הרצאה 4

1 ביטויים רגולריים

עוד דרך לתאר שפה רגולרית.

1.1 הגדרה

אוסף הביטויים הרגולריים מעל א"ב Σ מסומן R_Σ, ומוגדר באינדוקציה מבנית באופן הבא:

אטומים:

- $\phi, \epsilon \in R$. הקבוצה הריקה והתו הריק.
- $\forall \sigma \in \Sigma, \sigma \in R$. כל אות בא"ב.

פעולות יצירה:

- אם $r_1, r_2 \in R$, אזי: $(r_1 + r_2) \in R$, $(r_1 \cdot r_2) \in R$.
- אם $r \in R$ אז $r^* \in R$.

דוגמאות: הביטויים הבאים הם ביטויים רגולריים מעל $\Sigma = \{a, b\}$:

$$\phi, \quad \epsilon, \quad a, \quad b, \quad (\epsilon + b), \quad \big((\epsilon + b) \cdot b\big), \quad \phi^*[= \epsilon], \quad \big((\epsilon + a) \cdot b^*\big)$$

1.2 שפה של ביטוי רגולרי

תהי $L[r]$ השפה שמציין הביטוי r. נגדיר את הפונקציה: $L : R \to 2^{\Sigma^*}$:

- $L[\phi] = \phi$
- $L[\epsilon] = \epsilon$
- $\forall \sigma \in \Sigma : L[\sigma] = \sigma$
- אם $r_1, r_2 \in R$, אז:
  - $L[(r_1 + r_2)] = L[r_1] \cup L[r_2]$
  - $L[(r_1 \cdot r_2)] = L[r_1] \cdot L[r_2]$
- אם $r \in R$, אז $L[(r^*)] = (L[r])^*$

דוגמה: $r = \Big(\big((a + b) + c\big) + d\Big)^*$.

$$L[r] = L\left[\Big(\big((a + b) + c\big) + d\Big)^*\right] = \left(L\left[\Big(\big((a + b) + c\big) + d\Big)\right]\right)^* = \Big(L\big[\big((a + b) + c\big)\big] \cup L[d]\Big)^*$$
$$= (L[(a + b)] \cup L[c] \cup L[d])^* = \cdots = (a + b + c + d)^*$$

1.3 קיצורי כתיבה של ביטויים רגולריים

אם r ביטוי רגולרי, נסמן ב- r^+ את הביטוי הרגולרי $\big(r \cdot (r^*)\big)$. ("איטרציה לא ריקה").

נקבע סדר קדימויות כדי שנוכל להשמיט סוגריים:

1. איטרציה: $*$ בקדימות גבוהה.
2. שרשור: $\cdot$ בקדימות בינונית.
3. איחוד: $+$ בקדימות נמוכה.

בנוסף, בדרך כלל נשמיט את האופרטור של השרשור.

לעיתים נשתמש בביטוי הרגולרי לציון השפה שהוא מייצג.

1.4 דוגמאות נוספות

- $L = \{w \in \{a,b,c\}^* : w = a^i b^j c^k, 0 \le i,j,k\}$.$a^*b^*c^*$
- שפת כל המילים מעל $\Sigma = \{a,b\}$. $\Sigma^* = (a+b)^*$.
- שפת כל המילים באורך זוגי מעל $\Sigma = \{a,b\}$. $(\Sigma\Sigma)^* = \big((a+b)(a+b)\big)^*$.

2 שקילות ביטויים רגולריים לאוטומטים

2.1 כיוון ראשון

משפט: לכל $r \in R$ מעל Σ מתקיים כי $L[r]$ היא שפה רגולרית. הוכחה באינדוקציה מבנית על r:

בסיס: $L[\phi] = \phi,\ L[\epsilon] = \epsilon,\ \ \forall\sigma \in \Sigma : L[\sigma] = \{\sigma\}$ כולן רגולריות.

