Extracción de datos o scraping

Saber encontrar datos se está convirtiendo en una habilidad cada vez más valiosa en el periodismo.

Ya mencionamos cómo buscar información en la web profunda a través de búsquedas avanzadas. Y también cómo encontrar datos en Bases de Datos.

Ahora veremos algunas herramientas para la extracción de datos.

El ejemplo más sencillo de la extracción de datos es el de obtener el contenido de una tabla alojada en un PDF, cuando lo que se requiere para procesar los datos es disponerlos en una hoja de cálculo.

Para lograr ese objetivo, existen recursos sofisticados y software de descarga pagos; este es el plano en el que muchas veces necesitamos de un programador para que nos asista en la tarea.

Pero no siempre: hay una larga lista de herramientas online y sin costo alguno que nos sirven para extraer datos, tanto de un PDF como de una página web.

Vamos desde lo más sencillo a lo que aplicaremos en la práctica.

1. **Conversores**: A veces copiar y pegar funciona, pero otras veces no y hay que recurrir a conversores como Zamzar.com, que es gratuito y no requiere suscripción.

2. No olvides que las tablas y los gráficos pueden estar subidos a la web en formato de imagen o estar encerradas en un PDF. En estos casos se recurre a programas de **reconocimiento óptico de caracteres**. Hay algunos que pueden funcionar bien si el tamaño del archivo no supera cierto tamaño.

Uno es este:

https://www.sodapdf.com/es/ocr-pdf/

Y el otro: https://www.onlineocr.net/es/

Pero no siempre funcionan. Solo son eficaces para archivos pequeños y si no tienen sellos ni firmas manuscritas.

Para texto, funciona muy bien Easy PDF to Text

Y tiene opciones de conversión en varios formatos:

https://easypdf.com/

3. Un recurso muy bueno para extracción de datos es Document Cloud: https://www.documentcloud.org/ ; al mismo tiempo que sube un documento PDF (una vez que se obtiene una cuenta), corre un OCR (sistema de reconocimiento óptico de caracteres), que permite extraer texto de imágenes escaneadas. El texto extraído se lee en la pestaña "texto". Sirve para texto, más no para tablas.

Vea un ejemplo aquí:
https://www.documentcloud.org/documents/5772210-4478819A01

Abajo a la izquierda, puede seleccionar ver el documento (document) o Plain text (la versión con la extracción ya realizada)

4. Una de las mejores herramientas gratis para extraer tablas de PDFs y obtener los datos en formato tabla es Tabula.

https://tabula.technology/

Deberán descargar el archivo a su computadora. Es seguro. Y se abre sobre la Web, es decir, si tienen Chrome abierto como navegador. No funciona sobre otros navegadores.

El procedimiento es sencillo: se sube un PDF; luego se selecciona la tabla y es posible repetir esta acción en diferentes páginas para finalmente descargar el documento en CSV, que es formato de datos separados por comas: este se puede estructurar nuevamente en columnas, marcando la primera columna, luego yendo a la pestaña datos, ordenar en columnas, delimitados por comas y finalizar.

También se puede copiar y pegar el resultado desde el cuadro de datos a una hoja de cálculo en blanco.

Cómo armar un set de datos

Los datos ya han sido recuperados de la Web o descargados en la computadora. Tal vez el periodista haya tenido que recurrir a la entrada de datos manual, ya que esto es frecuente cuando los documentos, vía petición, son entregados en formato papel y no resulta posible la extracción de la información y su posterior reestructuración en tablas, considerando que no siempre funcionan con eficiencia los programas de reconocimiento óptico de caracteres (OCR)

En periodismo de datos consideramos que un set de datos es una colección de elementos de datos agrupados, que permiten su fácil registro. Dentro de los mismos, los datos se encuentran organizados en variables, por lo general medidas a lo largo del tiempo y en soportes descargables.

