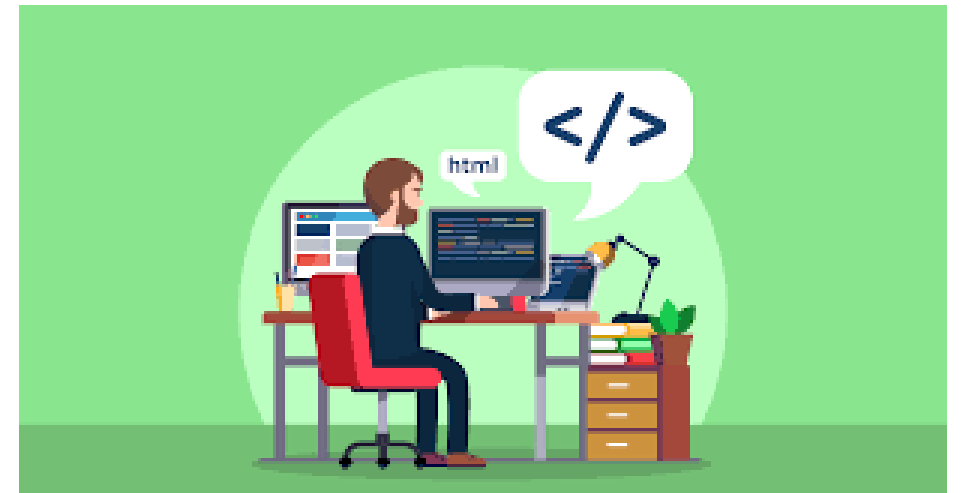


Busqueda y Recuperación de la Información

Dr. Cristian Muñoz Villalobos

HTML (HyperText Markup Language)

- HTML, o Lenguaje de Marcación de Hipertexto, es el lenguaje base de internet.
- Fue creado para ser entendido por seres humanos y máquinas, como por ejemplo el buscador de Google y otros sistemas que recorren la internet capturando información.
- La anotación utilizada por HTML son llamados de Tag.



HTML (HyperText Markup Language)

Que son los tag en HTML?

Son marcaciones que permiten indicar que informaciones serán exhibidas.

Por ejemplo:

- Un titulo importante marcamos con la tag H1:

```
<h1>Aqui va el texto del título</h1>
```

- Un párrafo es marcado con P así:

```
<p>Aqui va el texto del parrafo.  
Generalmente parrafos tienen muchas palabras,  
letras menores que las del título</p>
```

HTML (HyperText Markup Language)

Las paginas web mantienen una estructura base.

```
<!DOCTYPE html>  
<html lang="pt-br">
```

```
</html>
```

HTML (HyperText Markup Language)

Las paginas web mantienen una estructura base.

```
<!DOCTYPE html>  
<html lang="pt-br">  
<head>  
  
</head>  
  
<body>  
  
</body>  
</html>
```

HTML (HyperText Markup Language)

Las paginas web mantienen uma estrutura base.

```
<!DOCTYPE html>
<html lang="pt-br">
<head>
<meta charset="utf-8">
<title>Este es el titulo de la pagina</title>
</head>

<body>

</body>
</html>
```

HTML (HyperText Markup Language)

Las paginas web mantienen una estructura base.

```
<!DOCTYPE html>
<html lang="pt-br">
<head>
<meta charset="utf-8">
<title>Este es el titulo de la pagina</title>
</head>

<body>
<h1>Hola a todos</h1>
<p> HTML es un lenguaje facil de entender para
humanos y máquinas.!!</p>
</body>
</html>
```

Web Scrapping

Que sucede cuando Bob abre el navegador y escribe www.elperuano.pe/index.html ?



Web Scrapping

Que sucede cuando Bob abre el navegador y escribe www.elperuano.pe/index.html ?



Web Scrapping

Que sucede cuando Bob abre el navegador y escribe www.elperuano.pe/index.html ?



Web Scrapping

Que sucede cuando Bob abre el navegador y escribe www.elperuano.pe/index.html ?



Web Scrapping

Que sucede cuando Bob abre el navegador y escribe www.elperuano.pe/index.html ?



Web Scrapping

Que sucede cuando Bob abre el navegador y escribe www.elperuano.pe/index.html ?