צעד האינדוקציה נובע ישירות מסגירות השפות הרגולריות תחת פעולות רגולריות:

נניח שהטענה נכונה עבור ביטויים r_1, r_2 ונוכיח שהיא נכונה עבור: $r_1 + r_2,\ r_1r_2,\ (r_1^*)$. לפי סגירות לאיחוד, שרשור, ואיטרציה:

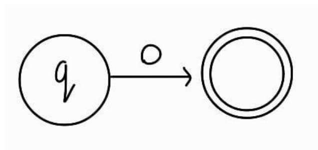
$L[r_1 + r_2] = L[r_1] \cup L[r_2],\ L[r_1r_2] = L[r_1]L[r_2],\ L[r_1^*] = (L[r_1])^*$ כולן רגולריות.

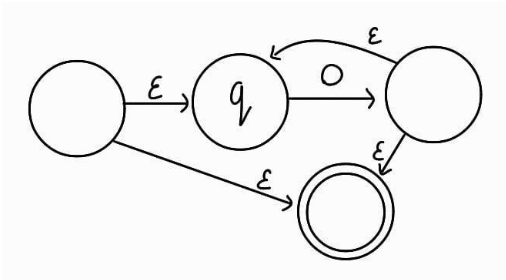
2.2 בניית אוטומט מתוך ביטוי רגולרי

נבנה אוטומט המקבל את השפה שמציין ביטוי רגולרי נתון לפי המשפט הקודם:

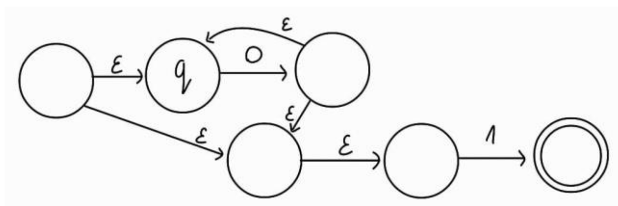
לדוגמה, תהי $L = 0^*1 + 1^*0$

נבנה אוטומט שמקבל את 0, ונרחיב אותו בשביל 0^*:

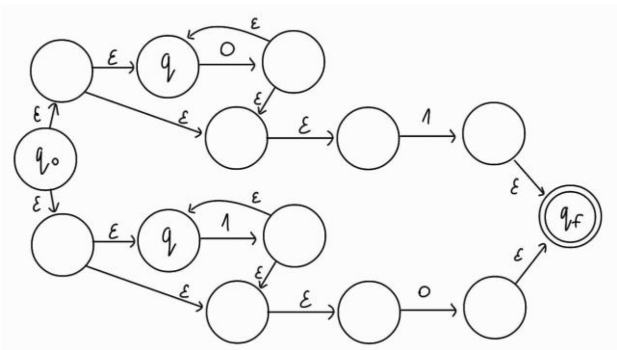




נוסיף את השרשור עם 1:



נבנה אחד דומה בשביל 1^*0, ונחבר ביניהם:



2.3 כיוון שני

משפט: לכל שפה רגולרית $L \subseteq \Sigma^*$ קיים ביטוי רגולרי r כך ש- $L[r] = L$.

הוכחה: L רגולרית, לכן קיים לה אס"ד $A = (\Sigma, \{q_1 \dots q_m\}, q_1, \delta, F)$ כך ש-$L(A) = L$.

לכל i,j,k נסמן $L_{i,j}^k$ את השפה שכוללת את המילים שמובילות את האוטומט מ-q_i ל-q_j בלי לעבור דרך מצב שמספרו גדול מ-k. ("לעבור דרך" אינו כולל את המצב שממנו יוצאים והמצב שאליו מגיעים. זוכרים "קודקודי ביניים" במסלולים קצרים באלגו 1?)

פורמלית:

$$L_{i,j}^k = \{w :\ \hat{\delta}(q_i,w)=q_j, \forall u,v \neq w, uv=w : \delta(q_i,u)=q_\ell \Rightarrow n \leq k\}$$

כל המילים w כך ש: w מובילה מ-q_i ל-q_j. ולכל שתי מילים (שהן לא w) שהשרשור שלהן הוא w, אם u מובילה מ-q_i ל-q_ℓ זה אומר ש $\ell \leq k$.

