# 编译原理 - 作业(1)：词法分析

**Q1:** (p114, Exercise 3.1.2) Tagged languages like HTML or XML are different from conventional programming languages in that the punctuation (tags) are either very numerous (as in HTML) or a user-defined set (as in XML). Further, tags can often have parameters. Suggest how to divide the following HTML document:

```
Here is a photo of <B>my house</B>:
<P><IMG SRC = "house.gif"><BR>
See <A HREF = "morePix.html">More Pictures</A> if you liked that
one.<P>
```

into appropriate lexemes. Which lexemes should get associated lexical values, and what should those values be?

text：文本内容 'Here is a photo of', 'my house', 'See', 'More Pictures', 'if you liked that one'

start tag：开始标签 <B>, <P>, <IMG>, <BR>, <A>

end tag：结束标签 </B>, </A>

attribute：标签参数 SRC, HREF

value：标签参数值 "house.gif", "morePix.html"

op：操作符 =

**Q2:** (p125, Exercises 3.3.2) Describe the languages denoted by the following regular expressions:

1) a(a|b)*a
2) ((ε|a)b*)*
3) (a|b)*a(a|b)(a|b)
4) a*ba*ba*ba*
5) (aa|bb)*((ab|ba)(aa|bb)*(ab|ba)(aa|bb)*)*

1) 开头结束是 a，中间是任意个 a 和 b 组成的字符串

2) 任意个 a 和 b 组成的字符串

3) 任意个 a 和 b 组成，倒数第 3 个是 a 的字符串

4) 任意个 a 和 3 个 b 组成的字符串

5) 偶数个 a 和偶数个 b 组成的字符串

**Q3:** Write regular expressions for the following languages, or indicate that there exists no such expression:

1) Strings over the alphabet {a, b, c} in which no a's appear after the first b (if one exists).
2) Binary numbers that are multiples of 4.
3) All strings over the alphabet {x, y} that contain no consecutive x's (including the empty string).
4) Strings over the alphabet {x, y} that have exactly as many x's as y's.
5) Identifiers that start with an uppercase letter and then have one or more alphanumeric characters, ending in a number.

1) (a|c)* (b|c)*

2) (0|1)* 0

3) (X?Y)* X?

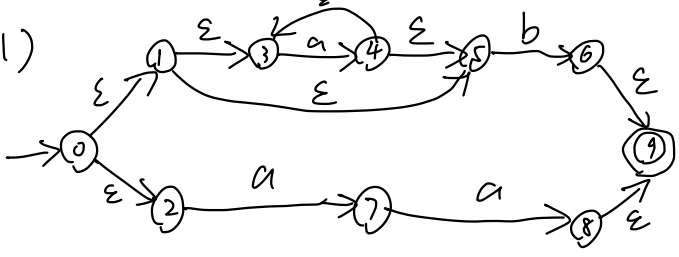
4) not exists

5) [A-Z][A-Za-z0-9]+[0-9]

**Q4:** Consider the following regular expression over the alphabet {a, b}: a*b | aa

1) Use M-Y-T algorithm to make an NFA from the regular expression (show it as a state diagram).
2) Use subset construction to create a DFA equivalent to the NFA you gave for part 1). Show the construction steps and final transition table.
3) Is the DFA in 2) minimized? If yes, explain; otherwise, do the minimization.

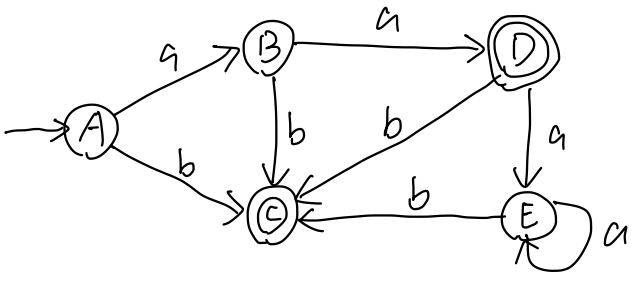
1)




2) ε-closure(0) = 01235

| | a | b |
|---|---|---|
| 01235 | 3457 | 69 |
| 3457 | 34589 | 69 |
| 34589 | 345 | 69 |
| 345 | 345 | 69 |
| 69 | | |

令 01235 为 A，3457 为 B，69 为 C，24589 为 D，345 为 E。DFA 如下：




3) 已经最小化

虽然 D 和 E 都是 $\xrightarrow{a}E$，$\xrightarrow{b}C$

但 D 是终止态不能与非终止态 E 合并