Módulo 3: Extracción de datos usando Tabula

[00:00:12] **Sandra Crucianelli** Hola a todos. Bienvenidos a este segundo video del Módulo 3, donde estamos aprendiendo técnicas para scraping de datos. Recuerden que scraping significa extraer datos, obtener datos a través de un formato que se nos ha dado con el cual es imposible trabajar para el procesamiento de grandes volúmenes de información.

[00:00:35] **Sandra Crucianelli** Hoy les voy a presentar una herramienta muy usada por todos los periodistas de datos. Hace muchos años comenzamos a tener problemas con los PDF y nos dimos cuenta que no había una herramienta gratuita mediante la cual podamos extraer tablas que están encerradas en un PDF. Un desarrollador, Manuel Aristarán, que es argentino, trabajó bastante tiempo en este problema para poder crear una herramienta finalmente que nos dio soluciones a todos. Se llama Tabula.

[00:01:10] **Sandra Crucianelli** Permítanme presentarles esta herramienta. Voy a compartir mi pantalla para que puedan ver de qué se trata. Tabula cuya url, su link o su enlace es tabula.technology, es una herramienta que ha sido creada para extraer tablas encerradas dentro de PDF. Tiene una versión de descarga en Windows y una opción de descarga para Mac. Es una herramienta completamente segura y la pueden utilizar en ambos sistemas operativos. También tienen una cuenta en Twitter mediante la cual ustedes se pueden poner en contacto con el desarrollador. Y en esa cuenta en Twitter van a poder ver algunos ejemplos de usuarios que han utilizado esta tecnología para extraer documentos muy voluminosos. Esto les va a permitir interactuar con otros usuarios y me parece un buen canal de comunicación entre el desarrollador y la herramienta. Una vez que la descargaron, ustedes se van a encontrar con una carpeta que van a tener que descomprimir de un archivo zip, donde van a tener el ícono de acceso, que lo pueden fijar en su escritorio, el ícono de acceso a esta herramienta para scraping de datos.

[00:02:36] **Sandra Crucianelli** Una vez que ya descargaron la herramienta, se van a concentrar en su homepage, en la página de inicio, y ahí hay un buscador mediante el cual ustedes pueden importar uno o más PDF. Yo por ejemplo, voy a tomar un PDF que tengo cargado, descargado en mi computadora y lo voy a importar. Cuando hago clic sobre el botón de "import" va a aparecer el documento que yo cargué perfectamente visualizado y de esta manera. En la barra lateral izquierda van a tener opciones de visualizar cada una de las hojas del documento. Y vamos a suponer que yo quiero extraer esta parte de esta tabla. Entonces lo que tengo que hacer es con el botón del mouse marcarla, ¿sí?, y una vez que la marqué voy a ir aquí arriba a la derecha, donde este "preview and export extracted data", es decir, previsualice y exporte los datos extraídos. Una vez que hago clic automáticamente la herramienta ha importado esta información, la ha extraído. Y ¿cómo la voy a exportar? Bueno tengo distintas opciones para exportar este archivo. Lo más frecuente que usamos en periodismo de datos de CSV y por eso yo tuve en todo este curso siempre insisto con la necesidad de que aprendan a manejar archivos CSV, que son archivos en los que los atributos de una tabla, los datos están separados por comas. Yo voy a hacer clic en Exportar y lo que voy a visualizar va a ser esto. Yo lo abrí como Excel para separar este archivo que está en CSV y convertirlo en un Excel sólo tengo que marcar como siempre la primera columna, ir a la pestaña Datos, hacer clic en el ícono donde hice texto en columnas, delimitar, siguiente, marco la coma, siguiente y finalizar. Una vez que hice eso, tengo perfectamente estructurada la tabla. Este es un caso sencillo, pero puede haber casos más complicados.

