Fuentes de Datos - 2

Otras fuentes de datos: archivos PDF

Los archivos PDF (formato de documento portátil, por sus siglas en inglés) son un tipo de archivo para almacenar documentos, que es independiente de la plataforma (Windows, Linux, etc.). Comenzó a utilizarse en 1993 (propiedad de Adobe Systems) y se publicó como un estándar abierto por la Organización Internacional de Estandarización (ISO) en 2008.

En muchos casos en los documentos PDF hay datos, en formato de tablas, que necesitamos. Además, muchas empresas u organismos presentan sus datos con este formato, por eso es necesario que podamos trabajar directamente con este tipo de archivos desde nuestro entorno de trabajo, en este caso RStudio.

Para obtener los datos, la forma de hacerlo es similar al scraping de una página web, pero trabajando directamente con texto plano, no sobre etiquetas del lenguaje en el que está programado, como es el caso del HTML.

Implementación en R

Para trabajar con archivos PDF vamos a utilizar las funciones del paquete **pdftools** [1]. Para abrir el archivo, la función *pdf_text()*,toma todos los elementos que contienen texto dentro del archivo y nos devuelve un vector de tipo “character”, donde el número de elementos corresponde al número de páginas del documento.

El parámetro necesario para la función es la ruta del archivo (**path**) y de forma opcional, la contraseña de propietario (**opw**) o de usuario (**upw**) si el archivo está protegido:

pdf_text(pdf = “dirección de archivo”, opw = “”, upw = “”)

La dirección, puede ser local o una URL para cargar un documento PDF disponible en una página web directamente.

Una vez que tenemos el archivo cargado, elegimos la página del documento que contiene la tabla que nos interesa, accediendo al elemento correspondiente del vector.

A partir de aquí, se empieza a trabajar con el texto plano que contiene el elemento, para darle la forma adecuada y obtener un conjunto de datos que podamos utilizar. Para ello primero serán útiles las funciones del paquete base *strsplit()*, *trimws()* y *grep()*.

strsplit(vector, split): divide el **vector** de caracteres en subcadenas de acuerdo a la coincidencia con la subcadena de caracteres **split**. Devuelve una

lista de la misma longitud que el vector original. **Split** puede ser una cadena o una expresión regular (una "expresión regular" es un patrón que describe un conjunto de cadenas) [2].

trimws(cadena): elimina los espacios en blanco al inicio y/o al final de la cadena de caracteres.

grep(patrón, vector): busca coincidencias con el **patrón** del argumento dentro de cada elemento del **vector** de caracteres. En patrón puede ser una cadena o una expresión regular. Devuelve un vector con los índices del vector donde hay coincidencias.

Además de las funciones del paquete base mencionadas, también vamos a utilizar funciones específicas para tratar texto del paquete **stringr** [3]. Este paquete tiene muchas funciones útiles, pero para los ejemplos presentados solamente utilizaremos dos: *str_split_fixed()* y *str_remove_all()*. La primera para dividir una cadena de caracteres según un patrón de interés en vectores de caracteres y la segunda para remover un patrón determinado de una cadena.

str_split_fixed(cadena, patrón, n): **n** es el número de vectores que queremos que nos devuelva.

str_remove_all(cadena, patrón)

En ambos casos, el patrón puede ser una cadena de caracteres o una expresión regular.

Actividades

1 – Cargue el archivo usbp_stats_fy2017_sector_profile.pdf que contiene datos de crímenes en Estados Unidos [4] y extraiga la tabla de datos de la primer página en formato de data.frame.

2 – Cargue el archivo de situación epidemiológica del Ministerio de Salud de Entre Ríos rep_covid-19_er_01_06_2022.pdf y obtenga del archivo PDF la tabla de “Localidad de Residencia” (páginas 5 a 8) y guarde los datos en un data.frame.

3 – Utilice el archivo pdf disponible en la URL: https://bancos.salud.gob.ar/sites/default/files/2022-10/BEN_621_SE_39.pdf para conseguir la tabla de datos de notificaciones de casos notificados y fallecidos por Región de la OMS, al 27-09-22.

Guarde los datos en un archivo CSV.

Referencias

1. Package pdftools. Disponible en: https://cran.r-project.org/web/packages/pdftools/pdftools.pdf
2. “Regular Expressions in R”, Albert Y. Kim. Disponible en: https://rstudio-pubs-static.s3.amazonaws.com/74603_76cd14d5983f47408fdf0b323550b846.html
3. Package stringr. Disponible en: https://cran.r-project.org/web/packages/stringr/stringr.pdf
4. Disponible en: https://github.com/jacobkap/crimebythenumbers/raw/master/data/usbp_stats_fy2017_sector_profile.pdf