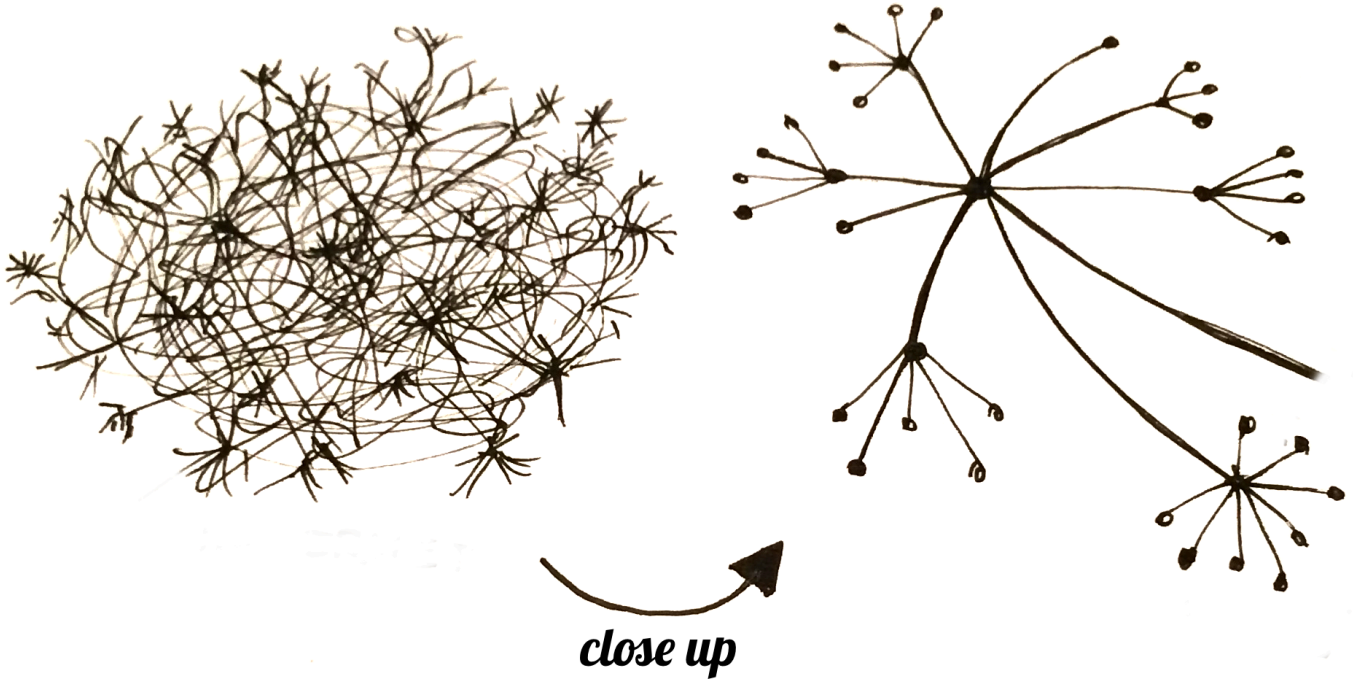
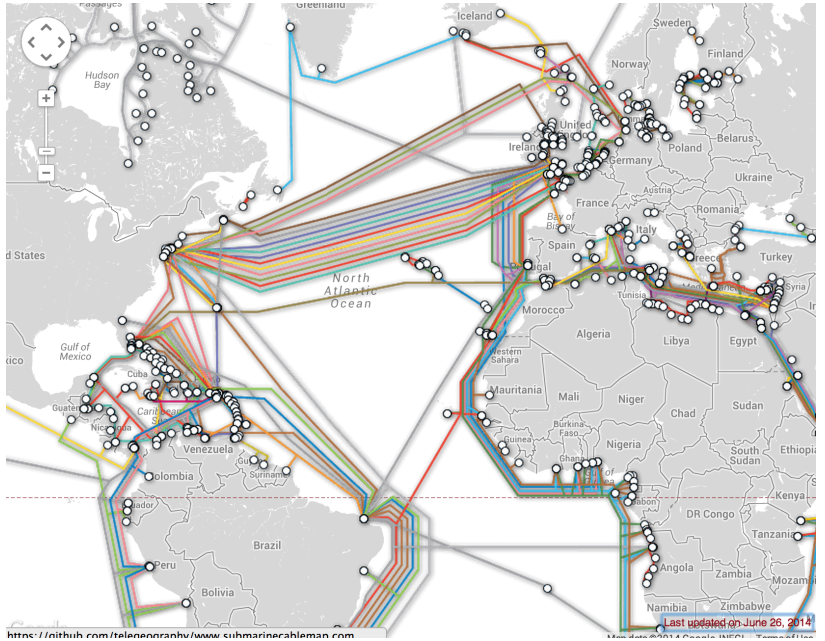
INTRO to WEB SCRAPING

What is Scraping?

