作业 2-正则表达式参考答案

1. 教材 3.1.1 写出表示下列语言的正则表达式
   a) 字母表{a, b, c}上包含至少一个 a 和至少一个 b 的串的集合

   $$c^*a(a+c)^*b(a+b+c)^*+c^*b(b+c)^*a(a+b+c)^*$$

   b) 倒数第 10 个符号是 1 的 0 和 1 的串的集合

   $$(0+1)^*1(0+1)^9$$

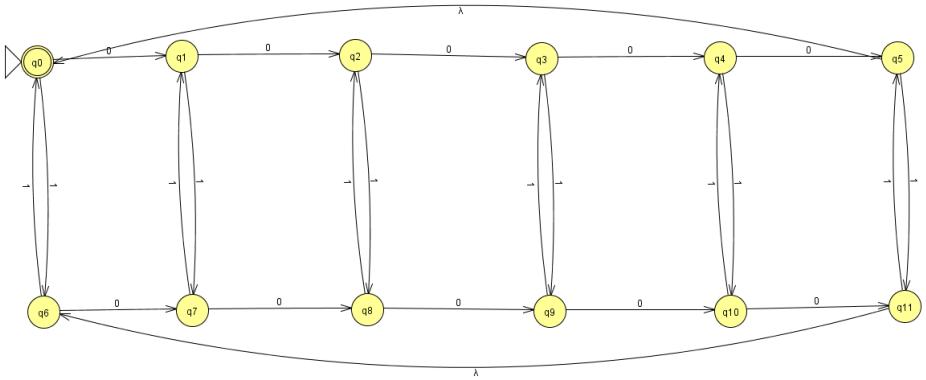
   c) 至多只有一对连续 1 的 0 和 1 的串的集合

   $$(10+0)^*11(01+0)^*+(0+10)^*(1+\varepsilon)$$

2. 教材 3.1.3 写出表示下列语言的正则表达式
   a) 不包含 101 作为子串的所有 0 和 1 的串的集合
   (1 后面只能跟 00 或 1)
   $0^*(1^*000^*)^*1^*0^*$
   b) 具有相同个数的 0 和 1，使得在任何前缀中，0 的个数不比 1 的个数多 2，1 的个数也不比 0 的个数多 2，所有这种 0 和 1 的串的集合

   $$(01+10)^*$$

   c) 0 的个数被 5 整除且 1 的个数是偶数的所有 0 和 1 的串的集合
   画出 DFA



生成正则语言如下：
(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)∗10+(000001+(01+10) (11)∗100001+(001+(01+10) (11)∗(0+101)) (11)∗10001+(0001+(01+10) (11)∗1001+(001+(01+10) (11)∗(0+101)) (11)∗(0+101)) (11)∗1001+(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+0(11)∗100001+0(11)∗(0+101) (11)∗10001+(0(11)∗1001+0(11)∗(0+101) (11)∗(0+101)) (11)∗1001+(0(11)∗10001+0(11)∗(0+101) (11)∗1001+(0(11)∗1001+0(11)∗(0+101) (11)∗

(0+101))(11)*(0+101))(11)*(0+101))*(1+1+0(11)*10000+0(11)*(0+101)(11)*1000+(0(11)*1001+0(11)*(0+101)(11)*(0+101))(11)*100+(0(11)*10001+0(11)*(0+101)(11)*1001+(0(11)*1001+0(11)*(0+101)(11)*(0+101))(11)*(0+101))(11)*10))*

