TI 3001 C

Analítica de datos y herramientas de inteligencia artificial

Expresiones regulares

Tecnológico de Monterrey



¿Qué son las expresiones regulares?

En este tema vamos a ver como extraer información de texto usando las expresiones regulares.

Las expresiones regulares son:

- Son secuencias de caracteres que especifican un patrón de búsqueda.
- Son una fórmula para buscar dentro de texto ciertas coincidencias.
- Son patrones utilizados para encontrar una determinada combinación de caracteres dentro de una cadena de texto, gracias a ellos se puede extraer información importante.

¿Qué son las expresiones regulares?

Las **expresiones regulares** se usan para buscar, contar, reemplazar y / o validar ciertos patrones de texto.

Ejemplos:

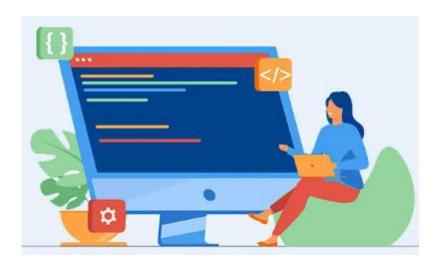
- Validar un correo electrónico, eliminar espacios dobles.
- Extraer información como correos, teléfonos, fechas, etc.



¿Cómo usar las expresiones regulares?

Para usar las expresiones regulares se necesita:

- 1. La expresión regular
- 2. El texto a manipular.



¿Qué pasa si quiero extraer de un directorio de 100 personas sus correos?

Los correos son distintos, cómo podríamos crear una fórmula para extraer todos los correos o modificar la información.

Laura Sanchez
442 153-32-42
lasanch@hotmail.com
http://www.laurasan.com

Pedro Lopez
448 342 33 12
pedrolo54@gmail.com
https://www.pedro.lopez.com.mx

Violeta Perez 443 214 43 72 violeta45@outlook.com violeta.net (*) Exp[resio]nes
regul[are]s

/[a-zA-Z0-9._-]+@[a-zA-Z0-9._-]+\.[a-zA-Z]+/g

¿Qué pasa si quiero extraer de un directorio de 100 personas sus correos?

Las expresiones regulares nos van a permitir crear fórmulas para extraer información. No solamente podemos extraer información de los correos, sino los teléfonos o las páginas web.

```
Carlos Arturo
449 123 45 67

carlos @hotmail.com
www.carlos.com

Manuel Alejandro
448-234-56-78

alejandro@outlook.com
https://www.manuel.alejandro.com.mx
http://alejandro.com.mx
```

Expresiones regulares

Las expresiones regulares no solamente nos sirven para extraer información, sino también son importantes para validar información. Por ejemplo: Un correo electrónico.

Página para validar expresiones regulares, me permite escribirlas y validar las coincidencias que encuentra:

regexr.com

Exp[resio]nes
regul[are]s

Módulo regex

El lenguaje de programación Python, en su librería estándar, nos proporciona el modulo regex el cual es utilizado para trabajar con expresiones regulares.

findall(): Encuentra todos los subtextos donde coincide la expresión regular y devuelve estas coincidencias como una lista. (patrón, texto)

Ejemplo:

```
import re
texto = input("Introduce un mensaje: ")
print(re.findall("is", texto)) # (Expresión regular, texto)

Shell *

>>>> %Run ER_Match.py
Introduce un mensaje: She is my sister Lis
['is', 'is', 'is']
```

¿Cómo funcionan las expresiones regulares?

