# Lab: Text Processing and Regular Expressions

Problems for exercises and homework for the ["Technology Fundamentals" course @ SoftUni](https://softuni.bg/courses/technology-fundamentals)

You can check your solutions in [Judge.](https://judge.softuni.bg/Contests/1330)

# Text Processing

## Repeat strings

### Write a program that reads an array of strings. Each string is repeated n times, where n is the length of the string. Print the concatenated string.

### Examples

|  |  |
| --- | --- |
| **Input** | **Output** |
| hi abc add | hihiabcabcabcaddaddadd |
| work | workworkworkwork |
| ball | ballballballball |

### Solution

Read a string array.



Initialize StringBuilder.



Iterate through elements in the array. Find the length of the word at each iteration and append it to the StringBuilder.



Print the StringBuilder.

## Substring

On the first line you will receive a string. On the second line you will receive a second string. Write a program that removes all of the occurrences of the first string in the second until there is no match. At the end print the remaining string.

### Examples

|  |  |  |
| --- | --- | --- |
| **Input** | **Output** | **Comment** |
| ice  kicegiciceeb | kgb | We remove ice once and we get "kgiciceeb"  We match "ice" one more time and we get "kgiceb"  There is one more match. The finam result is "kgb" |

### Hints

* Read the input.
* Find the first index where the key appears.
  + Use the built-in method indexOf()
* Remove the match.
  + Use the built-in method replace(String oldValue, String newValue)
* Repeat it until the text doesn't contain the key anymore.

## Reverse Strings

You will be given series of strings until you receive an **"end"** command. Write a program that reverses strings and printseach pair on separate line in format **"{word} = {reversed word}"**.

### Examples

|  |  |
| --- | --- |
| **Input** | **Output** |
| helLo  Softuni  bottle  end | helLo = oLleh  Softuni = inutfoS  bottle = elttob |
| Dog  caT  chAir  end | Dog = goD  caT = Tac  chAir = riAhc |

### Solution

Use while loop and read strings until you receive **"end"**.



Reverse the string with for loop. Start from the last index and append each symbol to the new string.



Print the reversed string in the specified format.



## Text Filter

Write a program that takes a **text** and a **string of banned words**. All words included in the ban list should be replaced with **asterisks** "**\***", equal to the word's length. The entries in the ban list will be separated by a **comma** and **space** "**,** ".

The ban list should be entered on the first input line and the text on the second input line.

### Examples

|  |  |
| --- | --- |
| **Input** | **Output** |
| Linux, Windows  It is not **Linux**, it is GNU/**Linux**. **Linux** is merely the kernel, while GNU adds the functionality. Therefore we owe it to them by calling the OS GNU/**Linux**! Sincerely, a **Windows** client | It is not \*\*\*\*\*, it is GNU/\*\*\*\*\*. \*\*\*\*\* is merely the kernel, while GNU adds the functionality. Therefore we owe it to them by calling the OS GNU/\*\*\*\*\*! Sincerely, a \*\*\*\*\*\*\* client |

### Hints

* Read the input.
* Replace all ban words in the text with asterisk (\*).
  + Use the built-in method replace(banWord, replacement).

## Digits, Letters and Other

Write a program that receives a single string and on the first line prints all the digits, on the second – all the letters, and on the third – all the other characters. There will always be at least one digit, one letter and one other characters.

### Examples

|  |  |
| --- | --- |
| **Input** | **Output** |
| Agd#53Dfg^&4F53 | 53453  AgdDfgF  #^& |

### Hints

* Read the input.
* Use loop to iterate through all characters in the text. If the char is digit print it, otherwise ignore it.
  + Use Character.isDigit(char symbol)
* Do the same for the letters and other chars
  + Find something like isDigit method for the letters.

# Regular Expressions

## Match Full Name

Write a Java Program to **match full names** from a list of names and **print** them on the console.

### Writing the Regular Expression

First, write a regular expression to match a valid full name, according to these conditions:

* A valid full name has the following characteristics:
  + It consists of **two words**.
  + Each word **starts** with a **capital letter**.
  + After the first letter, it **only contains lowercase letters afterwards**.
  + **Each** of the **two words** should be **at least two letters long**.
  + The **two words** are **separated** by a **single space**.

To help you out, we've outlined several steps:

1. Use an online regex tester like <https://regex101.com/>
2. Check out how to use **character sets** (denoted with square brackets - "[]")
3. Specify that you want **two words** with a space between them (the **space character '** **'**, and **not** any whitespace symbol)
4. For each word, specify that it should begin with an uppercase letter using a **character set**. The desired characters are in a range – **from** ‘**A**’ **to** ‘**Z**’.
5. For each word, specify that what follows the first letter are only **lowercase letters**, one or more – use another character set and the correct **quantifier**.
6. To prevent capturing of letters across new lines, put "\b" at the beginning and at the end of your regex. This will ensure that what precedes and what follows the match is a word boundary (like a new line).