3. 教材 3.1.4 给出下列正则表达式语言的自然语言描述：

a) $(1+\varepsilon)(00^*1)^*0^*$

表示不包含相邻的 1 的串的集合

b) $(0^*1^*)^*000(0+1)^*$

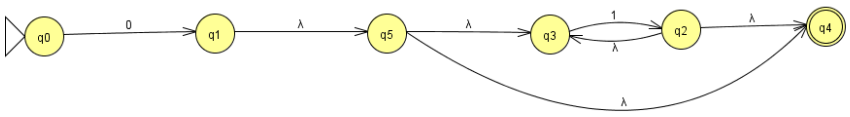
表示包含三个连续的 0 的串的集合

c) $(0+10)^*1^*$

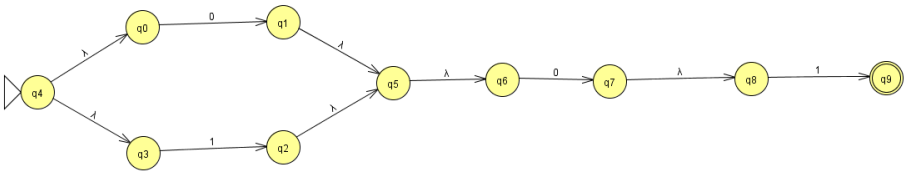
表示除了在串的末尾，不存在连续的 1 的串的集合

4. 教材 3.2.4 把下列正则表达式转化成带ε转移的 NFA

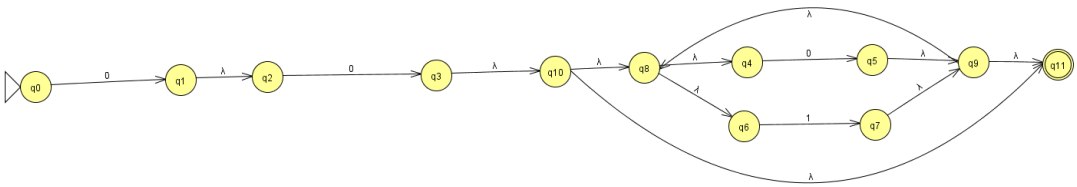
a) 01^*



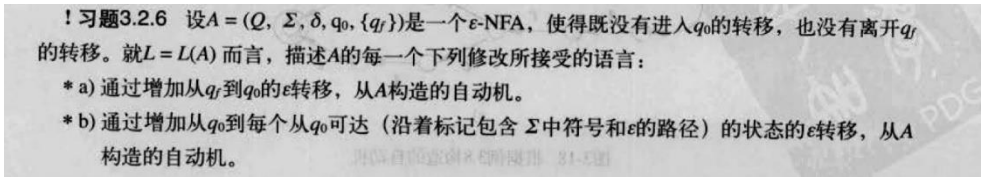
b) $(0+1)01$



c) $00(0+1)^*$



5. 教材 3.2.6



!习题3.2.6 设$A=(Q,\Sigma,\delta,q_0,\{q_f\})$是一个$\varepsilon$-NFA，使得既没有进入$q_0$的转移，也没有离开$q_f$的转移。就$L=L(A)$而言，描述$A$的每一个下列修改所接受的语言：

*a) 通过增加从q_f到q_0的ε转移，从A构造的自动机。

*b) 通过增加从q_0到每个从q_0可达（沿着标记包含Σ中符号和ε的路径）的状态的ε转移，从A构造的自动机。

c) 通过增加从每个能沿着某条路径到达q_f的状态到q_f的ε转移，从A构造的自动机。
d) 通过同时做(b)和(c)的修改从A构造的自动机。

a) 修改后的语言为LL^*

b) 修改后的语言表示L包含的字符串的所有后缀的集合

c) 修改后的语言表示L包含的字符串的所有前缀的集合

d) 修改后的语言表示L包含的字符串的所有子串的集合

6. 教材 3.2.7

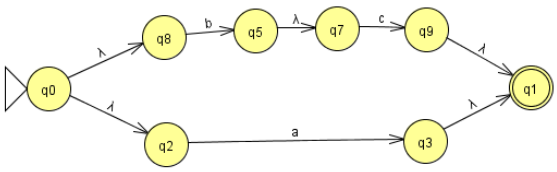
!! 习题3.2.7 把正则表达式转化为ε-NFA的定理3.7的构造，有些地方可以化简。这里有三处：