Las expresiones regulares son una fórmula va a buscar dentro de nuestro texto coincidencias. Que es una coincidencia, que el texto de arriba, sea el mismo de abajo.

```
Carlos Arturo
449 123 45 67

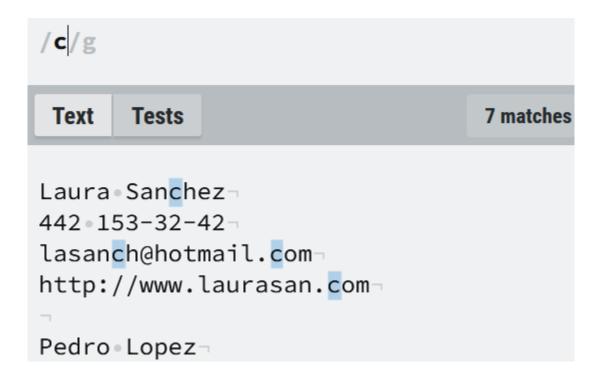
carlos @hotmail.com
www.carlos.com

Manuel Alejandro
448-234-56-78

alejandro@outlook.com
https://www.manuel.alejandro.com.mx
http://alejandro.com.mx
```

Expresiones regulares

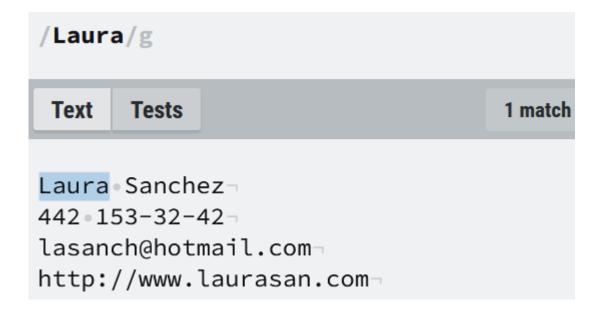
Por ejemplo, coloco la letra c. Lo que hace la expresión regular es buscar todas las coincidencias de la letra c. Encuentra 7 coincidencias.



Expresiones regulares

Por ejemplo, coloco Laura. Lo que hace la expresión regular es buscar la letra L seguida de la letra a, luego la letra u , r y a. Encuentra 1 coincidencia. Busca carácter por carácter.

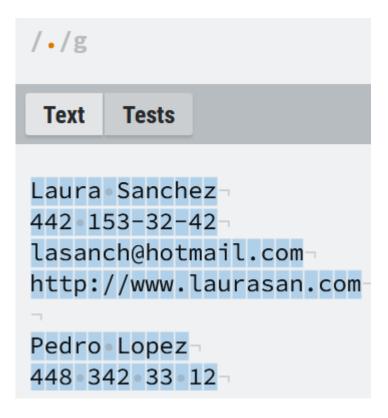
* g es una bandera, estamos haciendo una búsqueda global. Si le quitamos el global, solamente busca la primera coincidencia.



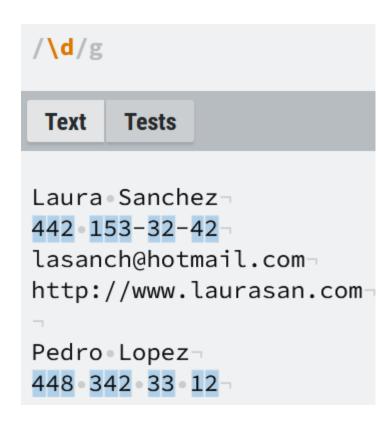
- Cualquier caracter excepto nueva línea.
- **\d** Dígitos (0-9)
- **\D** No dígitos (0-9)
- **\w** Caracter de palabra (a-z, A-Z, 0-9, _)
- **\W** No caracter de palabra
- **\s** Espacio en blanco (espacio, tab, nueva línea)
- **\S** No espacio en blanco (espacio, tab, nueva línea)
- \ Cancela caracteres especiales
- ^ Inicio de una cadena de caracteres (string)
- \$ Fin de una cadena de caracteres

Punto(.) Cualquier carácter excepto el salto de línea

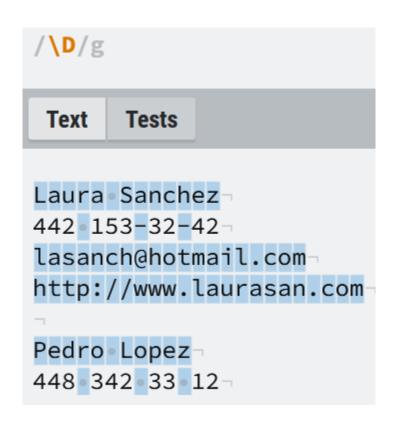
Encuentra todos los caracteres.



