



Chapter 5: Maintainability-Oriented Software Construction Approaches

5.3 Maintainability-Oriented Construction Techniques 面向可维护性的构造技术

April 22, 2020

Outline

■ State-based construction

- Automata-based programming
- **Design Pattern: Memento** provides the ability to restore an object to its previous state (undo).
- **Design Pattern: State** allows an object to alter its behavior when its internal state changes.

■ Grammar-based construction

- Grammar and Parser
- Regular Expression (regexp)

学了这么多OO设计模式，不外乎都是 delegation + subtyping，万变不离其宗

除了OO，还有什么其他能够提升软件可维护性的构造技术？——本节从委派+子类型跳出来，学习以下两个方面：

- (1) 基于状态的构造技术
- (2) 基于语法的构造技术

Reading

- MIT 6.031: 17、18
- Java编程思想：第13.6节





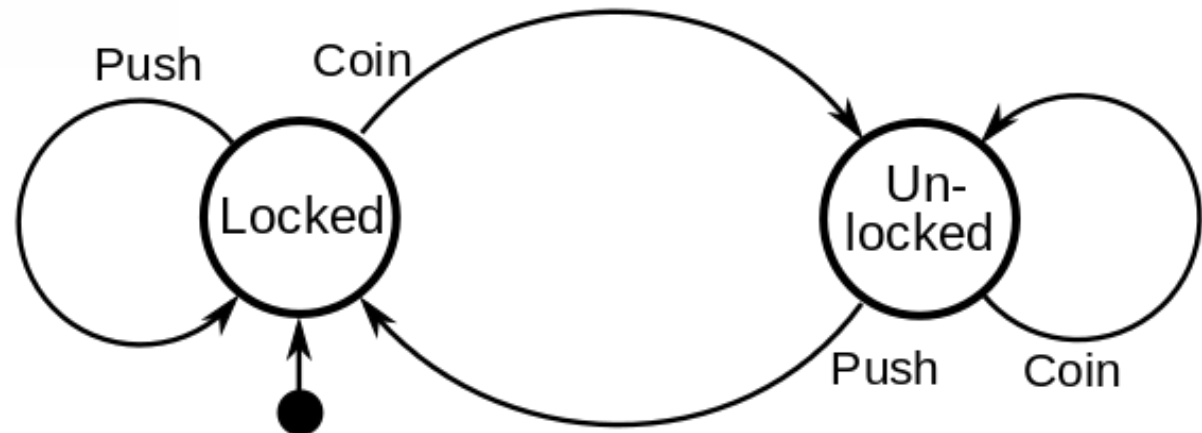
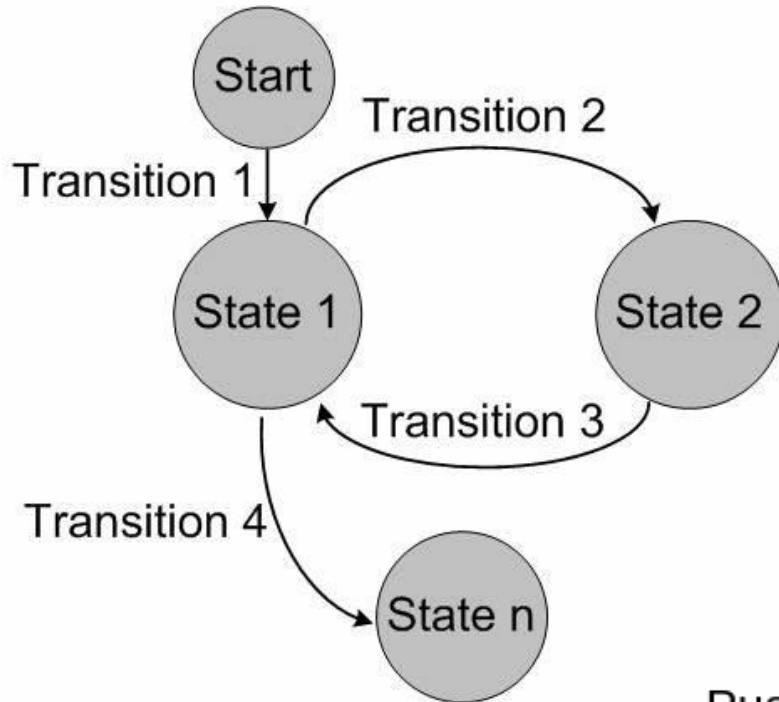
1 State-based construction



State-based programming

- **State-based programming** is a programming technology using finite state machines (**FSM**) to describe program behaviors, i.e., the use of “states” to control the flow of your program. 使用有限状态机来定义程序的行为、使用状态来控制程序的执行
 - For example, in the case of an **elevator**, it could be **stop**, **moving up**, **moving down**, **stopping**, **closing the doors**, and **opening the doors**.
- Each of these are considered a **state**, and what happens next is determined by the elevator’s current state. 根据当前状态，决定下一步要执行什么操作、执行操作之后要转移到什么新的状态
 - If the elevator has just **closed** its doors, what are the possibilities that can happen next? It can **stop**, **move up**, or **move down**.
 - When an elevator **stops**, you expect the next action to be the **doors opening**, **moving up**, or **moving down**.

State transitions



If you write the code...

```
public enum ElevatorState {  
    OPEN, CLOSED, MOVING_UP, MOVING_DOWN, STOP  
}
```

在ADT内部自行管理状态的转换，需要大量的if-else

```
public class Elevator  
{  
    ElevatorState currentState;  
  
    public Elevator(){  
        currentState = ElevatorState.CLOSED;  
    }  
  
    public void changeState(){  
  
        if(currentState == ElevatorState.OPEN){  
            currentState = ElevatorState.CLOSED;  
            closeDoors();  
        }  
  
        if(currentState == ElevatorState.CLOSED  
            && upButtonIsPressed()){  
            currentState = ElevatorState.MOVING_UP;  
            moveElevatorUp();  
        }  
  
        if(currentState == ElevatorState.CLOSED  
            && downButtonIsPressed()){  
            currentState = ElevatorState.MOVING_DOWN;  
            moveElevatorDown();  
        }  
  
        if((currentState == ElevatorState.MOVING_UP  
            || currentState == ElevatorState.MOVING_DOWN)  
            && reachedDestination()){  
            currentState = ElevatorState.STOP;  
            stopElevator();  
        }  
  
        if(currentState == ElevatorState.STOP){  
            currentState = ElevatorState.OPEN;  
            openDoors();  
        }  
    }  
}
```

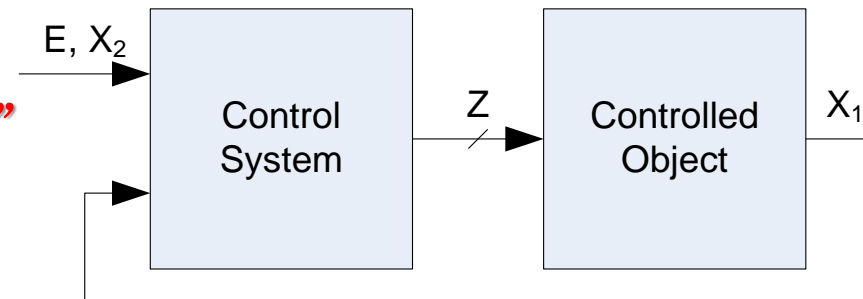


(1) Automata-based programming

基于自动机的编程

Automata-based programming

- **Automata-based programming** is a programming paradigm in which the program or part of it is thought of as a model of a finite state machine (FSM) or any other formal automaton.
 - Treat a program as a finite automata.
 - Each automaton can take one "step" at a time, and the execution of the program is broken down into individual steps.
 - The steps communicate with each other by changing the value of a variable representing "the state".
 - Control flow of the program is determined by the value of that variable.
- Application design approach should be similar to the design of control systems (Automata System).
- 核心思想：将程序看作是一个有限状态自动机，侧重于对“状态”及“状态转换”的抽象和编程



Automata-based programming

- The time period of the program's execution is clearly separated down to the *steps of the automaton*. 程序的执行被分解为一组自动执行的步骤
 - Each of the *steps* is effectively an execution of a code section (same for all the steps), which has a single entry point. Such a section can be a function or other routine, or just a cycle body.
- Any communication between the steps is only possible via the explicitly noted set of variables named *the state*. 各步骤之间的通讯通过“状态变量”进行
 - Between any two steps, the program can not have implicit components of its state, such as local (stack) variables' values, return addresses, the current instruction pointer, etc.
 - The state of the whole program, taken at any two moments of entering the step of the automaton, can only differ in the values of the variables being considered as the state of the automaton.

How to implement?

- The whole execution of the automata-based code is a (possibly explicit) cycle of the automaton's steps. 程序执行就可看作是各自动步骤的不断循环
- The "state" variable can be a simple `enum` data type, but more complex data structures may be used. 使用枚举类型`enum`定义状态
- A common technique is to create a state transition table, a two-dimensional array comprising rows representing every possible state, and columns representing input parameter. 使用二维数组定义状态转换表
 - The value of the table where the row and column meet is the next state the machine should transition to if both conditions are met.

```
State transition[][] = {  
    { State.Initial, State.Final, State.Error },  
    { State.Final, State.Initial, State.Error }  
};
```

See Wikipedia: https://en.wikipedia.org/wiki/State_transition_table



(2) State Pattern

状态模式 (behavioral pattern)

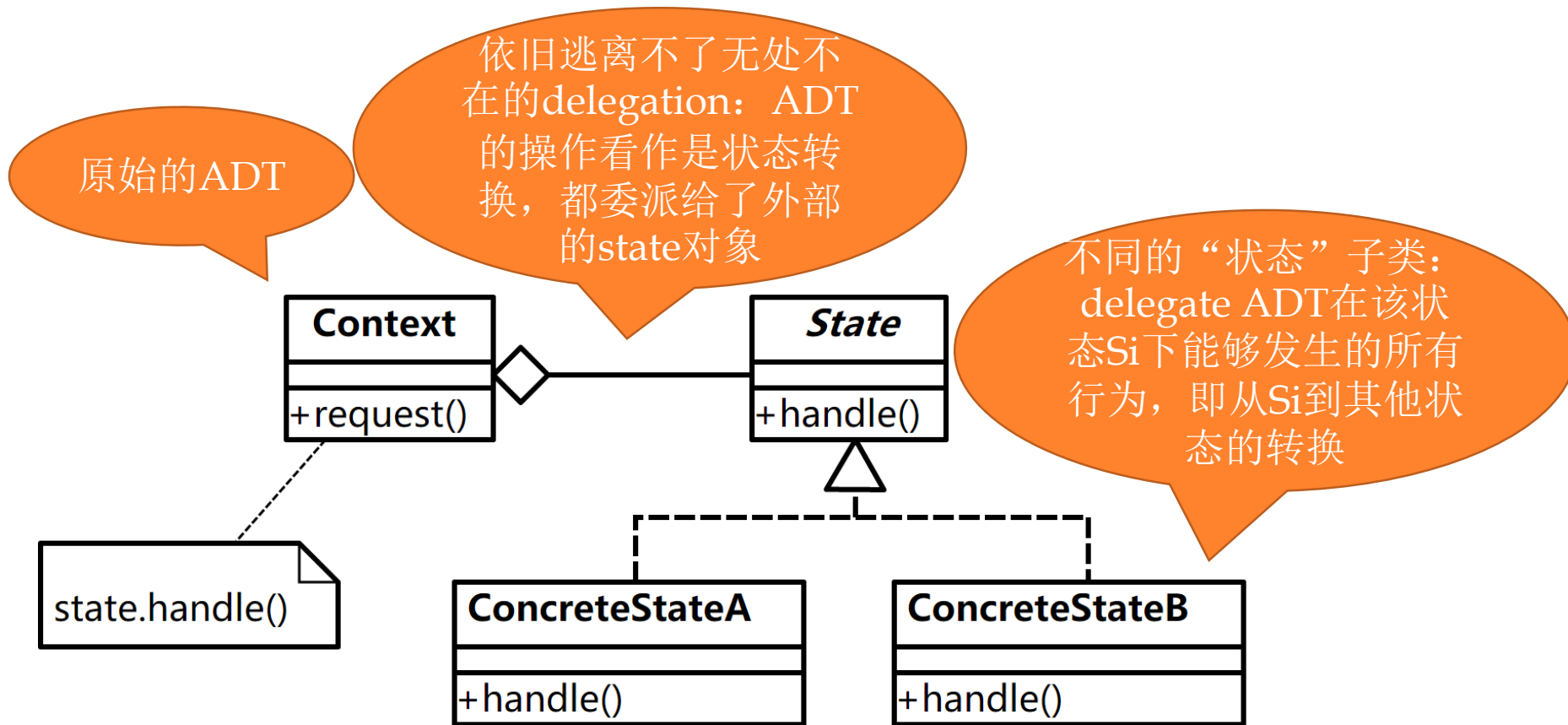
State pattern

- Suppose an object is always in one of several known states
- The state an object is in determines the behavior of several methods
- Could use **if/case** statements in each method
- Better solution: state pattern
- Have a reference to a state object
 - Normally, state object doesn't contain any fields
 - Change state: change state object
 - Methods delegate to state object