- Web scraping is the process of automatically collecting information from the World Wide Web
- We all have done it somehow, but *copy & paste* is not so efficient..

How does the Internet work?





What is a link?

A **hyperlink**, or simply a **link**, is a reference to data that the reader can directly follow either by clicking, tapping, or hovering. A hyperlink points to a whole document or to a specific element within a document.

Why is it useful?

The New York Times

Thursday, October 11, 2018

Today's Paper

S.&P. 500	-0.40% ↓
Dow	-0.45% ↓
Nasdaq	+0.08% ↑

24°C
24° 8°
Budapest, Hungary

Hurricane Michael Cuts Path of Devastation Through Florida's Panhandle

- Hurricane Michael landed as the most powerful storm to hit the continental United States in decades, turning homes into splinters of lumber.
- At least two people were killed, and the authorities feared they would find more bodies in the rubble in a vast search-and-rescue operation.

2m ago

Wind and rain roared through the region, flooding homes and sending debris flying at dangerous speeds.
Eric Thayer for The New York Times

The storm was on a path to strike cities in the Carolinas that have yet to fully recover from Hurricane Florence.

1h ago

Why a Storm Death in One State Might Not Count in Another

Inconsistencies in classifying deaths have made it difficult to compare the magnitude of different disasters.

58m ago

Hurricane Michael Intensified Quickly, Surprising Many

The storm's increase in strength was due in part to its low barometric pressure and warmer-than-average waters in the Gulf of Mexico.

2h ago

We're tracking the path of the storm and its impact.

https://www.nytimes.com/es/?URI=https%3A%2F%2Fwww.nytimes.com%2F

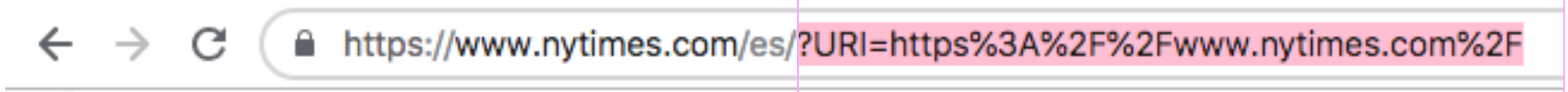
SECCIONES INICIO

EDICIÓN ESPAÑOL

The New York Times

jueves, 11 de octubre de 2018 | 24°C

NOTICIAS CULTURA OPINIÓN AMÉRICA LATINA ★ REPOSADO



What is that?

https://www.nytimes.com/es/?URI=https%3A%2F%2Fwww.nytimes.com%2F

https://www.nytimes.com/es/

It redirects you to the same place…

SECCIONES INICIO

EDICIÓN ESPAÑOL

The New York Times

jueves, 11 de octubre de 2018 | 24°C

NOTICIAS CULTURA OPINIÓN AMÉRICA LATINA ★ REPOSADO

JOÃO FAZENDA

COMENTARIO

El mundo (post)mundial: la muerte del gol

El campo de fútbol era un lugar de certezas. Ya no: a consecuencia del videoarbitraje, cuando una pelota entra a un arco no es seguro que sea una anotación. Se trata de un cambio radical en el deporte que está

TOM JAMIESON PARA THE NEW YORK TIMES

LIBROS

Las nuevas autoras de la distopía feminista

Un creciente canon de ciencia ficción que se centra en la mujer analiza temas como la desigualdad de género, la misoginia y el sexismo institucionalizado. Naomi Alderman, Sophie Mackintosh y Christina Dalcher son algunas de las nuevas voces de este movimiento literario.

Por ALEXANDRA ALTER

NEGOCIOS

La guerra contra las noticias falsas en Facebook desborda a los verificadores

Para acabar con las noticias falsas en Filipinas, la red social social se asoció con Rappler, un medio local de entretenimiento e investigación. Pero la desinformación surge rápido y ha resultado muy difícil de eliminar. Muchos señalan que Facebook no ofrece el apoyo necesario.

Por ALEXANDRA STEVENSON

PERÚ

La detención de Keiko Fujimori, un nuevo golpe al fujimorismo

https://www.nytimes.com/es/category/opinion/

SECCIONES INICIO

The New York Times ES

Opinión

JOÃO FAZENDA

COMENTARIO

El mundo (post)mundial: la muerte del gol

El campo de fútbol era un lugar de certezas. Ya no: a consecuencia del videoarbitraje, cuando una pelota entra a un arco no es seguro que sea una anotación. Se trata de un cambio radical en el deporte que está sustituyendo la emoción por una forma torpe de justicia.

