Automata and Formal Languages — Exercise Sheet 6

Exercise 6.1

1. Build the automata B_p and C_p for the word pattern $p = mammamia$.
2. How many transitions are taken when reading $t = mami$ in B_p and C_p?
3. Let $n > 0$. Find a text $t \in \{a, b\}^*$ and a word pattern $p \in \{a, b\}^n$ such that testing whether p occurs in t takes n transitions in B_p and $2n - 1$ transitions in C_p.

Exercise 6.2

In order to make pattern-matching robust to typos we want to include also "similar" words in our results. For this we consider words with a small Levenshtein-distance (edit-distance) "similar".

We transform a word w to a new word w' using the following operations (with $a_i, b \in \Sigma$):

- *replace* (R): $a_1 \dots a_{i-1}a_ia_{i+1} \dots a_l \to a_1 \dots a_{i-1}ba_{i+1} \dots a_l$
- *delete* (D): $a_1 \dots a_{i-1}a_ia_{i+1} \dots a_l \to a_1 \dots a_{i-1}\varepsilon a_{i+1} \dots a_l$
- *insert* (I): $a_1 \dots a_{i-1}a_ia_{i+1} \dots a_l \to a_1 \dots a_{i-1}a_iba_{i+1} \dots a_l$

The Levenshtein-distance (denoted $\Delta(w, w')$) of w and w' is the minimal number of operations (R,D,I) needed to transform w into w'. We denote with $\Delta_{L,i} = \{w \in \Sigma^* \mid \exists w' \in L.\, \Delta(w', w) \leq i\}$ the language of all words with edit-distance at most i to some word of L.

(a) Compute $\Delta(abcde, accd)$.

(b) Prove the following statement: If L is a regular language, then $\Delta_{L,n}$ is a regular language.

(c) Let p be the pattern $ABBA$. Construct an NFA-ϵ locating the pattern or variations of it with edit-distance 1.

Exercise 6.3

(a) Let $n \in \mathbb{N}$ be such that $n \geq 2$. Show that $L_n = \{w \in \{a, b\}^* \mid |w| \equiv 0 \pmod{n}\}$ has exactly n residuals, without constructing any automaton for L_n.

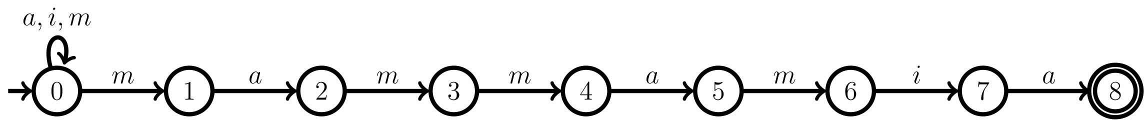
(b) Consider the following "proof" showing that L_2 is non regular:

> Let $i, j \in \mathbb{N}$ be such that i is even and j is odd. By definition of L_2, we have $\varepsilon \in (L_2)^{a^i}$ and $\varepsilon \notin (L_2)^{a^j}$. Therefore, the a^i-residual and a^j-residual of L_2 are distinct. Since there are infinitely many even numbers i and odd numbers j, this implies that L_2 has infinitely many residuals, and hence that L_2 is not regular. □