最好不要使用if/else结构在ADT内部实现状态转换（考虑将来的扩展和修改）

使用delegation，将状态转换的行为委派到独立的state对象去完成

Structure of State pattern



Example – Finite State Machine

```
class Context {  
    State state; //保存对象的状态  
    //设置初始状态  
    public Context(State s) {state = s;}  
    //接收外部输入, 开启状态转换  
    public void move(char c) { state = state.move(c); }
```

设置
delegation
关系

将改变状态的
“动作”
delegate到
state对象

每次状态转换之后, 形
成新状态, 替换原状态

//判断是否达到合法的最终状态

```
    public boolean accept() { return state.accept(); }  
    public State getState() { return this.state; }  
}
```

//状态接口

```
public interface State {  
    State move(char c);  
    boolean accept();  
}
```

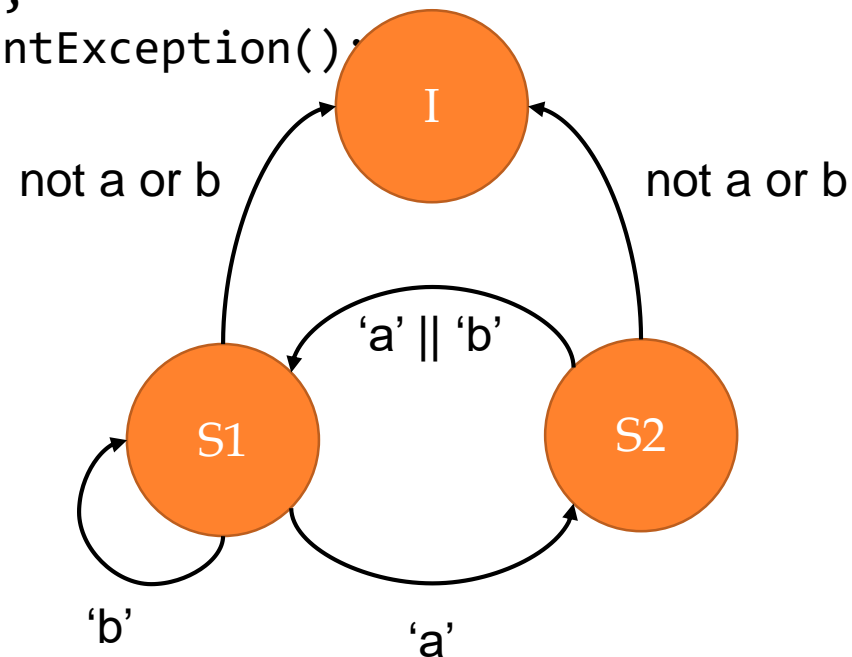
Delegate到当前状态的
accept()方法, 判断
是否达到最终状态

FSM Example – cont.

```

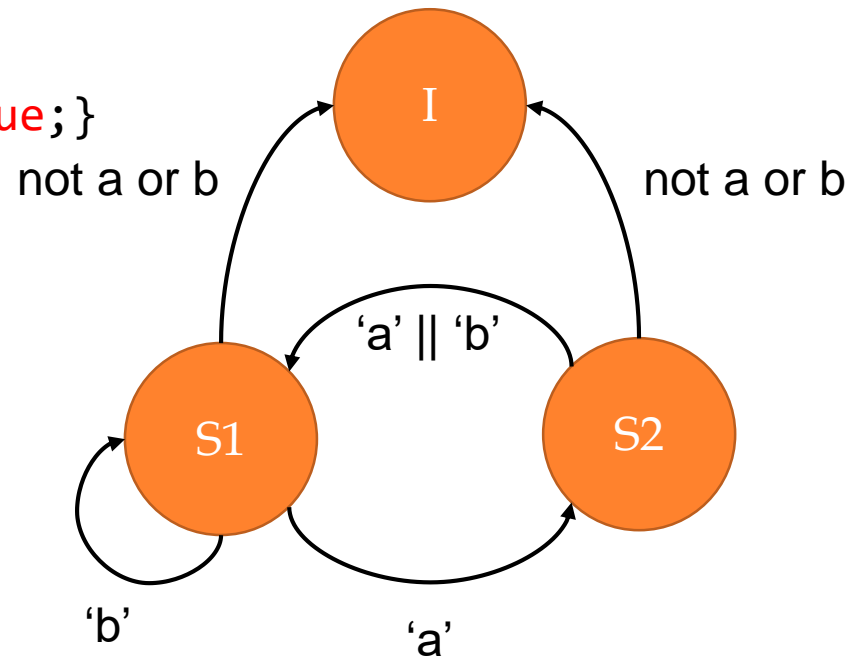
class State1 implements State {
    static State1 instance = new State1(); //singleton模式 (see 8-3)
    private State1() {}
    public State move (char c) {
        switch (c) {
            case 'a': return State2.instance; //返回新状态的singleton实例
            case 'b': return State1.instance;
            default: throw new IllegalArgumentException();
        }
    }
    public boolean accept() {
        return false;
    } //该状态非可接受状态
}

```



FSM Example – cont.

```
class State2 implements State {  
    static State2 instance = new State2();  
    private State2() {}  
    public State move (char c) {  
        switch (c) {  
            case 'a': return State1.instance;  
            case 'b': return State1.instance;  
            default: throw new IllegalArgumentException();  
        }  
    }  
    public boolean accept() {return true;}  
}
```

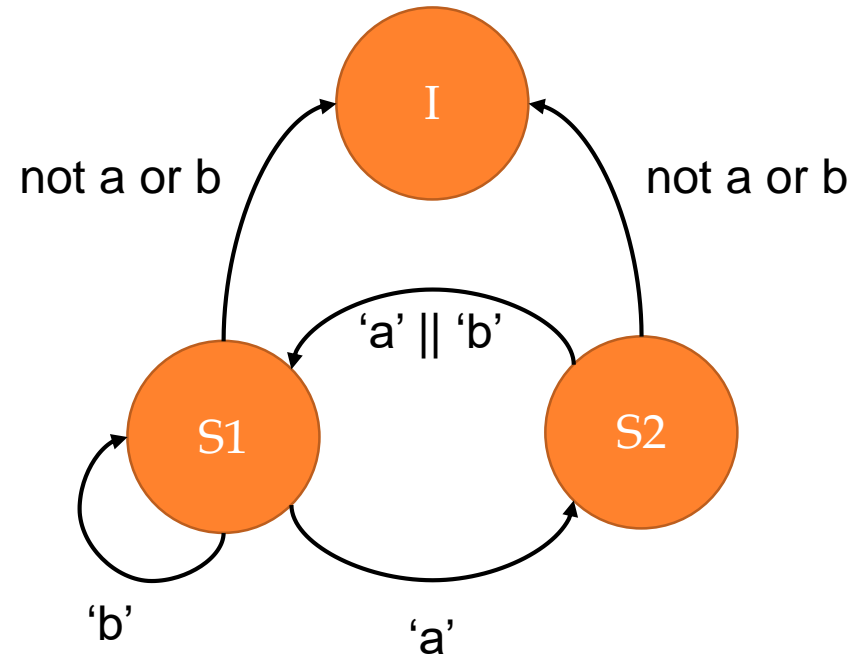


Example

```
public static void main(String[] args) {  
    Context context = new Context(State1.instance);  
    for (int i = 0; i < args.length; i++) {  
        context.move(args[i]);  
        if(context.accept())  
            break;  
    }  
}
```

给ADT初始状态

根据输入的一组字符，
每次用一个字符使状态
发生变迁，直到达到最
终状态，程序结束。





(3) Memento Pattern

备忘录模式 (behavioral)

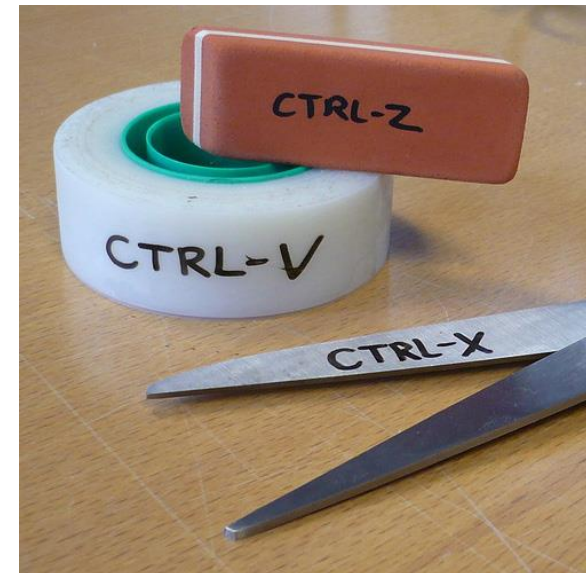
Memento Pattern

■ Intent

- Without violating encapsulation, capture and externalize an object's internal state so that the object can be returned to this state later.
- A magic cookie that encapsulates a "check point" capability.
- Promote **undo** or **rollback** to full object status.

- **Problem:** Need to restore an object back to its previous state (e.g. "undo" or "rollback" operations).

