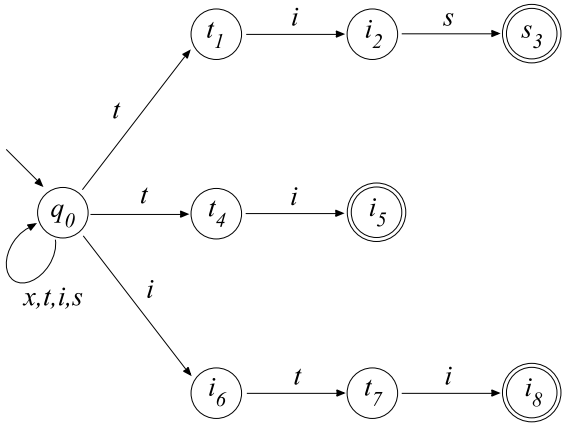
PV030 – Textové informační systémy

Úkoly 2013 – vzorové řešení (upravené řešení M. Matláka)

A) Mějme množinu vzorků P={tis, ti, iti}.

- Vytvořte NKA pro vyhledávání P.

  $P_{NKA} = \{Q, \Sigma, \delta, q_0, F\}$, $F = \{s_3, i_5, i_8\}$



	δ	t	i	s	x
→	q_0	q_0, t_1, t_4	q_0, i_6	q_0	q_0
	t_1		i_2		
	i_2			s_3	
←	s_3				
	t_4		i_5		
←	i_5				
	i_6	t_7			
	t_7		i_8		
←	i_8				

kde $x \in \Sigma \backslash \{t, i, s\}$.

- Vytvořte DKA příslušný tomuto NKA a zminimalizujte jej. Nakreslete přechodové diagramy obou automatů (DKA a minimálního DKA) a popište postup minimalizace.

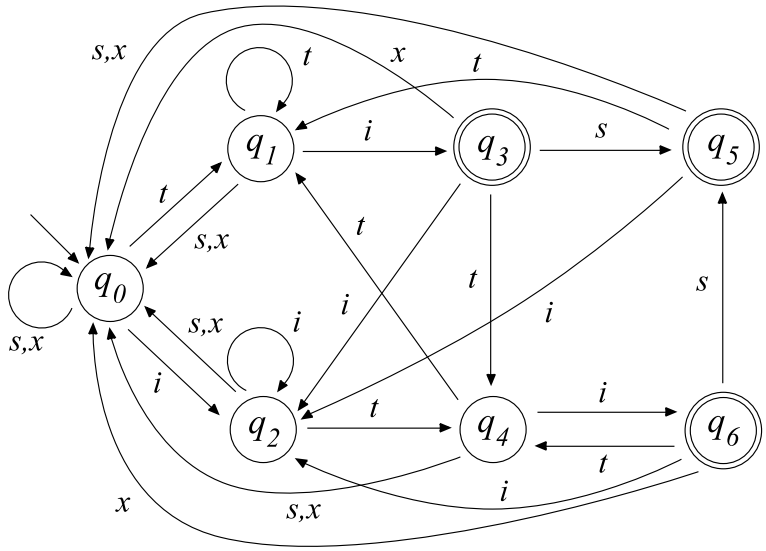
  Podle algoritmu z přednášky vytvoříme jazykově ekvivalentní deterministický automat:
  $P_{DKA} = \{Q', \Sigma, \delta', < q_0 >, F'\}$, $F' = \{< q_0 i_2 i_5 i_6 >, < q_0 s_3 >, < q_0 i_5 i_6 i_8 >\}$

	δ'	t	i	s	$x \in \Sigma \backslash \{t, i, s\}$
→	$< q_0 >$	$< q_0 t_1 t_4 >$	$< q_0 i_6 >$	$< q_0 >$	$< q_0 >$
	$< q_0 t_1 t_4 >$	$< q_0 t_1 t_4 >$	$< q_0 i_2 i_5 i_6 >$	$< q_0 >$	$< q_0 >$
	$< q_0 i_6 >$	$< q_0 t_1 t_4 t_7 >$	$< q_0 i_6 >$	$< q_0 >$	$< q_0 >$
←	$< q_0 i_2 i_5 i_6 >$	$< q_0 t_1 t_4 t_7 >$	$< q_0 i_6 >$	$< q_0 s_3 >$	$< q_0 >$
	$< q_0 t_1 t_4 t_7 >$	$< q_0 t_1 t_4 >$	$< q_0 i_2 i_5 i_6 i_8 >$	$< q_0 >$	$< q_0 >$
←	$< q_0 s_3 >$	$< q_0 t_1 t_4 >$	$< q_0 i_6 >$	$< q_0 >$	$< q_0 >$
←	$< q_0 i_2 i_5 i_6 i_8 >$	$< q_0 t_1 t_4 t_7 >$	$< q_0 i_6 >$	$< q_0 s_3 >$	$< q_0 >$

