Módulo 3: Extracción de datos y conversión de formato

[00:00:12] **Sandra Crucianelli** Bienvenidos a todos al Módulo 3 de este curso sobre Introducción al periodismo de datos. Bueno ya hablamos en las semanas pasadas sobre qué eran los datos abiertos, los datos cerrados, también les mostré distintas bases de datos donde poder hacer búsquedas, vimos los principales buscadores y técnicas de búsqueda avanzada.

[00:00:36] **Sandra Crucianelli** Por lo tanto, ahora tenemos que seguir adelante con lo que sería un proceso vital en materia de periodismo de datos, que es el scraping. ¿Qué es el scraping? Es la extracción de datos. Como yo les comenté en uno de los videos de este curso, hay formatos cerrados sobre los cuales tenemos datos que son importantes para nosotros procesar y tenemos que liberar esa información. Tenemos que extraerlos de esos formatos cerrados para convertirlos a datos abiertos. Estos procesos de conversión no son siempre sencillos. Es decir, tratar de llegar a un Excel, que sería el formato ideal sobre el cual podemos trabajar, no es sencillo. A veces en los sitios de gobierno abierto tenemos, como ya vimos en clase pasadas en formato en CSV, en valores separados por comas. A veces tenemos esos datos en Excel. Pero a veces, o la mayoría de las veces, los tenemos en PDF, esto pasa con mucha frecuencia.

[00:01:39] **Sandra Crucianelli** Déjenme mostrarles algunos ejemplos de trabajos que hemos hecho en la unidad de datos de Infobae. Este trabajo sobre los decretos secretos de la dictadura que publicamos en 2019, sobre un conjunto de 9.037 decretos que estaban contenidos en PDF y eran palabras, acá no había números, no había números que estructurar en tablas. Eran palabras. Era todo texto, salvo unos pocos casos en los que había unas pocas tablas y hubo que hacer el scraping. O sea, la extracción de estos textos que estaban atrapados en un PDF para llevarlos a una base de datos propia y poder analizarlos. Como ustedes podrán ver al final del artículo, como siempre hacemos en las notas de la Unidad de Datos de Infobae, compartimos el enlace con lo que sería la apertura de datos, los decretos publicados --estos son del 2013--, todos compartidos en formatos abiertos. Fomatos que cualquier persona puede descargarse y llevarse a su computadora.

[00:02:59] **Sandra Crucianelli** ¿Cuáles son las técnicas para llegar de estos formatos cerrados donde tenemos estas fotos --en algunos casos son fotos, provienen de imágenes, en otros casos no--, cómo podemos hacer para liberar esta información? Mediante técnicas de scraping, de extracción de datos y también usando conversores. Este, que comienzo compartiéndoles, es uno de los más antiguos, yo ya lo usaba hace diez años, Zamzar es una herramienta de conversión de formatos. Aquí ustedes pueden usarlo para convertir una gran variedad de formatos: audios, documentos, imágenes, videos y otros, como pueden ser los archivos comprimidos. Nosotros nos vamos a concentrar en lo que es la conversión de los PDFs, porque hay mucho PDF en nuestras bases de datos que requiere herramientas de conversión. En este caso lo que se hace es agregar archivos. El paso 1 es agregar archivos, ¿sí? Yo voy a agregar un archivo que quiero obtenerlo en este caso voy a elegir un formato de documentos, un formato de Word, algo que está en un PDF encerrado a un formato de DOC. Lo que sería un Word en el Office, ¿sí? Y le doy convertir. Podría haber elegido cualquier otro formato. La herramienta comienza el proceso de conversión. Y finalmente me da una opción de descarga. Las opciones de descarga por ejemplo en este caso yo utilicé un documento a dos columnas de un PDF. Yo decidí que la mejor manera de tener la información era en txt y me liberó la información de una manera bastante buena para un, para un conversor que no requiere que estés logueado y que no tiene pago, digamos la versión gratis. Por

supuesto, que tiene una versión paga. Pero esto que yo les estoy mostrando es la versión gratis.

[00:05:15] **Sandra Crucianelli** Después hay otras herramientas como Easy PDF. Easy PDF lo hemos estado probando en el último año, funciona bastante bien. Permite convertir PDF en Word, PDF en Excel que es de mucha utilidad. Y también otras conversiones, pero da resultados bastante buenos. Incluso tiene algo que se llama el OCR Online. Y ustedes me dirán, pero ¿qué significa OCR? Bueno, OCR son las siglas de Optical Character Recognition que serían Reconocimiento Óptico de Caracteres. Y esto se hace en línea, ustedes no tienen más que subir un archivo en donde esté encerrada la imagen que proviene de una foto. Lo seleccionan, lo suben, ¿sí?, y comienza el proceso de conversión. Pueden pedir de acuerdo a la estructura que tenga el contenido, pueden convertirlo en un Word, en un Excel o en un txt plano, ¿sí? Pero esta herramienta funciona bastante bien, da muy buenos resultados. Esta extracción que hice, que les mostré recién en txt, está hecha con PDF Easy Convert --, no lo hago ahora porque toma bastante tiempo, si la subida toma bastante tiempo. Pero sólo tenía un documento en PDF que tenía dos columnas, estaba provenía de una imagen y este fue el resultado que me dio la conversión. Por lo tanto Easy PDF es una alternativa para que tengan en cuenta para poder convertir archivos en PDF en formatos más abiertos.