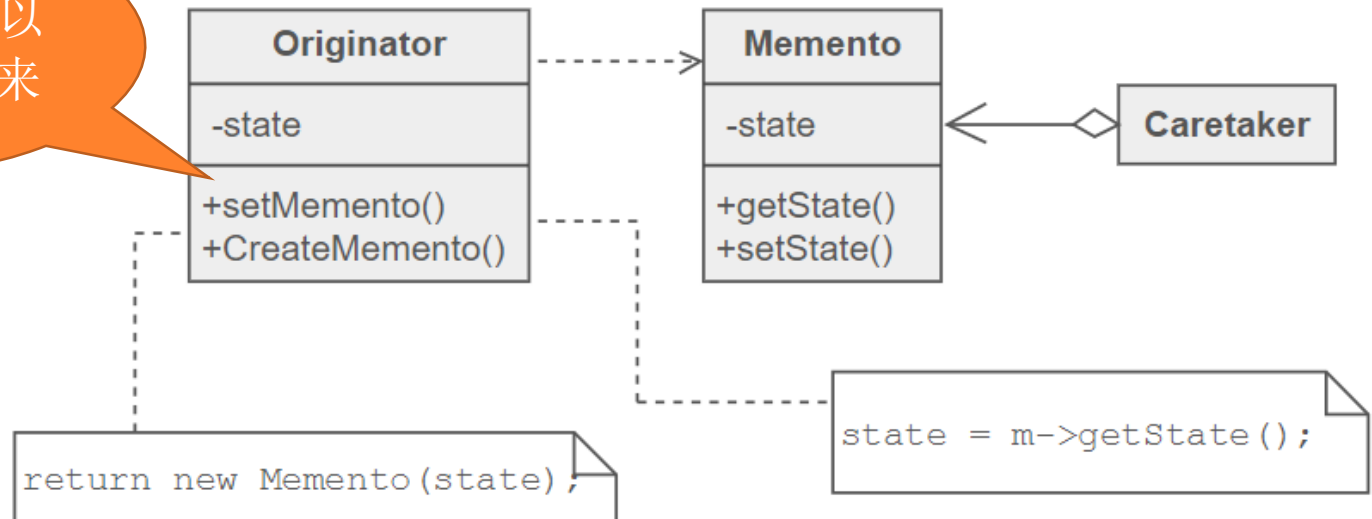
- 记住对象的历史状态，以便于“回滚”



Memento Pattern

- **Memento design pattern defines three distinct roles:**
 - **Originator** - the object that knows how to save itself. 需要“备忘”的类
 - **Caretaker** - the object that knows why and when the Originator needs to save and restore itself. 添加originator的备忘录和恢复
 - **Memento** - the lock box that is written and read by the Originator, and shepherded by the Caretaker. 备忘录，记录originator对象的历史状态

需要“备份”的
ADT，完全可以
自己增加方法来
支持备份

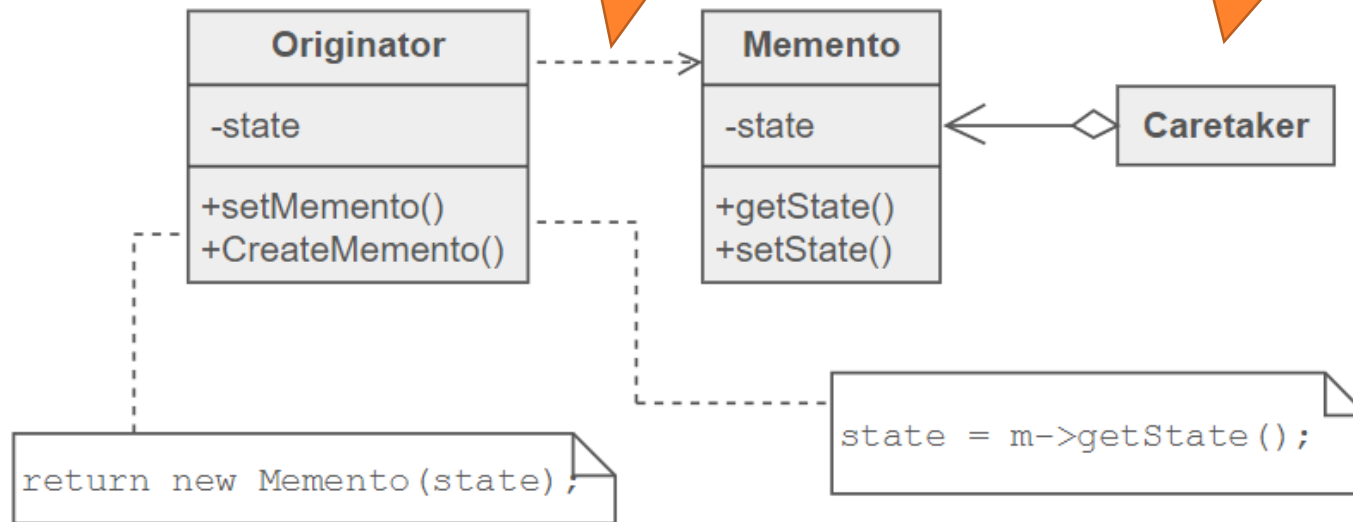


Memento Pattern

需要“备份”的
ADT
rep中只记录当前
状态

每次“备份”都
生成一个外部的
Memento对象

负责掌控全部的状态
备份，客户端通过
它来操纵ADT的
状态备份与恢复



Memento Pattern

Memento: 非常简单的类，只记录一个历史状态

```
class Memento {  
    private State state;  
  
    public Memento(State state) {  
        this.state = state;  
    }  
  
    public State getState() {  
        return state;  
    }  
}
```

ADT原本的状态转换功能，可能更复杂（例如State模式）

```
class Originator {  
    private State state;  
  
    public void setState(State state) {  
        System.out.println("Originator: Setting state to " + state.toString());  
        this.state = state;  
    }  
  
    public Memento save() {  
        System.out.println("Originator: Saving to Memento.");  
        return new Memento(state);  
    }  
  
    public void restore(Memento m) {  
        state = m.getState();  
        System.out.println("Originator: State after restoring from Memento: " + state);  
    }  
}
```

保存历史状态，delegate到memento去实现

利用传入的Memento对象来恢复历史状态

Memento Pattern

```
class Caretaker {
    private List<Memento> mementos
        = new ArrayList<>();

    public void addMemento(Memento m) {
        mementos.add(m);
    }

    public Memento getMemento() {
        return mementos.get(?);
    }
}
```

保留一系列
历史状态

添加一个新的
历史状态

取出需要回
滚的状态

```
Originator: Setting state to State1
Originator: Setting state to State2
Originator: Saving to Memento.
Originator: Setting state to State3
Originator: Saving to Memento.
Originator: Setting state to State4
Originator: State after restoring from
Memento: State3
```

如果要恢复最近备
份的状态，这里应
该是？

```
public class Demonstration {
    public static void main(String[] args) {
        Caretaker caretaker = new Caretaker();
        Originator originator = new Originator();
        originator.setState("State1");
        originator.setState("State2");
        caretaker.addMemento( originator.save() );
        originator.setState("State3");
        caretaker.addMemento( originator.save() );
        originator.setState("State4");
        originator.restore( caretaker.getMemento() );
    }
}
```

如何rollback两
步、三步、...

Memento Pattern

```
class Caretaker {  
    private List<Memento> mementos = new ArrayList<>();  
    public void addMemento(Memento m) { mementos.add(m); }  
    public Memento getMemento(int i) {  
        if(mementos.size()-i < 0)  
            throw new RuntimeException("Cannot rollback so many back!");  
        return mementos.get(mementos.size()-i);  
    }  
}
```

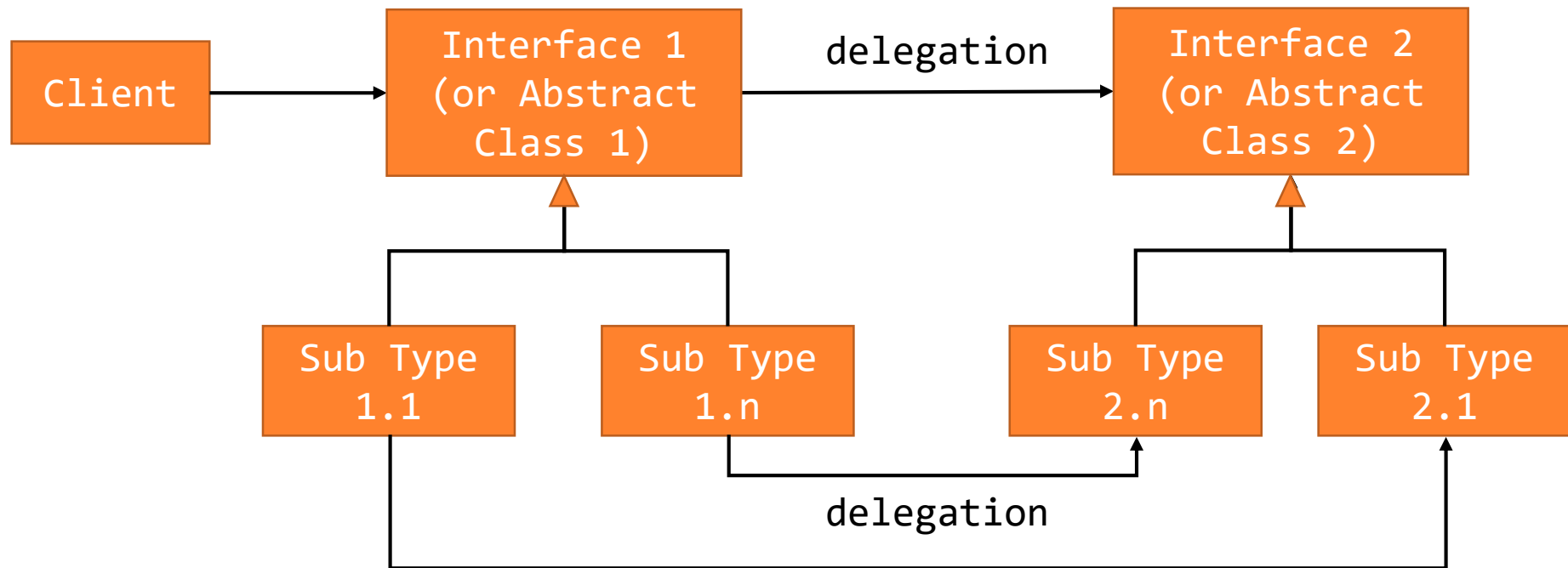
```
public class Demonstration {  
    public static void main(String[] args) {  
        Caretaker caretaker = new Caretaker();  
        Originator originator = new Originator();  
        originator.setState("State2");  
        caretaker.addMemento( originator.save() );  
        originator.setState("State3");  
        caretaker.addMemento( originator.save() );  
        originator.setState("State4");  
        originator.restore( caretaker.getMemento(2) );  
    }  
}
```

但这里有个潜在bug: 每次restore之后, 是否应删除某些备忘录? 如何继续修改该代码?

如果你需要支持repeat功能, 就不要删除memento对象

设计模式的对比：共性样式2

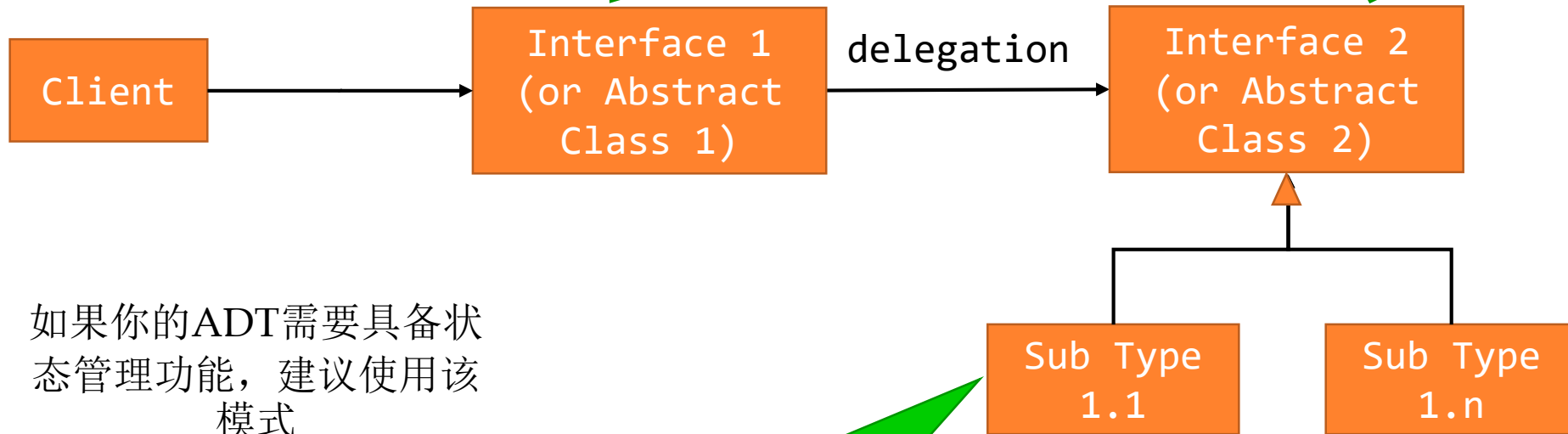
两棵“继承树”，两个层次的“delegation”