לפי הגדרה הקודמת, כיוון שאין מצב גדול מ-m הרי ש-$L_{i,j}^m$ כוללת את כל המילים המובילות את האוטומט מ-q_i ל-q_j. ובפרט:

$$L(A) = \bigcup_{q_j \in F} L_{1,j}^m$$

שימו לב שסימנו את המצב ההתחלתי ב-q_1.

השפה $L(A)$ היא איחוד של מספר סופי של שפות. לכן, אם נמצא לכל $L_{i,j}^k$ ביטוי רגולרי, נוכל למצוא ביטוי רגולרי עבור $L(A)$.

נוכיח באינדוקציה כי לכל i,j,k ניתן לבנות ביטוי רגולרי ל-$L_{i,j}^k$. יהיו i,j.

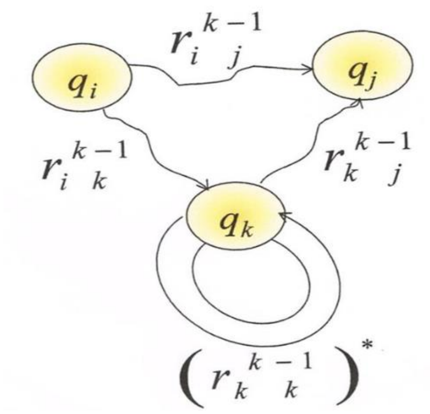
בסיס: עבור $k=0$, יתכן רק מעבר ישיר מ-q_i ל-q_j. (צעד אחד אם הם שונים, אפס צעדים אם הם שווים).

לכן כל מילה $w \in L_{i,j}^0$ היא בעלת אורך לכל היותר 1. קל לראות כי לשפה זו יש ביטוי רגולרי. לדוגמה $\epsilon, (\sigma_1+\sigma_2+\cdots+\sigma_n)$.
אם לא קיים מעבר מ-q_i ל-q_j בצעד יחיד, הרי שהביטוי הרגולרי המתאים הוא ϕ.

הגדרה רקורסיבית של $L_{i,j}^k$: נניח כי עבור $k-1$ ניתן לבנות ביטוי רגולרי ל- $L_{i,j}^{k-1}$, ונבנה ביטוי רגולרי - $L_{i,j}^k$.

עבור $k>0$, ב-$L_{i,j}^k$ קיימים שני סוגי מילים:

1. מילים שלא גורמות ל-A לעבור דרך q_k – אלה מילים ששייכות גם ל-$L_{i,j}^{k-1}$.
2. מילים שכן גורמות ל-A לעבור דרך q_k – ולכן לא שייכות ל-$L_{i,j}^{k-1}$.



את המילים הגורמות ל-A לעבור דרך q_k אפשר לחלק ל-3 חלקים uvw באופן הבא:

רישא u המובילה את A לביקור ראשון ב-q_k. $u \in L_{i,k}^{k-1}$.

חלק אמצעי v הגורם ל-A לבצע מספר סיבובים תוך חזרה ל-q_k. $v \in \left(L_{k,k}^{k-1}\right)^*$.

סיפא w המובילה את A מביקורו האחרון ב-q_k ל-q_j. $w \in L_{k,j}^{k-1}$.

סה"כ בניית הביטוי הרגולרי עבור האוטומט:

$$r_{i,j}^k = r_{i,j}^{k-1} + r_{i,k}^{k-1}\left(r_{k,k}^{k-1}\right)^* + r_{k,j}^{k-1}$$

2.4 משפט קליני

משפחת השפות הרגולריות היא הקבוצה הקטנה ביותר המכילה את כל השפות הסופיות והסגורה תחת הפעולות הרגולריות. הוכחה: בעצם נוכיח שקבוצת שפות מכילה את כל השפות הסופיות וסגורה לפעולות רגולריות אמ"מ היא משפחת השפות הרגולריות.

