az indexet, ahol a két szó először eltér. Ha $i \leq \sqrt{n\log(n)}$, akkor a 2.20 állítás miatt a két szó megkülönböztethető egy $\mathcal{O}(\sqrt{n\log(n)})$ állapotú automatával.

Ellenkező esetben az 3.14 lemma miatt az $a_{i-\sqrt{n\log(n)}} \dots a_i$ és $b_{i-\sqrt{n\log(n)}} \dots b_i$ szavak közül legalább az egyik nem periodikus. Tegyük fel, hogy az a-ban lévő nem periodikus. Válasszunk egy $j \leq c\frac{n\log(n)}{\sqrt{n\log(n)}+1} = \mathcal{O}(\sqrt{n\log(n)})$ prímet az 3.17 lemma alapján. Ekkor minden $k \equiv i \mod j, k \neq i$ számra $a_{k-\sqrt{n\log(n)}} \dots a_k \neq a_{i-\sqrt{n\log(n)}} \dots a_i$.

Készítsünk egy automatát, ami az $a_{i-\sqrt{n\log(n)}} \dots a_i$ részszó olyan előfordulását keresi, amelynek p kezdőpozíciója igaz, hogy $p \equiv i - \sqrt{n\log(n)} \mod j$. Az automata két részből fog állni. Az A rész az eddig beolvasott karakterek számát tartja számon

modulo j. A $count \equiv i - \sqrt{n\log(n)} \mod j$ állapot megegyezik a B rész kezdőállapotával (nulladik állapot). A B automata x-edik állapotában akkor vagyunk, ha az előző x karakter a részszó x hosszú prefixe volt, és a megfelelő pozícióban kezdődött modulo j. Az A rész j állapotból és a B rész további $\sqrt{n\log(n)} + 1$ állapotból áll, így az egész automata $\mathcal{O}(\sqrt{n\log(n)})$ állapotú. Az $a_0 \dots a_i$ kezdőszeletet megtalálja és a $b_0 \dots b_i$ kezdőszelet esetén sosem lép a q^+ elfogadó állapotba, ezért a 3.19 Állítás miatt készen vagyunk. □



1. ábra. Példa a bizonyítás során készített automatára a 011001 részszó és $j = 3$ esetén, ha a részszó kezdőpozíciójára azt írjuk elő, hogy 1-gyel legyen kongruens modulo 3. Forrás: [7]

1989-ben Robson ezt a gondolatot javítva bizonyította be az $\tilde{\mathcal{O}}(n^{2/5})$ felső korlátot, amely a [7] cikkben olvasható.

