第二次作业：正则表达式

3.1.1. 写出表示下列语言的正则表达式:

a)

(a+b+c)*(a(a+b+c)*b+b(a+b+c)*a)(a+b+c)*

b)

(0+1)*1(0+1)(0+1)(0+1)(0+1)(0+1)(0+1)(0+1)(0+1)(0+1)(0+1)(0+1)

c)

(0+10)*(11+1+ε)(0+01)*

3.1.3. 写出表示下列语言的正则表达式

a)

((ε+000*)1(ε+000*))*

b)

(10)*+(01)*

c)

(00000+11+(01+10)(11)*10000+(001+(01+10)(11)*(0+101))(11)*1000+(0001+(01+10)(11)*1001+(001+(01+10)(11)*(0+101))(11)*(0+101))(11)*100+(00001+(01+10)(11)*10001+(001+(01+10)(11)*(0+101))(11)*1001+(0001+(01+10)(11)*1001+(001+(01+10)(11)*(0+101))(11)*(0+101))(11)*(0+101))(11+00(11)*10001+00(11)*(0+101)(11)*1001+(00(11)*1001+00(11)*(0+101)(11)*(0+101))(11)*(0+101))*(10+01+00(11)*10000+00(11)*(0+101)(11)*1000+(00(11)*1001+00(11)*(0+101)(11)*(0+101))(11)*100))*

3.1.4. 给出下列正则表达式语言的自然语言描述:

a)

不包含连续的 1 的所有 0 和 1 的串的集合

b)

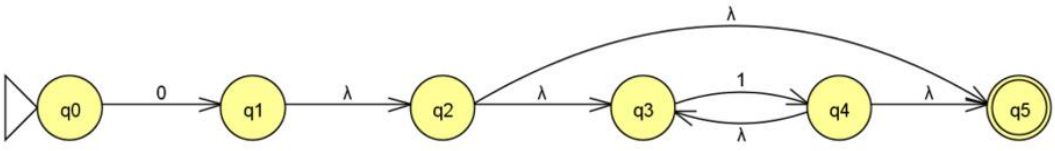
包含三个连续的 0 作为子串的所有 0 和 1 的串的集合

c)

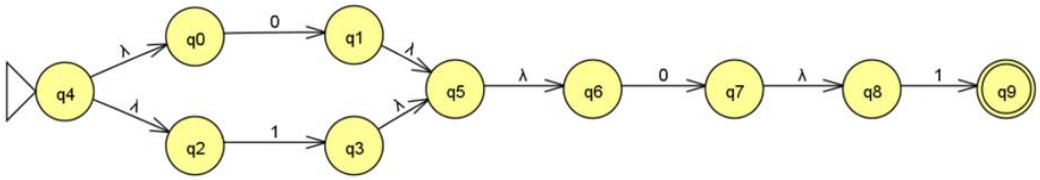
除了字符串末尾外不包含连续的 1 的所有 0 和 1 的串的集合

3.2.4. 把下列正则表达式转化成带 ε 转移的 NFA

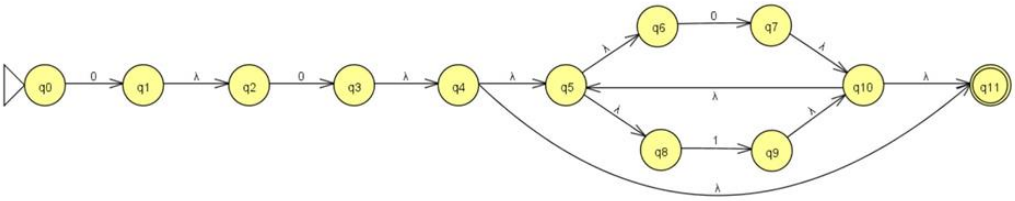
a)



b)



c)



3.2.8.

设对于某个字母集 S，A 的状态是 {1, 2, ..., s}。用 $N_{jlt}^{(k)}$ 表示字符串的个数，其中字符串 ω 满足：ω 是 A 中从状态 i 到状态 j 的路径的标记，其长度为 l，而且这条路径没有遍历大于 k 的中间顶点。

基础：

基础是 $k = 0$ 的情况，有如下表达式：

$$N_{jlt}^{(0)} = \begin{cases} 1 & i = j, l = 0 \\ 0 & i \neq j, l = 0 \\ 0 & l \geq 1 \end{cases}$$

归纳：

假设在从 i 到 j 的路径中经过比 k 大的中间状态。有两种可能的情形需要考虑：

1. 这条路径根本不经过 k。这这种情况下，路径的标记属于 $N_{jl}^{(k-1)}$ 的字符串个数。
2. 路径经过 k 至少一次。在这种情况下，$i \to k \to j$ 需要经过比 k 小的顶点从 i 到状态 k，最后一段不经过 k 而从 k 到 j，所有中间路径都不经过 k 而从 k 到自身。这部分字符串的个数可表示为：

$$\sum_{l_1+l_2=l} (N_{i,k,l_1}^{(k-1)} N_{k,k,l_2}^{(k-1)} \dots N_{k,j,n}^{(k-1)})$$

其中 l_i 满足 $\sum l_i = l (n \geq 2)$。

因此结合前两种情况有表达式：

$$N_{ijl}^{(k)} = N_{ijl}^{(k-1)} + \sum_{l_1+l_2=l} N_{ikl_1}^{(k-1)} N_{kl_2}^{(k-1)} \dots N_{kjl_n}^{(k-1)}$$

结论：

对于 DFA A 所接受的长度为 n 的串的个数可以表示为：

$$\sum_j N_{ijn}^{(s)}$$

其中，s 表示 DFA 的总状态数，1 为起始状态，j 为终态的接收状态。

3.3.1

1. 国际号码的电话格式如下：**国际冠码 + 国际电话区号 + 电话号码**
    - **国际冠码**：不同国家有不同的国际冠码（如中国大陆是“00”），拨打国际电话的时候要根据拨出地区确定国际冠码，但可以统一用“+”表示；
    - **国际电话区号**：每个国家分配的一个代码，如中国大陆为“86”；
2. 以中国大陆为例的国内电话号码格式如下：**长途冠码 + 省市区号 + 电话号码**
    - **长途冠码**：在国内拨打长途需加拨长途冠码“0”；
    - **省市区号**：不同省、直辖市、大型城市分配的代号，比如北京市为“10”；

总的来说，电话号码的形式十分多样。从号码的长度来看，最短的是7位的加拿大等国家，最长的是11位的中国等国家；号码的总长度大概在7位到16位之间。一般来说，电话号码可以是连续的数字，也可能被“-”分割成几个为单位的多个字段。

由于号码的形式十分多样，因此我希望表达式能尽可能识别更多的电话号码，这就不可能避免的识别到一些非电话号码的数字序列。综合上，给出的UNIX风格的表达式如下：

正则表达式：`\+?([0-9]( |-)?){6,15}[0-9]`