כיוון ראשון: אנחנו כבר יודעים שכל שפה סופית היא רגולרית. הוכחנו שקבוצת השפות הרגולריות סגורה לפעולות רגולריות.

כיוון שני: נובע מהמשפט האחרון, לפיו לכל שפה רגולרית קיים ביטוי רגולרי המציין אותה. כל קבוצה המכילה את השפות הסופיות והסגורה לפעולות רגולריות חייבת להכיל את כל השפות המצוינות ע"י ביטויים רגולריים, כלומר את השפות הרגולריות.

3 זהויות בין ביטויים רגולריים

לכל שפה קיימים הרבה ביטויים רגולריים המציינים אותה, ולכן נרצה לדעת מתי ביטויים רגולריים הם שקולים.

3.1 דוגמה 1

יהיו $r_1 = (0^*1)^*, \quad r_2 = \epsilon + (0+1)^*1$. נוכיח כי ביטויים אלה מייצגים את אותה השפה, ע"י הכלה דו כיוונית.

אינטואיטיבית: השפה הראשונה היא: איטרציה על: 0 איטרציה, משורשר עם 1.
השפה השנייה היא: אפשר לקחת: אפסילון, או שניקח מצד ימין: איטרציה על 0 או 1, ושרשור עם 1.

כיוון ראשון: נניח כי $w \in L[r_1]$. אזי $w = \epsilon$, או:

$$\exists w_1, \dots, w_n \in L[0^*1] : w = w_1 w_2 \cdots w_n$$

אם $w = \epsilon$ א בוודאי $w \in \{\epsilon\} \cup \{0,1\}^*\{1\} = L[r_2]$.

אחרת, w מסתיימת ב-1, ולכן $w \in \{\{0,1\}^*\{1\}\} \subseteq L[r_2]$.

כיוון שני: נניח כי $w \in L[r_2] = \{\epsilon\} \cup \{0,1\}^*\{1\}$.

אם $w = \epsilon$ אז בוודאי $w \in (\{0\}\{1\})^* = L[r_1]$.

אחרת, נוכל לכתוב $w = x1$ כאשר $x \in L[(0+1)^*]$. נניח כי ב-x יש k מופעים של 1.

במקרה זה נוכל לכתוב $x = y_1 1 y_2 1 \cdots y_k 1 y_{k+1}$, כאשר לכל $1 \le n \le k+1$ מתקיים $y_i \in \{0\}^*$.

ואז בעצם $w = (y_1 1)(y_2 1) \cdots (y_k 1)(y_{k+1} 1)$ ולכן $w \in L[r_1]$.

3.2 דוגמה 2

נראה שהביטויים $(0+1)^*, \ 0^* + 1^*$ אינם שקולים. אינטואיטיבית, כי הראשון זה כל המילים שהן רק 0 או רק 1, והשני זה כל המחרוזות הבינאריות. דוגמה נגדית פורמלית:

$01 \in L[(0+1)^*]$ כי $(0+1)^*$ היא שפת כל המילים מעל $\{0,1\}$. לעומת זאת, $01 \notin L[0^*]$ וגם $01 \notin L[1^*]$ ולכן $01 \notin L[0^*] \cup L[1^*] = L[0^* + 1^*]$.

3.3 הוכיחו/הפריכו

$(1+0)^* = (01^*)^* + (0^*1)^*$ – הפרכה: $10 \in L[(1+0)^*]$, אבל $10 \notin L[(0^*1)^*]$ וגם $10 \notin L[(01^*)^*]$.

$1(01)^* = (10)^*1$ – הוכחה: נראה הכלה דו-כיוונית:

כיוון ראשון: תהי $w \in L[1(01)^*]$. נראה באינדוקציה על $|w|$ ש $w \in L[(10)^*1]$.

בסיס: עבור $|w| = 1$, $w = 1$ ואכן $1 \in L[(10)^*1]$.

