Extracción de tablas de un PDF

Documento PDF

Un documento en PDF (Portable Document Format) es un archivo que está escrito bajo un formato de almacenamiento para documentos digitales independiente de plataformas de software o hardware. Este formato es de tipo compuesto (imagen vectorial, mapa de bits y texto). Debido a esta estructura el mismo posee una gran catidad de ventajas,

- Es multiplataforma, es decir, puede ser visto en los principales sistemas operativos (GNU/Linux, MacOS, Unix, Windows).
- Puede contener cualquier combinación de texto, elementos multimedia como vídeos o sonido.
- Es uno de los formatos más utilizados para compatir información.
- Su información puede ser cifrada con el fin de garantizar su seguridad.

Sin embargo, debido a esta misma estructura la tarea de extraer alguna información, ya sea texto o tablas, a partir de estos documentos se complica. Existen diferentes alternativas para lograr este fin, como lo son el uso de Adobe Acrobat o algún programa para este fin en internet, sin embargo en esta ocasión se explicará el proceso de extraer tablas de datos a partir de un PDF utilizando el software estadístico R, esto con el fin de poder guardarlos en un documento excel, o hacer cualquier tipo de análisis con los mismos.

Paquete Tabulizer

A continuación se explicará en detalle el proceso que permite la extracción de una tabla de datos de un PDF, con el fin de obtener esa información en el programa R, para así poder realizar cálculos o diferentes procesos sobre él. En principio este proceso podria realizarse manualmente creando un dataframe y transcribiendo los datos, pero esto no sería para nada óptimo.

Afortunadamente, existe un paquete que nos facilita esta tarea, el mismo lleva por nombre **Tabulizer**. Dentro de las funciones que cuenta este paquete están,

- **extract_metadata**: esta función permite extraer la metadata del PDF, es decir, información sobre el autor, fecha de creación, fecha de modificación entre otras.
- **extract_tables**: esta función permite extraer tablas que se encuentran en el PDF.
- **extract_text**: esta función permite extraer texto de un PDF y llevarlo a un string ó cadena de caracteres.
- **get_page_dims**: esta función permite conocer las dimensiones de la página del PDF en cuestión.
- **locate_areas**: esta función me permite localizar un área específica dentro del PDF.
- **make_thumbnails**: esta función me permite generar imágenes en miniatura de cada hoja del PDF.
- **split_pdf**: esta función me permite unir o separar las páginas de uno o varios PDF.

Cabe destacar que en esta ocasión sólo se va a explicar el uso de la funcion **"extract_tables"**.

Uso de la función extract_tables

Esta función cuenta con los siguientes argumentos,

- **file**: caracter que indica el nombre y la ruta del PDF.
- **pages**: entero que especifica la ó las páginas a extraer (opcional).

- **area**: lista que debe contener las coordenadas de la tabla a extraer (opcional).
- **columns**: lista que contiene un vector numérico de coordenas (x) sobre las cuales irán las columnas (opcional).
- **guess**: valor lógico (TRUE ó FALSE), que indica si se quiere o no adivinar la ubicación de las tablas en el PDF. Por defecto es TRUE.
- **method**: caracter que indica el método a aplicar para extraer la tabla ("decide", "lattice" ó "stream"). Por defecto se calcula usando el método "decide".
- **output**: caracter ("character","data.frame","csv","tsv","json","asis") que indica la salida o estructura que tendrán los datos de la tabla.
- **outdir**: caracter que indica la ruta del archivo en caso de que se seleccione "csv","tsv" ó "json" en el argumento anterior.
- **password**: caracter que indica la contraseña en caso de tener un archivo PDF con seguridad.
- **encoding**: caracter que especifica la codificación del PDF (opcional).
- **copy**: valor lógico, que indica si el archivo original se debe o no copiar antes de realizar los procesos sobre él.

