Preguntas

# ¿Cuán complejos son, en comparación con los scrappers analizamos en clase, tener este tipo de representación de la información? Piense en cuan estandarizados esta la representación de los datos.

El método de extracción por JSON-LD tiene la gran ventaja de tener un overhead mucho más chico que utilizar un web scrapper, ay que solo consiste en buscar un tag dentro de la pagina para acceder a este.

En el caso del web scrappers se necesita proveerle a este mucho mas contexto sobre como identificar la información relevante dentro de un sitio, esto es un gran problema cuando se busca extraer información de internet de manera masiva, debido al tiempo que hay que invertir configurando dichos scrappers y revisando la información recolectada.

La estandarización que brinda JSON-LD y schema.org facilita mucho la recolección de información, buscando terminología estándar y formatos para ordenar los datos de una manera mucho más fácil y rápida de entender y procesar.

# ¿Cuáles cree que son las ventajas y desventajas de utilizar JSON-LD?

Ayuda a localizar dentro del código de la página un resumen de la información que describe. Esto ahorra mucho tiempo en la recolección de datos y el análisis del contenido de la página.

Como desventaja tenemos que el JSON-LD no contiene toda la información que se renderiza de la página. Como ejemplo el puntaje de la audiencia en Rotten Tomatoes, este no está contenido dentro del JSON-LD. Para poder acceder a esta información habría que complementar la extracción de información con otros métodos como web scrapping.

# ¿Cuáles cree que son las ventajas y desventajas de utilizar el vocabulario de Schema.org?

Las ventajas de utilizar el vocabulario de schema.org es que se intenta implementar una estandarización de la información de una página web para su fácil acceso, esto ahorra mucho tiempo en l extracción y procesamiento de la información.

Las desventajas observadas en el desarrollo de este trabajo práctico, es que aún con la estandarización de los campos no existen convenciones de como se debe llamar determinado dato, por nombrar un ejemplo, “creator” y “author” los dos se refieran al escritor de la película, en IMDB el escritor de la película se encuentra en el campo “creator” y en Rotten Tomatoes se encuentra en “author”.

Tampoco hay una convención sobre que información debe llenarse, en el caso de eCartelera no se recolectan reseñas, por lo tanto, este campo estará vacío. O el caso de utilización de nombres de campos que no pertenecen al schema, como @id y actors.

La utilización de este esquema da un orden a la información en internet, aún así, cuando se extrae la información debe revisarse, para adaptarse a las particularidades de cada sitio.

# ¿Cuáles cree que son los desafíos que todavía quedan por resolver?

Se han dado grandes pasos en el ordenamiento y acceso a la información de internet, dicho esto, todavía existe mucho trabajo en la estandarización de terminología, forma de almacenado de la información en internet y la difusión de estos esquemas por los desarrolladores web.

Aun cuando los estándares existen, cada sitio web es libre de aplicarlo o no, tampoco de aplicarlos de manera estricta.

Lo observado en este trabajo práctico sugiere que algún sitio puede utilizar su nomenclatura propia, ya sea para integrarlo a sus sistemas existentes o por que considera que otras nomenclaturas describen mejor la información que desea transmitir.

Describir proceso de selección de datos

**Las decisiones en la selección y tratamiento de la información parten de que todas las fuentes de información son traducibles e intercambiables entre español e inglés**

* Se cargaron los jsons en un dataset para verificar la información que contienen
* Se comparan y se evidencia que no hay una convención en el nobramiento de las variables, tambien se detectan nombres de variables que no pertenecen a lo indicado en schema.org
* Se seleccionan variables tomando el criterio:
  + Para los casos en los tres sitios tenían la misma información se seleccionaba indistintamente.
  + En los casos de que existan dos iguales y otro distinto se tomó por votación
  + IMDB: es conocido como la referencia para información “específica” sobre las películas por lo tanto se tomaran de este: name, duration, image, genre, keywords, author.
  + Rotten Tomatoes es a referencia para conocer la opinión de la ausiencia y cinéfilos sobre una película, por lo tanto se tomarán los reviews de este.
  + Se normalizaron los puntajes de los 3 sitios en una escala de 1 a 10 y se promediaron.
  + Se almacenan los url de cada fuente de datos
* Se exporta el json final para el consumo.

Esta información estaría disponible para el sitio de comparación, adicionalmente se podrá desplegar la información original y enlaces de cada sitio.