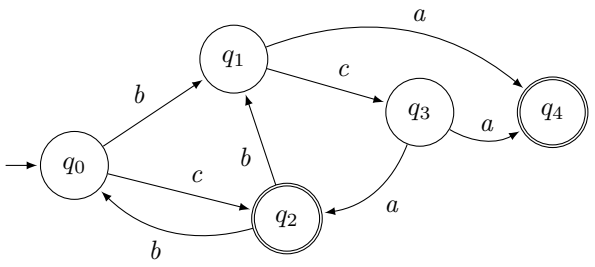
Sommersemester 2018

Einführung in die Computerlinguistik
Übungsaufgaben für die Zwischenklausur

Laura Kallmeyer

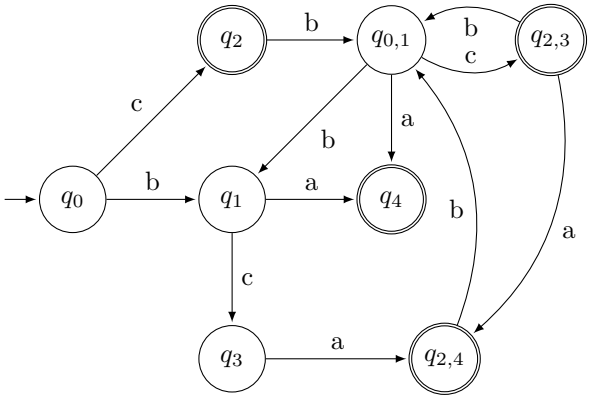
Heinrich-Heine-Universität Düsseldorf

Aufgabe 1 *Consider the following NFA:*



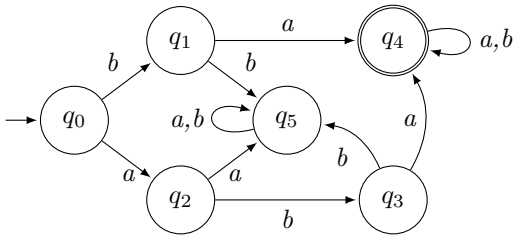
1. *Give the description of this automaton as a tuple* $\langle Q, \Sigma, \delta, q_0, F\rangle$.
2. *Why is this automaton non-deterministic?*
3. *Build (and draw) an equivalent DFA for the automaton.*

Solution:

1. $\langle\{q_0, q_1, q_2, q_3, q_4\}, \{a, b, c\}, \delta, q_0, \{q_2, q_4\}\rangle$ with
   $\delta(q_0, b) = \{q_1\}$ $\quad$ $\delta(q_1, a) = \{q_4\}$
   $\delta(q_0, c) = \{q_2\}$ $\quad$ $\delta(q_1, c) = \{q_3\}$
   $\delta(q_2, b) = \{q_0, q_1\}$ $\quad$ $\delta(q_3, a) = \{q_2, q_4\}$
   All other transitions lead to $\emptyset$
2. Some states have transitions to more than one state with a single symbol, e.g. q_2 with b.
3. 

Aufgabe 2 *Consider the following DFA.*

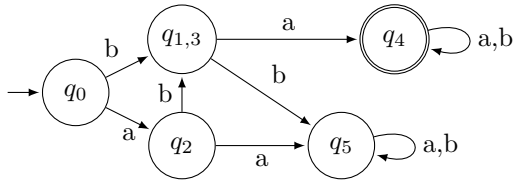
- *Minimize the DFA. Please detail the construction steps (matrix) and draw the minimized automaton.*
- *Give a regular expression denoting the language accepted by this automaton.*



Solution:

1.