 (\d) Dígitos (0-9)

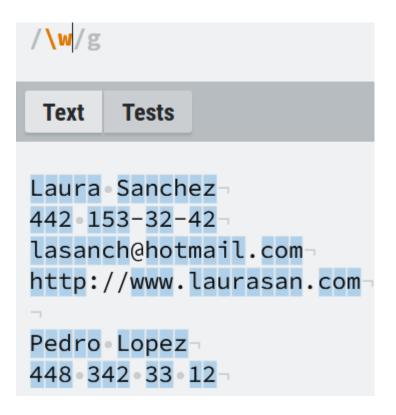


(\D) Todo lo que no sea un número



(\w) Caracter de palabra (a-z, A-Z, 0-9, _)

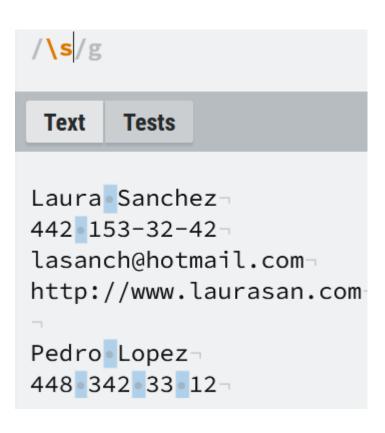
Busca de la a-z, A-Z, 0-9 y guion bajo.



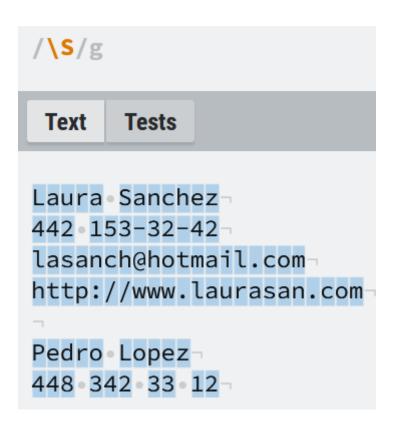
(\W) No es un caracter de palabra

```
/\W/g
Text
      Tests
Laura Sanchez-
442 153 32 42
lasanch@hotmail.com-
http://www.laurasan.com
Pedro Lopez-
448 342 33 12
```

(\s) Espacios, tabulaciones y nuevas líneas

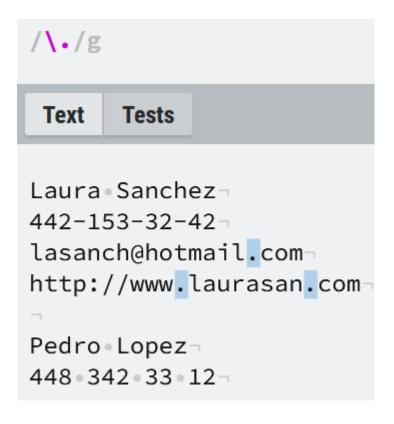


(\S) No espacio en blanco, tab y nueva línea



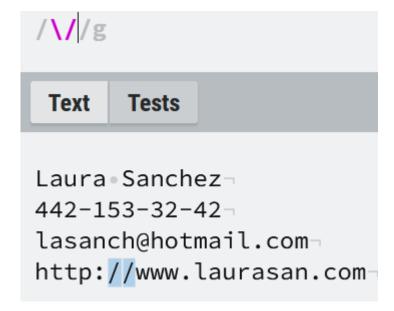
(\) Cancela caracteres especiales

Si quiero encontrar un punto en el texto. Cancela carácter especial punto.



