scraping de datos de la pagina ‘cuevana3’

Jose Antonio Herrera Ramírez

Contenido

[Resumen Ejecutivo 2](#_Toc478933641)

[Introducción 2](#_Toc478933642)

[Desarrollo 3](#_Toc478933643)

[Conclusiones y Recomendaciones 5](#_Toc478933644)

[Bibliografía 6](#_Toc478933645)

# Resumen Ejecutivo

Como parte de la implementación de la metodología previamente establecida para hacer scraping de datos en una página web, se implementó este scraper para extraer los datos de la página ‘Cuevana3’, los datos extraídos son de las películas que la pagina ofrece. El programa realiza la obtención de datos de las películas recorriendo todas las paginas disponibles y va guardando esta información en una base de datos no relacional debido a la comodidad con el manejo de datos orientados a documentos y Javascript.

En esta recolección se obtienen diferentes tipos de datos los cuales pueden servir luego para realizar consultas a los datos, filtrarlos por algún criterio o usarlos de cualquier manera.

# Introducción

Hoy en día el web scraping es una técnica muy utilizada entre los campos de la minería de datos, esta técnica consiste en extraer la información de páginas webs utilizando bots(programas) que realizan este proceso automáticamente.

Este scraper pretende dar una introducción al desarrollo e implementación de esta técnica, abarcando la extracción y manipulación de datos en un entorno como NodeJS. El entorno y las librerías usadas pueden ser familiares para aquellos usuarios Javascript ya que todo se desarrolla bajo este lenguaje.

Este trabajo es de interés para aquellos que desean envolverse de alguna manera en técnicas de minería de datos o talvez para algún usuario que necesite un conjunto de datos para manipular en su aplicación, etc. El punto es que hoy en día es una de las técnicas más utilizadas para extraer la información y muchos sistemas o navegadores la implementan.

Sin más preámbulo, a continuación, se detalla todo el proceso de desarrollo de esta herramienta aplicada a la extracción de datos sobre una página web de películas.

# Desarrollo

Como lenguaje de programación se utilizó Javascript en el entorno NodeJS, mediante npm se instalaron paquetes para cumplir diversas funcionalidades, a continuación, se mencionan y explican cada uno de los paquetes usados en el programa:

* **Express:** un framework del lado del servidor, bastante robusto, ofrece muchas funciones para web apps, utilizado para el lado del servidor.
* **Request:** una librería para simplificar las peticiones http, utilizada para hacer las peticiones a las páginas web y obtener el html de estas.
* **Cheerio:** una librería construida con el core de jQuery diseñada para el lado del servidor a fines de la manipulación de elementos del DOM.
* **Mongoose:** El ODM(Object-Document Mapping) con soporte oficial para MongoDB, utilizado para interactuar con la base de datos.

Como se mencionó anteriormente, la información a extraer es sobre las películas que el sitio web tiene en su catálogo, por lo que se seleccionaron algunas características para extraer, en la siguiente imagen se ilustran y luego se detallan cada una de estas:



|  |  |  |
| --- | --- | --- |
| **Nombre de atributo** | **Descripción** | **XPATH** |
| *Imagen* | El primer atributo a guardar, contiene la referencia a la imagen(portada) de la película. | //\*[@id="tab\_1"]/div[1]/img |
| Calificación IMDb(Rating) | Especifica el rating de la película de acuerdo a una cantidad de votaciones. | //\*[@id="tab\_1"]/div[1]/p[1]/span |
| Titulo Original | El titulo o nombre de la película | //\*[@id="tab\_1"]/div[1]/p[2]/span |
| Duración | Duración en minutos de la película | //\*[@id="tab\_1"]/div[1]/p[3]/span |
| Año | Año de estreno de la película | //\*[@id="tab\_1"]/div[1]/p[4]/span |
| Lanzamiento | Fecha en la que se estrenó la película | //\*[@id="tab\_1"]/div[1]/p[5]/span |
| País | País donde se produjo | //\*[@id="tab\_1"]/div[1]/p[6]/span |
| Idiomas | Idiomas disponibles de la película | //\*[@id="tab\_1"]/div[1]/p[7]/span |
| Calidad | Calidad del video | //\*[@id="tab\_1"]/div[1]/p[8]/span |
| Director | Director a cargo de la filmación | //\*[@id="tab\_1"]/div[1]/p[9]/span |
| Guion | Escritores o participantes en el guion | //\*[@id="tab\_1"]/div[1]/p[10]/span |
| Premios | Nominaciones o premios que la película ha obtenido | //\*[@id="tab\_1"]/div[1]/p[11]/span |
| Genero | Categoría de la película(Terror,Drama,etc) | //\*[@id="tab\_1"]/div[1]/p[12]/span |
| Reparto | Actores principales de la película | //\*[@id="tab\_1"]/div[1]/p[13] |
| Sinopsis | Descripción general de la película | //\*[@id="tab\_1"]/div[2]/p |

# Conclusiones y Recomendaciones

Se concluyó que mediante esta herramienta se pueden obtener cualquier cantidad de datos que luego pudiesen ser analizados o implementados en otro sistema.

Asimismo, se definió un entorno para desarrollar esta herramienta en la plataforma NodeJS y se analizó cada uno de los componentes de este scraper, esto puede servir de base para implementar esta herramienta en otras aplicaciones.

Se recomienda leer la documentación de cada uno de los paquetes utilizados para tener una mejor perspectiva del papel que cumple cada uno en el desarrollo de la herramienta.

# Bibliografía

GitHub. (2017). cheeriojs/cheerio. [online] Available at: <https://github.com/cheeriojs/cheerio> [Accessed 3 Apr. 2017].

GitHub. (2017). request/request. [online] Available at: <https://github.com/request/request> [Accessed 3 Apr. 2017].

Mongoosejs.com. (2017). Mongoose ODM v4.9.3. [online] Available at: <http://mongoosejs.com>/ [Accessed 3 Apr. 2017].