Ejemplo 1

El PDF que se usará en este ejemplo se encuentra en la carpeta "examples" del paquete "tabulizer", el mismo se descarga automáticamente cuando se instala el paquete. Este PDF consta de tres páginas donde cada una cuenta con una tabla. Los siguientes comandos permiten extraer esas tablas.

```
#Cargo paquete
library(tabulizer)

#Genero ruta del archivo
f <- system.file("examples", "data.pdf", package = "tabulizer")

#Uso función para extraer las tablas
f1 <-  extract_tables(f,output = "data.frame")

#Muestro dataframe de la primera hoja
f1[[1]]
```

```
##     mpg cyl  disp  hp drat    wt  qsec vs am gear
## 1  21.0   6 160.0 110 3.90 2.620 16.46  0  1    4
## 2  21.0   6 160.0 110 3.90 2.875 17.02  0  1    4
## 3  22.8   4 108.0  93 3.85 2.320 18.61  1  1    4
## 4  21.4   6 258.0 110 3.08 3.215 19.44  1  0    3
## 5  18.7   8 360.0 175 3.15 3.440 17.02  0  0    3
## 6  18.1   6 225.0 105 2.76 3.460 20.22  1  0    3
## 7  14.3   8 360.0 245 3.21 3.570 15.84  0  0    3
## 8  24.4   4 146.7  62 3.69 3.190 20.00  1  0    4
## 9  22.8   4 140.8  95 3.92 3.150 22.90  1  0    4
## 10 19.2   6 167.6 123 3.92 3.440 18.30  1  0    4
## 11 17.8   6 167.6 123 3.92 3.440 18.90  1  0    4
## 12 16.4   8 275.8 180 3.07 4.070 17.40  0  0    3
## 13 17.3   8 275.8 180 3.07 3.730 17.60  0  0    3
## 14 15.2   8 275.8 180 3.07 3.780 18.00  0  0    3
## 15 10.4   8 472.0 205 2.93 5.250 17.98  0  0    3
## 16 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3
## 17 14.7   8 440.0 230 3.23 5.345 17.42  0  0    3
## 18 32.4   4  78.7  66 4.08 2.200 19.47  1  1    4
```

```
## 19 30.4   4  75.7  52 4.93 1.615 18.52  1  1    4
## 20 33.9   4  71.1  65 4.22 1.835 19.90  1  1    4
## 21 21.5   4 120.1  97 3.70 2.465 20.01  1  0    3
## 22 15.5   8 318.0 150 2.76 3.520 16.87  0  0    3
## 23 15.2   8 304.0 150 3.15 3.435 17.30  0  0    3
## 24 13.3   8 350.0 245 3.73 3.840 15.41  0  0    3
## 25 19.2   8 400.0 175 3.08 3.845 17.05  0  0    3
## 26 27.3   4  79.0  66 4.08 1.935 18.90  1  1    4
## 27 26.0   4 120.3  91 4.43 2.140 16.70  0  1    5
## 28 30.4   4  95.1 113 3.77 1.513 16.90  1  1    5
## 29 15.8   8 351.0 264 4.22 3.170 14.50  0  1    5
## 30 19.7   6 145.0 175 3.62 2.770 15.50  0  1    5
## 31 15.0   8 301.0 335 3.54 3.570 14.60  0  1    5
```

```
#Muestro dataframe de la segunda hoja
f1[[2]]
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
#Muestro dataframe de la tercera hoja
f1[[3]]
```

```
##     X Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
## 1 145          6.7         3.3          5.7         2.5 virginica
## 2 146          6.7         3.0          5.2         2.3 virginica
## 3 147          6.3         2.5          5.0         1.9 virginica
## 4 148          6.5         3.0          5.2         2.0 virginica
## 5 149          6.2         3.4          5.4         2.3 virginica
## 6 150          5.9         3.0          5.1         1.8 virginica
```

Es importante señalar que la función “extract_tables” me devuelve una lista donde en cada elemento se encontrará una tabla. Con el fin de ver la estructura y clase de cada columna se usa el comando “str”,

```
#estructura tabla 1
str(f1[[1]])
```

```
## 'data.frame':    31 obs. of  10 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : int  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : int  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : int  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : int  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: int  4 4 4 3 3 3 3 4 4 4 ...
```

```r
#estructura tabla 2
str(f1[[2]])
```

```
## 'data.frame':    6 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4
##  $ Species     : chr  "setosa" "setosa" "setosa" "setosa" ...
```

```r
#estructura tabla 3
str(f1[[3]])
```

```
## 'data.frame':    6 obs. of  6 variables:
##  $ X           : int  145 146 147 148 149 150
##  $ Sepal.Length: num  6.7 6.7 6.3 6.5 6.2 5.9
##  $ Sepal.Width : num  3.3 3 2.5 3 3.4 3
##  $ Petal.Length: num  5.7 5.2 5 5.2 5.4 5.1
##  $ Petal.Width : num  2.5 2.3 1.9 2 2.3 1.8
##  $ Species     : chr  "virginica" "virginica" "virginica" "virginica" ...
```

de esta manera se han extraido las tres tablas que posee este documento PDF.