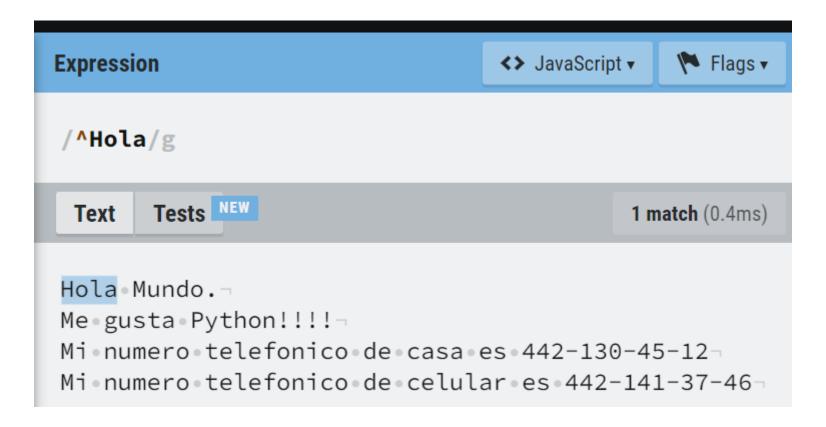
(\) Cancela caracteres especiales

- Si quiero encontrar una diagonal.
- Casi todos los símbolos le tenemos que poner backslash \,, ya que son metacaracteres o símbolos especiales. Ya que los símbolos significan algo.



(^) Inicio de una cadena de caracteres

• Si quiero que una línea empiece con la palabra "Hola"



(\$) Fin de una cadena de caracteres

• Si quiero que una línea termine con "Mundo."

```
Text Tests

Hola Mundo. 
Me gusta Python!!!!

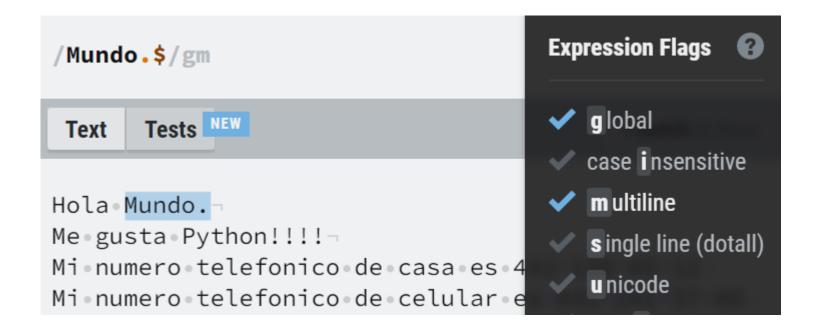
Mi numero telefonico de casa es 442-130-45-12

Mi numero telefonico de celular es 442-241-37-46

Mi numero telefonico de oficina es 442-380-14-22
```

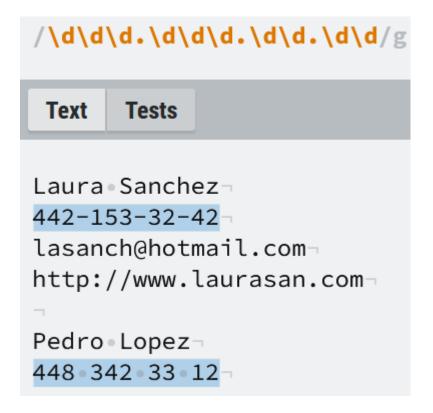
(\$) Fin de cadena de caracteres

- Si quiero que una línea termine con "Mundo."
- Agregar flag m para tomar texto multilínea. Que lea cada línea por separado.



Ejemplo: Extraer todos los números telefónicos.

Primero tres dígitos juntos, luego espacio o guion, otros tres dígitos, espacio o guion, luego dos y dos. El punto involucra cualquier carácter excepto salto de línea. \d\d\d\d\d\d\d\d\d\d\d\d\d\d\d\



Estos símbolos representan cuantas veces se repiten los caracteres

```
* 0 o más
```

- + 1 o más
- ? 0 o 1
- **Numero exacto**
- {n,} Numero n+
- {3,4} Rango de números (Mínimo, Máximo)