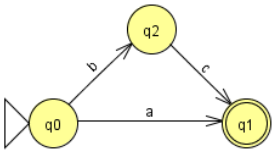
1. 对于并运算符，不是构造新的初始状态和接受状态，而是把两个初始状态合并成一个具备两个初始状态的所有转移的状态。同样，合并两个接受状态，让所有的转移相应地进入合并状态。
2. 对于连接运算符，把第一个自动机与第二个自动机的接受状态合并。
3. 对于闭包运算符，只是增加从接受状态到初始状态的以及反方向的ε转移。

每一个这种化简本身仍然产生正确的构造；也就是说，对于任何正则表达式，得出的ε-NFA接受这个表达式的语言。变化(1)、(2)和(3)的哪些子集可以一起用在构造中，对于每一个正则表达式，仍然产生正确的自动机？

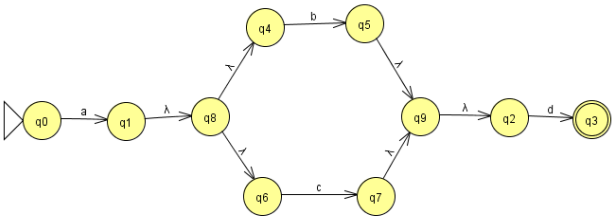
从定理的角度来看，定理 3.7 的前提是没有箭弧进入初始状态，没有箭弧离开接受状态，(3)本身就与前提矛盾，因此只需考虑{(1)(2)}正确，下面说明{(1)(2)}正确

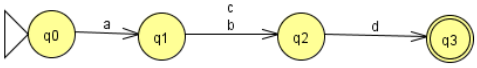
首先在并运算中嵌套连接运算，得到原图(左)和化简(右)，说明并运算中嵌套连接运算正确





其次在连接运算中嵌套并运算，得到原图(左)和化简(右)，说明连接运算中嵌套并运算正确





在此基础上不断叠加嵌套，运算都将保持正确，因此{(1)(2)}正确

若不考虑前提的矛盾，从另一个角度来看，首先{(1)(2)}正确；{(1)(3)}错误，例如 a*+b*，化简后将导致能够识别 ababab 之类的字符串；{(2)(3)}错误，例如 a*b*，化简后将导致能够识别 ababab 之类的字符串；{(1)(2)(3)}错误，理由同前两个例子

综上，正确的子集仅为{(1)(2)}

7. 教材 3.2.8

*!! 习题3.2.8 给出一个算法：输入一个DFA A，对于给定的n（与A的状态个数无关），计算出A所接受的长度为n的串的个数。这个算法应当对于n和A的状态数来说都是多项式的。提示：使用定理3.4的构造所提示的技术。

用$R_{ijl}^{(k)}$表示从状态 i 到状态 j 的路径长度为 l，而且这条路径上没有编号大于 k 的中间顶点，满足

这两个条件的路径的条数。

当$k=0$时，有$R_{ii0}^{(0)}=1$，$R_{ijl}^{(0)}=0(l=0\ and\ i\neq j\ or\ l\neq 0)$

当$k\neq 0$时，有递推式$R_{ijl}^{(k)}=R_{ijl}^{(k-1)}+\left(R_{ikp_1}^{(k-1)}+R_{kkp_2}^{(k-1)}+...+R_{kkp_{q-1}}^{(k-1)}+R_{kjp_q}^{(k-1)}\right)$，其中$\sum_{i=1}^{q}p_i=l$

令 A 的初始状态为 s，接受状态为数组 a，接受状态的个数为 t，总状态个数为 m，则根据递推式计算得到长度为 n 的串的个数为$\sum_{i=1}^{t}R_{sa_in}^{(m)}$