  Pro snazší určení minimálního $P_{DKA}/_{\equiv}$ přejmenujeme stavy v Q' automatu P_{DKA}. Tedy $P_{DKA} = \{Q', \Sigma, \delta', q_0, F'\}$, $F' = \{q_3, q_5, q_6\}$

	δ'	t	i	s	$x \in \Sigma \backslash \{t, i, s\}$
→	q_0	q_1	q_2	q_0	q_0
	q_1	q_1	q_3	q_0	q_0
	q_2	q_4	q_2	q_0	q_0
←	q_3	q_4	q_2	q_5	q_0
	q_4	q_1	q_6	q_0	q_0
←	q_5	q_1	q_2	q_0	q_0
←	q_6	q_4	q_2	q_5	q_0

Takto vypadá přechodový diagram P_{DKA}.



Automat P_{DKA} nyní zminimalizujeme (ztotožněním jazykově ekvivalentních stavů):

		$\equiv_0$	t	i	s	$x \in \Sigma\backslash\{t,i,s\}$
1	$\rightarrow$	q_0	1	1	1	1
		q_1	1	2	1	1
		q_2	1	1	1	1
		q_4	1	2	1	1
2	$\leftarrow$	q_3	1	1	2	1
	$\leftarrow$	q_5	1	1	1	1
	$\leftarrow$	q_6	1	1	2	1

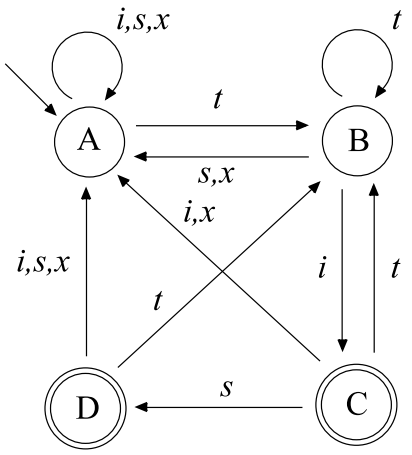
$\sim$

		$\equiv_1$	t	i	s	$x \in \Sigma\backslash\{t,i,s\}$
1	$\rightarrow$	q_0	2	1	1	1
		q_2	2	1	1	1
2		q_1	2	4	1	1
		q_4	2	4	1	1
3	$\leftarrow$	q_5	2	1	1	1
4	$\leftarrow$	q_3	2	1	3	1
	$\leftarrow$	q_6	2	1	3	1

Sestrojíme nyní finální přechodovou funkci (v normalizovaném tvaru) automatu $P_{DKA}/_{\equiv} = \{Q'/_{\equiv}, \Sigma, \eta, A, F'/_{\equiv}\}$, $F'/_{\equiv} = \{C, D\}$.

	η	t	i	s	$x \in \Sigma\backslash\{t,i,s\}$
$\rightarrow$	A	B	A	A	A
	B	B	C	A	A
$\leftarrow$	C	B	A	D	A
$\leftarrow$	D	B	A	A	A

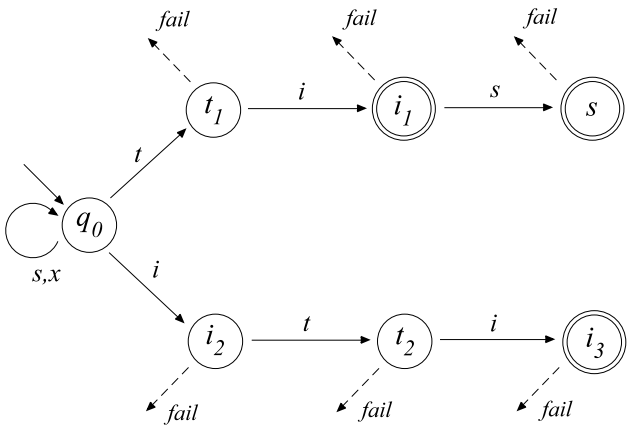
Vysledný diagram minimálního automatu $P_{DKA}/_{\equiv}$ ekvivalentního s P_{DKA} vypadá takto:



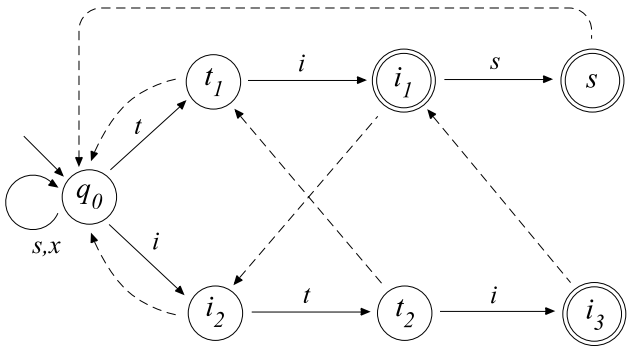
- Srovnejte s výsledkem vyhledávacího stroje AC.

  Sestrojíme tedy Aho a Corasickové (AC) vyhledávací stroj (VS):

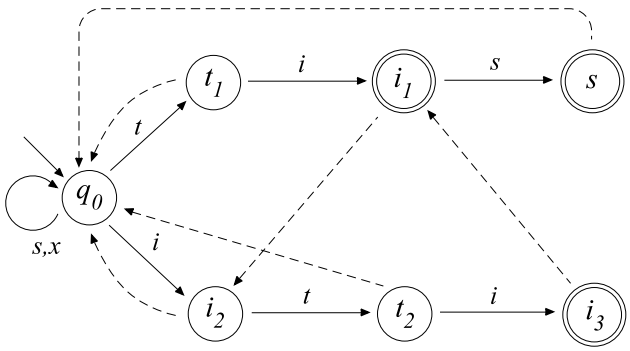
  Dopředná přechodová funkce g a její znázornění pomocí přechodového diagramu:



  Chybová funkce f a její znázornění (přerušovanou čarou) pomocí přechodového diagramu:



  Optimalizovaná chybová funkce h a její znázornění (přerušovanou čarou) pomocí přechodového diagramu:



Funkce f, g, h jsou uvedeny v následující tabulce:

		g				f	h
		t	i	s	x		
$\rightarrow$	q_0	t_1	i_2	q_0	q_0	–	–
	t_1		i_1			q_0	q_0
$\leftarrow$	i_1			s		q_0	q_0
$\leftarrow$	s					q_0	q_0
	i_2	t_2				q_0	q_0
	t_2		i_3			t_1	q_0
$\leftarrow$	i_3					i_1	i_1

Srovnání: VS AC je defacto *NKA* s ε-přechody. Po odstranění těchto přechodů determinizací získáme *DKA* ekvivalentní s P_{DKA} viz výše.

- Řešte úlohu také algoritmem přímé konstrukce *DKA* (derivováním) a diskutujte, zda výsledkem jsou izomorfní automaty.

$Q = Q_0 = \{tis + ti + iti\}$

$\underline{Q_1}$:

$\frac{d(tis+ti+iti)}{dt} = \frac{d(tis)}{dt} + \frac{d(ti)}{dt} + \frac{d(iti)}{dt} = \frac{dt}{dt}is + \frac{dt}{dt}i + \frac{di}{dt}ti = \varepsilon.is + \varepsilon.i + \emptyset.ti = \boxed{is + i}$

$\frac{d(tis+ti+iti)}{di} = \boxed{ti}$ [1] $\qquad \frac{d(tis+ti+iti)}{ds} = \emptyset$

$Q_1 = \{is + i, ti\}$

$\underline{Q_2}$:

$\frac{d(is+i)}{dt} = \emptyset \qquad \frac{d(is+i)}{di} = \boxed{s + \varepsilon} \qquad \frac{d(is+i)}{ds} = \emptyset$

$\frac{d(ti)}{dt} = \boxed{\text{i}} \qquad \frac{d(ti)}{di} = \emptyset \qquad \frac{d(ti)}{ds} = \emptyset$

$Q_2 = \{s + \varepsilon, i\}$

$\underline{Q_3}$:

$\frac{d(s+\varepsilon)}{dt} = \emptyset \qquad \frac{d(s+\varepsilon)}{di} = \emptyset \qquad \frac{d(s+\varepsilon)}{ds} = \boxed{\varepsilon}$