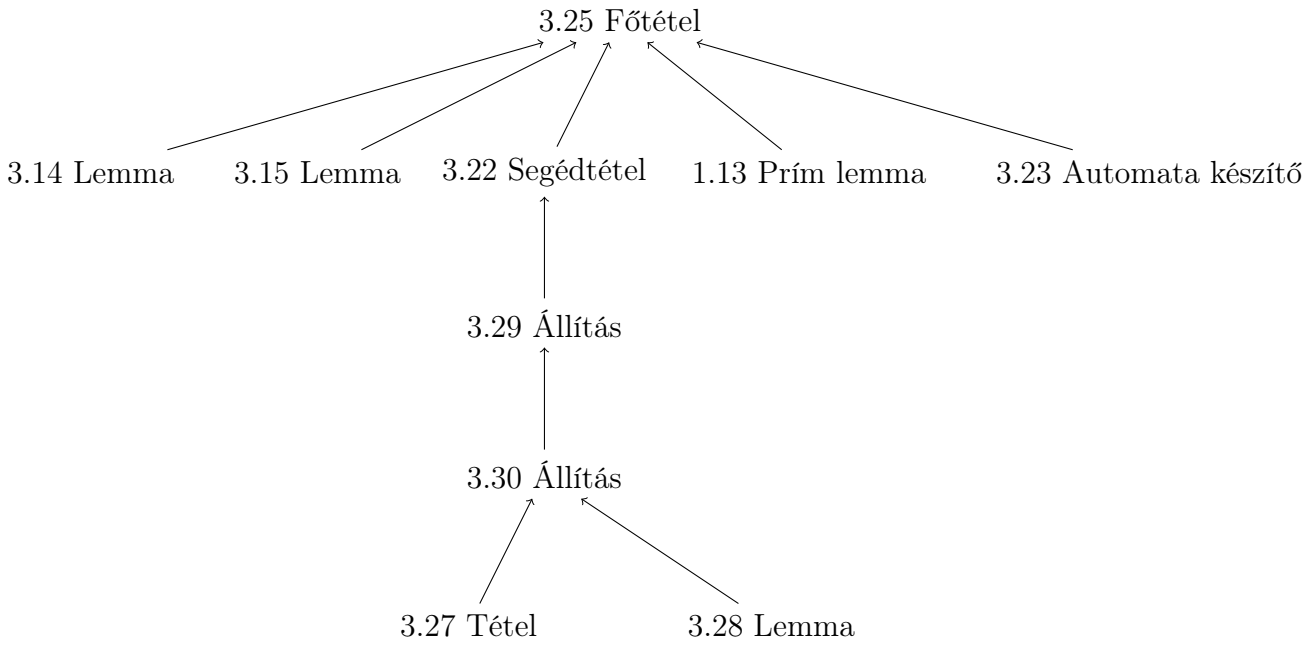
3.4. A legerősebb ismert becslés

Az első gyökös becslés ötletének javításával Zachary Chase 2020-ban bebizonyította az eddig ismert legjobb felső korlátot [4], amely szerint $S_2(n) = \tilde{\mathcal{O}}(n^{1/3})$. Ebben a fejezetben ezen eredmény bizonyításának lépéseit fogom bemutatni.

3.21. Definíció. Az $A \subseteq \{1, \dots n\}$ halmazt **d-szeparált**nak nevezzük, ha minden $i \neq j \in A$-ra $|j - i| \geq d$.

3.22. Tétel (Chase [4])**.** Ha $A \neq B \subseteq \{1, \dots n\}$ és $n^{1/3}$-szeparáltak, akkor létezik egy $p = \mathcal{O}(n^{1/3}\log^6(n))$ prím és i maradékosztály modulo p, melyekre $|A_{p,i}| \neq |B_{p,i}|$.

Először ezen tétel felhasználásával bebizonyítom a fő eredményt, majd a következő részben bemutatom a segédtétel bizonyításának lépéseit.



2. ábra. Bemutatja, hogy a felhasznált tételekből hogyan jutunk el a fő eredményhez. Segítséget nyújthat az egyes lépések közötti összefüggések megértéséhez és a teljes kép könnyebb átlátásához.

3.23. Lemma (Chase [4])**.** Legyen $m \in \mathbb{Z}^+$, $i \in [m]_0$ egy maradékosztály modulo m, q prímszám és $a \in [q]_0$ maradékosztály modulo q. Legyen $w \in \{0,1\}^l$, ahol $l \leq m$. Ekkor létezik egy $2mq$ állapotú automata, amely pontosan azokat az $x \in \{0,1\}^n$ szavakat fogadja el, amelyekre $|\mathrm{pos}_w(x)_{m,i}| \equiv a \mod q$.