Ejemplo 2

Para este ejemplo se usará un PDF localizado en una dirección específica, el mismo consta de dos hojas en las cuales hay información sobre las tablas de la distribución t-student.

```r
#Fijo ruta
ruta <- "http://cms.dm.uba.ar/academico/materias/1ercuat2015/probabilidades_y_estadistica_C/tabla_tstud

#Extraigo tablas
tabla <- extract_tables(ruta,output = "data.frame")

#Veo la estructura de la tabla
str(tabla)
```

```
## List of 2
##  $ :'data.frame':    51 obs. of  8 variables:
##   ..$ X  : chr [1:51] "rados de" "libertad" "1" "2" ...
##   ..$ X.1: num [1:51] NA 0.25 1 0.816 0.765 ...
##   ..$ X.2: num [1:51] NA 0.1 3.08 1.89 1.64 ...
##   ..$ X.3: num [1:51] NA 0.05 6.31 2.92 2.35 ...
##   ..$ t0 : num [1:51] NA 0.025 12.706 4.303 3.182 ...
##   ..$ X.4: logi [1:51] NA NA NA NA NA NA ...
##   ..$ X.5: num [1:51] NA 0.01 31.82 6.96 4.54 ...
##   ..$ X.6: num [1:51] NA 0.005 63.656 9.925 5.841 ...
##  $ :'data.frame':    51 obs. of  7 variables:
##   ..$ X50    : chr [1:51] "51" "52" "53" "54" ...
##   ..$ X0.6794: num [1:51] 0.679 0.679 0.679 0.679 0.679 ...
##   ..$ X1.2987: num [1:51] 1.3 1.3 1.3 1.3 1.3 ...
```

```
##   ..$ X1.6759: num [1:51] 1.68 1.67 1.67 1.67 1.67 ...
##   ..$ X2.0086: num [1:51] 2.01 2.01 2.01 2 2 ...
##   ..$ X2.4033: num [1:51] 2.4 2.4 2.4 2.4 2.4 ...
##   ..$ X2.6778: num [1:51] 2.68 2.67 2.67 2.67 2.67 ...
```

```
#Muestro tabla 1 extraida
head(tabla[[1]])
```

```
##          X    X.1    X.2    X.3      t0 X.4     X.5     X.6
## 1 rados de     NA     NA     NA      NA  NA      NA      NA
## 2 libertad 0.2500 0.1000 0.0500  0.0250  NA  0.0100  0.0050
## 3        1 1.0000 3.0777 6.3137 12.7062  NA 31.8210 63.6559
## 4        2 0.8165 1.8856 2.9200  4.3027  NA  6.9645  9.9250
## 5        3 0.7649 1.6377 2.3534  3.1824  NA  4.5407  5.8408
## 6        4 0.7407 1.5332 2.1318  2.7765  NA  3.7469  4.6041
```

```
#Muestro tabla 2 extraida
head(tabla[[2]])
```

```
##   X50 X0.6794 X1.2987 X1.6759 X2.0086 X2.4033 X2.6778
## 1  51  0.6793  1.2984  1.6753  2.0076  2.4017  2.6757
## 2  52  0.6792  1.2980  1.6747  2.0066  2.4002  2.6737
## 3  53  0.6791  1.2977  1.6741  2.0057  2.3988  2.6718
## 4  54  0.6791  1.2974  1.6736  2.0049  2.3974  2.6700
## 5  55  0.6790  1.2971  1.6730  2.0040  2.3961  2.6682
## 6  56  0.6789  1.2969  1.6725  2.0032  2.3948  2.6665
```

En este caso debido al título de la primera tabla que posee dos filas, se deben realizar algunas modificaciones sobre el resultado de la primera extracción,

```
#Asigno primera tabla
tabla1 <- tabla[[1]]

#Asigno nombre a tabla 1
names(tabla1) <- tabla1[2,]
names(tabla1)[1] <- "Grados de Libertad"

#Elinimo primera y segunda fila
tabla1 <- tabla1[-c(1,2),-6]

#Asigno segunda tabla
tabla2 <- tabla[[2]]
names(tabla2) <- names(tabla1)

#Muestro tablas
head(tabla1)
```

```
##   Grados de Libertad   0.25    0.1   0.05   0.025    0.01   0.005
## 3                  1 1.0000 3.0777 6.3137 12.7062 31.8210 63.6559
## 4                  2 0.8165 1.8856 2.9200  4.3027  6.9645  9.9250
## 5                  3 0.7649 1.6377 2.3534  3.1824  4.5407  5.8408
## 6                  4 0.7407 1.5332 2.1318  2.7765  3.7469  4.6041
## 7                  5 0.7267 1.4759 2.0150  2.5706  3.3649  4.0321
## 8                  6 0.7176 1.4398 1.9432  2.4469  3.1427  3.7074
```

```r
head(tabla2)
```

```
##   Grados de Libertad   0.25    0.1   0.05  0.025   0.01  0.005
## 1                 51 0.6793 1.2984 1.6753 2.0076 2.4017 2.6757
## 2                 52 0.6792 1.2980 1.6747 2.0066 2.4002 2.6737
## 3                 53 0.6791 1.2977 1.6741 2.0057 2.3988 2.6718
## 4                 54 0.6791 1.2974 1.6736 2.0049 2.3974 2.6700
## 5                 55 0.6790 1.2971 1.6730 2.0040 2.3961 2.6682
## 6                 56 0.6789 1.2969 1.6725 2.0032 2.3948 2.6665
```

De esta forma se ha extraido de manera sencilla las tablas provenientes del PDF en cuestión, sin tener la necesidad de utilizar un software pago o alguna herramienta de internet. Para este ejemplo sólo se usó la función "extract_tables", sin embargo mediante el uso de este paquete es posible realizar más cosas interesantes, como lo son,

- Extraer texto de una página en específico.
- Separar o unir páginas de uno o varios PDF.
- Obtener metadata de un documento en específico.
- Obtener el número de páginas, así como abtener el ancho y el alto de una página de un PDF.
- Convertir en imagen una página en específico del documento PDF.

Como próximos pasos, se propone extraer texto desde una página en específico, ó la conversión a imagen de una página de un PDF. Para mayor comodidad puede revisar mi repositorio en el siguiente enlace.