Por MARTÍN CAPARRÓS

ILUSTRACIÓN DE MIKEL JASO; FOTOGRAFÍAS DE ZOOM-ZOOM Y MONTICELLO, VÍA GETTY IMAGES

COMENTARIO

Cómo evitar la turistificación de Madrid

El turismo tiene dos rostros: puede contribuir al crecimiento económico de las ciudades, pero también puede desbordarlas, encarecer zonas para sus habitantes y generar condiciones precarias de trabajos temporales. La capital de España corre el riesgo de sufrir las consecuencias de la turistificación, pero es un problema que puede solucionarse.

PERISCOPIO ELECTORAL

Ni Haddad ni Bolsonaro tienen lo necesario para evitar una crisis en Brasil

La economía más grande de América Latina tiene muy poco que festejar. El país necesita reformas difíciles de implementar en una economía frágil y con una política desacreditada. Ni el candidato ultraderechista ni el izquierdista están capacitados para lidiar con una crisis económica inminente.

Por MONICA DE BOLLE

COMENTARIO

La aristocracia del fraude que dirige Estados Unidos

Hasta hace poco, se pensaba que la evasión fiscal no era tan generalizada. Luego llegó

Chrome File Edit View History Bookmarks People Window Help

Always Show Bookmarks Bar ⇧⌘B
Always Show Toolbar in Full Screen ⇧⌘F

Stop ⌘.
Force Reload This Page ⇧⌘R

Enter Full Screen ^⌘F
Actual Size ⌘0
Zoom In ⌘+
Zoom Out ⌘−

Cast...

Developer ▶

View Source ⌥⌘U
Developer Tools ⌥⌘I
JavaScript Console ⌥⌘J
Allow JavaScript from Apple Events

Opinión – Espa

https://w

SECCIONES

INICIO

The New York Times ES

Op

Chrome File Edit View History Bookmarks People Window Help

Always Show Bookmarks Bar ⇧⌘B
Always Show Toolbar in Full Screen ⇧⌘F

Stop ⌘.
Force Reload This Page ⇧⌘R

Enter Full Screen ^⌘F
Actual Size ⌘0
Zoom In ⌘+
Zoom Out ⌘−

Cast...

Developer ▶

View Source ⌥⌘U
Developer Tools ⌥⌘I
JavaScript Console ⌥⌘J
Allow JavaScript from Apple Events