Bizonyítás. Készítsünk egy q darab m hosszú körből álló automatát, melyben minden állapotból kettő van, egy 0-ás és egy 1-es. Mindegyik kör azt ellenőrzi, hogy a w szót olvassuk-e a megfelelő pozícióban. Azaz ha a kör i-edik állapotában a w_1-et olvassuk, akkor a kör $i+1 \mod m$-edik állapotának az 1-es változatába kerülünk. Ameddig a w következő betűjét olvassuk a (modulo m értve) következő állapot 1-es változatába kerülünk. Ha valamelyik lépésben nem a w megfelelő betűje következik, akkor a következő állapot 0-ás változatába lépünk. Ha egy kör $i+l-1 \mod m$-edik állapotának az 1-es példányában a w utolsó betűjét olvassuk, akkor átkerülünk a következő kör $i+l \mod m$-es állapotába, ellenkező esetben ugyanazon a körön maradunk. Amikor újra az i-edik állapothoz érünk (akár a régi vagy az új körön), újra leellenőrizzük, hogy azon a helyen w szó áll-e az előbbi eljárást ismételve. Az hogy éppen hanyadik körön vagyunk azt jelenti, hogy a w részszót hányszor olvastuk az x megfelelő pozícióiból kezdődően (modulo q számolva). Tehát az a-adik kör állapotai lesznek az elfogadó állapotok. □

Fel fogjuk használni a Robson-féle gyökös becslésnél bebizonyított 3.14 és 3.15 lemmákat:

3.14. Lemma (Robson [7])**.** Az $w0$ és $w1$ szavak közül legalább az egyik nem periodikus. (Ahol $w0$ azt jelöli, hogy az w mögé írunk egy 0-t.)

3.15. Lemma (Robson [7]). Legyen $w \in \{0,1\}^l$, $a \in \{0,1\}^n$ és $l < n$. Ha w periódusa p és $a_i \dots a_{i+l-1} = w = a_j \dots a_{j+l-1}$, akkor $|j-i| \geq p$.

3.24. Következmény. Ha w szó periódusa p, akkor a $\mathrm{pos}_w(a)$ halmaz p-szeparált.

3.25. Tétel (Chase [4]). $S_2(n) = \mathcal{O}(n^{1/3}\log^7(n))$.

Bizonyítás. Legyen $a, b \in \{0,1\}^n$ két különböző bináris szó. Ha az első $2n^{1/3}$ pozíció valamelyikén eltér a két szó, akkor a 2.17 állítás miatt készen vagyunk. Különben legyen $k > 2n^{1/3}$ az a pozíció, ahol a két szó először eltér. Legyen $w' = a_{k-2n^{1/3}+1} \dots a_{k-1} = b_{k-2n^{1/3}+1} \dots b_{k-1}$ az ezen pozíció előtti közös $2n^{1/3}-1$ hosszú részszavuk.

A 3.14 Lemma miatt választható egy $w \in \{w'0, w'1\}$ szó, amely nem periodikus. A w hossza $2n^{1/3}$, ezért a periódusa legalább $n^{1/3}$. A 3.24 Következmény miatt a $\mathrm{pos}_w(a), \mathrm{pos}_w(b)$ halmazok $n^{1/3}$-szeparáltak. Emellett $\mathrm{pos}_w(a) \neq \mathrm{pos}_w(b)$ is igaz lesz.

A 3.22 segédtételt felhasználva azt kapjuk, hogy létezik egy $p = \mathcal{O}(n^{1/3}\log^6(n))$ prím és $i \in [p]_0$ maradékosztály, melyekre $|\mathrm{pos}_w(a)_{p,i}| \neq |\mathrm{pos}_w(b)_{p,i}|$. A 1.13 következmény miatt létezik egy $q = \mathcal{O}(\log(n))$ prím, melyre $|\mathrm{pos}_w(a)_{p,i}| \not\equiv |\mathrm{pos}_w(b)_{p,i}| \mod q$.

Ezután felhasználva a 3.23 lemmát kapunk egy $2pq = \mathcal{O}(n^{1/3}\log^7(n))$ állapotú automatát, amely megkülönbözteti a két szót. □

A 3.22 segédtétel bizonyítása

3.26. Definíció. Definiáljuk a komplex polinomok egy részhalmazát az alábbi módon: $\mathcal{P}_\mathbf{n} = \{p(x) = 1 - \sigma x^d + \sum_{j=n^{1/3}}^{n} a_j x^j \in \mathbb{C}[x] : 1 \leq d \leq n^{1/3}, \sigma \in \{0,1\}, |a_j| \leq 1 \forall j\}$

3.27. Tétel (Chase [4]). Létezik egy C_1 abszolút konstans, melyre igaz, hogy ha $n \geq 2$ és $p \in \mathcal{P}_n$, akkor $\max_{x\in[1-n^{-2/3},1]}|p(x)| \geq \exp(-C_1 n^{1/3}\log^5(n))$.