Urllib

Como el protocolo HTTP es tan común, existe una biblioteca que hace todo el trabajo de socket para nosotros y hace que las paginas web parezcan un archivo.

```
from urllib.request import urlopen  
  
html = urlopen("http://www.pucp.edu.pe")  
text = html.read()
```

Urllib

Como el protocolo HTTP es tan común, existe una biblioteca que hace todo el trabajo de socket para nosotros y hace que las paginas web parezcan un archivo.

```
from urllib.request import urlopen

html = urlopen("http://www.pucp.edu.pe")
text = html.read()

print(text)

with open("index.html", "w+") as file:
    file.write(text)
```



BeautifulSoup

Es una biblioteca de Python para extraer datos de archivos HTML y XML. Permite navegar, buscar y modificar las estructuras con facilidad. Reduciendo horas de programación.

Back in 2004 most parsers could only parse well-formed XML and HTML. The poorly-formed stuff you saw on the Web was referred to as "tag soup", and only a web browser could parse it. BeautifulSoup started out as an HTML parser that would take tag soup and make it beautiful, or at least workable.



Análisis de HTML avanzado

Al preguntar a Michelangelo como el consiguió esculpir una obra de arte tan majestuosa como su David, dicen que el respondió: “Fácil, es solo sacar las partes de piedra que no se parecen a David.” (*It is easy, You just chip away the stone that doesn't look like David*)



Análisis de HTML avanzado

Busqueda:

- **find e findAll**

```
findAll(tag, attributes, recursive, text, limit, keywords)
```

```
find(tag, attributes, recursive, text, keywords)
```

Lidiar con hijos y decendientes:

- `children`

Lidiar padres:

- `parent`

Caso de Estúdio

Análisis de noticias del día del diario Gestión

1) Bajar un conjunto de noticias por fecha

<https://gestion.pe>

2) Analizar las noticias



Selenium

Es un framework (open-source) de testes automatizados de aplicaciones web.

Selenium WebDriver: Interacciona con la aplicación web por medio del browser (Chrome, Firefox, etc.).



Selenium

Busca los elementos por la tag

- `find_element_by_tag_name(tag)`

Obtiene um atributo:

- `get_attribute(att_name)`

Selenium

Descargar documentos relativos ao Poder Judicial.

1) Guardar la lista de urls de documentos PDF en la tabla:

https://www.pj.gob.pe/wps/wcm/connect/cortesuprema/s_cortes_suprema_home/as_inicio/as_enlaces_destacados/as_imagen_prensa/as_actualidad_importante

2) Descargar PDFs



Expresiones Regulares

Expresiones Regulares es una texto especial, en lenguaje formal, utilizada para extraer o buscar partes de un texto.

Ex: Como podemos procurar por?

cachorro

cachorros

Cachorro

Cachorros



Expresiones Regulares

Una de las operaciones principales en expresiones regulares es la **disyunción**, que permite escoger entre dos o mas caracteres.

La Expresion Regular ...	Encuentra ...
[cC]achorro	cachorro, Cachorro
[123456789]	Cualquier digito

Por ejemplo, si queremos encontrar la palabra cachorro o Cachorro, podemos representarla con la expresión [cC]achorro, donde el termino entre corchetes representa la disyunción del carácter. Algunas herramientas web de ayuda para explorar Regex:

- <https://regex101.com/>
- <https://regexr.com/>
- <https://unicode-table.com/pt>

Expresiones Regulares

Rango de términos [A-Z]:

Expresion Regular	Encuentra	Ejemplo
[A-Z]	una letra mayuscula	Cristian Muñoz
[a-z]	una letra minúscula	... Si queremos encontrar...
[0-9]	Um único dígito	Parte 1: Introducción

También se puede aplicar negación a una disyunción. Esto nos permite identificar cual carácter o rango de caracteres no deseamos encontrar en el texto.

Expresion Regular	Encuentra	Ejemplo
[^A-Z]	Una letra no mayúscula	Cristian Muñoz
[^0-9]	Que no un digito	
[x^y]	x^y	En la ecuación $z=x^y$

El acento circunflexo solo indica la negación solo si es el primer termino dentro de la disyunción, cualquier otro caso es tratado como carácter normal

Expresiones Regulares

Una barra vertical también representa una disyunción entre palabras. Por ejemplo, si queremos buscar las palabras “crear” y “cachorro”, podemos utilizar la expresión `crear|cachorro`.