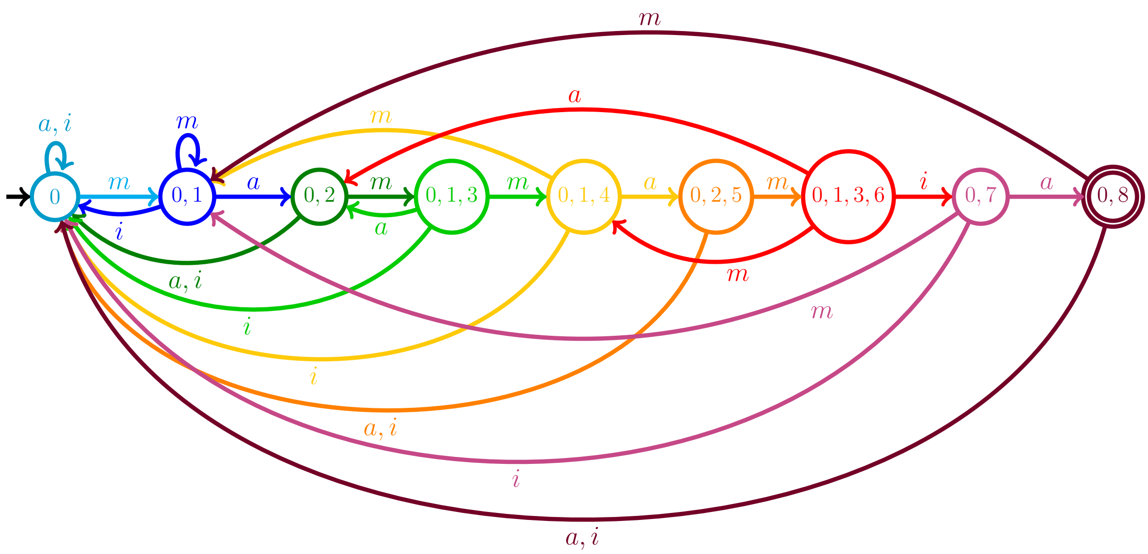
Language L_2 is regular, so this "proof" must be incorrect. Explain what is wrong with the "proof".

Solution 6.1

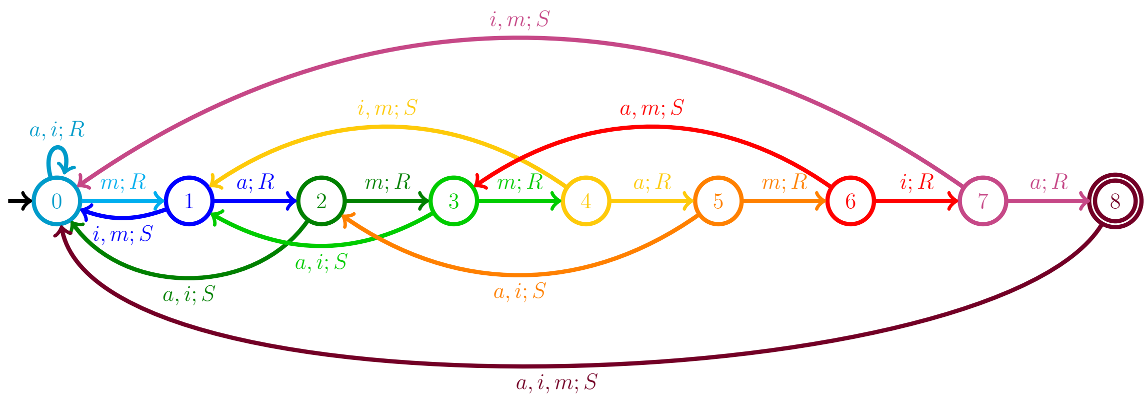
1. A_p :



B_p :



C_p :

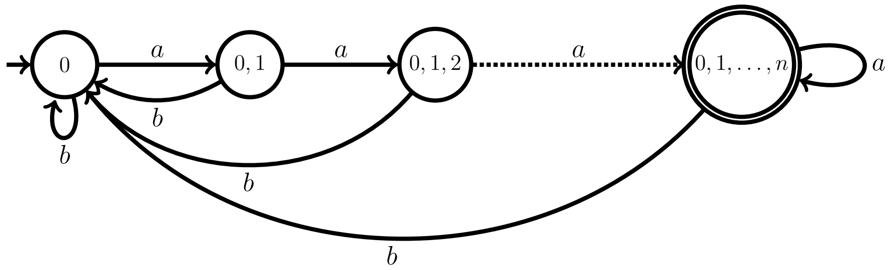


2. Four transitions taken in B_p: $\{0\} \xrightarrow{m} \{0,1\} \xrightarrow{a} \{0,2\} \xrightarrow{m} \{0,1,3\} \xrightarrow{i} \{0\}$.