La forma más sencilla de gestionar el set de datos es usando Google Drive, la herramienta que reemplazó a Google Docs por sus mayores y mejores funcionalidades.

Ejemplo: patrimonio de legisladores en hoja de cálculo (los datos originales estaban contenidos en un PDF. Se usó Tabula para la extracción)

https://docs.google.com/spreadsheets/d/1SzSZBZQBdNx_CIBcGmQP5C-X6gzFd5kkNs2zl152rLE/edit#gid=0

Las formas más comunes de gestionar y alojar sets de datos son a través de planillas de Excel o mediante un formato denominado CSV, igualmente abierto y reutilizable, en el que las columnas están separadas por comas.

Cuando se descarga un fichero de datos en CSV la imagen que se obtiene es del tipo que se visualiza en este enlace:

Datos en CSV

https://docs.google.com/spreadsheets/d/1D4yxfdzXs93rchbfYA_8nBphhgcZUd-5qRkA_Q9vpR8/edit#gid=597761278

Los mismos datos estructurados

https://docs.google.com/spreadsheets/d/1D4yxfdzXs93rchbfYA_8nBphhgcZUd-5qRkA_Q9vpR8/edit#gid=1145391279

El archivo es público: **no necesitan solicitar acceso**; solo descargarlo a su computadora, y una vez descargado, abrir con Excel

En este caso, para convertir CSV a Excel hay que volver a estructurar las columnas.

El procedimiento es:

Sin mover el cursor de la Fila A, márquela o píntela.

1) Ir a la pestaña “Datos”

2) Luego hacer clic en “Texto en Columnas”

3) Se abrirá un cuadro de diálogo, marcar “delimitados” + “siguiente”

4) Luego marcar “Tabulación” + “comas”

5) Nótese que durante el procedimiento la columna A permanece marcada o “pintada”. Luego hacer clic en siguiente y lo que se obtendrá el resultado en la pestaña 2

6) Para visualizar mejor los datos a veces se centran los valores, se mejoran los títulos, se colocan colores de fondo...eso es a elección.

Con este último paso el procedimiento se completa.

Hay una lista de herramientas de scraping de datos en:

https://docs.google.com/spreadsheets/d/1RNuyJbt2Mz9KIZ_HWpfAg-u4vJNbUJcD4fsWiHMXByw/edit#gid=1060952292

que pueden descargar a la computadora.

Resumiendo:

Web scraping es una técnica utilizada mediante programas de software para extraer información de sitios web

Scraping es un método que te permite extraer datos escondidos en un documento, como páginas web y PDF, y los hace útiles para usarlos después

Finalmente no olviden herramientas de seguridad digital a tener en cuenta.

En español, el mejor recurso que disponemos por el momento es este:

https://securityinabox.org/es

Y si desean recorrer los enlaces que seleccioné en LiveBinders sobre el tema, pueden ingresar a esta URL:

http://www.livebinders.com/play/play?id=2004543

(Recuerden que en Livebinders a cada carpeta (azul) le corresponde un enlace, que pueden abrir y dentro de cada carpeta hay subcarpetas (gris) cada una con la URL de la herramienta)

Otras herramientas para extracción de datos que pueden probar:

https://www.cometdocs.com/

https://www.octoparse.com/

https://smallpdf.com/es

https://chrome.google.com/webstore/detail/data-scraper-easy-web-scr/ndknepjnldbdbepjfgmncbggmopgden (extensión para Chrome)

https://chrome.google.com/webstore/detail/table-capture/iebpjdmgckacbodjpijphcplhebcmeop (extensión para Chrome)

https://chrome.google.com/webstore/detail/web-scraper-free-web-scra/jnhgnonknehpejjnehehllkliplmbmhn (extensión para Chrome)

https://support.google.com/docs/answer/3093339?hl=es-419
La función IMPORT permite llevar datos de una tabla publicada en la Web a una hoja de cálculo en Google Spreadsheet