# String Processing

## Tópicos

* Búsqueda, análisis y construcción de Strings
* Búsqueda, análisis y remplazo de Strings usando expresiones regulares, patrones de expresión para coincidir limitado por punto, asterisco, más, ?, \d, \D, \s, \S, \w, \W, \b, \B, [], ()
* Formatear Strings usando los parámetros de formato %b, %c, %d, %f y %s

## Introducción

Los objetos String son usados constantemente en aplicaciones Java. Los programas usan los String no solo en la lógica de negocio, también se utilizan para implementar funcionalidad de Logging. Por lo tanto, es muy útil entender los Strings y los conceptos relacionados.

Aparte de las operaciones comunes, buscar y analizar son dos operaciones muy útiles que son ejecutadas en los Strings. Java provee un gran soporte para lograr estas operaciones eficiente y fácilmente. Empezando por la versión 4, Java soporta expresiones regulares y ofrece dos clases útiles para explotar el poder de las expresiones regulares. Desde la versión 5, Java también soporta el método como en C printf() para formatear Strings.

El examen tiene tres temas relacionados con Strings: “Crear y manipular Strings”; “Manipular datos usando la clase StringBuilder y sus métodos”; y “Probar la igualdad entre Strings y otros objetos usando == y el método equals()”: Asumienddo que ya tienes los prerrequisitos para el examen y estás familiarizado con el uso de Strings, procederemos a los temas que cubre el OCPJP7.

Como candidato para el examen, debes de estar familiarizado con la búsqueda, análisis y construcción de Strings; expresiones regulares y su uso; y el formateo de Strings así como los especificadores de formato. En este capítulo vamos a cubrir esos temas a detalle.

## Procesando Strings

En esta sección aprenderás cómo puedes procesar caracteres almacenados en un objeto String.

### Búsqueda de String

Cuando trabajas con Strings, puedes necesitar buscar un String dado. Hay muchas versiones sobrecargadas del método indexOf() disponibles en la clase String, aparte del método lastIndexOf() para buscar un String al final, y el método regionMatches() para comparar una región de texto dentro de una String. Vamos a discutir cada uno de estos métodos uno a uno.

#### indexOf()

¿Cómo puedes buscar un carácter en particular en un String, y si está, cuál es la posición? Por ejemplo, cuál sería la posición del carácter “J” ocupado en el String “OCPJP”:

El método indexOf() busca el carácter especificado y retorna la PRIMERA ocurrencia del carácter (el índice empieza con 0 como en los arrays). Ahora vamos a buscar una sub-string a una String dada. El método indexOf() está sobre-cargado y una definición del método toma una String como argumento de búsqueda. Veamos un ejemplo:

// SearchString2.java

public class SearchString2 {

public static void main(String... args) {

String str = “I am preparing for OCPJP”;

System.out.println(“La sub-string \”for\” ocurre al índice: ” + str.indexOf(“for”));

}

}

Imprime

La sub-string “for” ocurre al índice: 15

// SearchString1.java

public class SearchString1 {

public static void main(String... args) {

String str = “OCPJP”;

System.out.println(“El carácter J ocurre en el índice: ” + str.indexOf(‘J’));

}

}

Imprime

El carácter J ocurre en el índice 3

¿Qué pasaría si la string buscada no existe en el String? Entonces retorna -1 que significa “no encontrada”.

Estos dos son problemas fáciles; vamos a intentar con uno un poco más difícil. Dada una String grande, ¿Cómo puedes encontrar cuántas veces ocurre la string “am” en esa String? Nada de qué preocuparte, en este caso hay otra versión del método indexOf(). En este método puedes especificar un índice que indica desde dónde va a comenzar la búsqueda. Vamos a ver una implementación:

// SearchString3.java

public class SearchString3 {

public static void main(String... args) {

String str = “I am a student. I am preparing for OCPJP”;

int fromIndex = 0;

while(str.indexOf(“am”, fromIndex > -1)) {

formIdex = str.indexOf(“am”, fromIndex);

System.out.println(“Substring \”am\” ocurre al índice: ” + fromIndex);

fromIndex++;

}

}

}

Imprime

Substring “am” ocurre al índice: 2

Substring “am” ocurre al índice 18

Usas un while-loop para checar cuando hay más ocurrencias de la string buscada. Esto también mantiene un índice (fromIndex) de donde iniciarás la búsqueda. Se incrementa el fromIndex después de cada ocurrencia, por lo tanto la siguiente ocurrencia puede ser encontrada.