Itt $p(x)$ egy komplex polinom, aminek a maximumát egy olyan intervallumon nézzük, ami a valós számok részhalmaza.

3.28. Lemma (Borwein, Erdélyi, Kós [3]). Tegyük fel, hogy $p(x) = \sum_{j=0}^{n} a_j x^j \in \mathbb{C}[n]$ komplex polinomra igaz, hogy $|a_j| \leq 1$ minden j-re. Ekkor ha $(x-1)^k | p(x)$, akkor $\max_{x\in[1-\frac{k}{9n},1]}|p(x)| \leq (n+1)\left(\frac{e}{9}\right)^k$.

A lemma azt a gondolatot használja fel, hogy ha egy polinomnak az 1 többszörös gyöke, akkor a polinom deriváltjai eltűnnek az 1 helyen, azaz itt sima. A $p(1) = \sum_{j=0}^{n} a_j \leq n+1$ és az 1 pont körüli simaság miatt az 1 közelében a polinom értékei nem lehetnek túl magasak.

A továbbiakban a 3.27 Tétel és a 3.28 Lemma felhasználásával bebizonyítjuk a segédállítást. Ezen felhasznált két tétel bizonyítására nem térek ki, a megjelölt forrásokban megtalálhatóak.

3.29. Állítás (Chase [4]). Létezik egy $C > 0$ abszolút konstans, hogy $\forall n \geq 2$ és $p \in \mathcal{P}_n$-re igaz $(x-1)^{\lfloor Cn^{1/3}\log^5(n)\rfloor}$ nem osztja $p(x)$-et.

Bizonyítás. Legyen C egy megfelelően nagy konstans. Indirekt tegyük fel, hogy valamely n-re és $p \in \mathcal{P}_n$-re $(x-1)^{\lfloor Cn^{1/3}\log^5(n)\rfloor}$ osztja $p(x)$-et. A 3.28 Lemma és a 3.27 Tétel felhasználásával az alábbi egyenlőtlenség lánc vezethető le:

$$\begin{aligned}(n+1)\left(\frac{e}{9}\right)^{\lfloor Cn^{1/3}\log^5(n)\rfloor} &\geq \max_{x\in[1-\frac{C}{9}n^{-2/3}\log^5(n),1]}|p(x)|\\ &\geq \max_{x\in[1-n^{-2/3},1]}|p(x)|\\ &\geq \exp(-Cn^{1/3}\log^5(n))\\ &= \left(\frac{1}{e}\right)^{Cn^{1/3}\log^5(n)}\end{aligned}$$

Ha C elég nagy, akkor ez ellentmondás, ugyanis $\frac{e}{9} < \frac{1}{e}$. □

3.30. Állítás (Chase [4])**.** Legyenek az $A \neq B \subseteq \{1,\dots n\}$ halmazok $n^{1/3}$-szeparáltak. Ekkor létezik egy $m = \mathcal{O}(n^{1/3}\log^5(n))$ egész szám, amelyre $\sum_{a\in A} a^m \neq \sum_{b\in B} b^m$.