State

需要具备状态转换功能的ADT，维持一个State对象表征其当前状态，并提供一系列状态转换方法，各方法内部调用State对象的相应操作来完成。

State接口，定义一系列状态转换操作



如果你的ADT需要具备状态管理功能，建议使用该模式

每个子类型用不同方式实现状态转换操作，返回新状态的实例

Memento

5

iques

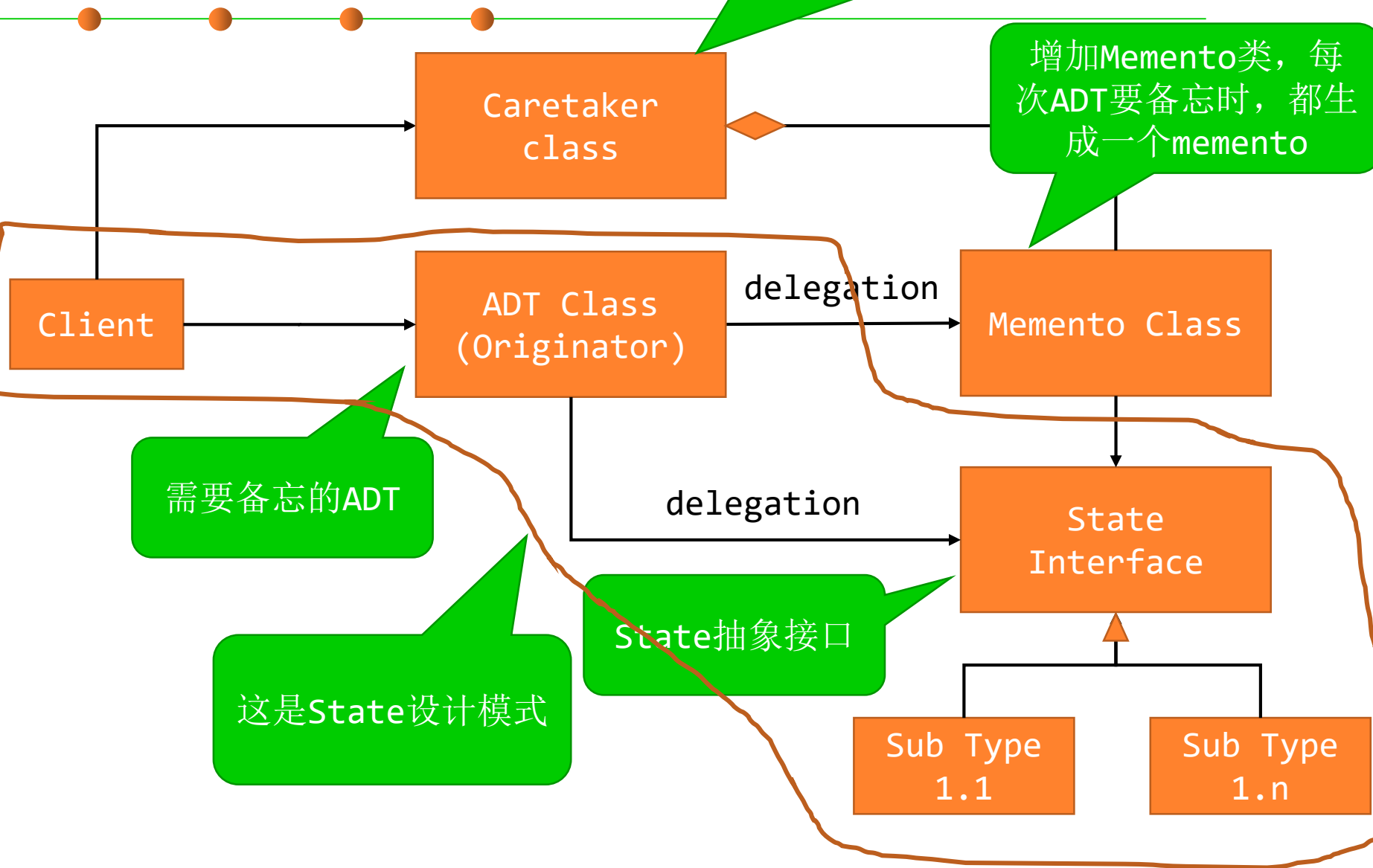
但是ADT自己不维护各个备忘录，而是交给Caretaker来负责，它管理着所有的备忘录（加入新的、从中取出旧的）

增加Memento类，每次ADT要备忘时，都生成一个memento

需要备忘的ADT

这是State设计模式

State抽象接口





2 Grammar-based construction



语法驱动的构造

String/Stream based I/O

- Some program modules take input or produce output in the form of a sequence of bytes or a sequence of characters, which is called a *string* when it's simply stored in memory, or a *stream* when it flows into or out of a module. 有一类应用，从外部读取文本数据，在应用中做进一步处理。
- Concretely, a sequence of bytes or characters might be:
 - A file on disk, in which case the specification is called the *file format* 输入文件有特定格式，程序需读取文件并从中抽取正确内容
 - Messages sent over a network, in which case the specification is a *wire protocol* 从网络上传输过来的消息，遵循特定的协议
 - A command typed by the user on the console, in which case the specification is a *command line interface* 用户在命令行输入的指令，遵循特定的格式
 - A string stored in memory 内存中存储的字符串，也有格式需要

String/Stream based I/O

```

22  UsageLog ::= <2019-01-02,15:30:00,Wechat,1>
23  UsageLog ::= <2019-01-03,11:00:00,Weibo,10>
24  UsageLog ::= <2019-01-03,09:00:00,BaiduMap,400>
25
26
27  App ::= <Wechat,Tencent,13.2,"The most popular social networking App in China","Social network">
28  App ::= <QQ,Tencent,29.2,"The second popular social networking App in China","Social network">
29  App ::= <Weibo,Sina,v0.2.3.4,"The third popular social networking App in China","Social network">
30  App ::= <Didi,Didi,ver03.32,"The most popular car sharing App in China","Travel">
31  App ::= <Eleme,Eleme,20V0.03,"The most popular online food ordering App in China","Food">
32  App ::= <BaiduMap,Baidu,2.9000000.20v03,"The most popular map App in China","Travel">
33
34  Period ::= Day
35
36
37  Relation ::= <Wechat,QQ>
38  Relation ::= <Wechat,Eleme>
39  Relation ::= <Didi,BaiduMap>

```

```

C:\>curl --HEAD http://cms.hit.edu.cn
HTTP/1.1 200 OK
Date: Fri, 20 Apr 2019 14:00:00 GMT
Server: Apache/2.4.33 (Unix) mod_jk/1.2.43
Accept-Ranges: bytes
Vary: Accept-Encoding
Connection: close
Content-Type: text/html

```

```

1  CentralUser ::= <TommyWong,30,M>
2
3  SocialTie ::= <TommyWong, LisaWong, 0.98>
4  SocialTie ::= <TommyWong, TomWong, 0.2>
5  SocialTie ::= <TomWong, FrankLee, 0.71>
6  SocialTie ::= <FrankLee, DavidChen, 0.02>
7  SocialTie ::= <TommyWong, DavidChen, 0.342>
8  SocialTie ::= <JackMa, PonyMa, 0.999>
9
10 Friend ::= <LisaWong, 25, F>
11 Friend ::= <TomWong, 61, M>
12 Friend ::= <FrankLee, 42, M>
13 Friend ::= <DavidChen, 55, M>
14 Friend ::= <JackMa, 58, M>
15 Friend ::= <PonyMa, 47, M>

```

The notion of a grammar

- For these kinds of sequences, the notion of a grammar is a good choice for design:
 - It can not only help to distinguish between legal and illegal sequences, but also to parse a sequence into a data structure a program can work with. 使用grammar判断字符串是否合法，并解析成程序里使用的数据结构
 - The data structure produced from a grammar will often be a recursive data type. 通常是递归的数据结构
- Regular expression 正则表达式
 - It is a widely-used tool for many string-processing tasks that need to disassemble a string, extract information from it, or transform it.
- A parser generator is a kind of tool that translate a grammar automatically into a parser for that grammar. 根据语法，开发一个它的解析器，用于后续的解析



(1) Constituents of a Grammar



Terminals: Literal Strings in a Grammar

- To describe a string of symbols, whether they are bytes, characters, or some other kind of symbol drawn from a fixed set, we use a **compact representation called a grammar**.
- A **grammar** defines a set of **strings**. 用语法定义一个“字符串”
 - For example, the grammar for URLs will specify the set of strings that are legal URLs in the HTTP protocol.
- The **literal strings** in a grammar are called **terminals** 终止节点、叶节点
 - They're called terminals because they are the leaves of a parse tree that represents the structure of the string. 语法解析树的叶子节点
 - They don't have any children, and can't be expanded any further. 无法再往下扩展
 - We generally write terminals in quotes, like 'http' or ':'. 通常表示为字符串

Nonterminals and Productions in a Grammar

- A grammar is described by a set of **productions** 产生式节点, where each production defines a **nonterminal** 非终止节点
 - A nonterminal is like a variable that stands for a set of strings, and the production as the definition of that variable in terms of other variables (nonterminals), operators, and constants (terminals). 遵循特定规则, 利用操作符、终止节点和其他非终止节点, 构造新的字符串
 - Nonterminals are internal nodes of the tree representing a string.
- A **production** in a grammar has the form
 - **nonterminal ::= expression of terminals, nonterminals, and operators**
- One of the nonterminals of the grammar is designated as the **root**.
 - The set of strings that the grammar recognizes are the ones that match the root nonterminal.
 - This nonterminal is often called **root or start**. 根节点



(2) Operators in a Grammar



Three Basic Grammar Operators

- **The three most important operators in a production expression are:**

- **Concatenation** 连接, represented not by a symbol, but just a space:

$x ::= y\ z$ x matches y followed by z

- **Repetition** 重复, represented by $*$:

$x ::= y^*$ x matches zero or more y

- **Union**, also called alternation 选择, represented by $|$:

$x ::= y\ |\ z$ x matches either y or z

Grouping operators using parentheses

- By convention, the postfix operators *****, **?**, and **+** have highest precedence, which means they are applied first.
- **Concatenation** is applied next.
- **Alternation** **|** has lowest precedence, which means it is applied last.
- **Parentheses can be used to override precedence:**
 - $x ::= (y z \mid a b)^*$ x matches zero or more yz or ab pairs
 - $m ::= a (b \mid c) d$ m matches a , followed by either b or c , followed by d

A small example

- `url ::= 'http://mit.edu/'`
 - Use these operators to generalize our url grammar to match some other hostnames, such as `http://stanford.edu/` and `http://google.com/`.
- `url ::= 'http://' hostname '/'`
 - The `url` nonterminal matches strings that start with the literal string `http://`, followed by a match to the `hostname` nonterminal, followed by the literal string `/`.
- `hostname ::= 'mit.edu' | 'stanford.edu' | 'google.com'`
 - A `hostname` can match one of the three literal strings, `mit.edu` or `stanford.edu` or `google.com`.
- So this grammar represents the set of three strings.

A small example

- `hostname ::= 'mit.edu' | 'stanford.edu' | 'google.com'`
- To make it represent more URLs, we allow any lowercase word in place of `mit`, `stanford`, `google`, `com` and `edu`:

```
url ::= 'http://' hostname '/'  
hostname ::= word '.' word  
word ::= ('a' | 'b' | 'c' | 'd' | 'e' | 'f' | 'g' | 'h' | 'i'  
          | 'j' | 'k' | 'l' | 'm' | 'n' | 'o' | 'p' | 'q'  
          | 'r' | 's' | 't' | 'u' | 'v' | 'w' | 'x' | 'y' | 'z')*
```

- The new word rule matches a string of zero or more lowercase letters, so the overall grammar can now match `http://alibaba.com/` and `http://zyxw.edu/` as well.
- Unfortunately word can also match an empty string, so this `url` grammar also matches `http://./`, which is not a legal URL.