Opinión – Espa

https://w

SECCIONES

INICIO

The New York Times ES

Op

Opinión – Español

view-source:https://www.nytim

view-source:https://www.nytimes.com/es/category/opinion/

```
<!DOCTYPE html>
<!--[if (gt IE 9)|!(IE)]> <!--> <html lang="es" class="tone-opinion no-js" itemscope xmlns:og="http://ogp.me/ns#"> <!--<![endif]-->
<!--[if IE 9]> <html lang="es" class="tone-opinion no-js ie9 lt-ie10" xmlns:og="http://ogp.me/ns#"> <![endif]-->
<!--[if IE 8]> <html lang="es" class="tone-opinion no-js ie8 lt-ie10 lt-ie9" xmlns:og="http://ogp.me/ns#"> <![endif]-->
<!--[if (lt IE 8)]> <html lang="es" class="tone-opinion no-js lt-ie10 lt-ie9 lt-ie8" xmlns:og="http://ogp.me/ns#"> <![endif]-->
<head>
<!-- APP STATS: {"NYT\\Media\\Controller\\Shortcodes":{"cacheHits":40},"NYT\\Database\\SQLLogger":{"queries":4}} -->
<meta charset="utf-8"/>
<meta http-equiv="X-UA-Compatible" content="IE=edge,chrome=1" />
<meta name="viewport" content="width=device-width, initial-scale=1, maximum-scale=1" />
<link rel="shortcut icon" href="https://static01.nyt.com/favicon.ico" />
<link rel="apple-touch-icon-precomposed" sizes="144×144" href="https://static01.nyt.com/images/icons/ios-ipad-144x144.png" />
<link rel="apple-touch-icon-precomposed" sizes="114×114" href="https://static01.nyt.com/images/icons/ios-iphone-114x144.png" />
<link rel="apple-touch-icon-precomposed" href="https://static01.nyt.com/images/icons/ios-default-homescreen-57x57.png" />
<meta name="sourceApp" content="nyt-international" />
<meta property="fb:pages" content="751965821517958" />
<meta id="applicationName" name="applicationName" content="nyt-espanol" />
<meta id="foundation-build-id" name="foundation-build-id" content="" />
<meta itemprop="inLanguage" content="es-MX" />
<meta property="collection:language" content="es">
<meta name="msapplication-starturl" content="https://www.nytimes.com" />
<meta property="fb:app_id" content="9869919170" />
<meta name="google-site-verification" content="z-X9yT4qiS9XgWx9_wZQn3CD8uOup_UnqyfwVIP-jt8" />
<link rel="manifest" href="https://www.nytimes.com/es/wp-json/nyt/v1/app-manifest/">
<meta name="collection:section" content="opinion" />

<meta property="og:description" content="Editoriales y columnas de opinión de The New York Times en Español." />
<link rel="stylesheet" type="text/css" media="screen" href="https://storage.googleapis.com/nyt-intl/css/prd/shell-9a7293a2ce.min.css" />


<link rel="stylesheet" type="text/css" media="screen" href="https://storage.googleapis.com/nyt-intl/css/prd/collection-03f04b657c.min.css" />

<link rel="stylesheet" type="text/css" media="screen" href="https://storage.googleapis.com/nyt-intl/css/prd/theme-d34f8118bb.min.css" />
<script>
window._nyt = {"locale":"es_MX","mailToLink":{"path":"\/es","message":"De The New York Times en Espa\u00f1ol:"},"dataEdition":"es","dataEditionCookie":"spanish","editic
{"heading":"Edici\u00f3n","links":{"us":{"url":"https:\/\/www.nytimes.com","label":"English","slug":"us","data-edition":"us"},"chinese":
{"url":"http:\/\/cn.nytimes.com\/","label":"\u4e2d\u6587 (Chinese)","slug":"chinese","data-edition":"chinese"},"spanish":
{"url":"https:\/\/www.nytimes.com\/es\/","label":"Espa\u00f1ol","slug":"spanish","data-edition":"es"}}},"helpMenu":{"heading":"Ayuda","links":[{"label":"Preguntas
Frecuentes","url":"\/es\/preguntas-frecuentes\/"},{"label":"Cont\u00e1ctanos ayuda@nytimes.com","url":"mailto:ayuda@nytimes.com?
subject=Comentario%20sobre%20el%20New%20York%20Times"}]},"helpNavItem":{"address":"ayuda@nytimes.com","subject":"Comentario sobre el New York
Times"},"collectionsEndpoint":"https:\/\/www.nytimes.com\/es\/wp-
json\/nyt\/v1\/collections\/?","tagxHost":"https:\/\/tagx.nytimes.com","dfpProp":"esnyt","dfpHost":"\/29390238\/esnyt\/","emailSubscriberEndpoint":"https:\/\/www.nytime
/int\/email-subscriber\/subscribe","latestViewCollectionPage":2};
</script>
<meta name="nyt-collection:identifier" content="nyt-es">
<meta name="nyt-collection:type" content="sectioncollection">
<meta name="nyt-collection:url" content="https://www.nytimes.com/es/category/opinion/">
<meta name="nyt-collection:uri" content="category_name=opinion">
```

What is HTML?

- HTML stands for "HyperText Markup Language".
- HyperText means it's a type of text that supports hyperlinks between pages.
- Markup means we have taken a document and marked it up with code to tell something (in this case, a browser) how to interpret the page.
- HTML code is built with tags, each one starting with < and ending with >. These tags represent markup elements.

Most important HTML tags

Tag	Description
<!DOCTYPE>	Defines the document type
<html>	Defines an HTML document
<head>	Defines information about the document
<title>	Defines a title for the document
<body>	Defines the document's body
<h1> to <h6>	Defines HTML headings
<p>	Defines a paragraph
 	Inserts a single line break
<hr>	Defines a thematic change in the content
<!--...-->	Defines a comment

Styles and Semantics

Tag	Description
<style>	Defines style information for a document
	Defines a section in a document
	Defines a section in a document
<header>	Defines a header for a document or section
<footer>	Defines a footer for a document or section
<main>	Specifies the main content of a document
<section>	Defines a section in a document
<article>	Defines an article
<aside>	Defines content aside from the page content
<details>	Defines additional details that the user can view or hide
<dialog>	Defines a dialog box or window
<summary>	Defines a visible heading for a <details> element
<data>	Links the given content with a machine-readable translation

More styling elements here

Tag helps to navigate through the website, and that is what we use when we decide which part of the html file we would like to store. Let’s see an example!

More about scraping..

https://github.com/velf/Data_Analysis_with_Python/blob/master/Lecture2/
https://tutorial.djangogirls.org/en/how_the_internet_works/
https://www.analyticsvidhya.com/blog/2015/10/beginner-guide-web-scraping-beautiful-soup-python/
https://towardsdatascience.com/an-introduction-to-web-scraping-with-python-bc9563fe8860