Bizonyítás. Definiáljuk az f polinomot a következő módon: $f(x) = \sum_{j=0}^{n}\epsilon_j x^j$, ahol $\epsilon_j = \mathbb{1}_A(j) - \mathbb{1}_B(j)$. Legyen $r \in \mathbb{Z}^+$ az a legnagyobb szám, amelyre $\epsilon_0 = \dots = \epsilon_{r-1} = 0$. Ekkor az $\tilde{f}(x) = \frac{f(x)}{x^j} = \epsilon_r + \epsilon_{r+1}x + \dots \epsilon_n x^{n-r}$ polinom főegyütthatója 1 vagy -1. Feltehető, hogy $\epsilon_r = 1$ (ellenkező esetben az A és B halmazt megcseréljük). Az A és B halmazok $n^{1/3}$-szeparálhatóak, amiből az következik, hogy $\tilde{f}(x) \in \mathcal{P}_n$.

Az 3.29 Állítás miatt $(x-1)^{\lfloor Cn^{1/3}\log^5(n)\rfloor}$ nem osztja az $\tilde{f}(x)$ polinomot, emiatt $f(x)$-et sem. Ez azt jelenti, hogy létezik egy $0 \leq k < \lfloor Cn^{1/3}\log^5(n)\rfloor$ egész szám, amelyre $f(x) = (x-1)^k g(x)$ és $(x-1) \nmid g(x)$, azaz $g(1) \neq 0$. Ekkor az f függvény k-adik deriváltja $f^{(k)}(x) = (k!)g(x) + (x-1)(...)$, tehát $f^{(k)}(1) \neq 0$ és k a legkisebb ilyen szám. Ha $k = 0$, akkor $0 \neq f^{(0)}(1) = f(1) = |A| - |B|$, tehát $m = 0$-val igaz az állítás.

Ha $k > 0$, akkor indukcióval belátjuk, hogy minden $m < k$-ra $\sum_{a\in A} a^m = \sum_{b\in B} b^m$ és $\sum_{a\in A} a^k \neq \sum_{b\in B} b^k$. A kezdőlépéshez be kell látni, hogy $m = 0$-ra igaz az egyenlőség, azaz $|A| = |B|$. Ez igaz, mert $0 = f^{(0)}(1) = f(1) = |A| - |B|$.

Az indukciós lépéshez tegyük fel, hogy $0,\dots m-1$-re igaz az egyenlőség és be szeretnénk látni, hogy $\sum_{a\in A} a^m = \sum_{b\in B} b^m$, ha $m < k$ vagy $\sum_{a\in A} a^m \neq \sum_{b\in B} b^m$ ha $m = k$. Tudjuk, hogy $c = f^{(m)}(1) = \sum_{j=0}^{n} j(j-1)\dots(j-m+1)\epsilon_j$ és $c = 0$, ha $m < k$, $c \neq 0$, ha $m = k$. Rendezzük a j^m-es tagokat a bal oldalra, ekkor azt kapjuk, hogy $\sum_{j=0}^{n} j^m \mathbb{1}_A(j) - \sum_{j=0}^{n} j^m \mathbb{1}_B(j) = \sum_{i=0}^{m-1}\alpha_i\left(\sum_{j=0}^{n} j^i\mathbb{1}_A(j) - \sum_{j=0}^{n} j^i\mathbb{1}_B(j)\right) - c = -c$. Itt az α_i-k a megfelelő együtthatók és felhasználjuk, hogy $\sum_{j=0}^{n} j^i\mathbb{1}_A(j) = \sum_{a\in A} a^m$. Ezzel beláttuk az allítást.

□

3.22. Tétel (Chase [4])**.** Ha $A \neq B \subseteq \{1,\dots n\}$ és $n^{1/3}$-szeparáltak, akkor létezik egy $p = \mathcal{O}(n^{1/3}\log^6(n))$ prím és i maradékosztály modulo p, melyekre $|A_{p,i}| \neq |B_{p,i}|$.

Bizonyítás. Az előző állítás alapján választható egy olyan $m = \mathcal{O}(n^{1/3}\log^5(n))$ egész szám, amelyre $\sum_{a\in A} a^m \neq \sum_{b\in B} b^m$. A 1.13 következmény alkalmazásához felhasz-

náljuk, hogy $\sum_{a\in A} a^m \leq n \cdot n^m$ és $\sum_{b\in B} b^m \leq n \cdot n^m$. Így azt kapjuk, hogy létezik egy $p = \mathcal{O}(\log(n^{m+1})) = \mathcal{O}((m+1)\log(n)) = \mathcal{O}(n^{1/3}\log^6(n))$ prím, amelyre $\sum_{a\in A} a^m \not\equiv \sum_{b\in B} b^m \mod p$.