A small example

```
word ::= ('a' | 'b' | 'c' | 'd' | 'e' | 'f' | 'g' | 'h' | 'i'  
          | 'j' | 'k' | 'l' | 'm' | 'n' | 'o' | 'p' | 'q'  
          | 'r' | 's' | 't' | 'u' | 'v' | 'w' | 'x' | 'y' | 'z')  
       ('a' | 'b' | 'c' | 'd' | 'e' | 'f' | 'g' | 'h' | 'i'  
          | 'j' | 'k' | 'l' | 'm' | 'n' | 'o' | 'p' | 'q'  
          | 'r' | 's' | 't' | 'u' | 'v' | 'w' | 'x' | 'y' | 'z')*
```

Too complicated!

More grammar operators

- **Additional operators are just syntactic sugar (i.e., they're equivalent to combinations of the big three operators):**

- **Optional** (0 or 1 occurrence), represented by **?**:

$x ::= y?$ an x is a y or is the empty string

- **1 or more occurrences**: represented by **+**:

$x ::= y+$ an x is one or more y (equivalent to $x ::= y y^*$)

- **A character class [...]**, representing the length-1 strings containing any of the characters listed in the square brackets:

$x ::= [a-c]$ is equivalent to $x ::= 'a' \mid 'b' \mid 'c'$

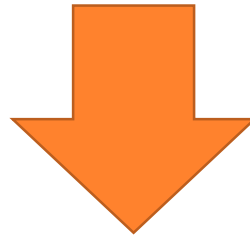
$x ::= [aeiou]$ is equivalent to $x ::= 'a' \mid 'e' \mid 'i' \mid 'o' \mid 'u'$

- **An inverted character class [^...]**, representing the length-1 strings containing any character not listed in the brackets:

$x ::= [^a-c]$ is equivalent to $x ::= 'd' \mid 'e' \mid 'f' \mid \dots$
(all other characters)

Go back to the example

```
url ::= 'http://' hostname '/'
hostname ::= word '.' word
word ::= ('a' | 'b' | 'c' | 'd' | 'e' | 'f' | 'g' | 'h' | 'i'
          | 'j' | 'k' | 'l' | 'm' | 'n' | 'o' | 'p' | 'q'
          | 'r' | 's' | 't' | 'u' | 'v' | 'w' | 'x' | 'y' | 'z')
        ('a' | 'b' | 'c' | 'd' | 'e' | 'f' | 'g' | 'h' | 'i'
          | 'j' | 'k' | 'l' | 'm' | 'n' | 'o' | 'p' | 'q'
          | 'r' | 's' | 't' | 'u' | 'v' | 'w' | 'x' | 'y' | 'z')*
```



```
url ::= 'http://' hostname '/'
hostname ::= word '.' word
word ::= [a-z]+
```



(3) Recursion in grammars



Recursion in grammars

- Hostnames can have more than two components, and there can be an optional port number:

`http://didit.csail.mit.edu:4949/`

- To handle this kind of string, the grammar is now:

```
url ::= 'http://' hostname (':' port)? '/'  
hostname ::= word '.' hostname | word '.' word  
port ::= [0-9]+  
word ::= [a-z]+
```

- hostname is now defined **recursively** in terms of itself.
- Using the repetition operator, we could also write hostname without recursion, like this:

`hostname ::= (word '.')+ word`

Exercise

- **Consider this grammar:**

$$S ::= (B \ C)^* \ T$$
$$B ::= M+ \mid P \ B \ P$$
$$C ::= B \mid E+$$

- **What are the nonterminals in this grammar?**
- **What are the terminals in this grammar?**
- **Which productions are recursive?**

Exercise

- Which strings match the root nonterminal of this grammar?

`root ::= 'a'+ 'b'* 'c'?`

- **Strings**

`aabcc`

`bbbc`

`aaaaaaaaa`

`abc`

`abab`

`aac`

Exercise

- Which strings match the root nonterminal of this grammar?

```
root      ::= integer ('-' integer)+  
integer ::= [0-9]+
```

- Strings

617

617-253

617-253-1000

integer-integer-integer

5--5

3-6-293-1

Exercise

- Which strings match the root nonterminal of this grammar?

root ::= (A B)+

A ::= [Aa]

B ::= [Bb]

- Strings

aaaBBB

abababab

aBAbabAB

AbAbAbA



(4) Parse Trees

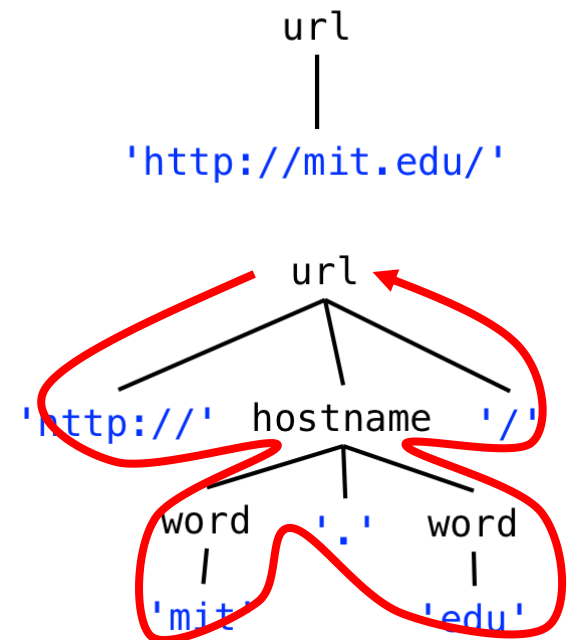


Parse Tree

- Matching a grammar against a string can generate a *parse tree* that shows how parts of the string correspond to parts of the grammar.
 - The leaves of the parse tree are labeled with terminals, representing the parts of the string that have been parsed.
 - They don't have any children, and can't be expanded any further.
 - If we concatenate the leaves together, we get back the original string.

`url ::= 'http://mit.edu/'`

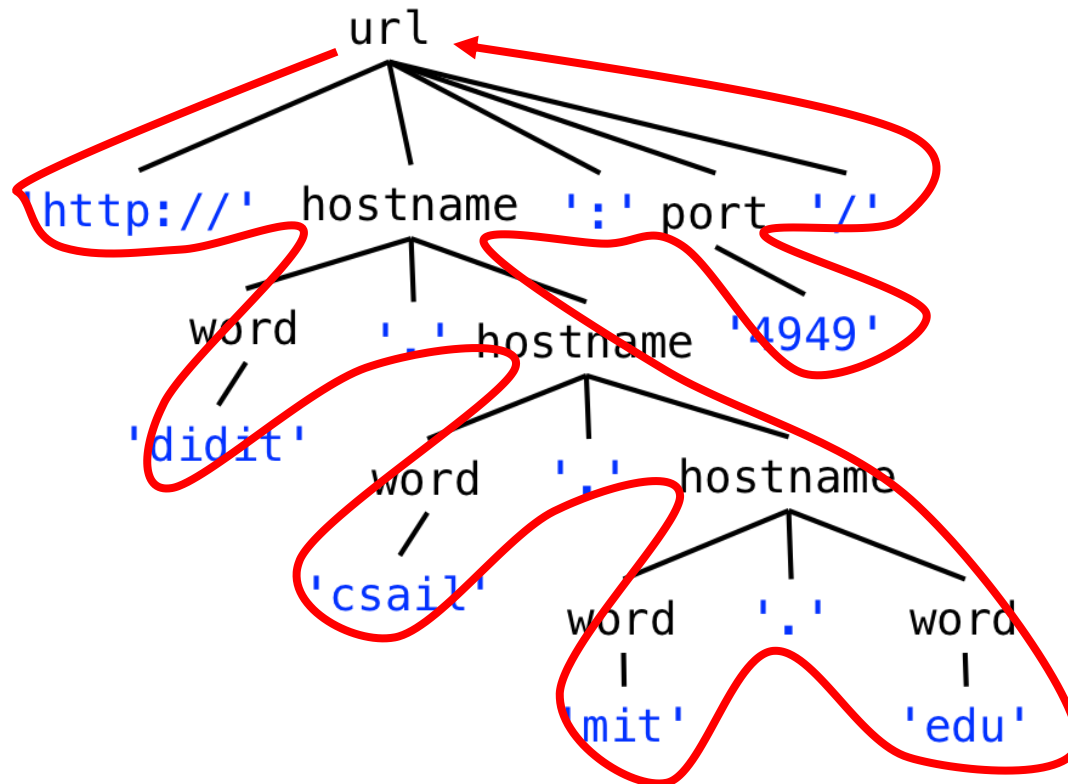
`url ::= 'http://' hostname '/'`
`hostname ::= word '.' word`
`word ::= [a-z]+`



Parse Tree

`url ::= 'http://' hostname (':' port)? '/'`
`hostname ::= word '.' hostname | word '.' word`
`port ::= [0-9]+`
`word ::= [a-z]+`

`http://didit.csail.mit.edu:4949/`



Parse Tree

- If the same string was matched against this grammar with a non-recursive `hostname` rule:

`url ::= 'http://' hostname (':' port)? '/'`

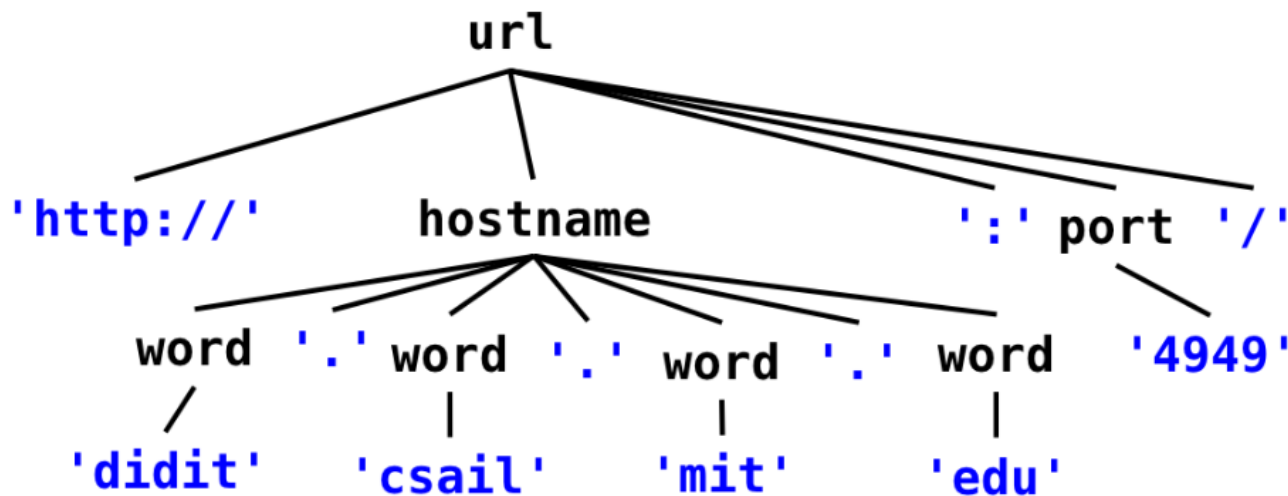
`hostname ::= (word '.')+ word`

`port ::= [0-9]+`

`word ::= [a-z]+`

`http://didit.csail.mit.edu:4949/`

- What does its parse tree look like?



More generalizations...

- **There are more things we should do to go farther:**
 - Generalizing http to support the additional protocols that URLs can have, such as ftp, https, ...
 - Generalizing the / at the end to a slash-separated path, such as `http://didit.csail.mit.edu:4949/homework/lab1/`
 - Allowing hostnames with the full set of legal characters instead of just a-z such as `http://ou812.com/`

- **Can you do these?**

```
url ::= protocol '://' hostname (':' port)? '/' (word '/')*  
protocol ::= 'ftp' | 'http' | 'https'  
hostname ::= (word '.')+ word  
port ::= [0-9]+  
word ::= [a-z 0-9]+
```

Exercise

- We want the URL grammar to also match strings of the form:

- https://webis.mit.edu/
- ftp://ftp.athena.mit.edu/

- but not strings of the form:

- ptth://web.mit.edu/
- mailto:bitdiddle@mit.edu

- So we change the grammar to:

- What could you put in place of **TODO** to match the desirable URLs but not the undesirable ones?