Si quieres buscar la última ocurrencia de la string buscada, puedes usar las versiones sobre cargadas del método lastIndexOf definidas en la clase String. La clase String también tiene métodos simples para checar el principio o el final de las strings y estos métodos están en la siguiente tabla:

|  |  |
| --- | --- |
| Método | Descripción |
| boolean startsWith(String prefixString, int offset) | Empezando por el offset, checa si esta String contiene la prefixString |
| boolean startsWith(String prefixString) | Es igual que startsWith(prefixString, 0) |
| boolean endsWith(String suffixString) | Checa si esta string tiene suffixString |

regionMatches()

Considera la string “Tarzan: Hi Jane, wanna ride an elephant?\n Jane: No thanks!” ¿Cómo puedes checar si la string “Jane: No thanks!” está presente como la primera parte en la respuesta de Jane? (Nota: la respuesta de Jane es la región de texto que viene después de la nueva línea \n)

Hay muchas maneras de resolver este problema, y puedes usar el método regionMatches() aquí en función de aprender a usarlo. En la siguiente tabla se listan los dos métodos sobrecargados de regionMatches(). Para resolver este problema en la string dada puedes empezar por localizar el carácter “\n” y de esta posición, puedes buscar la string “Jane: No thanks!”, como se ve en el siguiente código muestra cómo.

|  |  |
| --- | --- |
| Método | Descripción |
| boolean regionMatches(int start, String matchingStr, int matchStartOffset, int matchLen) | Empezando por start en esta String, checa si la región del texto dato por matchingStr coincide. En matchStr, checa por el número de caracteres matchLen empezando por matchStartOffset |
| boolean regionMatches(boolean ignoreCase, int start, String matchString, int matchStartOffset, int matchLen) | Lo mismo que un método previo, pero con un argumento adicional al inicio que si es true, ignora si son altas o bajas las letras |

### Análisis de cadenas

public class MatchRegionInString {

public static void main(String[] args) {

String chat = “Tarzan: Hi Jane, wanna ride an Elephant? \n Jane: No: thanks! I’m” +

“ preparing for OCPJP now!”);

String matchString = “Jane: No thanks!”;

// primero obtenemos la posición de donde inicia la región a buscar

int startIndex = chat.indexOf(‘\n’);

System.out.println(“La respuesta de Jane empieza en el índice: ” + startIndex);

// si encuentra a \n intenta buscar la String coincidente “Jane: no thanks!”

// desde startIndex

if(startIndex > -1) {

boolean doesMatch = chat.regionMatches(startIndex + 1, matchString, 0, matchString.length());

if(doesMatch)

System.out.println(“El string coincide” + matchString);

}

}

}

Analizar cadenas es una operación útil e interesante en los Strings. (La palabra parse significa analizar para descomponer en partes constituyentes basadas en una estructura dada) Ahora te presentaremos la operación básica de parsear en esta sección; después discutiremos tópicos avanzados de parseo en este capítulo cuando revisemos expresiones regulares.

#### Conversión de cadenas

En tus programas, encontrarás constantemente la necesidad de convertir cadenas de y hacia tipos primitivos como float, int, boolean. Para convertir de un tipo primitivo a un tipo String, puedes usar el método sobrecargado valueOf() de la clase String.

Vamos a empezar con la conversión del valor entero 10 a String.