Expresion Regular	Encuentra
<code>crear cachorro</code>	
<code>tuyo mio</code>	
<code>x y c</code>	
<code>[cC]rear [cC]achorro</code>	

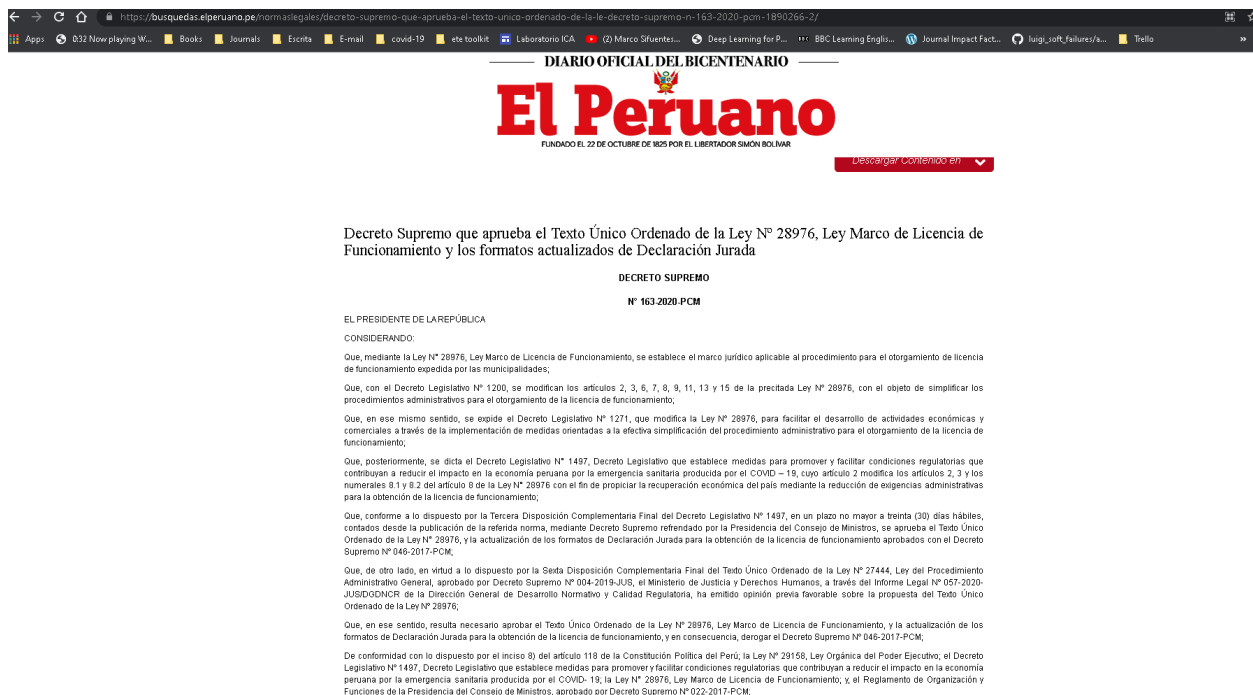
Expresiones Regulares

Finalmente existe un conjunto de caracteres especiales que son muy importantes en expresiones regulares (?,.,+,*)

Expresion Regular	Encuentra	Ejemplo
cachorros?	cachorro y cachorros	?: indica si el carácter es opcional
caball.o	caballero e caballo	.: representa cualquier caracter que pude ocupar esse espacio
k+	k,kkk,kkkk,...	+: el carácter predecesor aparece por lo menos 1 vez
Kk*	k,kkk,kkkk,....	*: El carácter predecesor puede aparecer em la expresión. Em cualquier cantidad, inclusive 0.

Extraer información de um decreto supremo

<https://busquedas.elperuano.pe/normaslegales/decreto-supremo-que-aprueba-el-texto-unico-ordenado-de-la-le-decreto-supremo-n-163-2020-pcm-1890266-2/>



The screenshot shows a web browser displaying the official website of 'El Peruano', the Diario Oficial del Bicentenario. The page features the site's logo at the top, followed by a navigation bar with a 'Descargar contenido en' button. The main content area is titled 'Decreto Supremo que aprueba el Texto Único Ordenado de la Ley N° 28976, Ley Marco de Licencia de Funcionamiento y los formatos actualizados de Declaración Jurada'. Below the title, the text 'DECRETO SUPREMO N° 163-2020-PCM' is displayed. The body of the document begins with 'EL PRESIDENTE DE LA REPÚBLICA CONSIDERANDO:' and lists several paragraphs detailing the legal basis and purpose of the decree, including references to the Ley Marco de Licencia de Funcionamiento and the Ley N° 28976. The text is presented in a clean, professional layout with clear headings and subheadings.