Megfigyelhetjük, hogy $\sum_{a\in A} a^m \equiv \sum_{i=1}^{p-1} |A_{p,i}| i^m \mod p$ és ugyanez igaz B-re is. Ezt felhasználva adódik, hogy $\sum_{i=1}^{p-1} |A_{p,i}| i^m \not\equiv \sum_{i=1}^{p-1} |B_{p,i}| i^m \mod p$. Ebből következik, hogy valamely $i \in \{0, \ldots p-1\}$-re $|A_{p,i}| \neq |B_{p,i}|$. □

4. Kapcsolódó kérdések

A separating words problem többféle változata, módosítása ismert, amelyek között még sok megválaszolatlan kérdés van. A következőkben megemlítek néhány változatot az eredmények teljeskörű ismertetése nélkül.

4.1. Permutáció automaták

A separating words problem egyik megszorítása, amikor az összes DFA helyett csak a permutáció automaták közül szeretnénk olyat keresni, amely elválaszt két szót.

4.1. Definíció. Az $A = (Q, \Sigma, \delta, q_0, F)$ automatát **permutáció automatá**nak nevezzük, ha minden $\sigma \in \Sigma$ inputhoz tartozó $\delta(\cdot, \sigma) : Q \mapsto Q$ átmenet az állapotok egy permutációja.

A definíció ekvivalens azzal, hogy minden $q \in Q$ állapotra és $\sigma \in \Sigma$ betűre pontosan egy olyan q' állapot van, hogy ott a σ-t olvasva q-ba lépünk. Ebből következik, hogy minden állapot befoka (és kifoka is) $|\Sigma|$. A következtetés visszafele nem igaz, azaz létezik olyan automata aminek befoka és kifoka is $|\Sigma|$, de nem permutáció automata.

4.2. Definíció. Definiáljuk az $\mathbf{M_{x,m}}, x < m$ permutáció automatákat a $\Sigma = \{0, 1\}$ abc felett a következő módon. Az állapotok száma $2m$ és halmazuk a $Q = \{(c, p) : 0 \leq c < m, 0 \leq p \leq 1\}$. Az átmenetfüggvény legyen $(c, p) \to (c + 1 \mod m, p)$, ha 0-t olvasunk és $(c, p) \to (c + 1 \mod m, 1 - p \text{ if } c = x \text{ else } p)$, ha 1-et olvasunk. A kezdőállapot legyen $q_0 = (0, 0)$ és az elfogadó állapotok a $(c, 1)$ állapotok. Megfigyelhető, hogy egy (c, p) állapotban a c a beolvasott karakterek számát határozza meg modulo m és p az olyan 1-esek számát modulo 2, amelyek indexe kongruens x-el modulo m. Tehát az $M_{x,m}$ automata pontosan azokat a szavakat fogadja el, amelyekben páratlan sok 1-es van x-el kongruens pozícióban modulo m.

1996-ban Robson bebizonyította, hogy az $\tilde{\mathcal{O}}(\sqrt{n})$-es felső korlát a permutáció automaták között is igaz. (A separating words problem olyan változatai, ahol speciális automatákat használunk nehezebb az eredeti feladatnál, ugyanis a szavak megkülönböztetéséhez kevesebb automata közül választhatunk.)

4.3. Tétel (Robson [8])**.** Ha $a, b \in \{0, 1\}^n$ különböző szavak, akkor egy $M_{x,m}$ permutáció automata szétválasztja őket úgy, hogy $m = \mathcal{O}(\sqrt{n})$.