(*) Cero o más veces

La admiración se repita 0 o más veces

```
Text Tests NEW 1 match (0.5ms)

Hola • Mundo . ¬
Me • gusta • Python!!!! ¬
Mi • numero • telefonico • de • casa • es • 442-130-45-12¬
Mi • numero • telefonico • de • celular • es • 442-141-37-46¬
```

(+) Una o más veces

La admiración se repita 1 o más veces

```
Text Tests NEW 1 match (0.6ms)

Hola • Mundo. ¬
Me • gusta • Python!!!! ¬
Mi • numero • telefonico • de • casa • es • 442-130-45-12¬
Mi • numero • telefonico • de • celular • es • 442-141-37-46¬
```

(?) 0 o 1 vez

La admiración se repita 0 o 1 vez

```
Text Tests NEW 1 match (0.3ms)

Hola Mundo.

Me gusta Python!!!!

Mi numero telefonico de casa es 442-130-45-12

Mi numero telefonico de celular es 442-141-37-46
```

{n} Número n exacto

Busca la cantidad exacta de elementos. Va a identificar a la palabra **Python mas dos signos de admiración**.

```
/Python! {2}/gm

Text Tests NEW 1 match (0.5ms)

Hola Mundo. Me gusta Python!!!!

Mi numero telefonico de casa es 442-130-45-12

Mi numero telefonico de celular es 442-141-37-46
```

{n, } Número n o más elementos

Busca n o más elementos del caracter.

```
/Python!{2,}/gm

Text Tests NEW 1 match (0.5ms)

Hola Mundo. Me gusta Python!!!!

Mi numero telefonico de casa es 442-130-45-12

Mi numero telefonico de celular es 442-141-37-46
```

{min, max } Rango de números mínimo y máximo

Siempre va a ir por la mayor cantidad

```
/Python!{2,3}/gm

Text Tests NEW 1 match (0.4ms)

Hola Mundo. Me gusta Python!!!!

Mi numero telefonico de casa es 442-130-45-12

Mi numero telefonico de celular es 442-141-37-46
```

Ejemplo: Extraer los nombres del directorio con dos textos

\w Caracter, digito o guion bajo.

- + Uno o más caracteres
- **\s** Espacio
- ^ Al inicio de la línea
- \$ Al final de la línea

Al inicio de la cadena de texto, encuentres caracteres de palabra, uno o más, un espacio, caracteres de palabra, uno o más al final de la cadena de texto.

```
/^\w+\s\w+$/gim
Text
      Tests
Laura Sanchez
442-153-32-42
lasanch@hotmail.com¬
http://www.laurasan.com
Pedro Lopez
448 • 342 • 33 • 12
```

Ejemplo 2: Extraer los nombres del directorio con dos textos y espacio.

\w Caracter o guion bajo.

- + Uno o más caracteres
- \s Espacio, tabulador o salto de línea
- ^ Al inicio de la línea
- \$ Al final de la línea
- ? O o más

Al inicio de la cadena de texto, encuentres caracteres de palabra, uno o más, un espacio, caracteres de palabra, uno o más, un espacio cero o más veces al final de la cadena de texto.

```
/^\w+\s\w+\s?$/gim
 Text
      Tests
Laura Sanchez
442-153-32-42
lasanch@hotmail.com¬
http://www.laurasan.com
Pedro Lopez
448 • 342 • 33 • 12
```

Ejemplo 3: Extraer los nombres del directorio con dos textos o uno.

\w Caracter o guion bajo.

+ Uno o más caracteres

\s Espacio

- ^ Al inicio de la línea
- \$ Al final de la línea
- ? O o más

Al inicio de la cadena de texto, encuentres caracteres de palabra, uno o más, un espacio, caracteres de palabra, uno o más, un espacio cero o más al final de la cadena de texto.