- word
- 'ftp' | 'http' | 'https'
- ('http' 's'?) | 'ftp'
- ('f' | 'ht') 'tp' 's'?

```
url ::= protocol '://' hostname (':'  
    port)? '/'  
protocol ::= TODO  
hostname ::= word '.' hostname |  
    word '.' word  
port ::= [0-9]+  
word ::= [a-z]+
```



(5) Markdown and HTML



Markdown and HTML



- **Markup languages:** represents typographic style in text.

<https://daringfireball.net/projects/markdown/syntax>

- **Markdown example for italics**

This is *_italic_*.

- **HTML example for italics**

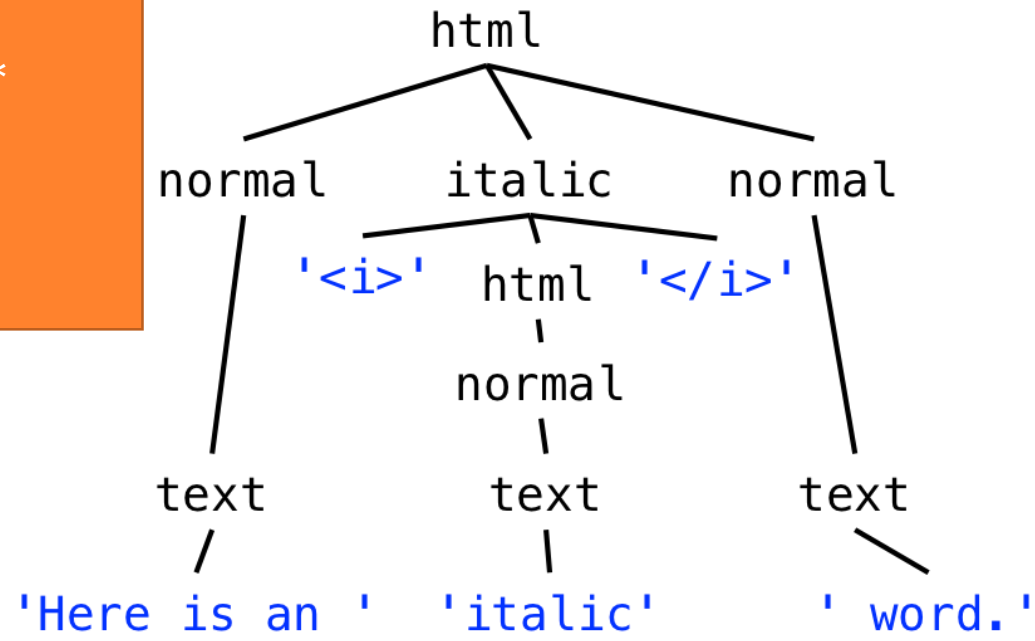
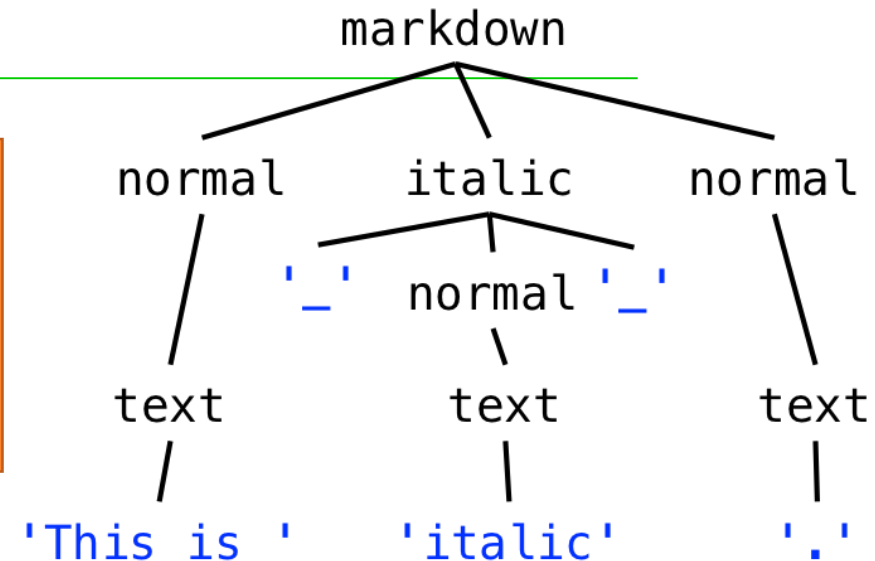
Here is an `<i>italic</i>` word.

- For simplicity, we assume the plain text between the formatting delimiters is not allowed to use any formatting punctuation, like `_` or `<>`.
- Can you write down their grammars?

Markdown and HTML

```
markdown ::= ( normal | italic ) *
italic   ::= '_' normal '_'
normal   ::= text
text     ::= [^_]*
```

```
html ::= ( normal | italic ) *
italic ::= '<i>' html '</i>'
normal ::= text
text   ::= [^<>]*
```



Markdown and HTML

```
markdown ::= ( normal | italic ) *  
italic   ::= '_' normal '_'  
normal   ::= text  
text     ::= [^_]*
```

markdown:

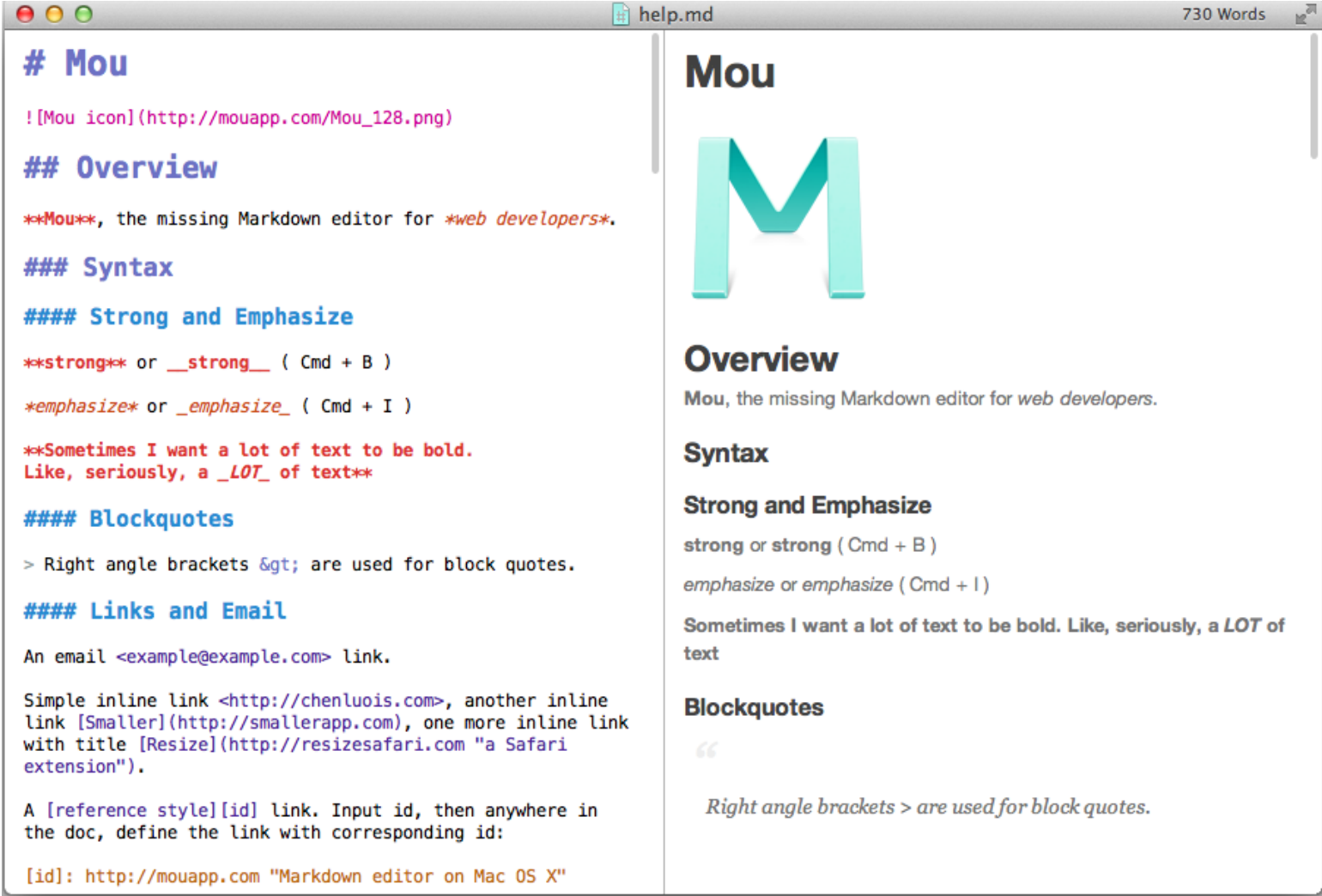
a_b_c_d_e

html:

a<i>b<i>c</i>d</i>e

```
html ::= ( normal | italic ) *  
italic ::= '<i>' html '</i>'  
normal ::= text  
text    ::= [^<>]*
```

If you match the specified grammar against it, which letters are inside matches to the italic nonterminal?



Donald E. Knuth (高德纳)

- Donald E. Knuth (1938-)
- Stanford University
- 1974年图灵奖获得者
史上最年轻的图灵奖获得者
- 被誉为现代计算机科学的鼻祖
- 算法分析之父，为理论计算机科学的发展做出重要贡献
- 《计算机程序设计艺术》(The Art of Computer Programming)，计算机科学理论与技术的经典巨著，其作用与地位可与《几何原本》相比。
- TeX的发明者
- “计算机老顽童”



创造TEX和METAFONT

The image shows a side-by-side comparison of a LaTeX source file and its rendered PDF output. The left window, titled 'David_Grant.tex - TeXworks', displays the source code. The right window, titled 'David_Grant.pdf - TeXworks', shows the rendered document. An orange arrow points from the 'Education' section in the PDF to the corresponding LaTeX code in the source file.

Source File (David_Grant.tex):

```
%
% resume.tex
%
% (c) 2002 Matthew Boedicker <mboedick@mboedick.org> (original author)
% http://mboedick.org
% (c) 2003-2007 David J. Grant <davidgrant-at-gmail.com> http://www.davidgrant.ca
%
% This work is licensed under the Creative Commons Attribution-ShareAlike 3.0 Unported
% License. To view a copy of this license, visit http://creativecommons.org/licenses/by-
% sa/3.0/ or send a letter to Creative Commons, 171 Second Street, Suite 300, San
% Francisco, California, 94105, USA.

\documentclass[letterpaper,11pt]{article}

%-----
%Margin setup

\setlength{\voffset}{0.1in}
\setlength{\paperwidth}{8.5in}
\setlength{\paperheight}{11in}
\setlength{\headheight}{0in}
\setlength{\headsep}{0in}
\setlength{\textheight}{11in}
\setlength{\textwidth}{9.5in}
\setlength{\topmargin}{-0.25in}
\setlength{\textwidth}{7in}
\setlength{\topskip}{0in}
\setlength{\oddsidemargin}{-0.25in}
\setlength{\evensidemargin}{-0.25in}

%-----
\usepackage{fullpage}
\usepackage{shading}
%\textheight=9.0in
\pagestyle{empty}
\raggedbottom
\raggedright
\setlength{\tabcolsep}{0in}

%-----
%Custom commands
\newcommand{\resitem}[1]{\item #1 \hspace{-2pt}}
\newcommand{\resheading}[1]{\large \parashade[.9]{sharpcorners}{\textbf{#1}}
\phantom{p\^{}{E}}}}
\newcommand{\ressubheading}[4]{
\begin{tabular*}{6.5in}{l@{\extracolsep{\fill}}}{r}
```

Rendered PDF (David_Grant.pdf):