	0	1	2	3	4
5	X	X	X	X	X
4	X	X	X	X	
3	X		X		
2	X	X			
1	X				



2. $(b|ab)a(a|b)^*$

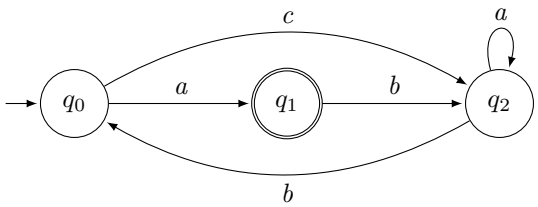
Aufgabe 3

1. *Which languages are denoted by the following regular expressions?*
   (a) $b|ac$ *(b)* $((a|b|\emptyset)c)^*$ *(c)* $((a|b|\varepsilon)c)^*$
2. *give a regular expression denoting each of the following languages:*
   (a) $\{a, aab, ab\}$
   (b) $\{w \in \{a,b,c\}^* \mid$ *no more than 3 c can come in a row*$\}$
   (c) $\{w \in \{a,b\}^* \mid$ *the number of a is even and between each two a there must be at least one b*$\}$

Solution:

1. (a) $L(b|ac) = \{b, ac\}$
   (b) $L(((a|b|\emptyset)c)^*) = \{ac, bc\}^* = \{w \in \{a,b,c\}^* \mid$ each a and each b in w are immediately followed by a c and there are no two c's in a row$\}$
   (c) $L(((a|b|\varepsilon)c)^*) = \{ac, bc, c\}^* = \{w \in \{a,b,c\}^* \mid$ each a and each b in w are immediately followed by a $c\}$
2. (a) $(a|ab|aab)$
   (b) $(a|b)^*|(a|b)^*((c|cc|ccc)(a|b)^+)^*((c|cc|ccc)(a|b)^*)$
   (c) $b^*(\varepsilon|(ab^+ab^+)^*ab^+ab^*)$

Aufgabe 4 *Gegeben sei der folgende DFA:*



1. *Geben Sie einen regulären Ausdruck an, der die Sprache beschreibt, die von dem Automaten akzeptiert wird.*
2. *Beschreiben Sie in Einzelschritten, wie Sie den regulären Ausdruck berechnen.*

Lösung:

1. $a|(ab|c)(a|b(ab|c))^*ba$
2. $r^2_{0,1} = r^1_{0,1}|r^1_{0,2}(r^1_{22})^*r^1_{21}$
   $r^1_{0,1} = a$
   $r^1_{0,2} = (ab|c)$
   $r^1_{2,2} = (\varepsilon|a|b(ab|c))$
   $r^1_{2,1} = ba$
   $r^2_{0,1} = a|(ab|c)(\varepsilon|a|b(ab|c))^*ba$

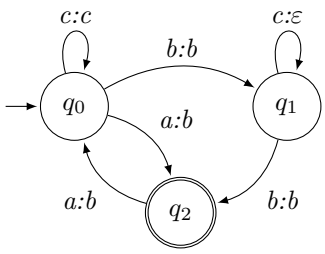
Aufgabe 5

1. *Which language is generated by the following left-regular grammar?*
   $G = \langle\{S, A, B\}, \{a, b\}, P, S\rangle$ *with the following productions in* P*:*
   $S \to Ab, S \to Bb, A \to Ab, A \to \varepsilon, B \to Baa, B \to aa$
   You can either describe the language or give a regular expression for it.
2. *Give a right-regular grammar for the following language:*
   $L = \{w \in \{a, b\}^* \mid$ *w contains an even (0,2,4,6,etc) number of as and any number of bs (possibly 0)*$\}$

Solution:

1. Regular expression: $b^+|(aa)^+b$.
   Language: $\{b^n \mid n \geq 1\} \cup \{(aa)b^m \mid m \geq 1\}$
2. Grammar: $\langle\{S, A\}, \{a, b\}, P, S\rangle$ with the following productions in P:
   $S \to bS, S \to \varepsilon, S \to aA, A \to bA, A \to aS$

Aufgabe 6 *Consider the following FST:*



1. *Which strings does the FST build for the following inputs (a) bcb (b) caacbb (c) aabcbaca*
2. *Which strings are accepted by this FST and how are they transformed? (a description in your own words is enough)*

Solution:

1. (a) *bb* (b) *cbbcbb* (c) *bbbbbcb*

2. The strings accepted are denoted by the regular expression $(c^*(bc^*b|a)a)^*c^*(a|bc^*b)$. The FST transforms every a into b , copies b into the output and for all odd numbers i, it removes cs that occur between the ith b and the $(i+1)$th b in the input. All other cs are copied into the output.