נניח שהטענה נכונה עבור $1w$ כאשר $|w| = t$, ונראה עבור $1x$ כאשר $|x| = t+2$. היות ו- $1x \in L[1(01)^*]$, ניתן לרשום $1x = 1w01$. לפי הנ"א, $1w \in L[(10)^*1]$ ולכן:

$$1x = 1w01 = (10)^k101 = (10)^{k+1}1 \in L[(10)^*1]$$

כיוון שני: בתרגול.

4 מעבר מאוטומט לביטוי רגולרי

למדנו את הוכחת השקילות בין אוטומט לביטוי רגולרי. לפעמים, כדי לכתוב ביטוי לשפה רגולרית, יהיה נוח לבצע את המעבר הזה בפועל, אבל הפעלת השקילות עלולה להיות ארוכה.

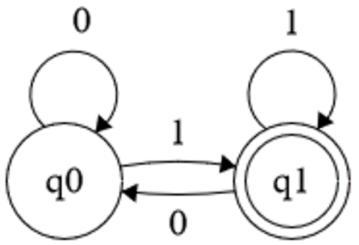
באופן כללי, המעבר יתבצע בצורה הבאה: $L(A) = \bigcup_{q\in F} L(q)$

ולכן נרצה לאפיין את השפה של כל מצב מקבל, ולבצע "איחוד" ביניהם.

בנוסף, באוטומטים יהיו מעגלים, ולכן נאפיין כיצד מבצעים את המעגלים האלו (דומה ברעיון להוכחה של $L_{k,k}^{k-1}$) בלי לעבור ב-q_0. כי אם חזרנו ל-q_0, חזרנו לרישא, אז המסלול הזה מיותר).

$$L[A] = (q_0 \to q_0)^*(q_0 \to q_f)(q_f \to q_f)^*$$

4.1 דוגמה 3



בנו ביטוי רגולרי לשפה: $L = \{w1 : w \in \{0,1\}^*\}$.

תחילה נבנה אוטומט:

ואז נבנה ממנו ביטוי רגולרי: נאפיין את המעגלים והמסלולים:

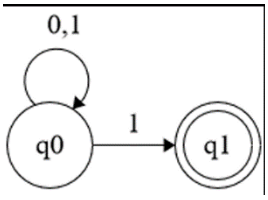
$$q_0 \to q_1 = 1, \qquad q_0 \to q_0 = 0 + 1^*0, \qquad q_1 \to q_1 = 1^*$$

(במעגל $q_1 \to q_1$, בלי לעבור ב-q_0).

ונרשום את השפה:

$$L = L[r] = (q_0 \to q_0)^*(q_0 \to q_1)(q_1 \to q_1)^* = (0^* + 1(1^*)0)^*1 \cdot (1^*)$$

4.2 דוגמה 3 שוב



בדוגמה הקודמת, היינו מגיעים לתשובה מהר יותר אם היינו בונים ב"ר מהאסל"ד של השפה:

כעת המעגלים פשוטים יותר: $q_0 \to q_0 = (0+1)^*$.

ולכן השפה היא $L = L[r] = (0+1)^*1$

וניתן להוכיח שקילות בין הביטוי מהדוגמה הקודמת לביטוי הנוכחי.

4.3 שיטת הבלוקים

לפעמים ציור האוטומט יהיה ארוך, והפקת הביטוי ממנו תהיה עוד יותר ארוכה כי יהיו הרבה מעגלים. לכן, יש שיטה נוספת להפקת ביטוי רגולרי, "שיטת הבלוקים". השיטה תעבוד על שפות בסגנון "כל המילים ללא תת מחרוזת x". "נפריד" את חלקי המילה לבלוקים של "מה כן מותר", ונראה מה קורה לפני הבלוק הראשון, בין הבלוקים, ולאחר הבלוק האחרון. ומשם נאפיין את השפה.

דוגמה: נבנה ב"ר לשפה $L = \{w \in \{a,b,c\}^* : w \text{ does not contain } 'abc'\}$.