[00:07:02] **Sandra Crucianelli** Ahora hay una herramienta que es superior en cuanto al nivel de funcionalidad que tiene para todos los periodistas y que es Document Cloud. Cualquier periodista puede tener una cuenta gratuita en Document Cloud. Y lo interesante es que ustedes en este tipo de herramienta cuando suben un documento -- este es el mismo documento que yo les compartí cuando les mostraba formatos abiertos--, bueno, en el momento en que suben un documento, corre. Fíjense que este es un documento largo que tiene 39 páginas y en el que si ustedes van marcando, pueden ver que ha hecho una conversión bastante exitosa del texto. El documento original, como nosotros lo teníamos, logra hacer una conversión, ¿por qué? Porque cuando ustedes lo suben a documentcloud.org -- esa es la url de la herramienta -- corre un OCR, es decir, corre un programa de reconocimiento óptico de caracteres. Entonces va a poder extraer con cierta eficacia el texto. Por supuesto que estas extracciones no son perfectas, es decir, tienen un nivel de error porque por ejemplo, en este caso el documento que está con sellos, con firmas, con logos o caracterizaciones que podrían entorpecer la tarea de extracción. No obstante, Document Cloud da una buena respuesta. Fíjense que yo tengo alojados una gran cantidad de documentos aquí, por ejemplo, los boletines oficiales, los boletines oficiales de mi país, ustedes podrían hacer lo mismo si tuvieran la necesidad de extraer la información que hay en la Gaceta Oficial, ¿sí? Y tiene un nivel de precisión bastante alto. Vean que aquí está la línea de texto. Acá abajo a la izquierda tienen tres niveles de visualización: lo que sería la pestaña de documento donde ustedes ven el documento original; lo que sería plain text, que es la extracción que me hizo del texto que estaba atrapado en una imagen; y las miniaturas de la extensión completa del archivo.

[00:09:26] **Sandra Crucianelli** Por lo tanto, estas -- Zamzar, Easy PDF y Document Cloud -- son las primeras tres herramientas que deberían empezar a usar y a probar, a practicar. Lo que pueden hacer es descargarse cualquier PDF y comenzar tanto en Zamzar como en Easy PDF no necesitan estar logueados. Si van a usar Documento Cloud, van a necesitar tener una cuenta en Document Cloud. No es difícil, no es automático, a eso voy. Sí, no es automático, tener una cuenta en Document Clud. Toma un poco de tiempo porque tienen que verificar que ustedes sí trabajen en un medio de comunicación y que sí sean periodistas que usan documentos regularmente en su trabajo cotidiano. Pero lo fabuloso de esta herramienta es que como les digo además de subir documentos, pueden

extraer el texto de esos documentos, analizarlos, incluso aplica buscadores y permite búsquedas internas dentro de documentos que son muy voluminosos, incluso editarlos. Por eso la importancia de esta herramienta.

[00:10:36] **Sandra Crucianelli** Estas son tres herramientas básicas el periodismo de datos, digamos los conversores de formato, el que nos ayudan a liberar información que está atrapada adentro de un PDF. Por supuesto que hay herramientas más sofisticadas que usamos en la sala de redacción, como los lenguajes de programación. Hoy por hoy los periodistas de datos trabajamos mano a mano junto con los programadores o desarrolladores o son ingenieros de sistemas en muchos casos. Pero trabajamos con personas que manejan lenguajes de programación porque se requieren estas habilidades para poder hacer un scraping de datos. Especialmente cuando trabajamos scraping web: es cuando tenemos una página web y necesitamos rescatar de esa página web elementos que son importantes para nuestro trabajo, como pueden ser los contenidos de una tabla. En muchos casos funciona el copiar y pegar. En otros casos no. Para Chrome, el navegador Chrome, que es el más usado, hay extensiones que les permiten extraer los datos, pero no funcionan con mucha eficiencia. Siempre la tarea del programador va a ser mucho más eficiente que la de una herramienta automática en este sentido. Pero, pero como les decía, esto de incorporar un programador, un desarrollador a la sala de noticias para que trabaje mano a mano con el periodista es algo relativamente nuevo, los últimos diez años, sí, en los últimos diez años empezamos a ver que el programador y el desarrollador se unen al periodista. No obstante eso, estas tres herramientas les permiten abrir datos que ustedes solos no podrían abrir, habría que hacer entrada de datos manual. Muchas veces, cuando tenemos algo en un papel, si nos entregan un documento en un papel y queremos digitalizarlo, la única forma que tenemos es hacer lo que se llama como data entry manual, es la cargada manual de los datos. Vamos a evitar ese proceso. Ese proceso arrastra un 10 -15 por ciento de error, muy especialmente si están cargando una hoja de cálculo. Es algo que yo quiero evitar siempre, la cargada manual de datos. Entonces es mejor que se vayan familiarizando con estas herramientas como Zamzar, como Easy PDF -- hay otras tantas, por supuesto, hay conversores de todo tipo. Las mejores herramientas generalmente son las pagas, yo trato de mostrarles las que son gratuitas y tiene una interfaz en español para que se sientan más cómodos y puedan utilizarlas con facilidad.

[00:13:07] **Sandra Crucianelli** En el próximo video les voy a contar cómo funciona una herramienta muy importante que se llama Tabula.