David Grant
 #666-1234 Main Street
 Vancouver, BC A1B 2C3

604-555-5555
 davidgrant-at-gmail.com
 http://www.davidgrant.ca

Education

- University of Waterloo** Waterloo, ON
M.A.Sc., Electrical Engineering (Grades: 80%) Sep. 2002 - May. 2004
 – Relevant courses: Semiconductor Devices: Physics and Modelling, Digital VLSI Design, Amorphous Silicon, Mixed-signal modelling with VHDL-AMS
- University of British Columbia** Vancouver, BC
B.A.Sc. Engineering Physics (Electrical Engineering Option) 1997-2002
 – Graduated with Honors, **86%** cumulative average, and Dean's Honour List each year.
 – Relevant courses: Solid-state physics, Quantum Mechanics, Semiconductor Devices (BJT, HBT, FET, analog IC layout and simulation), Digital Systems Design using VHDL, Waveguides and Photonics, RF, Analog/Digital Communications Systems, Analog Hardware Design

Work Experience

- D-Wave Systems** Vancouver, BC
Junior Research Scientist and Software Engineer May 2002 - Aug. 2002
 – Implemented quantum computing algorithms in a JAVA quantum computer simulator, such as

100% page 1 of 3

创造TEX和METAFONT

- 除了对排版美的追求，TEX使人们像编程一样写作文。

- TEX的版本号不是自然数列，也不是年份，而是从3开始，不断逼近圆周率(目前最新版本是3.14159265)，意思是说：这个东西趋近完美，不可能再有什么大的改进了
- 设立了一个奖项：谁发现TEX的一个错误，就付他2.56美元，第二个错误5.12美元，第三个10.24美元……以此类推

TEX

Developer(s)	Donald Knuth
Initial release	1978; 41 years ago
Stable release	3.14159265 / January 2014; 5 years ago
Repository	www.tug.org/svn/texlive/
Written in	WEB/Pascal
Operating system	Cross-platform
Type	Typesetting
License	Permissive free software
Website	tug.org



(6) Regular Grammars and Regular Expressions



Regular grammar

- A **regular grammar** has a special property: by substituting every nonterminal (except the root one) with its righthand side, you can reduce it down to a single production for the root, with only terminals and operators on the right-hand side. 正则语法：简化之后可以表达为一个产生式而不包含任何非终止节点
- Which of them are regular grammars?

```
url ::= 'http://' hostname (':' port)? '/'  
hostname ::= word '.' hostname | word '.' word  
port ::= [0-9]+  
word ::= [a-z]+
```

```
markdown ::= ( normal | italic ) *  
italic ::= '_' normal '_'  
normal ::= text  
text ::= [^_]*
```

```
html ::= ( normal | italic ) *  
italic ::= '<i>' html '</i>'  
normal ::= text  
text ::= [^<>]*
```

Regular grammar

```
url ::= 'http://' hostname (':' port)? '/'
hostname ::= word '.' hostname | word '.' word
port ::= [0-9]+
word ::= [a-z]+
```

```
url ::= 'http://' ([a-z]+ '.' )+ [a-z]+ (':' [0-9]+)? '/'
```

Regular!

```
markdown ::= ( normal | italic ) *
italic ::= '_' normal '_'
normal ::= text
text ::= [^_]*
```

```
markdown ::= ([^_]* | '_' [^_]* '_' )*
```

Regular!

```
html ::= ( normal | italic ) *
italic ::= '<i>' html '</i>'
normal ::= text
text ::= [^<>]*
```

```
html ::= ( [^<>]* | '<i>' html '</i>' )*
```

Not regular!

Regular Expressions (*regex*)

- The reduced expression of terminals and operators can be written in an even more **compact** form, called a **regular expression**. 正则表达式
- A regular expression **does away with the quotes** around the terminals, and **the spaces between terminals and operators**, so that it consists just of terminal characters, parentheses for grouping, and operator characters. 去除引号和空格，从而表达更简洁(更难懂)

$$\text{markdown} ::= ([^_]* \mid '[_]' [^_]* '[_]')^*$$

$$\text{markdown} ::= ([^_]* | _[^_]*_)^*$$

- Regular expressions are also called ***regex*** for short.
 - A regex is far less readable than the original grammar, because it lacks the nonterminal names that documented the meaning of each subexpression.
 - But a regex is fast to implement, and there are libraries in many programming languages that support regular expressions.

Some special operators in regex

- **.** any single character
- **\d** any digit, same as **[0-9]**
- **\s** any whitespace character, including space, tab, newline
- **\w** any word character, including underscore, same as **[a-zA-Z_0-9]**
- **\., \(\, \), *, \+, ...**

escapes an operator or special character so that it matches literally

An example

- **Original:**

```
'http://' ([a-z]+ '.' )+ [a-z]+ (':' [0-9]+)? '/'
```

- **Compact:**

```
http://([a-z]+.)+[a-z]+(:[0-9]+)?/
```

- **With escape:**

```
http://([a-z]+\.)+[a-z]+(:[0-9]+)?/
```

Exercise

- Consider the following regular expression:

[A-G]+(b|#)?

- Which of the following strings match the regular expression?

- Ab
- C#
- ABKb
- AbB
- GFE

Context-Free Grammars

- In general, a language that can be expressed with a system of grammars is called **context-free**.
 - Not all context-free languages are also regular; that is, some grammars can't be reduced to single nonrecursive productions.
 - The HTML grammar is context-free but not regular.
- The grammars for most programming languages are also context-free.
- In general, any language with nested structure (like nesting parentheses or braces) is context-free but not regular.

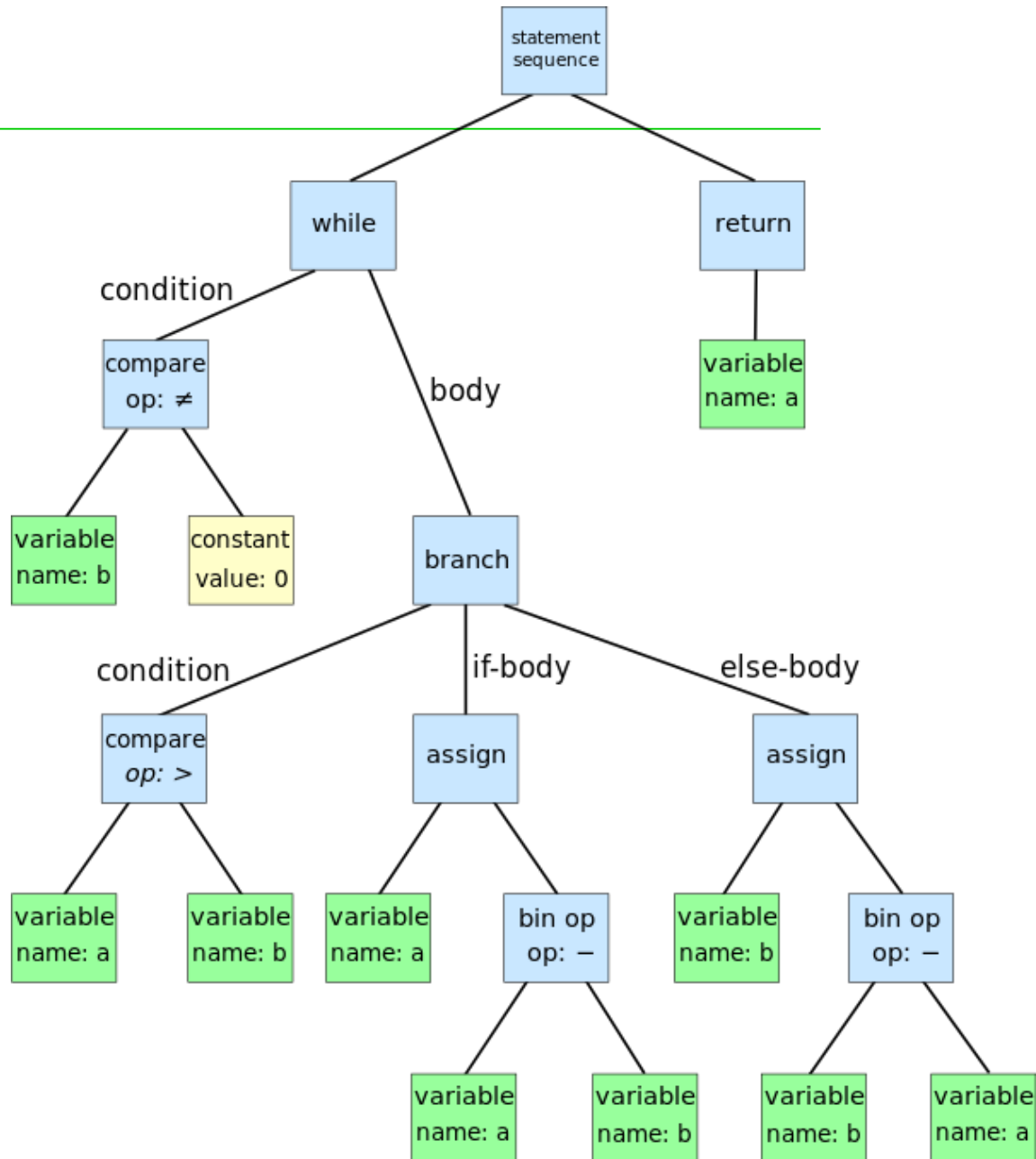
Java grammar

```
statement ::=
    '{' statement* '}'
  | 'if' '(' expression ')' statement ('else' statement)?
  | 'for' '(' forinit? ';' expression? ';' forupdate? ')' statement
  | 'while' '(' expression ')' statement
  | 'do' statement 'while' '(' expression ')' ';'
  | 'try' '{' statement* '}' ( catches | catches? 'finally' '{' statement* '}' )
  | 'switch' '(' expression ')' '{' switchgroups '}'
  | 'synchronized' '(' expression ')' '{' statement* '}'
  | 'return' expression? ';'
  | 'throw' expression ';'
  | 'break' identifier? ';'
  | 'continue' identifier? ';'
  | expression ';'
  | identifier ':' statement
  | ';'

```


Java AST

```
while (b != 0) {  
    if (a > b)  
        a = a - b;  
    else  
        b = b - a;  
}  
return a;
```





(7) * Parsers



Parser 将输入文本转为parse tree

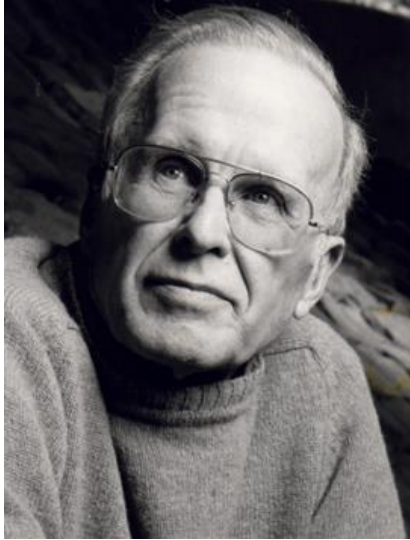
- A parser takes a sequence of characters and tries to match the sequence against the grammar. **parser:** 输入一段文本，与特定的语法规则建立匹配，输出结果
- The parser typically produces a **parse tree**, which shows how grammar productions are expanded into a sentence that matches the character sequence. **parser:** 将文本转化为**parse tree**
 - The root of the parse tree is the starting nonterminal of the grammar.
 - Each node of the parse tree expands into one production of the grammar.
- The final step of parsing is to do something useful with this parse tree. 利用产生的**parse tree**，进行下一步的处理
- A recursive abstract data type that represents a language expression is called an *abstract syntax tree (AST)*.