In order to check your RegEx, use these values for reference (paste all of them in the **Test String** field):

|  |  |
| --- | --- |
| **Match ALL of these** | **Match NONE of these** |
| Ivan Ivanov | ivan ivanov, Ivan ivanov, ivan Ivanov, IVan Ivanov, Ivan IvAnov, Ivan Ivanov |

By the end, the matches should look something like this:



After you’ve constructed your regular expression, it’s time to write the solution in Java.

### Implementing the Solution in Java

Create a new Java project and copy your **regular expression** into a String variable:



Now, it’s time to **read the input** and create two **classes** to help us work with **regular expressions**:

* **Pattern Class** − A Pattern object is a compiled representation of a regular expression.
* **Matcher Class** − A Matcher object is the engine that interprets the pattern and performs match operations against an input string.



Now, it’s time to **extract all the matches** from our input and print them. We use matcher method **find(),** witch **attempts to find the next subsequence of the input sequence that matches the pattern.** To get our matches, we need to use method **group().**



### Examples

|  |
| --- |
| **Input** |
| Ivan Ivanov, Ivan ivanov, ivan Ivanov, IVan Ivanov, Test Testov, Ivan Ivanov |
| **Output** |
| Ivan Ivanov Test Testov |

## Match Phone Number

Write a regular expression to match a **valid phone number** from **Sofia**. After you find all **valid phones**, **print** them on the console, separated by a **comma and a space** “, ”.

### Compose the Regular Expression

A valid number has the following characteristics:

* It starts with "**+359**"
* Then, it is followed by the area code (always **2**)
* After that, it’s followed by the **number** itself:
  + The number consists of **7 digits** (separated in **two** **groups** of **3** and **4** **digits** respectively).
* The different **parts** are **separated** by **either a space or a hyphen** ('**-**').

You can use the following RegEx properties to **help** with the matching:

* Use **quantifiers** to match a **specific number** of **digits**
* Use a **capturing group** to make sure the delimiter is **only one of the allowed characters** **(space or hyphen)** and **not** a **combination** of both (e.g. +359 2-111 111 has **mixed delimiters**, it is **invalid**). Use a **group backreference** to achieve this.
* Add a **word boundary** at the **end** of the match to avoid **partial matches** (the last example on the right-hand side).
* Ensure that before the **'+'** sign there is either a **space** or the **beginning of the string**.

You can use the following table of values to test your RegEx against:

|  |  |
| --- | --- |
| **Match ALL of these** | **Match NONE of these** |
| +359 2 222 2222  +359-2-222-2222 | 359-2-222-2222, +359/2/222/2222, +359-2 222 2222  +359 2-222-2222, +359-2-222-222, +359-2-222-22222 |

### Implement the Solution in Java

Now it’s time to write the solution, so let’s start writing!

First, just like in the previous problem, put your RegEx in a variable:



Again we need a **Pattern** and **Matcher.**



We can also save our matches in a List if we need.



After that, just print the valid phone number list, using string.Join():



### Examples

|  |
| --- |
| **Input** |
| +359 2 222 2222,359-2-222-2222, +359/2/222/2222, +359-2 222 2222 +359 2-222-2222, +359-2-222-222, +359-2-222-22222 +359-2-222-2222 |
| **Output** |
| +359 2 222 2222, +359-2-222-2222 |

## Match Dates

Write a program, which matches a date in the format “dd{separator}MMM{separator}yyyy”. Use **named** **capturing groups** in your regular expression.

### Compose the Regular Expression

Every valid date has the following characteristics:

* Always starts with **two digits**, followed by a **separator**
* After that, it has **one uppercase** and **two lowercase** letters (e.g. Jan, Mar).
* After that, it has a **separator** and **exactly 4 digits** (for the year).
* The separator could be either of three things: a period (“.”), a hyphen (“-“) or a forward slash (“/”)
* The separator needs to be **the same** for the whole date (e.g. 13.03.2016 is valid, 13.03/2016 is **NOT**). Use a **group backreference** to check for this.

You can follow the table below to help with composing your RegEx:

|  |  |
| --- | --- |
| **Match ALL of these** | **Match NONE of these** |
| 13/Jul/1928, 10-Nov-1934, 25.Dec.1937 | 01/Jan-1951, 23/sept/1973, 1/Feb/2016 |

Use **named capturing groups** for the **day**, **month** and **year**.