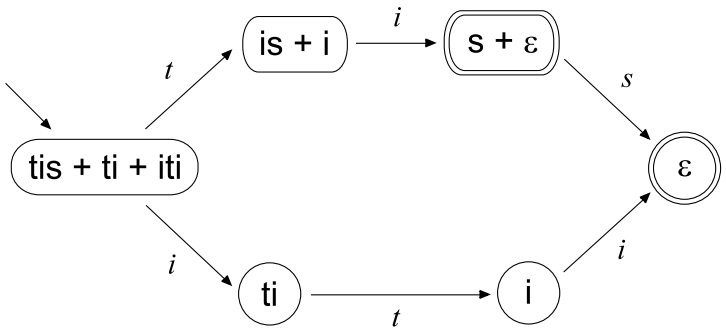
$\frac{di}{dt} = \emptyset \qquad \frac{di}{di} = \boxed{\varepsilon} \qquad \frac{di}{ds} = \emptyset$

$Q_3 = \{\varepsilon\}$

$Q = \{tis + ti + iti, is + i, ti, s + \varepsilon, i, \varepsilon\}$, $F = \{s + \varepsilon, \varepsilon\}$, $\Sigma = \{t, i, s\}$
$M = \{Q, \Sigma, \delta_Q, Q_0, F\}$ – minimální automat akceptující jazyk {tis,ti,iti}, který jsme získali výše jmenovaným algoritmem.

[1] Pro jednoduchost zápisu uvádím již bez elementárních úprav jen výsledky.

Přechodový graf výsledného automatu M je následující:



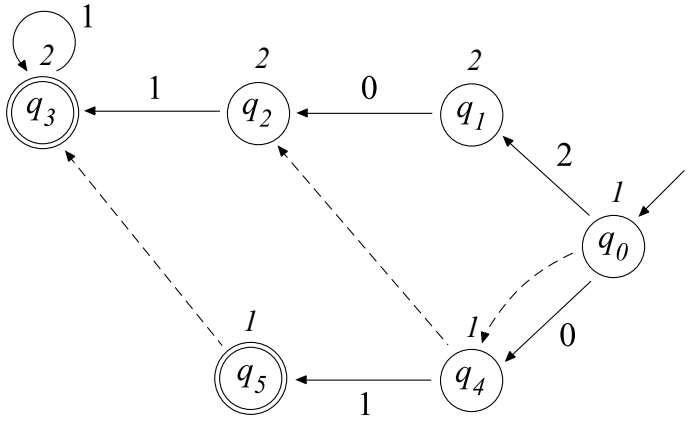
Nechť i_n je zobrazení $R_n \to Q$, kde R_n je jedna z množin stavů výše popsaných automatů ($R_1 = Q'$, $R_2 = Q'/_{\equiv}$, $R_3 = Q_{VS-AC}$), Q je množina stavů automatu M. Automat M není zřejmě izomorfní s automaty P_{DKA}, $P_{DKA}/_{\equiv}$ a $VS\text{-}AC$, neboť ani jedno ze zobrazení $i_1 \ldots i_3$ není bijektivní.

B) Mějme regulární výraz $R = 1(0 + 1^*02)$ nad abecedou $A = \{0, 1, 2\}$.

- Sestrojte *DKA* pro protisměrné vyhledávání R (Buczilowski) a spočtěte chybovou funkci. Nakreslete přechodový diagram tohoto automatu včetně vizualizace chybové funkce.

  $R^R = (0 + 201^*)1$

  Výsledný $R_{DKA} = \{Q, A, \delta, q_0, \{q_3, q_5\}\}$ je následující. Přerušované přechody vizualizují chybovou funkci. Čísla uvedená nad stavy označují velikost posunu při nesrovnání vzorku s textem.



	δ	0	1	2
$\to$	q_0	q_4		q_1
	q_1	q_2		
	q_2		q_2	
$\leftarrow$	q_3		q_3	
	q_4		q_5	
$\leftarrow$	q_5			

	$shift$	0	1	2
$\to$	q_0		1	
	q_1		2	2
	q_2	2		2
$\leftarrow$	q_3	2		2
	q_4	1		1
$\leftarrow$	q_5	1	1	1

- Zapište výsledný automat jako *2DKAS* a trasujte vyhledávání v textu 11201012102.