Decreto Supremo que aprueba el Texto Único Ordenado de la Ley N° 28976, Ley Marco de Licencia de Funcionamiento y los formatos actualizados de Declaración Jurada

DECRETO SUPREMO
N° 163-2020-PCM

EL PRESIDENTE DE LA REPÚBLICA
CONSIDERANDO:

Que, mediante la Ley N° 28976, Ley Marco de Licencia de Funcionamiento, se establece el marco jurídico aplicable al procedimiento para el otorgamiento de licencia de funcionamiento expedida por las municipalidades;

Que, con el Decreto Legislativo N° 1200, se modifican los artículos 2, 3, 6, 7, 8, 9, 11, 13 y 15 de la precitada Ley N° 28976, con el objeto de simplificar los procedimientos administrativos para el otorgamiento de la licencia de funcionamiento;

Que, en ese mismo sentido, se expide el Decreto Legislativo N° 1271, que modifica la Ley N° 28976, para facilitar el desarrollo de actividades económicas y comerciales a través de la implementación de medidas orientadas a la efectiva simplificación del procedimiento administrativo para el otorgamiento de la licencia de funcionamiento;

Que, posteriormente, se dicta el Decreto Legislativo N° 1497, Decreto Legislativo que establece medidas para promover y facilitar condiciones regulatorias que contribuyan a reducir el impacto en la economía peruana por la emergencia sanitaria producida por el COVID-19, cuyo artículo 2 modifica los artículos 2, 3 y los numerales 8.1 y 8.2 del artículo 8 de la Ley N° 28976 con el fin de propiciar la recuperación económica del país mediante la reducción de exigencias administrativas para la obtención de la licencia de funcionamiento;

Que, conforme a lo dispuesto por la Tercera Disposición Complementaria Final del Decreto Legislativo N° 1497, en un plazo no mayor a treinta (30) días hábiles, contados desde la publicación de la referida norma, mediante Decreto Supremo referendado por la Presidencia del Consejo de Ministros, se aprueba el Texto Único Ordenado de la Ley N° 28976, y la actualización de los formatos de Declaración Jurada para la obtención de la licencia de funcionamiento aprobados con el Decreto Supremo N° 046-2017-PCM;

Que, de otro lado, en virtud a lo dispuesto por la Sexta Disposición Complementaria Final del Texto Único Ordenado de la Ley N° 27444, Ley del Procedimiento Administrativo General, aprobado por Decreto Supremo N° 004-2019-JUS, el Ministerio de Justicia y Derechos Humanos, a través del Informe Legal N° 057-2020-JUS/DOCNOR de la Dirección General de Desarrollo Normativo y Calidad Regulatoria, ha emitido opinión previa favorable sobre la propuesta del Texto Único Ordenado de la Ley N° 28976;

Que, en ese sentido, resulta necesario aprobar el Texto Único Ordenado de la Ley N° 28976, Ley Marco de Licencia de Funcionamiento, y la actualización de los formatos de Declaración Jurada para la obtención de la licencia de funcionamiento, derogar el Decreto Supremo N° 046-2017-PCM;

De conformidad con lo dispuesto por el inciso 9) del artículo 119 de la Constitución Política del Perú; la Ley N° 29150, Ley Orgánica del Poder Ejecutivo; el Decreto Legislativo N° 1497, Decreto Legislativo que establece medidas para promover y facilitar condiciones regulatorias que contribuyan a reducir el impacto en la economía peruana por la emergencia sanitaria producida por el COVID-19; la Ley N° 28976, Ley Marco de Licencia de Funcionamiento; y, el Reglamento de Organización y Funciones de la Presidencia del Consejo de Ministros, aprobado por Decreto Supremo N° 022-2017-PCM;

Stemming / Lemmatization

- Por razones gramaticales, los documentos utilizan diferentes formas de una misma palabra como por ejemplo: organizar, organización, organizando. El objetivo de Stemming y Lemmatization es transformar palabras en una versión base. Por ejemplo:
- Stemming: corta al inicio o final de la palabra, toma en cuenta una lista de prefijos y sufijos.
- Lemmatization: Considera el análisis morfológica de la palabra. Para esto utiliza un diccionario.

