Taller de Introducción a la Minería de Datos

Universidad de Córdoba

16/09/2015

1. La American Community Survey distribuye los datos de manera descargables sobre las Comunidades de Estados Unidos. Descargue la encuesta de 2006 microdatos sobre la vivienda para el estado de Idaho utilizando `download.file ()` desde aquí:

https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv.

y cargue los datos en R. El libro de códigos, que describe los nombres de variables está aquí:

https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FPUMSDataDict06.pdf

¿Cuántas propiedades valen $ 1.000.000 o más?

2. Leer los datos XML sobre restaurantes de Baltimore desde aquí:

https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Frestaurants.xml

¿Cuántos restaurantes tienen código postal 21231?

3. Cuántos caracteres se encuentran en las líneas 10, 20, 30 y 100 del HTML de esta página:

http://biostat.jhsph.edu/~jleek/contact.html

(Sugerencia: la función `nchar()` de R puede ser útil)

4. Leer este conjunto de datos en R y reportar la suma de los números en la cuarta de las nueve columnas.

https://d396qusza40orc.cloudfront.net/getdata%2Fwksst8110.for

Fuente original de los datos: http://www.cpc.ncep.noaa.gov/data/indices/wksst8110.for

(Sugerencia este es un formato de archivo de ancho fijo)

5. La American Community Survey distribuye los datos de manera descargables sobre las Comunidades de Estados Unidos. Descargue la encuesta de 2006 microdatos sobre la vivienda para el estado de Idaho utilizando `download.file ()` desde aquí:

https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv.

y cargue los datos en R. El libro de códigos, que describe los nombres de variables está aquí:

https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FPUMSDataDict06.pdf

Cree un vector lógico que identifique los hogares de más de 10 acres que vendieron más de $ 10,000 dólares en productos agrícolas. Asignar ese vector lógico a la variable agricultureLogical. Aplique la función `which()` para identificar las filas del data frame donde el vector lógico es TRUE. `which(agricultureLogical)` ¿Cuáles son los 3 primeros valores que resultan?