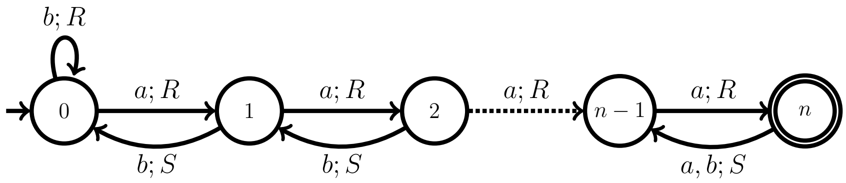
   Six transitions taken in C_p: $0 \xrightarrow{m} 1 \xrightarrow{a} 2 \xrightarrow{m} 3 \xrightarrow{i} 1 \xrightarrow{i} 0 \xrightarrow{i} 0$.

3. $t = a^{n-1}b$ and $p = a^n$. The automata B_p and C_p are as follows:

   B_p:



C_p:



The runs over t on B_p and C_p are respectively:

$$\{0\} \xrightarrow{a} \{0,1\} \xrightarrow{a} \{0,1,2\} \xrightarrow{a} \cdots \xrightarrow{a} \{0,1,\dots,n-1\} \xrightarrow{b} \{0\}\ ,$$

and

$$0 \xrightarrow{a} 1 \xrightarrow{a} 2 \xrightarrow{a} \cdots \xrightarrow{a} (n-1) \xrightarrow{b} (n-2) \xrightarrow{b} (n-3) \xrightarrow{b} \cdots \xrightarrow{b} 0\ .$$

Solution 6.2

(a) $\Delta(abcde, accd) = 2$.

(b) Let $M = (Q, \Sigma, \delta, q_0, F)$ be a DFA for L. We obtain an NFA-ϵ N for $\Delta_{L,n}$ by adding n "error-levels". Formally:

$$N = (Q \times [0,n], \Sigma, \delta', (q_0, 0), F \times [0,n])$$

with

$$\begin{aligned} \delta' &= \{((q,i), a, (p,i)) \mid q,p \in Q \wedge i \le n \wedge a \in \Sigma \wedge \delta(q,a) = p\} && \text{no change} \\ &\cup \{((q,i), \varepsilon, (p,i+1)) \mid q,p \in Q \wedge i < n \wedge (\exists a \in \Sigma.\, \delta(q,a) = p)\} && \text{delete} \\ &\cup \{((q,i), a, (q,i+1)) \mid q \in Q \wedge i < n \wedge a \in \Sigma\} && \text{insert} \\ &\cup \{((q,i), b, (p,i+1)) \mid q,p \in Q \wedge i < n \wedge (\exists a \in \Sigma \setminus \{b\}.\, \delta(q,a) = p)\} && \text{replace} \end{aligned}$$

Let us prove that $\Delta_{L,n} = L(N)$.

$\Delta_{L,n} \subseteq L(N)$. If $w \in \Delta_{L,n}$, it means that there is $w' \in L$ such that $\Delta(w', w) = k \le n$, or in other words, starting from the word w', we can obtain w by applying k "mistakes" (delete, insert, replace). As $w' \in L$ (accepted by M) and as the 0-level of N is a copy of M, note that w' has a run in N that reaches a final state $(q_f, 0)$. By construction of the automaton N, there is a run of the word w that follows the run of w' where each "mistake" can be seen as moving to the next error-level, using the corresponding transition from δ' (delete, insert, replace) depending on a mistake. It is easy to see that if the word w' reaches a final state $(q_f, 0)$ in N, then w can reach (q_f, k), and thus $w \in L(N)$.

