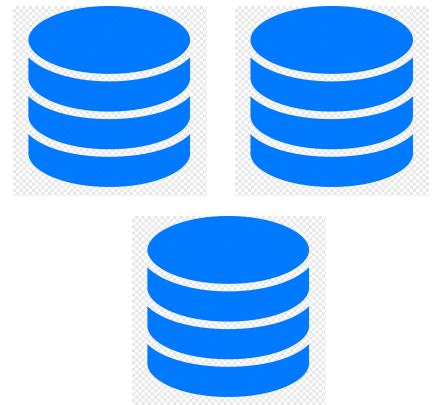
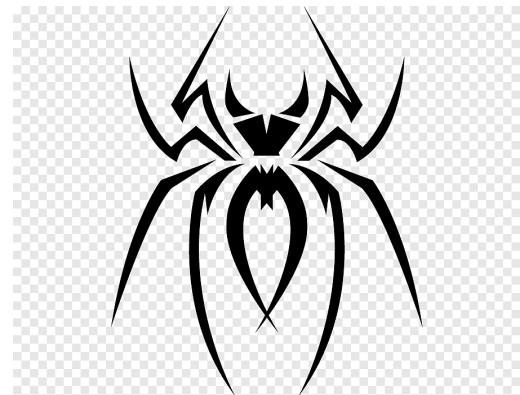
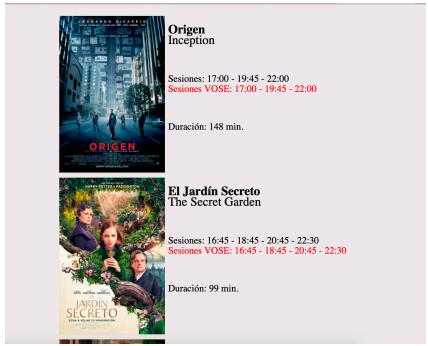


Lower-case semantic web

Datos estructurados embebidos en HTML...

Microdatos, JSON-LD...

Estrategia de scrapping



Páginas en la Web (Sitio)
Datos sin estructura

Tecnologías de scrapping

Datos estructurados.
Modelo de datos

Analizar las páginas web

The screenshot shows a web browser window with the following details:

- Menu Bar:** Chrome, Archivo, Editar, Ver, Historial, Favoritos, Personas, Pestaña, Ventana, Ayuda.
- Developer Tools:** The "Ver" menu is open, showing options like "Mostrar siempre la barra de favoritos" and "Herramientas del programador".
- Page Content:** A movie theater website for "Cines Broadway". The main navigation includes "Inicio", "Broadway", and "Estrenos". The "Estrenos" section features a movie poster for "Origen (Inception, 2010) Película" with session times: "Sesiones: 17:00 - 19:45 - 22:00" and "Sesiones VOSE: 17:00 - 19:45 - 22:00". It also lists the duration as "Duración: 148 min".
- Elements Panel:** Shows the DOM structure of the page, focusing on the "Ficha Técnica" section which contains the movie's technical details: País: Estados Unidos, Año: 2010, Duración: 148 min.
- Console Panel:** Displays a message about new issues and accessibility information.

Alternativas con python

- Dos bibliotecas básicas
 - Request (<https://requests.readthedocs.io/en/master/>)
 - hace sencilla la manipulacion del protocolo HTTP.
 - Utilizaremos get.
 - BeautifulSoup (<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>)
 - Permite manipular en forma sencilla documentos html y xml. Entre las funcionalides estan:
 - Buscar un elemento.
 - Buscar todos los elementos que corresponden con un tag.

Desafíos enfrentados en el TP01

- Encontrar la url de la página donde está la información del recurso que necesitamos (suponiendo que está todo en una)
- Por cada atributo de nuestro recurso:
 - Encontrar el nodo del dom donde está el atributo que necesitamos
 - Extraer el texto del atributo del contenido del nodo (suponiendo que contienen algo más que lo que nosotros necesitamos)
 - Normalizar/convertir el valor del atributo a lo que necesitamos (fecha, booleano, numérico)
- Por cada recurso
 - Definir un identificador único y/o una función de identidad.
 - Agregarlo si es nuevo. Si ya lo tenemos, no repetirlo sino actualizarlo/completarlo

Desafíos latentes

- Es necesario actualizar el scrapper cada vez que la página cambie (poco robusto)
- El código resultante es poco claro (incluso cuando hago el mejor esfuerzo) y difícil de mantener. Al menos requiere analizar en paralelo el código y el html.
- Debemos escribir un scrapper por cada sitio en el que hay datos que nos interesan.
- El código escrito no es reutilizable en otros sitios o otras páginas del mismo sitio con diferente estructura.
- Es imposible automatizarlo. Escribir los scrappers requiere que un humano interprete el contexto en el que se encuentra un texto para definir su semántica

“The web has made people smarter. We need to understand how to use it to make machines smarter, too.”

-- Michael I. Jordan, paraphrased from a talk at
AAAI, July 2002 by Michael Jordan (UC Berkeley)

“The multi-agent systems paradigm and the web both emerged around 1990. One has succeeded beyond imagination and the other has not yet made it out of the lab.”