נפריד בין כל שני a בעזרת בלוק, ונראה מה יכול להיות רשום בו: $\_\_a\_\_a\_\_a\_\_a\_\_$. בין כל שני a אסור שיהיה רשום bc. מה כן מותר? אם נהרוס את הרצף abc, הכל מותר. כלומר נראה c או bb, ואז באופן חופשי $(b+c)^*$. בנוסף, נראה b בודד או ϵ ואז נמשיך ל a הבאה.

כלומר, לאחר שראינו a ניתן לראות $(bb+c)(b+c)^* + b + \epsilon$. נסמן $r = (bb+c)(b+c)^* + b + \epsilon$ ונמשיך. אותו דבר אחרי ה-a האחרונה. לפני ה-a הראשונה אפשר לראות $(b+c)^*$ באופן חופשי. אפשר לראות את ar אינסוף פעמים, או אפס.
בסה"כ קיבלנו: $L = (b+c)^*(ar)^*$

4.4 תרגיל ממבחן

יהיו L_1, L_2 שפות רגולריות עם ביטויים רגולריים r_1, r_2 בהתאמה. נניח ש $L_1 \cap L_2 \neq \emptyset$. בנו ביטויים רגולריים לשפות:

$$L_3 = \{w_1 w_2 \dots w_n : \forall i \in [n]: w_i \in (L_1 \cup L_2) \wedge \mathit{at\ most\ 4\ words\ are\ from\ } L_2\}$$

באופן מגושם, נבנה כך שלכל היותר יש 4 מילים מ r_2:

$$L_3 = (r_1)^* r_2 (r_1)^* + (r_1)^* r_2 (r_1)^* r_2 (r_1)^* + (r_1)^* r_2 (r_1)^* r_2 (r_1)^* r_2 (r_1)^* + (r_1)^* r_2 (r_1)^* r_2 (r_1)^* r_2 (r_1)^* r_2 (r_1)^*$$

או באופן אלגנטי יותר:

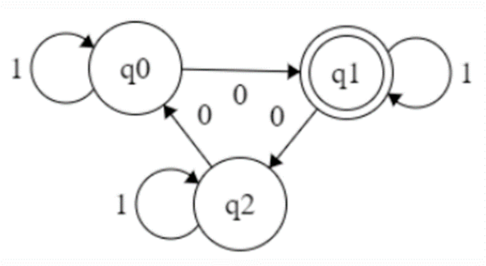
$$L_3 = (r_1)^* (r_2 + \epsilon)(r_1)^* (r_2 + \epsilon)(r_1)^* (r_2 + \epsilon)(r_1)^* (r_2 + \epsilon)(r_1)^*$$

$$L_4 = \left\{ \begin{matrix} w_1 w_2 \dots w_n : \forall i \in [n]: w_i \in (L_1 \cup L_2) \\ \wedge\ \mathit{no\ 3\ adjacent\ words\ are\ from\ } L_1, L_2, L_2\ (\mathit{in\ that\ order}) \end{matrix} \right\}$$

שיטת הבלוקים: יש בלוקים וביניהם r_1. בבלוק אמצעי, יכול להיות ϵ או r_2 בודד. כנ"ל באחרון. בבלוק הראשון זה $(r_2)^*$. כלומר:

$$L_4 = (r_2)^* \left(r_1 (r_2 + \epsilon) \right)^*$$

4.5 עוד דוגמה



בנו ב"ר לשפה: $L = \{w \in \{0,1\}^* : \#_0(w) \equiv 1 (mod\ 3)\}$. נבנה אוטומט:
ונבנה ממנו ב"ר:

$$L = L[A] = L[q_1] = (q_0 \to q_0)^* (q_0 \to q_1)(q_1 \to q_1)^*$$

$$(q_0 \to q_0) = 1 + 01^*01^*0, \qquad (q_0 \to q_1) = 0, \qquad (q_1 \to q_1) = 1$$

$$L = (1 + 01^*01^*0)^* 01^*$$