$R_{2DKAS} = \{\{start, q_0, q_1, q_2, q_3, q_4, q_5\}, \{0, 1, 2\}, \delta, start, 2, \uparrow, \{q_3, q_5\}\}$

	δ	0	1	2
$\rightarrow$	start	$(q_0, 1)$	$(q_0, 1)$	$(q_0, 1)$
	q_0	$(q_4, -1)$	$(q_0, 1)$	$(q_1, -1)$
	q_1	$(q_2, -1)$	$(q_0, 2)$	$(q_0, 2)$
	q_2	$(q_0, 2)$	$(q_2, -1)$	$(q_0, 2)$
$\leftarrow$	q_3	$(q_0, 2)$	$(q_3, -1)$	$(q_0, 2)$
	q_4	$(q_0, 1)$	$(q_5, -1)$	$(q_0, 1)$
$\leftarrow$	q_5	$(q_0, 1)$	$(q_0, 1)$	$(q_0, 1)$

Trasování vyhledávání vzorku R v zadaném textu:

$$\begin{aligned} \text{start} \uparrow \quad & 11201012102 \\ \vdash \quad & 1q_0\uparrow 1201012102 \\ \vdash \quad & 11q_0\uparrow 201012102 \\ \vdash \quad & 1q_1 1\uparrow 201012102 \\ \vdash \quad & 1120q_0\uparrow 1012102 \\ \vdash \quad & 11201q_0\uparrow 012102 \\ \vdash \quad & 1120q_4 1\uparrow 012102 \\ \vdash \quad & 112q_5 01\uparrow 012102 \end{aligned}$$

Stav q_5 je akceptující. R_{2DKAS} je ovšem navržen pro vyhledání všech vzorků obsažených v textu, proto doplňuji zbývající konfigurace až do konce textu.

$$\begin{aligned} \vdash \quad & 112010q_0\uparrow 12102 \\ \vdash \quad & 1120101q_0\uparrow 2102 \\ \vdash \quad & 112010q_1 1\uparrow 2102 \\ \vdash \quad & 112010121q_0\uparrow 02 \\ \vdash \quad & 11201012q_4 1\uparrow 02 \\ \vdash \quad & 1120101q_5 21\uparrow 02 \\ \vdash \quad & 1120101210q_0\uparrow 2 \\ \vdash \quad & 112010121q_1 0\uparrow 2 \\ \vdash \quad & 11201012q_2 10\uparrow 2 \\ \vdash \quad & 1120101q_3 210\uparrow 2 \\ \vdash \quad & 11201012102q_0\uparrow \end{aligned}$$

C) Dokažte:

$$h\left(\frac{dV}{dx}\right) = \{y : xy \in h(V)\}$$

Důkaz provedeme indukcí:

1. $|x| = 0 \quad (x = \varepsilon)$

   $h\left(\frac{dV}{dx}\right) = h(V) \Leftrightarrow \{y : \varepsilon y \in h(V)\}$

2. Předpokládejme platnost $h\left(\frac{dV}{dx}\right) = \{y : xy \in h(V)\}$ pro $|x| = k$, dokážeme platnost i pro $|x| = k + 1$.

$$h\left(\frac{dV}{dxa}\right) = h\left(d\,\frac{\frac{dV}{dx}}{a}\right) = h\left(d\,\frac{\{y : xy \in h(V)\}}{a}\right) = h\left(\frac{dy_1}{da} + \frac{dy_2}{da} + \cdots + \frac{dy_n}{da}\right) =$$

$$= h\left[\left(\frac{dy_{11}}{da} y_{12} \ldots y_{1m}\right) + \left(\frac{dy_{21}}{da} y_{22} \ldots y_{2m}\right) + \cdots + \left(\frac{dy_{n1}}{da} y_{n2} \ldots y_{nm}\right)\right] =$$

$$= h\left[(\emptyset + \varepsilon y_{12} \ldots y_{1m}) + (\emptyset + \varepsilon y_{22} \ldots y_{2m}) + \cdots + (\emptyset + \varepsilon y_{n2} \ldots y_{nm})\right] =$$

$$= \{y' : xay' \in h(V)\}, \text{ kde } y'_i = y_{i2} \ldots y_{im}$$

Důkaz strukturní indukcí je možný také.

D) Najděte takový příklad řetězců X a Y, že platí zároveň $R(X, Y) = 5$, $DIR(X, Y) = 4$ i $DIRT(X, Y) = 3$, nebo dokažte neexistenci takových řetězců.

Možné řešení je: X=aabba a Y=bbaab nebo X=abcde a Y=badcf.