String str1 = String.valueOf(10); // modo correcto de convertir un entero a un String

Nota que la asignación directa causará un error de compilación:

String str1 = 10; // genera error de compilación

String str1 = (String) 10; // genera error de compilación

¿Qué pasa con la conversión a la inversa: si el String tiene un valor de algún tipo primitivo, ¿Cómo puedes ejecutar la conversión? Obviamente , las siguientes dos sentencias, que intentan asignar o cambiar directamente, a través de un casteo explícito, resultan en un error de compilación:

int i = “10”;

int i = (int) “10”;

Para hacer esta conversión, necesitas usar el método estático parseInt() disponible en la clase Integer como sigue:

int i = Integer.parseInt(“10”); // modo correcto de convertir un String a int

Este método está sobrecargado. Hay otra versión de este método que toma un argumento adicional: la base (o radix) del valor integral como octal o hexadecimal. Las clases envoltorio Byte, Short, Long, Float y Double tienen los métodos equivalentes de parseo para convertir strings a sus correspondientes tipos primitivos. ¿Qué pasa si tú pasas un argumento inválido a uno de estos métodos? Por ejemplo:

float f = Float.parseFloat(“no such value”); // java.lang.NumberFormatException en Runtime

#### El método split()

El siguiente código es un ejemplo de cómo dividir una sentencia e imprimir todas las palabras en el string usando el método split() de la clase String:

public class ParseString1 {

public static void main(String... args) {

String quote = “Never lend books-nobody ever returns them!”

String[] words = quote.split(“ ”); // divide la cadena basado en el delimitador “ ”

for(String word: words)

System.out.println(word);

}

}

El método split() toma como delimitador una expresión regular (se verán después). En este ejemplo tú provees un espacio en blanco como delimitador, por lo que es capaz de extraer todas las palabras en la sentencia. Nota como los caracteres “-” y “!” son parte de las palabras, ya que tú no especificaste signos de puntuación como delimitadores. Puedes revisitar este problema cuando aprendas expresiones regulares.

NOTA: el argumento del método split() es una string delimitadora, cuando es una expresión regular y pasas una sintaxis inválida obtienes una excepción PatternSyntaxException.

Ahora, asume que tienes una cadena que contiene la ruta de un directorio y quieres parsearla para imprimir los nombres de cada folder individualmente. Observa la implementación:

public class ParseString2 {

public static void main(String... args) {

String str = “c:\\work\\programs\\parser”;

String[] dirList = str.split(“\\\\”);

for(int 1 = 0; i < dirList.length; i++)

System.out.println(dirList(i));

}

}

### Expresiones regulares

Una expresión regular define un patrón de búsqueda que puede ser utilizado para ejecutar operaciones como búsqueda de cadenas y manipulación de las mismas. Una expresión regular es más que una secuencia de símbolos predefinidos especificados en una sintaxis específica que te ayuda a buscar o manipular cadenas. Una expresión regular es especificada como una cadena que se aplica a otra, de izquierda a derecha.

Puedes preguntarte para qué necesitas regex cuando puedes ejecutar una búsqueda directamente usando la función de String como viste en los temas pasados, usando indexOf(), por ejemplo. Bueno, la respuesta es muy simple. Puedes usar el método indexOf() o cualquier otro método similar cuando conoces con exactitud la cadena que vas a buscar. Sin embargo, en algunos casos donde tú solo conoces el patrón de la cadena pero no la cadena específica necesitas usar regex. Regex es una muy poderosa herramienta que simplifica la búsqueda y manipulación de Strings. Por ejemplo, digamos que necesitas buscar todas las direcciones de e-mail en una cadena. No puedes lograrlo usando el método indexOf() ya que no sabes las direcciones exactas; sin embargo, puedes usar regex para especificar un patrón que encontrará todos los e-mails en la cadena dada.

#### Entender los símbolos de regex

Ahora nos enfocaremos en entender la sintaxis y la semántica de los símbolos usados para especificar regex. En la siguiente tabla se muestran los símbolos usados comúnmente:

|  |  |
| --- | --- |
| Símbolo | Descripción |
| ^expr | Coincide la expresión al principio de la línea |
| expr$ | Coincide la expresión al final de la línea |
| . | Coincide un carácter único con excepción de nueva línea |
| [xyz] | Coincide cualquiera: x o y o z |
| [p-z] | Especifica un rango. Coincide cualquier carácter de la p a la z |
| [p-z1-9] | Coincide tanto caracteres de la p a la z como cualquier dígito del 1 al 9 (recuerda, este no quiere coincidir p1) |
| [^p-z] | ^ como primer carácter dentro de las llaves niega el patrón; coincide cualquier carácter con excepción de la p a la z |
| Xy | Coincide x seguida de y |
| x|y | Coincide x o y |

