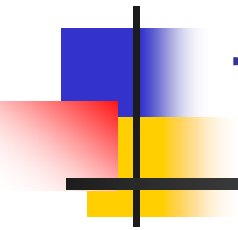


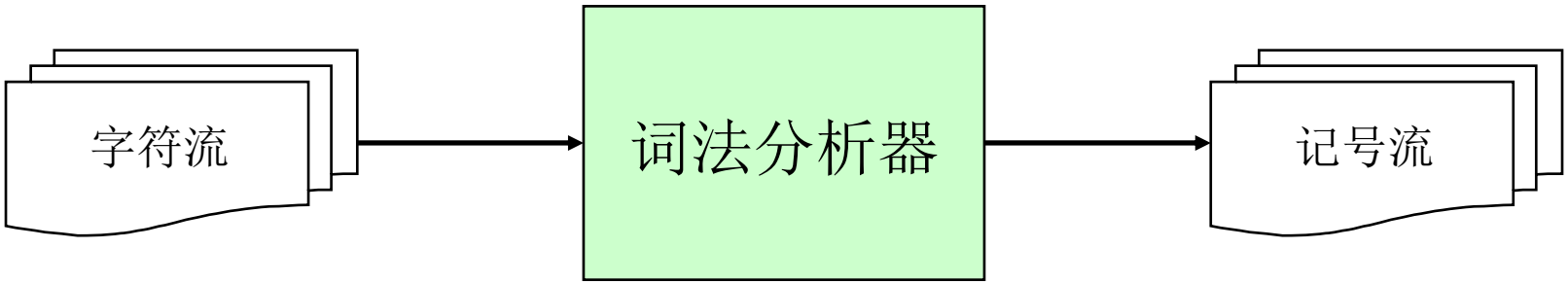
词法分析---正则表达式

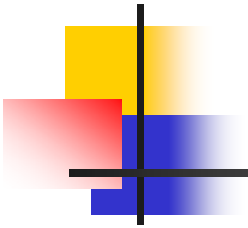
编译原理

华保健

bjhua@ustc.edu.cn

回顾





回顾：词法分析器的实现方法

- 至少两种实现方案:
  - 手工编码实现法
    - 相对复杂、且容易出错
    - 但是目前非常流行的实现方法
      - GCC, LLVM, …
  - 词法分析器的生成器
    - 可快速原型、代码量较少
    - 但较难控制细节
- 我们已经讨论了第一种实现方案
  - 从这一讲开始讨论第二种方案

自动生成

声明式的规范

词法分析器

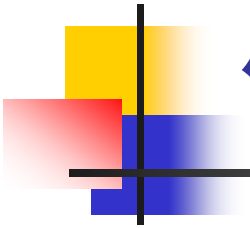
正则表达式

- 对给定的字符集$\Sigma=\{c1, c2, \cdots, cn\}$
- 归纳定义：
  - 空串ε是正则表达式
  - 对于任意$c\in\Sigma$，c是正则表达式
  - 如果M和N是正则表达式，则以下也是正则表达式
    - 选择　　$M \mid N = \{M, N\}$
    - 连接　　$MN = \{mn \mid m\in M, n\in N\}$
    - 闭包　　$M^* = \{\varepsilon, M, MM, MMM, \cdots\}$

正则表达式的形式表示

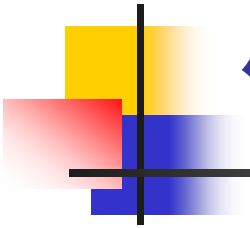
```
e -> ε
   | c
   | e | e
   | e e
   | e*
```

问题：对于给定字符集Σ={a, b}，可以写出哪些正则表达式？



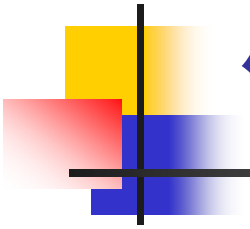
例子：关键字

- C语言中的关键字，例如if，while等
  - 如何用正则表达式表示？



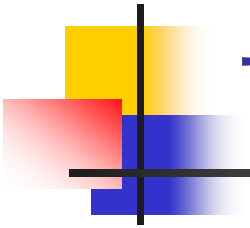
例子：标识符

- C语言中的标识符：以字母或下划线开头，后跟零个或多个字母、数字或下划线。
  - 如何用正则表达式表示？



例子：C语言中的无符号整数

- （十进制整型数）规则：或者是0；或者是以1到9开头，后跟零个或多个0到9
  - 如何用正则表达式表示？



语法糖

- 可以引入更多的语法糖，来简化构造
  - [c1-cn] == c1|c2|…|cn
  - e+ == 一个或多个 e
  - e? == 零个或一个 e
  - “a*” == a* 自身, 不是a的Kleen闭包
  - e{i, j} == i到j个e的连接
  - . == 除‘\n’外的任意字符