Továbbá az is igaz, hogy m választható úgy, hogy négyzetmentes, és m és x relatív prímek vagy $x = 0$ és $m = 2$.

4.2. Nemdeterminisztikus szeparáció

4.4. Definíció. nsep$(\mathbf{a}, \mathbf{b})$ az a legkisebb k szám, amire létezik egy k állapotú nemdeterminisztikus véges automata (NFA), amely az a szót elfogadja (van olyan lefutása az a-t olvasva, hogy elfogadó állapotba kerül) és a b szót elutasítja (a b-t olvasva semelyik lefutás esetén sem kerül elfogadó állapotba).

A separating words problem nemdeterminisztikus változatában eddig nem sok eredmény született, és több nyitott kérdés vár megválaszolásra, amelyek közül néhány megtalálható az [5] összefoglalóban. A továbbiakban kimondok a témához kapcsolódó két állítást, amelyek bizonyítása szintén a [5]-ban olvasható.

4.5. Állítás ([5])**.** Az nsep nem szimmetrikus, azaz létezik $a, b \in \Sigma^n$, hogy $\mathrm{nsep}(a, b) \neq \mathrm{nsep}(b, a)$.

4.6. Tétel ([5])**.** A $\frac{\mathrm{sep}(a,b)}{\mathrm{nsep}(a,b)}$ hányados nem korlátos.

4.3. Szeparálás minden kezdőállapotból

Az legfrissebb cikkekben (Tran [12], [11]) az alábbi változatokkal foglalkoztak:

1. $\exists$-szeparáció: Létezik olyan közös kezdőállapot, amelyből a két szót indítva azok különböző végállapotba kerülnek. Ez az eredeti separating words problem.
   - $S_\exists(n) = \mathcal{O}(n^{1/3} \log^7(n))$ (Chase[4])
   - $S_\exists(n) = \Omega(\log(n))$
2. $\forall$-szeparáció: Bármely állapotot választjuk közös kiindulási állapotnak, a két szót onnan indítva különböző végállapotba kell érkezniük.
   - $S_\forall(n) = \mathcal{O}(\sqrt{n} \log(n) \log\log(n))$ (Tran [11])
   - $S_\forall(n) = \Omega(\log(n))$
3. $\forall^2$-szeparáció: Bármely (q_0, q_0') kezdőállapot pár esetén az első szót q_0-ból, a másodikat q_0'-ből indítva különböző végállapotba kerülnek.
   - Pontosan azok a szópárok nem $\forall^2$-szeparálhatóak, amelyek közül az egyik valódi suffixe a másiknak. (Tran [12])
   - Egyforma hosszú szavak esetén: $S_{\forall^2}(n) = n + 1$ (Tran [12])
   - Nyitott kérdés a $\mathrm{sep}_{\forall^2}(a, b)$ kiszámolása polinomiális algoritmussal.
4. $\forall^2 01$-szeparáció: Léteznek $e \neq e' \in Q$ állapotok, hogy bármely (q_0, q_0') kezdőállapot párra az automata q_0-ból a-t olvasva az e állapotba érkezik és q_0'-ből b-t olvasva e'-be.
   - Ugyanaz a helyzet, mint a $\forall^2$ szeparációnál (Tran [12])

4.7. Megjegyzés. Definiálhatjuk a $\exists^2$-szeparációt: Olyan automatát keresünk, amelyben létezik olyan (q_0, q_0') kezdőállapot pár, hogy az első szót q_0-ból, a másodikat q_0'-ből indítva különböző végállapotokba kerülnek. Ez a kérdés azonban nagyon egyszerűen megválaszolható, ugyanis egy két állapotú automata, amely bármit olvasva a másik állapotba lép megfelelő lesz minden szópárhoz. Ha a két szó hossza különböző, akkor ugyanabból az állapotból indítva, ha a két szó hossza megegyezik, akkor különböző állapotból indítva megkülönbözteti őket az automata. Tehát $S_{\exists^2}(n) = 2$. (Tran [12])