Since this problem requires more complex RegEx, which includes **named capturing groups**, we’ll take a look at how to construct it:

* First off, we don’t want anything at the **start** of our date, so we’re going to use a **word boundary** “\b”:  
  
* Next, we’re going to match the **day**, by telling our RegEx to match **exactly two digits**,and since we want to **extract** the day from the match later, we’re going to put it in a **capturing group**:  
    
  We’re also going to give our group a **name**, since it’s easier to navigate by **group name** than by **group index**:  
  
* Next comes the separator – either a **hyphen**, **period** or **forward slash**. We can use a **character class** for this:  
    
  Since we want to use the separator we matched here to match the **same separator** further into the date, we’re going to put it in a **capturing group**:  
  
* Next comes the **month**, which consists of a **capital Latin letter** and **exactly two lowercase Latin letters**:
* Next, we’re going to match the **same separator** **we matched earlier**. We can use a **backreference** for that:  
  
* Next up, we’re going to match the year, which consists of **exactly 4 digits**:  
  
* Finally, since we don’t want to match the date if there’s anything else **glued to it**, we’re going to use another **word boundary** for the end:  
  

Now it’s time to find all the **valid dates** in the input and **print each date** in the following format: “Day: {day}, Month: {month}, Year: {year}”, each on a **new line**.

### Implement the Solution in Java

First off, we’re going to put our RegEx in a variable.



Next, we’re going to **iterate** over every single Match and **extract** the **day**, **month** and **year** from the **groups.**

### Examples

|  |
| --- |
| **Input** |
| 13/Jul/1928, 10-Nov-1934, , 01/Jan-1951,f 25.Dec.1937 23/09/1973, 1/Feb/2016 |
| **Output** |
| Day: 13, Month: Jul, Year: 1928  Day: 10, Month: Nov, Year: 1934  Day: 25, Month: Dec, Year: 1937 |

## Match Numbers

Write a program, which finds all **integer** and **floating-point numbers** in a string.

### Compose the Regular Expression

A number has the following characteristics:

* Has either **whitespace** before it or the **start** of the string (match either **^** or what’s called a [positive lookbehind](http://www.regular-expressions.info/lookaround.html)). The entire syntax for the **beginning** of your **RegEx** might look something like “(^|(?<=\s))”.
* The number might or might not be negative, so it might have a hyphen on its left side (“-“).
* Consists of **one or more digits**.
* Might or might not have **digits after the** **decimal point**
* The decimal part (if it exists) consists of a period (“.”) and **one or more digits** after it. Use a **capturing group**.
* Has either **whitespace** before it or the **end** of the string (match either **$** or what’s called a [positive lookahead](http://www.regular-expressions.info/lookaround.html)). The syntax for the **end** of the **RegEx** might look something like “($|(?=\s))”.

Let’s see how we would translate the above rules into a **regular expression**:

* First off, we need to establish what needs to exist **before** our number. We can’t use \b here, since it includes “-“, which we need to match **negative numbers**.   
  Instead, we’ll use a **positive lookbehind**, which **matches** if there’s something **immediately behind** it. We’ll match if we’re either at the **start** of the string (^), or if there’s any **whitespace** **behind** the string:  
  
* Next, we’ll check whether there’s a **hyphen**, signifying a **negative number**:  
  Since having a negative sign **isn’t required**, we’ll use the “?” quantifier, which means “**between 0 and 1 times**”.
* After that, we’ll match any integers – naturally, consisting **one or more digits**:  
  
* Next, we’ll match the **decimal** part of the number, which **might or might not exist** (note: we need to escape the **period** character, as it’s used for something else in RegEx):  
  
* Finally, we’re going to use the same logic for the end of our string as the start – we’re going to match **only** if the number has **either a whitespace or the end of the string (“**$**”)**:  
  

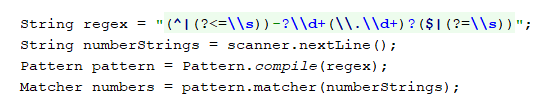
You can follow the table below to help with composing your RegEx:

|  |  |
| --- | --- |
| **Match ALL of these** | **Match NONE of these** |
| 1 -1 123 -123 123.456 -123.456 | 1s s2 s-s -1- \_55\_ s-2 s-3.5 s-1.1 |

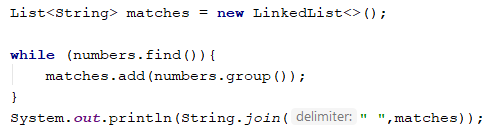
Find all the **numbers** from the string and **print them** on the **console**, separated by **spaces**.

### Implement the Solution in Java

Now that we’ve written our regular expression, we can start by putting it in a variable and extracting the matches:



After that, it’s only a matter of printing the numbers, separated by spaces:



### Examples

|  |  |
| --- | --- |
| **Input** | **Output** |
| 1 -1 1s 123 s-s -123 \_55\_ \_f 123.456 -123.456 s-1.1 s2 -1- zs-2 s-3.5 | 1 -1 123 -123 123.456 -123.456 |