[00:05:08] **Sandra Crucianelli** Por ejemplo, este. Esta es una tabla que está contenida en un documento. Y si yo quisiera procesarla, llevarla a una visualización interactiva, tendría que cargar manualmente todos estos datos. Pero con esta herramienta puedo extraer los datos. Lo que voy a hacer es marcar los datos que me interesan de esta tabla, y repito el procedimiento: con el botón derecho superior previsualizo los datos extraídos y una vez que tengo la visualización de los datos en la pantalla, le voy a pedir al sistema que me los exporte como un CSV. Aquí está la extracción ya hecha. El procedimiento es el mismo. Repito: marco la primera columna, voy a la pestaña de datos, clic en texto en columnas, delimitados -- obviamente, como es un CSV, están delimitados por comas. Siempre tienen que mirar cuál es la separación que tienen los datos. En este caso cada dato está separado por una coma, pero podría ser que estuvieran separados por un punto y coma o por un espacio. En ese caso marcarán punto y coma o marcarán espacio. Yo como aquí veo que están separados por comas, voy a marcar la coma--. Siguiente y le doy finalizar. De esta manera voy a tener, como ustedes ven yo voy a estirar el tamaño de las columnas para que puedan ver mejor que los datos ya están en una tabla, tienen un cierto orden. Por supuesto, esto..., estas tablas que quedan convertidas las voy a guardar, por ejemplo la tengo en CSV, la voy a guardar como Excel, lo guardo como Excel para que conserve todas, todas sus características. Y la extracción sale, como a veces decimos, sucia. Sucia es porque por ejemplo, las vocales que tienen acentos quedan expresadas como un símbolo en muchos casos. Y en este caso hay que hacer lo que se llama limpieza de datos, es decir, que hay que organizarlo, hay que limpiarlo, hay que pulirlo para tener perfectamente organizada la tabla a los efectos de lo que queremos hacer.

[00:07:24] **Sandra Crucianelli** Con otro ejemplo. En este caso tengo una hoja de trabajo bastante compleja, ¿por qué? Porque tiene tablas, pero también tiene texto. En este caso ven es un informe gubernamental del estado de Chihuahua, en México, donde cada tanto aparece una tabla. Y bueno, ¿qué pasa cuando queremos extraer estas tablas? El mismo procedimiento: marcamos la tabla, le pedimos a Tabula que haga una previsualización e inicie el proceso para poder exportar esta tabla--, aquí ya tengo lo que sería la previsualización del archivo -- y lo exporto como CSV. Como ustedes ven aquí ya está el archivo descargado en CSV. Procedimiento idéntico al anterior: marco la primera columna, voy a la pestaña datos, texto en columnas, le digo que sean delimitados, siguiente -- están delimitados por comas así que dejo el tilde en la coma-- siguiente y finalizar. Una vez que hice esto, voy a ver que los datos están perfectamente ordenados en una tabla.

[00:08:44] **Sandra Crucianelli** Este es uno de los procedimientos más sencillos en materia de extracción de datos. Por supuesto que hay otros más complejos, ¿sí? Y hay herramientas en que se utilizan programas específicos que permiten automatizar la descarga de los PDF. Y no solamente eso, sino que automatizar la extracción de datos en volúmenes grandes de información. Pero en esta primera aproximación a lo que es el periodismo de datos, para todos ustedes que no han manejado nunca estas herramientas, siempre pienso en quienes no han podido practicar mucho Excel, en los que quieren sumarse a esta disciplina del periodismo de datos o de periodismo de base de datos, siempre pienso en mostrarles herramientas que estén en español, básicamente en su interfase, pero además que sean gratuitas porque soy consciente, muy en especial en otros países de América Latina, que no todos tenemos la oportunidad de acceder a una licencia paga, a un plan de pagos para el uso de una herramienta que generalmente suele estar en dólares. Entonces, me parece que una buena manera de comenzar es empezar a practicar estas herramientas que son ¡gratuitas!, y de libre uso para todos.

[00:09:57] **Sandra Crucianelli** ¡Nos vemos en el próximo video!