Aufgabe 7 *Consider a bigram language model for a language consisting of sequences over* $\{0,1\}$*. The language is such that every sequence starts with a* 0*, every 100th* 0 *on average is followed by a* 1*, all other* 0*s are followed by* 0*, sequences never end with* 0*, and every second* 1 *on average is follwed by another* 1 *and every 4th* 1 *on average is followed by a* 0*.*

1. *What are the bigram probabilities for the language model, including the probabilities of having a sentence-initial* 0 *or* 1 *and having an end-of-sentence marker following a* 0 *or* 1*?*
2. *Given these probabilities, compute the perplexity of* 010101*. Include the probability of the end of sequence following the last* 1 *and take* $n = 6$ *to be the relevant sentence length.*

Solution:

1. Bigram probabilities:
   $P(1|\texttt{<s>}) = 0 \quad P(0|0) = 0.99 \quad P(0|1) = 0.25$
   $P(0|\texttt{<s>}) = 1 \quad P(1|0) = 0.01 \quad P(1|1) = 0.5$
   $P(\texttt{</s>}|0) = 0 \quad P(\texttt{</s>}|1) = 0.25$
2. Perplexity of 010101:
   $\frac{1}{\sqrt[6]{0.01 \cdot \frac{1}{4} \cdot 0.01 \cdot \frac{1}{4} \cdot 0.01 \cdot \frac{1}{4}}} = \frac{1}{\sqrt[6]{\frac{1}{2}^6 \cdot \frac{1}{10}^6}} = \frac{1}{\frac{1}{2 \cdot 10}} = 20$

Aufgabe 8 *Consider that you have a HMM-POS Tagger, with the following probabilities (N, V are the possible POS-Tags):*

Emission probabilities:

$P(I|N) = 1 \cdot 10^{-3} \quad P(like|N) = 3 \cdot 10^{-3} \quad P(like|V) = 4 \cdot 10^{-3}$
$P(likes|N) = 2 \cdot 10^{-3} \quad P(likes|V) = 3 \cdot 10^{-3}$

All the other probabilities for I, like *and* likes *are* 0*.*

Transition probabilities (amongst others):

$P(V|N) = 4 \cdot 10^{-1} \quad P(N|V) = 5 \cdot 10^{-1} \quad P(N|N) = 1 \cdot 10^{-1} \quad P(V|V) = 1 \cdot 10^{-1}$

The probability that a sentence starts with N *is* $2 \cdot 10^{-1}$*. The probability that the sentence end follows a* V *or a* N *is* $1 \cdot 10^{-1}$*.*

*Give the Viterbi Matrix which gives the probabilities for the input "*I like likes*". It is enough to give the probabilities which are* $\neq 0$*. For every (non-empty) cell, show the calculation.*

Solution:

q_F				$320 \cdot 10^{-13}$, N
V		$32 \cdot 10^{-8}$, N	$96 \cdot 10^{-12}$, V	
N	$2 \cdot 10^{-4}, q_0$	$6 \cdot 10^{-8}$, N	$320 \cdot 10^{-12}$, V	
	1 I	2 like	3 likes	

I, N: $1 \cdot 10^{-3} \cdot 2 \cdot 10^{-1}$
like, N: $2 \cdot 10^{-4} \cdot 1 \cdot 10^{-1} \cdot 3 \cdot 10^{-3}$
like, V: $2 \cdot 10^{-4} \cdot 4 \cdot 10^{-1} \cdot 4 \cdot 10^{-3}$
likes, N: $max\{6 \cdot 10^{-8} \cdot 1 \cdot 10^{-1} \cdot 2 \cdot 10^{-3}$ (previous N), $32 \cdot 10^{-8} \cdot 5 \cdot 10^{-1} \cdot 2 \cdot 10^{-3}$ (previous V)$\} = 320 \cdot 10^{-12}$ (previous V)
likes, V: $max\{6 \cdot 10^{-8} \cdot 4 \cdot 10^{-1} \cdot 3 \cdot 10^{-3}$ (previous N), $32 \cdot 10^{-8} \cdot 1 \cdot 10^{-1} \cdot 3 \cdot 10^{-3}$ (previous V)$\} = 96 \cdot 10^{-12}$ (previous V)
q_F: $320 \cdot 10^{-12} \cdot 1 \cdot 10^{-1}$