-- Anonymous, 2001

Dificultades de las máquinas para entender la web

- La web actual representa la información a través de:
 - Lenguaje natural (español, inglés, chino, ...)
 - Gráficos, multimedia, formatos de páginas, ...
- Nosotros podemos procesar esta información fácilmente
 - Podemos deducir nueva información a partir de información parcial
 - Podemos crear asociaciones mentales
 - Incluso si se usa diferente terminología (Metro, subterráneo, underground)
- En cambio, para las computadoras procesar esta información requiere “combinar” datos en la Web
 - búsquedas sobre diferentes repositorios de conocimientos
 - Buscar una casa para alquilar cerca de la red de metro
 - La información de casas en alquiler y red de metro proviene de diferentes sitios

Consultas en la web Sintáctica (1)

Quiero información
sobre Woody Allen



Consultas en la web Sintáctica (2)

paris - Busqueda Google de imágenes - Mozilla Firefox

Archivo Editar Ver Historial Marcadores Herramientas Ayuda del.icio.us

G paris - Busqueda Google de imág... Acceder

La Web Imágenes Maps Noticias Video Gmail Más ▾

Google™ paris Buscar imágenes

Imágenes Mostrando: Imágenes grandes

Se muestran sólo las imágenes grandes (mostrar imágenes de todos)

las fotos identifican por su contexto (tags, fileNames,...) no por su contenido

 Wallpaper París 1280 x 960 - 550 KB - jpg www.quadricula.com	 Paris 2 1656 x 500 - 197 KB - jpg www.ganso.org	 Plano París · plano paris 1120 x 828 - 331 KB - jpg europaestuya.wordpress.com	 París: La Catedral de Notre-Dame de ... 1024 x 1310 - 321 KB - jpg www.arikah.net	 Paris Hotel at Night 1024 x 768 - 264 KB - jpg www.mundo-descargas.com
 Por fin hacia París 1024 x 768 - 59 KB - jpg historiasflyter.blogspot.com	 Paris Hilton la actriz, cantante, ... 1024 x 768 - 132 KB - jpg img405.imageshack.us	 ... situación con el Club de París, ... 800 x 1200 - 126 KB - jpg elciudadanoargentino.blogspot.com	 Paris- Francia: Base de la Torre ... 1024 x 768 - 139 KB - jpg www.escartinlam.com	 Foto de París- Tronco de la ... 1024 x 768 - 135 KB - jpg www.escartinlam.com
<p>Listo</p>				

Web mas fácil de entender para las máquinas

- ¿por qué a alguien le interesaría que sus datos se extraigan y consuman con facilidad?
 - ¿El Gourmet TV?
 - ¿Universal Studios / Disney Channel?
 - ¿Ministerio de transporte?
 - ¿Nikon, Samsung, ... (camaras, teléfonos)?
 - BBC, History Channel, ...



¿y con un incentivo?

[Little Water Cantina - Eastlake - Seattle, WA](#)
www.yelp.com › Restaurants › Mexican
★★★★★ 90 reviews - Price range: \$\$
90 Reviews of Little Water Cantina "Three things are on my list when I eat out: great food, atmosphere, and...
[Vegetarian Vegan Pizza No Cheese\) Recipe - Food.com - 248865](#)
www.food.com/recipe/vegetarian-vegan-pizza-no-c...
★★★★★ 2 reviews - 1 hr 32 mins - 242.9 cal
Aug 26, 2007 - This is from my dad, who developed some **vegan recipes** that don't have any cheese, and you

[Leonard Cohen – Free listening, videos, concerts, stats, & pictures at...
www.last.fm/music/Leonard+Cohen
Watch videos & listen to Leonard Cohen: Suzanne, Hallelujah & more, plus 132 pictures. Leonard Cohen, \(born September 21, 1934 in Montréal, Quebec, ...
Track Duration
Suzanne 03:48
The Darkness 04:29
Going Home 03:51
Hallelujah 06:12](#)

Google batman v superman: dawn of justice

Todos Imágenes Noticias Videos Maps Más ▾ Herramientas de búsqueda

Películas que se proyectan cerca de New York, NY

Batman v Superman: Dawn of Justice Miracles from Heaven

The Divergent Series: Allegiant 10 Cloverfield Lane

Zootopia Deadpool

My Big Fat Greek Wedding 2 God's Not Dead 2

Horarios de Batman v Superman: Dawn of Ju...

lun., 28 mar. mar., 29 mar. mié., 30 mar. jue., 31 mar.