Puedes usar los símbolos dados en la tabla anterior y especificar un regex. Por ejemplo, puedes escribir “[0-9]” para coincidir todos los números (dígitos) o “[\t\r\f\n]” para coincidir todos los espacios en blanco. También puedes usar ciertos metasímbolos predefinidos para facilitar la especificación de un regex. Por ejemplo, puedes especificar “\d” en lugar de “[0-9]” para coincidir dígitos o “\s” en lugar de “[\t\r\f\n]” para coincidir con todos los espacios en blanco. La siguiente tabla muestra una lista de los metasímbolos más comunes:

|  |  |
| --- | --- |
| Símbolo | Descripción |
| \d | dígitos del 0 al 9 |
| \D | Coincide siempre y cuando no sea dígito |
| \w | Coincide caracteres en palabras |
| \W | Coincide caracteres que no sean palabras |
| \s | Espacios en blanco (equivalente a [\t\r\f\n]) |
| \S | Coincide con lo que no son espacios en blanco |
| \b | Coincide con el límite de la palabra cuando está fuera de las []\b. Coincide con el backslash cuando está dentro de [\b] |
| \B | Coincide con los no límites de las palabras |
| \A | Coincide con el principio del String |
| \Z | Coincide con el fin del String |

Ok, tan lejos de estar bien. Pero ¿Qué si quisieras especificar una regex cuando la coincidencia envuelve una ocurrencia de la cuenta de caracteres? Bueno, para estas situaciones puedes especificar símbolos de cuantificación provistos en la siguiente tabla:

|  |  |
| --- | --- |
| Símbolo | Descripción |
| expr? | Coincide con 0 o 1 ocurrencia de expr (equivalente a expr{0,1}) |
| expr\* | Coincide con 0 o más ocurrencias de expr (equivalente a expr{0,}) |
| expr+ | Coincide con 1 o más ocurrencias de expr (equivalente a expr{1,}) |
| expr{x} | Coincide con x ocurrencias de expr |
| expr{x,y} | Coincide ocurrencias entre x & y de expr |
| expr{x,} | Coincide con x o más ocurrencias de expr |

#### Expresiones regulares soportadas en Java

Java 1.4 presentó las expresiones regulares. El paquete java.util.regex soporta las expresiones regulares. Este consiste en dos clases importantes, Pattern y Matcher. Pattern representa una regex en una representación compilada, y Matcher interpreta la regex y coincide la respectiva subcadena en la cadena dada.

En este punto te puedes preguntar por qué es regex soportado por clases dedicadas como Matcher y Pattern, cuando otros métodos como split() en la clase String soportan regex. Una sola respuesta es: rendimiento. Las clases Pattern y Matcher están optimizadas mientras los métodos de la clase String no lo están.

Usa las clases Matcher y Pattern siempre que ejecutes búsqueda o reemplazo de cadenas de manera pesada: estas clases tienen métodos más eficientes que split() u otros métodos.

Ahora veremos cómo puedes usar las clases Pattern y Matcher. Primero necesitas llamar un método estático compile() en la clase Pattern para obtener una instancia de un pattern. El primer argumento de este método es una regex. Después necesitas llamar a otro método estático llamado matcher() de la clase Pattern para obtener una instancia de la clase Matcher. El método matcher() retorna un objeto Matcher. Este objeto es usado para ejecutar operaciones en la cadena de entrada.

Vamos a ver cómo usar las clases Pattern y Matcher. Asume que tienes una cadena que consiste en detalles personales (como nombre, dirección, teléfono...) de un conjunto de personas. Usarás la siguiente cadena para los ejemplos del resto de esta sección.