\s?\w+? Opcional espacio y carácter de palabra.

```
/^\w+\s?\w+?\s?$/gim
 Text
       Tests
Laura Sanchez -
442-153-32-42
lasanch@hotmail.com¬
http://www.laurasan.com-
Pedro
448 • 342 • 33 • 12 ¬
pedrolo54@gmail.com¬
https://www.pedro.lopez.com.mx
Violeta Perez-
443 • 214 • 43 • 72 ¬
violeta45@outlook.com¬
violeta.net
```

```
/^\w+\s?\w+?\s?$/gim
    End. Matches the end of the string, or
    the end of a line if the multiline flag
    (m) is enabled.
442-153-32-42
lasanch@hotmail.com¬
http://www.laurasan.com-
Pedro
448 • 342 • 33 • 12
pedrolo54@gmail.com¬
https://www.pedro.lopez.com.mx-
Violeta Perez-
443 • 214 • 43 • 72
violeta45@outlook.com¬
violeta.net
```

- [] Encuentra caracteres en corchetes
- [^] Encuentra caracteres que no están dentro de corchetes
- Condicional O
- () Grupos

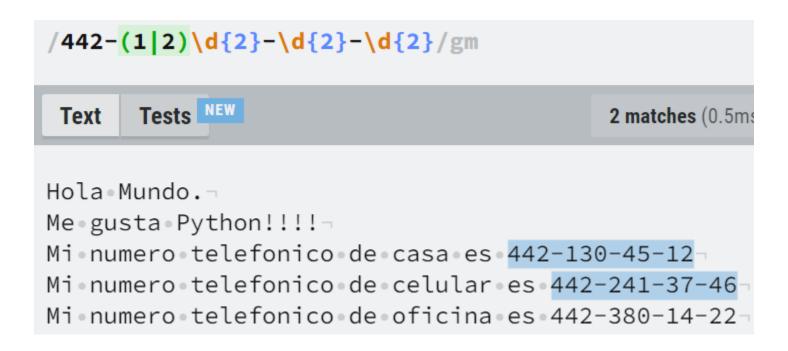
Grupos () Grupos

Ejemplo: Identificar números telefónicos y agrupar la lada.

```
/(d{3})-d{3}-d{2}-d{2}/gm
       Tests NEW
 Text
Hola∘Mundo.¬
Me gusta Python!!!!¬
Mi numero telefonico de casa es 442-130-45-12
Mi numero telefonico de celular es 442-141-37-46
                               match: 442-141-37-46
                               range: 113-125
Tools
                               group #1: 442
```

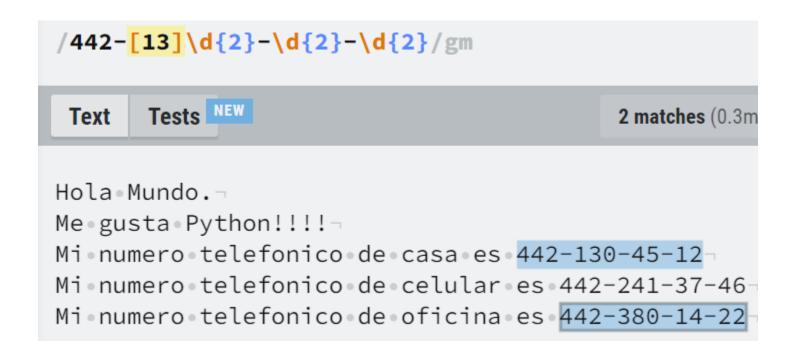
() Grupos

Ejemplo: Identificar los números telefónicos que comiencen con 1 o 2 después de la lada 442.



[] Encuentra caracteres en corchetes

Ejemplo: Identificar los números telefónicos que comiencen con 1 o 3 después de la lada 442.



