Strings and Regular Expressions

Object Oriented Programming
2024 First Semester
Shin-chi Tadaki (Saga University)

1. `String` class
2. Creation and concatenation
3. Regular expressions

Today's sample programs

- `https://github.com/oop-mc-saga/StringsAndRegularExpressions`

String class

- The `String` class stores a sequence of characters
- Major methods
  - `concat()`: concatenates the specified string at the end
  - `indexOf()`: returns the index of the specified sub-string
  - `replace()`: replaces the sub-strings into the new one
  - `substring()`: obtains the sub-strings by specifying indexes

Notes on `String` class

- `String` objects are *immutable* (it can not be changed).
- Comparing two `String` objects
  - `==`: equals as an objects
  - `equals()`: stores the same string as data

```java
public static void main(String[] args) {
    String a = "abc";
    String b = a;
    String c = "ab";
    String d = c;
    c = c + "c";
    if (a == b) {
        System.out.println("a is the same object of b");
    }
    if (a == c) {
        System.out.println("a is the same object of c");
    }
    if (c != d) {
        System.out.println("d is not the same object of c");
    }
    if (a.equals(b)) {
        System.out.println("a is the same string of b");
    }
    if (a.equals(c)) {
        System.out.println("a is the same string of c");
    }
}
```

example/StringTest.java

Searching in `String` class

- `char charAt(int index)`
  - returns a character at the specified `index`
- `int indexOf(String str)`
  - returns the index where the specified `str` first appears
- `int indexOf(String str, int from)`
  - returns the index where the specified `str` first appears after `from`
- `String substring(int begin, int end)`
  - returns the substring by specifying the begin and end indexes

`toString()` method

- `Object.toString()`
  - Converts the specified object into a string.
- You can define how to convert the object into a string by overriding the `toString()` method.

StringBuilder class

- The `StringBuilder` class supports to concatenate objects into a string
- `append(Object o)`: appends the object at the end of the string using `o.toString()`
- `delete(int start, int end)`: deletes the substring by specifying the range
- `insert(int offset, Object o)`: inserts the object `o` at the position `offset`
- `toString()`: converts the content of `StringBuilder` instance into a `String` instance

```java
public static <T> String list2String(List<T> list) {
    StringBuilder sb = new StringBuilder();
    sb.append("[");
    list.stream().forEachOrdered(
            p -> sb.append(p).append(",")
    );
    int k = sb.lastIndexOf(",");
    sb.deleteCharAt(k).append("]");
    return sb.toString();
}
```

example/BuilderSample.java

StringJoiner class

- Joining object with a separator
- Also can specify the prefix and the suffix.

```
1  public static <T> String list2String2(List<T> list) {
2      StringJoiner sj = new StringJoiner(",", "(", ")");
3      list.stream().forEachOrdered(t -> sj.add(t.toString()));
4      return sj.toString();
5  }
```

example/BuilderSample.java

Regular expressions

- Regular expressions are string patterns with repetitions of characters or strings
- special characters of positions
  - `^`: the beginning of the string
    - `^Java`: Strings starting `Java`
  - `$`: the end of the string
    - `Java$`: Strings end with `Java`

Parts of syntaxes

- `X?` : `X` appears 0 or 1 time
- `X+` : `X` repeats more than once
- `X*` : `X` repeats more than 0 times
- `X{n}`: `X` repeats n times
- `X{n,}`: `X` repeats more than n times
- `[abc]`: `a`, `b`, or `c`
- `\s`: whitespace characters (space, tab, etc)
- `\S` : non-whitespace characters
- `\d` : digit `[0-9]`
- `\D` : non-digit

Splitting strings using regular expressions

- Lines using various delimiters such as space, tab, comma, colon, etc.
- Regular expressions enables
- `String ss[] = s.split("\\s|,|:");`
- `x|y`: `x` or `y`

```java
public static void main(String[] args) {
    String input[] = {
        "a,b,c,d,e,f",
        "a b c d e f",
        "a\tb\tc\td\te\tf",
        "a:b:c:d:e:f"
    };
    for (String s : input) {
        String ss[] = s.split("\\s|,|:");
        for (String e : ss) {
            System.out.print(e + " ");
        }
        System.out.println();
    }
}
```

example/SplitExample.java

Find strings using regular expressions

- Define regular expression
  - `Pattern p = Pattern.compile(String regex);`
- Generate `matcher`
  - `Matcher m = p.matcher(input);`
- Search matched strings
  - `boolean m.find()`: find the next subsequence matched
  - `int m.start()`: the start index of the previous match
  - `String m.group()`: the input subsequence of the previous match.

```java
public static void main(String[] args) {
    String input = "00101110100011";
    //Define regular expression
    Pattern p = Pattern.compile("101+");
    Matcher m = p.matcher(input);
    int c = 0;//starting point of matching
    while (m.find(c)) {//matching by starting point
        c = m.start();//matched point
        String s = m.group();
        System.out.println("matches "+s + " at " + c);
        c++;//next matching
    }
}
```

example/RegexSample.java

Numbering matched patterns

- Matched patterns `((A)(B(C)))` are numbered as
  1. `((A)(B(C)))`
  2. `(A)`
  3. `(B(C))`
  4. `(C)`

```java
public static void main(String[] args) {
    String dates[] = {"20100401", "20110530", "20101109",
        "19991010", "19890321", "Aug5,2019", "2010Sep9"};
    Pattern p = Pattern.compile("((\\d{4})(\\d\\d)(\\d\\d))");
    for (String d : dates) {
        Matcher m = p.matcher(d);
        while (m.find()) {
            int n = m.groupCount();//number of matched positions
            StringJoiner sj = new StringJoiner("/");
            String str = m.group(1);
            for(int i=2;i<=n;i++){
                sj.add(m.group(i));
            }
            System.out.println(str + " -> " + sj.toString());
        }
    }
}
```

example/RegexSample2.java

Replacing strings using regular expressions

- Simple replacements
  - `m.replaceFirst()`
  - `m.replaceAll()`
- reusing matched string
  - matched strings are numbered as `$n`

```java
public static void main(String[] args) {
    String input = "0010111010011101";
    //Define regular expression
    Pattern p = Pattern.compile("101+");
    Matcher m = p.matcher(input);
    //Simple replacement
    System.out.println(m.replaceFirst("121"));
    System.out.println(m.replaceAll("121"));
    //Using matched string
    System.out.println(m.replaceAll("_$0_"));
    //Using matched position
    p = Pattern.compile("(10)(1+)");
    m = p.matcher(input);
    System.out.println(m.replaceAll("12$2"));
}
```

example/ReplaceExample.java

Exercise

Understand the role of variable `k` in `exercise/Simple.java`.