Parser Generator 根据语法定义生成parser

- A parser generator is a tool that reads a grammar specification and converts it to a Java program that can recognize matches to the grammar. **Parser generator**是一个工具，根据语法规则生成一个**parser**程序
 - Read <http://web.mit.edu/6.031/www/sp17/classes/18-parsers>
This is not the mandatory contents of this course.
- **More broadly:**
 - A parser generator is a programming tool that creates a parser, interpreter, or compiler from some form of formal description of a language.
 - The input may be a text file containing the grammar written in BNF or EBNF that defines the syntax of a programming language. **输入是遵循BNF或EBNF格式的文本文件**
 - The output is some source code of the parser for the grammar. **输出为parser的源代码**

Backus Normal Form (BNF) 巴克斯范式

- 1959年6月，Backus Normal Form (BNF)首次提出，以递归形式描述语言的各种成分，凡遵守其规则的程序就可保证语法上的正确性。
 - 经过Peter Naur的改进与完善以及Niklaus Wirth的扩充，形成了EBNF（Extended BNF），也就是目前使用的BNF。
 - 经Donald Knuth 的建议，BNF中的N变成了Naur (Backus-Naur Form)。



John Backus (1924-2007)
1977年图灵奖得主



Peter Naur (1928-2016)
2005年图灵奖得主



Niklaus Wirth (1934-)
1984年图灵奖得主

Grammar, Parser Generator, and Parser

- Grammar定义语法规则（BNF格式的文本），Parser generator根据语法规则产生一个parser，用户利用parser来解析文本，看其是否符合语法定义并对其做各种处理（例如转成parse tree）

Grammar

```
root ::= html;
html ::= ( italic | normal ) *;
italic ::= '<i>' html '</i>';
normal ::= text;
text ::= [^<>]+;
```

例如：
正则表达式语法
HTML语法
Java语法

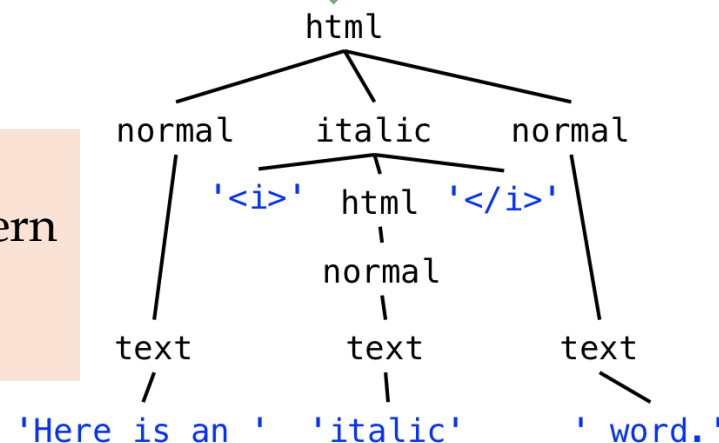
Parser Generator (Tool)

例如：
Regex pattern
HTML tree
Java AST

Here is an *italic* word.

Parser (API or tool)

例如：
Java regex parser
HTML parser
Java compiler



In your future study...

基础：形式语言
任务：设计一种语言

Grammar

```
root ::= html;
html ::= ( italic | normal ) *;
italic ::= '<i>' html '</i>';
normal ::= text;
text ::= [^<>]+;
```

例如：
正则表达式语法
HTML语法
Java语法

Parser Generator (Tool)

例如：
Regex pattern
HTML tree
Java AST

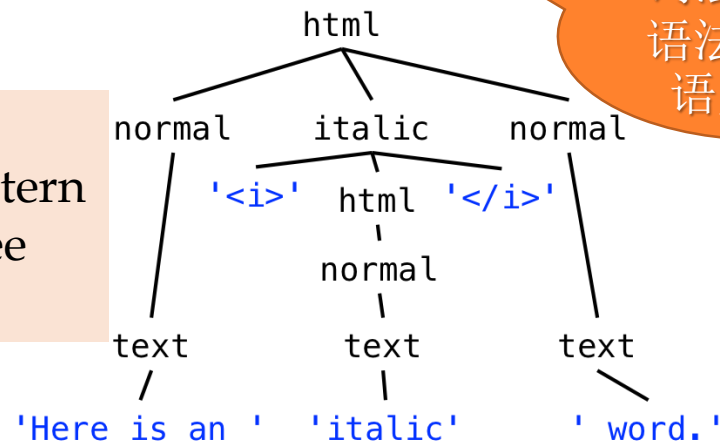
编译原理课程：
为某个语言设计
编译器

Here is an `<i>italic</i>` word.

Parser (API or tool)

例如：
Java regex parser
HTML parser
Java compiler

词法分析、
语法分析、
语义分析





(8) Using regular expressions in Java

在本课程里，只需要能够熟练掌握正则表达式regex这种“基本语法”，并熟练使用JDK提供的 regex parser进行数据处理即可

Using regular expressions in Java

- Regexes are widely used in programming.
- In Java, you can use regexes for manipulating strings (see `String.split`, `String.matches`, `java.util.regex.Pattern`).
- They're built-in as a first-class feature of modern scripting languages like Python, Ruby, and JavaScript, and you can use them in many text editors for find and replace.
- Regular expressions are your friend!

java.util.regex for Regex processing

- The **java.util.regex** package primarily consists of three classes:
 - A **Pattern** object is a compiled representation of a regular expression. The **Pattern** class provides no public constructors. To create a pattern, you must first invoke one of its public static compile methods, which will then return a **Pattern** object. These methods accept a regular expression as the first argument. **Pattern**是对regex正则表达式进行编译之后得到的结果
 - A **Matcher** object is the engine that interprets the pattern and performs match operations against an input string. Like the **Pattern** class, **Matcher** defines no public constructors. You obtain a **Matcher** object by invoking the matcher method on a **Pattern** object. **Matcher**: 利用**Pattern**对输入字符串进行解析
 - A **PatternSyntaxException** object is an unchecked exception that indicates a syntax error in a regular expression pattern.

java.util.regex for Regex processing

Package java.util.regex

Classes for matching character sequences against patterns specified by regular expressions.

See: [Description](#)

Interface Summary

Interface	Description
MatchResult	The result of a match operation.

Class Summary

Class	Description
Matcher	An engine that performs match operations on a character sequence by interpreting a Pattern .
Pattern	A compiled representation of a regular expression.

Exception Summary

Exception	Description
PatternSyntaxException	Unchecked exception thrown to indicate a syntax error in a regular-expression pattern.

Using regular expressions in Java

- Replace all runs of spaces with a single space:

```
String singleSpacedString = string.replaceAll(" +", " ");
```

- Match a URL:

```
Pattern regex = Pattern.compile("http://([a-z]+\\.)+[a-z]+(:[0-9]+)?/");  
Matcher m = regex.matcher(string);  
if (m.matches()) {  
    // then string is a url  
}
```

- Extract part of an HTML tag:

```
Pattern regex = Pattern.compile("<a href=\"([^\"]*)\">");  
Matcher m = regex.matcher(string);  
if (m.matches()) {  
    String url = m.group(1);  
    // Matcher.group(n) returns the nth parenthesized part of the regex  
}
```

An example

- Write the shortest regex you can to remove single-word, lowercase-letter-only HTML tags from a string:

```
String input = "The <b>Good</b>, the <i>Bad</i>, and the  
                <strong>Ugly</strong>";
```

```
String regex = "TODO";
```

```
String output = input.replaceAll(regex, "");
```

- If the desired output is "The Good, the Bad, and the Ugly", what is shortest regex you can put in place of **TODO**?

</?[a-z]+>

Character Classes

Construct	Description
[abc]	a, b, or c (simple class)
[^abc]	Any character except a, b, or c (negation)
[a-zA-Z]	a through z, or A through Z, inclusive (range)
[a-d[m-p]]	a through d, or m through p: [a-dm-p] (union)
[a-z&&[def]]	d, e, or f (intersection)
[a-z&&[^bc]]	a through z, except for b and c: [ad-z] (subtraction)
[a-z&&[^m-p]]	a through z, and not m through p: [a-lq-z] (subtraction)

Metacharacters: <([{\^-= \$! |]})? * + . >

Two ways to force a metacharacter to be treated as an ordinary character:

- Precede the metacharacter with a backslash \
- Enclose it within \Q (which starts the quote) and \E (which ends it).


Predefined Character Classes

Construct	Description
.	Any character (may or may not match line terminators)
\d	A digit: [0-9]
\D	A non-digit: [^0-9]
\s	A whitespace character: [\t\n\x0B\f\r]
\S	A non-whitespace character: [^\s]
\w	A word character: [a-zA-Z_0-9]
\W	A non-word character: [^\w]

Quantifiers

Greedy	Reluctant	Possessive	Meaning
$X?$	$X??$	$X?+$	X , once or not at all
X^*	$X^*?$	X^*+	X , zero or more times
X^+	$X^+?$	X^{++}	X , one or more times
$X\{n\}$	$X\{n\}?$	$X\{n\}^+$	X , exactly n times
$X\{n, \}$	$X\{n, \}?$	$X\{n, \}^+$	X , at least n times
$X\{n, m\}$	$X\{n, m\}?$	$X\{n, m\}^+$	X , at least n but not more than m times

Boundary Matchers



Boundary Construct	Description
<code>^</code>	The beginning of a line
<code>\$</code>	The end of a line
<code>\b</code>	A word boundary
<code>\B</code>	A non-word boundary
<code>\A</code>	The beginning of the input
<code>\G</code>	The end of the previous match
<code>\Z</code>	The end of the input but for the final terminator, if any
<code>\z</code>	The end of the input

Pattern method equivalents in `java.lang.String`

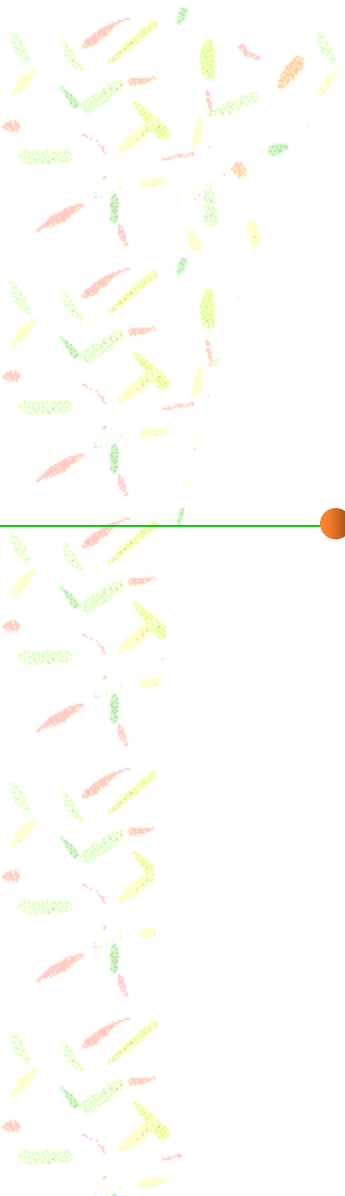
- `public boolean matches(String regex)`: Tells whether or not this string matches the given regular expression.
`Pattern.matches(regex, str).`
- `public String[] split(String regex, int limit)`: Splits this string around matches of the given regular expression.
`Pattern.compile(regex).split(str, n)`
- `public String[] split(String regex)`: Splits this string around matches of the given regular expression.
- `public String replace(CharSequence target, CharSequence replacement)`

Learn by yourself


- 
- <https://docs.oracle.com/javase/tutorial/essential/regex/index.html>
 - https://en.wikipedia.org/wiki/Regular_expression



Summary



Summary

- 
- Machine-processed textual languages are ubiquitous in computer science.
 - Grammars are the most popular formalism for describing such languages
 - Regular expressions are an important subclass of grammars that can be expressed without recursion.

Summary



■ Safe from bugs

- Grammars and regular expressions are declarative specifications for strings and streams, which can be used directly by libraries and tools.
- These specifications are often simpler, more direct, and less likely to be buggy than parsing code written by hand.

■ Easy to understand

- A grammar captures the shape of a sequence in a form that is easier to understand than hand-written parsing code.
- Regular expressions, alas, are often not easy to understand, because they are a one-line reduced form of what might have been a more understandable regular grammar.

■ Ready for change

- A grammar can be easily edited, but regular expressions, unfortunately, are much harder to change, because a complex regular expression is cryptic and hard to understand.



The end

April 22, 2020