Cualquier hora del día Mañana Tarde Anochecer Por la noche

AMC Loews Kips Bay 15 - [Mapa](#)

Microformats (en desuso)

- Define sus propios vocabularios
 - hCalendar, hAtom, hCard, hRecipe
 - Una comunidad (vocabularios cerrados)
- Utiliza las propiedades class y rel de los tags HTML/XHTML (no agrega tags ni propiedades nuevos)

```
<p class="vevent">
  The <span class="summary">English Wikipedia was launched</span>
  on 15 January 2001 with a party from
  <abbr class="dtstart" title="2001-01-15T14:00:00+06:00">2pm</abbr>-
  <abbr class="dtend" title="2001-01-15T16:00:00+06:00">4pm</abbr> at
  <span class="location">Jimmy Wales' house</span>
  (<a class="url" href="http://en.wikipedia.org/wiki/History_of_Wikipedia">more information</a>)
</p>
```

Schema.org

- 2011 : Esfuerzo colaborativo entre los grandes buscadores: Bing, Google, Yahoo, Yandex
- Definen vocabulario (Schema.org)
 - Un único lugar donde ir a buscar los vocabularios (control de los mismos) aunque inspirado en otras iniciativas (FOAF, SKOS, Microformats)
- Define un mecanismo para embeber esos vocabularios en las páginas (RDFa, Microdatos, y recientemente JSON-LD).

Schema.org

Browse the full hierarchy:

- [One page per type](#)
- [Full list of types, shown on one page](#)

Or you can jump directly to a commonly used type:

- Creative works: [CreativeWork](#), [Book](#), [Movie](#), [MusicRecording](#), [Recipe](#), [TVSeries](#) ...
- Embedded non-text objects: [AudioObject](#), [ImageObject](#), [VideoObject](#)
- [Event](#)
- [Health and medical types](#): notes on the health and medical types under [MedicalEntity](#).
- [Organization](#)
- [Person](#)
- [Place](#), [LocalBusiness](#), [Restaurant](#) ...
- [Product](#), [Offer](#), [AggregateOffer](#)
- [Review](#), [AggregateRating](#)
- [Action](#)

Microdatos (Microdata)

Agrega propiedades nuevas a los tags HTML

- **itemscope** – crea un item e indica que los descendientes de ese elemento contienen información al respecto
- **itemtype** – url de un elemento en el vocabulario schema.org
- **Itemprop** - indica que el elemento incluye el valor de una propiedad (también del vocabulario)

Microdatos (Microdata)

```
<section itemscope itemtype="http://schema.org/Person">
  Hello, my name is
  <span itemprop="name">John Doe</span>,
  I am a
  <span itemprop="jobTitle">graduate research assistant</span>
  at the
  <span itemprop="affiliation">University of Dreams</span>.
  My friends call me
  <span itemprop="additionalName">Johnny</span>.
  You can visit my homepage at
  <a href="http://www.JohnnyD.com" itemprop="url">www.JohnnyD.com</a>.
  <section itemprop="address" itemscope itemtype="http://schema.org/PostalAddress">
    I live at
    <span itemprop="streetAddress">1234 Peach Drive</span>,
    <span itemprop="addressLocality">Warner Robins</span>,
    <span itemprop="addressRegion">Georgia</span>.
  </section>
</section>
```

Algunos sitios que usan microdatos

- <https://www.telegraph.co.uk/>
- <https://www.ebay.com>

JSON-LD

- JSON (JavaScript Object Notation) es un formato cada vez más popular para publicar/almacenar información (con una estructura orientada a objetos)
- Por ahora no vamos a pensar en el LD, pero deriva de Linked Data
- Propone utilizar los vocabularios de Schema.org (aunque podrían ser otros), para incrustar datos semánticos en páginas web
 - en un

```
<script type="application/ld+json"></script>
```

en el encabezado.
 - Utiliza dos propiedades especiales @context y @type

JSON-LD

```
1  {
2      "@context": "http://www.schema.org",
3      "@type": "AutoDealer",
4      "name": "Fangio motors",
5      "url": "https://www.fangiomotors.com/",
6      "logo": "https://www.fangiomotors.com/wp-content/uploads/2017/02/logo-e1486411126955.jpg",
7      "description": "Fangio Motors specializes in pre-owned classic, Luxury and sport cars.",
8      "address": {
9          "@type": "PostalAddress",
10         "streetAddress": "1426 S.W 12 AVE",
11         "addressLocality": " POMPANO BEACH",
12         "addressRegion": "Florida",
13         "postalCode": " 33069",
14         "addressCountry": "United States"
15     },
16     "openingHours": "Mo 07:30-21:30",
17     "contactPoint": {
18         "@type": "ContactPoint",
19         "telephone": "305-322-4849 "
20     }
21 }
```

Algunos sitios que usan JSON-LD

Son cada vez más !

- <https://www.bestbuy.com>
- <https://www.imdb.com>
- <https://www.rottentomatoes.com>
- <https://www.fandango.lat>
- <https://www.metacritic.com>
- ...

Para probar...

- <http://schema-creator.org> (ver rápidamente como se puede combinar JSON-LD y html5 para Clases populares de Schema.org)
- <https://developers.google.com/structured-data/testing-tool/> (validador de google – extractor)
- <http://www.google.com/webmasters/tools/richsnippets> (demostrador de Rich Snippets de Google Search)

TP2

- TP1 pero ahora para sitios con Microdatos
 - IMDB
 - Fandango (programas de cines)
 - RottenTomatoes (críticas)
 - Metacritic (críticas)
 - ...
- ¿qué desafíos hace mas fáciles de atacar?
- ¿qué desafíos quedan pendientes?