https://datajournalism.com/read/newsletters/data-scraping-for-stories

Conversaciones con datos

El texto parte con la definición de lo que es “scraping” de datos: “el proceso que permite extraer contenido de una página web usando una herramienta especializada o con la escritura de un código”.

Las limitaciones que tiene la extracción de datos son:
-sitios que tienen el HTML mal formateado con muy poca información estructural.
-sistemas de autenticación que impiden tener un acceso automatizado.
-cambios en los marcadores de una página web.

Re- chequea tu código para que no pierdas ningún dato

El reportero de ciencias de BuzzFeed News, Peter Aldhous, usó Python Requests y Beautiful Soup para extraer de un sitio web, los registros de acciones disciplinarias contra doctores en el estado de Nueva York. Con esos datos hizo un reportaje sobre cómo médicos adictos y caídos en desgracia estaban dirigiendo las pruebas de medicamentos para enfermos.

Con las mismas herramientas, extrajo los nombres de autores, disciplina científica, número de citaciones y otra metadata de papers publicados durante 10 años en el Journal de la Academia Nacional de Ciencias, y descubrió que sus miembros usaban un acceso privilegiado para publicar, dando a conocer quiénes ejercían este privilegio.

Actualmente por el flujo de su trabajo usa mucho el paquete rvest R, de R tidyverse (una colección de paquetes de R). Como ejemplo de los reportajes que ha hecho con rvest R están:

- “Por qué las estrellas de atletismo no establecen récords mundiales como solían hacerlo (pero los nadadores sí)”, que descubrió recogiendo datos de las 100 mejores actuaciones al aire libre en muchos eventos de atletismo del sitio web de la Asociación Internacional de Federaciones de Atletismo.
- Aldhous extrajo del sitio web de “The American Presidency Project” los textos completos de todos los discursos del estado de la nación (State of the Union Address) y otros discursos presidenciales dados al Congreso para comparar las palabras que usaba Trump, con las que habían usado todos los presidentes desde que George Washington dio el primer discurso de este tipo en la historia de Estados Unidos en 1790.

Luego de los ejemplos, Aldhous ofrece links a algunos tutoriales que explican cómo usar rvest, porque lo encuentra más claro e intuitivo que las alternativas de Python.

Una advertencia del autor: es clave asegurarse que se está extrayendo todos los datos. Lo plantea porque pueden haber variaciones sutiles en la forma en que los sitios web están codificados, lo cual implica que

se deben llenar los vacíos de forma manual. Sugiere usar el browser inspector web, y estudiar cuidadosamente el código del sitio para saber cómo debe escribirse el extractor. Para Chrome se puede usar la extensión Selectgadget.

La extracción de datos de buena calidad lleva tiempo: comunícate de manera efectiva y busca las API existentes

Esta sección está escrita por el periodista Mikołaj Mierzejewski, del periódico más grande de Polonia -Gazeta Wyborcza- quien explica que al extraer datos se pueden dar tres situaciones:

1.- El que los datos estén en HTML simple y se pueda extraerlos fácilmente con herramientas como Portia.

2.- La segunda situación es más difícil, porque necesita cookies o mantener la sesión del usuario iniciada; o puede que sí se carguen los datos pero que se requiera de herramientas de desarrollador en el navegador para descargarlos.

3.- En el tercer nivel de dificultad, los datos se cargan sólo a medida que se interactúa con el sitio, por lo que se necesita un programador que desarrolle una pequeña aplicación que actúe como un navegador para descargar los datos.

Uno de los grandes desafíos es poder comunicar las necesidades a los miembros del equipo que no son técnicos, entre otras cosas porque deben entender que una buena extracción de datos requiere tiempo.

Mierzejewski ofrece ejemplos de lo que han hecho extrayendo datos de Instagram.

Finalmente, sugiere siempre buscar la API primero, antes de extraer datos. Recomienda Postman e Insomnia como herramientas para API.

9 cosas a recordar sobre la extracción de datos

Esta sección es de Paul Bradshaw, quien dirige el curso de la maestría en Periodismo de datos de la Universidad de la ciudad de Birmingham, y es autor de “Scraping for Journalists”.

1.- No necesita incluir siempre código. Su colega de la BBC usó hojas de cálculo de Google para una historia sobre reclamos por ruido.

2.- Piensa en la política de Términos y Condiciones del uso de la información. Dice que quería extraer datos de un sitio de propiedades y los T&C lo prohibía. Al final convencieron a los dueños del sitio web. Hay que consultar al departamento legal cuando los T&C prohiban extraer la información.

3.- Usa la extracción de datos como segunda opción. Para una historia de recortes en bibliotecas compararon información obtenida a través de solicitudes de acceso a información pública y la compararon con datos extraídos de 150 PDF que les entregó un auditor.

4.- Recuerda que si tiene una estructura o un patrón, probablemente puedes extraer los datos. En una historia sobre violaciones, extrajeron datos clave de los reportes de la policía porque estaban en un mismo formato.

5.- Chequea que la data no esté disponible de antemano, antes de extraerla. A veces ya está disponible en formato JSON para descargar, o puedes encontrarla en otro sitio con Google Inspector.

6.- Haz un control aleatorio de tus datos extraídos y compáralos con los datos en el sitio original.

7.- Usa filtros y tablas dinámicas para encontrar resultados inusuales. Cuando la extracción no funciona bien, lo hace de forma sistemática. Esas desviaciones pueden ayudarte a chequearlo.

8.- Extrae la mayor cantidad de datos primero, y después filtra y limpia. Son dos procesos diferentes.

9.- Extrae la información más de una vez, e identifica la información añadida o eliminada.

Usa scrapers como una herramienta de monitoreo

Sección a cargo de Maggie Lee, reportera freelance de Atlanta.

Aunque no tengas una gran historia, Lee promueve la extracción de datos sistemática, porque sirve para monitorear sectores de cobertura, por ejemplo policial. El extraer datos de forma continua y sistemática puede automatizar el proceso para identificar cambios relevantes, o ahorrar la necesidad de ir de forma constante al sitio web a chequear si hay nueva información publicada, ya que se puede recibir un correo electrónico, por ejemplo, cuando un documento nuevo es incluido al sitio. Así el reportero ahorra tiempo.

Asegúrate de que el scraper sea resistente y siempre ten un plan b

Esta sección está escrita por Gianna-Carina Grün, jefa de Periodismo de datos de la DW.

Si se extraen datos de forma regular con un extractor, este debe estar diseñado para identificar los cambios pequeños en escritura de una página o en fechas, por ejemplo, para que no rompan el código. El extractor debe ser lo más resistente posible.

Se puede anticipar alternativas para solucionar potenciales problemas futuros, por ejemplo pensando qué otras fuentes de información pueden tener los datos en el mismo formato para poder usar el extractor.

No olvides respaldar los datos en la extracción

Sección a cargo de Erika Panuccio, quien es asistente de comunicaciones en ALTIS.

Es importante respaldar los datos extraídos porque estos podrían ser eliminados. Da el ejemplo del análisis de las cuentas de Twitter de personas a favor de ISIS, con el que armó una base de datos de 30.000 tweets de 100 usuarios. El respaldo con una plataforma IFTTT automatizada de los tweets cuando se publicaban fue clave, porque hubo cuentas eliminadas por la política de Twitter en publicidad terrorista y tweets eliminados por los usuarios de las cuentas.