[] Encuentra caracteres en corchetes

Ejemplo: Identificar las letras y dígitos.

```
Text Tests NEW 136 matches (0.5ms

Hola Mundo.

Me gusta Python!!!!

Mi numero telefonico de casa es 442-130-45-12

Mi numero telefonico de celular es 442-241-37-46

Mi numero telefonico de oficina es 442-380-14-22
```

[^] Encuentra caracteres no están dentro de corchetes

Ejemplo: Identificar todos los caracteres que no son letras minúsculas.

```
Text Tests NEW 77 matches (0.6ms

Hola Mundo.

Me gusta Python!!!!

Mi numero telefonico de casa es 442-130-45-12

Mi numero telefonico de celular es 442-241-37-46

Mi numero telefonico de oficina es 442-380-14-22
```

[^] Encuentra caracteres que no están dentro de corchetes

Ejemplo: Identificar todos los caracteres que no son dígitos.

```
Text Tests NEW 146 matches (0.5mm)

Hola Mundo.

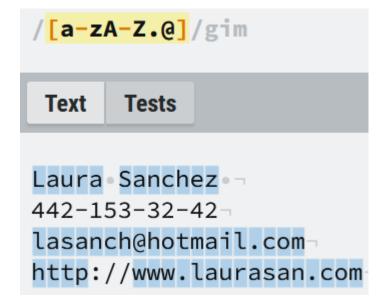
Me gusta Python!!!!

Mi numero telefonico de casa es 442-130-45-12

Mi numero telefonico de celular es 442-241-37-46

Mi numero telefonico de oficina es 442-380-14-22-
```

- [0-5]+ En los grupos **no** es necesario el uso del **slash ** para los metacaracteres como el punto.
- [a-zA-Z.@] Encuentra caracteres del abecedario en mayúsculas o minúsculas, puntos, arrobas.



 [^a-zA-Z.@] No caracteres del abecedario en mayúsculas o minúsculas, puntos, arrobas.

```
/[^a-zA-Z.@]/gim
Text
      Tests
Laura Sanchez -
442-153-32-42
lasanch@hotmail.com¬
http://www.laurasan.com-
Pedro Lopez-
448 342 33 12
pedrolo54@gmail.com¬
https://www.pedro.lopez.com.mx
```

() Nos permiten comparar entre algunos valores

Ejemplo 1: Extraer todos los números telefónicos con espacio o guion.

Primero tres dígitos juntos, luego espacio o guion, otros tres dígitos, espacio o guion, luego dos y dos.



Ejemplo 2: Extraer todos los números telefónicos haciendo uso de grupos y cuantificadores.

Tres dígitos, luego un espacio o guion, otros tres dígitos, espacio o guion, luego dos dígitos, espacio o guion y dos dígitos.

```
/\d{3}(\s|-)\d{3}(\s|-)\d{2}(\s|-)\d{2}/gim
 Text
       Tests
Laura • Sanchez • ¬
442-153-32-42
lasanch@hotmail.com¬
http://www.laurasan.com-
Pedro • Lopez ¬
448 • 342 • 33 • 12
pedrolo54@gmail.com¬
https://www.pedro.lopez.com.mx-
Violeta Perez-
443 214 43 72
violeta45@outlook.com¬
violeta.net
```

() Nos permiten comparar entre algunos valores

Ejemplo 3: Encuentre los número de teléfono con ladas 442, 443 y 448 solamente.

Conjunto de caracteres 2, 3 u 8.



```
Text Tests

Laura Sanchez

442 153 - 32 - 42

laura sanchez@hotmail - mx.com

http://www.laurasan.com

Pedro Lopez

448 342 33 12

pedro lopez 54@gmail.com

https://www.pedro.lopez.com.mx
```

Expresión regular

Ejercicio: Seleccionar correo electrónico.

[a-zA-Z] una o más letras.

```
/\w+@\w+\.com/gm
      Tests
Text
Laura Sanchez -
442-153-32-42
lasanch@hotmail.com-
http://www.laurasan.com-
Pedro Lopez
448 • 342 • 33 • 12
pedrolo54@gmail.com¬
https://www.pedro.lopez.com.mx
```

```
/\w+@\w+\.[a-zA-Z]+/gm
Text
       Tests
Laura Sanchez -
442-153-32-42
lasanch@hotmail.com¬
http://www.laurasan.com-
Pedro Lopez
448 • 342 • 33 • 12 ¬
pedrolo54@gmail.com¬
https://www.pedro.lopez.com.mx
```

Expresión regular

Ejercicio: Seleccionar correo electrónico.