5. Tesztek

Az $S_2(n)$ és $S_3(n)$ kis n értékek esetén való kiszámítására készítettem egy programot [2]. Ez az összes n hosszú szópár vagy szóhármas esetén megkeresi az őket szétválasztó legkisebb automatát. Ez a lehetséges nem izomorf automaták kipróbálásával történik. A nem izomorf autamaták generálásához az általam kitalált és korábban leírt 1.11 módszert használtam. A C++ ban írt programot Ubuntu22 operációs rendszer alatt egy Intel Core i5-1135G7, 2.40GHz processzorral és 8GB-os memóriával rendelkező laptopon futtattam. 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						

A számítások elvégzése után néhány hónappal találtam egy hasonló táblázatot, amelyet egy ugyanebben az évben megjelent cikkben [11] közöltek. Itt $S_2(n)$ értékeit $n \leq 18$-ig számolták ki egy jobb teljesítményű számítógépen. Ez megerősített abban, hogy az általam talált értékek helyesek, és hogy ennél nagyobb n-ekre az $S_2(n)$ meghatározása nem egyszerű feladat.

Hivatkozások

[1] URL: https://cp-algorithms.com/string/aho_corasick.html.

[2] URL: https://github.com/CsanyiDavid/separating_words/tree/original.

[3] Peter Borwein, Tamás Erdélyi és Géza Kós. „Littlewood-Type problems on [0,1]". *Proceedings of the London Mathematical Society* 79.1 (1999), 22–46. old. DOI: 10.1112/S0024611599011831.

[4] Zachary Chase. „A new upper bound for separating words". (2022). arXiv: 2007.12097 [math.CO].

[5] Erik D. Demaine, Sarah Eisenstat, Jeffrey O. Shallit és David A. Wilson. „Remarks on separating words". *CoRR* abs/1103.4513 (2011). arXiv: 1103.4513. URL: http://arxiv.org/abs/1103.4513.

[6] P. Goralčík és V. Koubek. „On discerning words by automata". *Automata, Languages and Programming*. Szerk. Laurent Kott. Berlin, Heidelberg: Springer Berlin Heidelberg, 1986, 116–122. old. ISBN: 978-3-540-39859-2.

[7] J.M. Robson. „Separating strings with small automata". *Information Processing Letters* 30.4 (1989), 209–214. old. ISSN: 0020-0190. URL: https://www.sciencedirect.com/science/article/pii/0020019089902159.

[8] John Robson. „Separating Words with Machines and Groups." *RAIRO. Informatique Théorique et Applications* 30 (1996. jan.). DOI: 10.1051/ita/1996300100811.

[9] J. Barkley Rosser és Lowell Schoenfeld. „Approximate formulas for some functions of prime numbers". *Illinois Journal of Mathematics* 6.1 (1962), 64–94. old. URL: https://doi.org/10.1215/ijm/1255631807.

[10] Jeffrey Shallit és Yuri Breitbart. „Automaticity I: Properties of a Measure of Descriptional Complexity". *Journal of Computer and System Sciences* 53.1 (1996), 10–25. old. ISSN: 0022-0000. URL: https://www.sciencedirect.com/science/article/pii/S002200009690046X.

[11] Nicholas Tran. „Separating Words from Every Start State with Horner Automata". *Electronic Proceedings in Theoretical Computer Science* 386 (2023. szept.), 243–252. old. ISSN: 2075-2180. DOI: 10.4204/eptcs.386.19.

[12] Nicholas Tran. „Variations of the Separating Words Problem". *Implementation and Application of Automata*. Szerk. Pascal Caron és Ludovic Mignot. Cham: Springer International Publishing, 2022, 165–176. old. ISBN: 978-3-031-07469-1.

[13] M. N Vyalyi és R. A Gimadeev. „Separating words by occurrences of subwords". *Journal of Applied and Industrial Mathematics* (2014). DOI: https://doi.org/10.1134/S1990478914020161.