$L(N) \subseteq \Delta_{L,n}$. If $w \in L(N)$, this means there is a run of w in N that reaches a final state $(q_f, k) \in F \times [0,n]$. Intuitively, for each transition of that run that changes the level, we modify w so that it "stays in the same level". Formally, we check the nature of the transition that changes the level and modify w as follows:

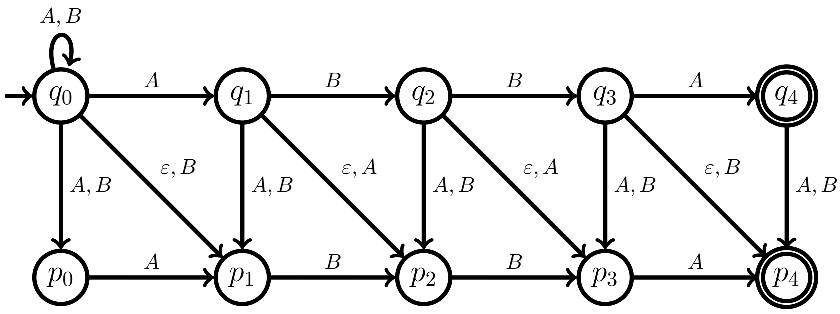
(i) If $(p,i) \xrightarrow{a} (p,i+1)$ is an insert edge, this occurrence of the letter a will be removed from w.

(ii) If $(p,i) \xrightarrow{a} (q,i+1)$ is a replace edge, and there exists a $(p,i) \xrightarrow{b} (q,i)$ edge, for some letter b, then we replace this occurrence of a in w with b.

(iii) If $(p,i) \xrightarrow{\epsilon} (q,i+1)$ is a delete edge, and there exists a $(p,i) \xrightarrow{a} (q,i)$ edge, for some letter a, then we add the letter a at this place in w.

Denote the obtained word by w'. It is easy to see that w' is obtained from w by applying mistakes (delete, insert, replace) k times, as in the run of w there are exactly k transitions that change the level. Therefore, $\Delta(w', w) \leq k \leq n$. Moreover, it is easy to see that if w reaches (q_f, k), then w' reaches $(q_f, 0)$. As the 0-level is a copy of M, then $w' \in L$. To summarize, there exists $w' \in L$ such that $\Delta(w', w) \leq n$, that is, $w \in \Delta_{L,n}$.

(c) We use the same construction as in (b) with the automaton A_p for pattern $p = ABBA$.



Solution 6.3

(a) We claim that the residuals of L_n are

$$(L_n)^{a^0}, (L_n)^{a^1}, \ldots, (L_n)^{a^{n-1}}. \tag{1}$$

Let us first show that for every word w we have $(L_n)^w = (L_n)^{a^{|w| \bmod n}}$. Let $w \in \{a, b\}^*$. For every $u \in \{a, b\}^*$, we have

$$\begin{aligned} u \in (L_n)^w &\iff wu \in L_n \\ &\iff |wu| \equiv 0 \pmod{n} \\ &\iff |w| + |u| \equiv 0 \pmod{n} \\ &\iff (|w| \bmod n) + |u| \equiv 0 \pmod{n} \\ &\iff |a^{|w| \bmod n}| + |u| \equiv 0 \pmod{n} \\ &\iff |a^{|w| \bmod n} u| \equiv 0 \pmod{n} \\ &\iff a^{|w| \bmod n} u \in L_n \\ &\iff u \in (L_n)^{a^{|w| \bmod n}}. \end{aligned}$$

It remains to show that the residuals of (1) are distinct. Let $0 \leq i, j < n$ be such that $i \neq j$. We have $a^{n-i} \in (L_n)^{a^i}$, and $a^{n-i} \notin (L_n)^{a^j}$ since $|a^j a^{n-i}| \bmod n = j - i \neq 0$. Therefore, $(L_n)^{a^i} \neq (L_n)^{a^j}$. □

(b) The part of the "proof" showing that $(L_2)^{a^i} \neq (L_2)^{a^j}$, for every even i and odd j, is correct. However, this only shows that L_2 has at least two residuals. Indeed, even if there are infinitely many even and odd numbers, the following is not ruled out:

$$\begin{aligned} (L_2)^{a^0} &= (L_2)^{a^2} = (L_2)^{a^4} = \cdots, \\ (L_2)^{a^1} &= (L_2)^{a^3} = (L_2)^{a^5} = \cdots. \end{aligned}$$

In order to show that a language has infinitely many residuals, one must exhibit an infinite subset of residuals that are *pairwise* distinct.