8. 教材 3.3.1

！习题3.3.1 给出一个正则表达式，来描述所能想到的所有不同形式的电话号码。考虑国际号码以及不同国家有不同位数的区号和本地电话号码。

约定符号 $X=0+1+2+3+4+5+6+7+8+9$，$A=3+4+5+6+7+8+9$，$B=2+3+4+5+6+7+8$，$C=1+2$。

一般公民可以接受或使用的电话号码有国际固定电话、国内固定电话和移动电话之分。国际固定电话前加国际冠码 00，国内长途电话前加长途冠码 0；国际电话在冠码后有 1 到 3 位不等的国际电话区号，和 11 位电话号码，而国内电话在冠码后有 2 或 3 位的区号，和 7 或 8 位的电话号码。移动电话号码是 1 开头的 11 位电话号码。

国内电话号也有一些规定：国内移动电话第 1 位为 1，第 2 位为 3-9；国内固定电话第 1 位为 2-8。此外，在区号方面，直辖市和特区区号以及部分省会，令 $Y=010+021+022+023+024+025+027+028+029+852+853$。

考虑到手机或者固定电话等为显示清晰，移动电话在 3，7 位后会增加空格或横杠作为分割，8 位固定电话在第 4 位后，7 为固定电话在第 3 位后增加空格或横杠(-)。令 $Z=-+blankspace$。此处空格用 $blankspace$ 代替。

在网络上，国内电话可能有国际冠码，则国内移动电话正则表达式：

$$R_1 = (\varepsilon + 0086 + 86)(1BX(Z + \varepsilon)XXXX(Z + \varepsilon)XXXX)$$

国内固定电话在本地拨打时，可以省略冠码和区号。正则表达式：

$$R_2 = \big((\varepsilon + 0086 + 86)(Y + 0AXX) + \varepsilon\big)BXXX(Z + \varepsilon)XXXX + \big((\varepsilon + 0086 + 86)0AXX + \varepsilon\big)BXX(Z + \varepsilon)XXXX$$

国际电话正则表达式：

$$R_3 = 00(B + C)(\varepsilon + B + C)(\varepsilon + B + C)XXX(\varepsilon + Z)XXXX(\varepsilon + Z)XXXX$$

将三个正则表达式合并得：

$$\begin{aligned} R = \ & R_1 + R_2 + R_3 \\ = \ & (\varepsilon + 0086 + 86)(1BX(Z + \varepsilon)XXXX(Z + \varepsilon)XXXX) + \\ & \big((\varepsilon + 0086 + 86)(Y + 0AXX) + \varepsilon\big)BXXX(Z + \varepsilon)XXXX + \\ & \big((\varepsilon + 0086 + 86)0AXX + \varepsilon\big)BXX(Z + \varepsilon)XXXX + \\ & 00(B + C)(\varepsilon + B + C)(\varepsilon + B + C)XXX(\varepsilon + Z)XXXX(\varepsilon + Z)XXXX \end{aligned}$$

9. 教材 3.3.2

!! 习题3.3.2 给出一个正则表达式，来表示在招聘广告中可能出现的薪水。考虑可能按小时、周、月或年发放的薪水。这些薪水可能有也可能没有$（如美元）符号或其他单位（如后面跟着的“K”）。可能有一个或多个邻近的单词标志着薪水。提示：查看报纸上的分类广告或在线职位列表，来获得一些关于什么样的模式可能有用的想法。

规定薪水数字 $X = 0 + 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9$，$Y = 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9$.

货币符号（包括符号和英文缩写，由于较多，用 ... 表示省略其他国家地区的货币符号或字母缩写）：$Z = ¥ + € + \$ + £ + ฿ + ₫ + ₭ + \ldots + JPY + CHF + CAD + HKD + FIM + IEP + RMB + SUR + \ldots$

工程技术单位（考虑到可能大写强调）：

$U = mn, bn, k, thousand, million, billion, MN, BN, K, THOUSAND, MILLION, BILLION$.

考虑到可能标注发放薪水时间，令 $W = per +/$ ，$V = m + y + h + d + month + year + hour + day$.

则表示薪水的正则表达式为：

$$RE = (\varepsilon + Z)(0 + YX^*)(\varepsilon + .XX^*)(\varepsilon + U)(\varepsilon + WV)$$