String str = “Danny Doo, Flat no 502, Big Apartment, Wide Road, Near Huge Milestone, Hugo-city 56010, Ph: 9876543210”, Email: danny@myworld.com. Maggy Myer, Post bag no 52, Big bank post office, Big bank city 56000, ph: 9876501234, Email: maggi07@myuniverse.com.”;

NOTA: necesitas especificar un regex usando el backslash (\); no uses el forwardslash (/) en su lugar. El compilador no te dará ningún error, sin embargo no obtendrás el comportamiento deseado.

#### Buscar y analizar con regex

Vamos a empezar con un ejemplo simple. Necesitas escribir un código que imprima todas las palabras de la string str. ¿Cómo puedes hacer esto? Bueno, ¿Recuerdas el metacaracter \w, el cual coincide con todos los símbolos que forman una palabra? Usarás “\w” junto con el cuantificador “+” para hacer “\w+”, que significa que quieres buscar todas las palabras de tamaño uno o más. Observa el siguiente código:

// Regex1.java

import java.util.regex.Matcher;

import java.util.regex.Pattern;

public class Regex1 {

public static void main(String[] args) {

String str = “Danny Doo... \\ observar el str de arriba

Pattern pattern = Pattern.compile(“\\w+”);

Matcher matcher = Pattern.matcher(str);

while(matcher.find()) {

System.out.println(matcher.group());

}

}

}

Puedes ver que la regex busca todas las palabras que consisten en al menos un carácter. ¿Qué pasa cuando invocas al método compile() junto con el patrón regex de la clase Pattern. Después de esto, obtienes una instancia del Matcher llamando al método matcher de la instancia pattern. Y finalmente obtienes el resultado usando los métodos find() y group() de la clase Matcher. El método find() retorna true si existe ahí un resultado de búsqueda. El métdo group() retorna los resultados de la búsqueda, uno a uno, como una cadena.

NOTA: Usamos dos backslash en la regex (“\\w+”) porque el backslash es un carácter de escape en regex. Sin embargo, backslash también es un carácter de escape en las cadenas, lo que significa literal “\\” se interpreta como un único backslash. Esto se traduce en una salida interesante: escribimos “\\” como un único backslash en una regex, cuando será escrito como “\\\\” en un programa de Java si queremos especificar un único backslash.

En el mismo sentido, puedes buscar todos los números usando la regex “\d+”. Ahora, vamos a buscar e imprimir todos los códigos postales que aparecen en la string. Asume que los códigos siempre tienen de medida 5. Puedes lograr esto usando regex. Observa el siguiente programa:

// Regex2.java

import java.util.regex.Matcher;

import java.util.regex.Pattern;

public class Regex2 {

public static void main(String[] args) {

String str ...

Pattern pattern = Pattern.compile(“\\d{5}”);

Matcher matcher = pattern.matcher(str);

while(matcher.find()) {

System.out.println(matcher.group());

}

}

}

Opsss!!! el código anterior imprime los zip pero también imprime tres números de teléfono de manera parcial, lo que era inesperado. Hmmm, para obtener solo el código zip debes especificar más propiedades para el regex. Observa el siguiente código

En el código anterior usaste el regex “\D\d{5}\D” y trabajó bien. Lo que hiciste fue especificar un no-dígito precedente seguido por un número de seis dígitos. Fácil cierto!! Bueno, aquí hay un problema en esta solución. Ek programa está imprimiendo un espacio en blanco justo antes del número de seis dígitos ya que coinciden con “\D”. Aquí te mostramos una solución elegante para este problema: puedes usar “\b” (usado para detectar los límites de las palabras). Observa si funciona el siguiente código:

// Regex4.java

import java.util.regex.Matcher;

import java.util.regex.Pattern;

public class Regex4 {

public static void main(String... args) {

String str ...

Pattern pattern = Pattern.compile(“\\b\\d{5}\\b”);

Matcher matcher = pattern.matcher(str);

while(matcher.find())

System.out.println(matcher.group);

}

}

// Regex3.java

import java.util.regex.Matcher;

import java.util.regex.Pattern;

public class Regex3 {

public static void main(String... args) {

String str ...

Pattern pattern = Pattern.compile(“\\D\\d{5}\\D”);

Matcher matcher = pattern.matcher(str);

while(matcher.find())

System.out.println(matcher.group);

}

}

todo: 216