Preparación de datos con R: ejercicios

Carlos J. Gil Bellosta

2014-12-19

Los ejercicios tienen que resolverse individualmente. No son sencillos: parte de ellos están inspirados en problemas prácticos reales. Por eso puedes ayudarte de cualquier tipo de instrumento (Google, blogs, libros, etc.) que estaría a tu alcance en tu trabajo. Eso sí, en las soluciones que envíes, indica los recursos que utilices y deja clara cuál es tu aportación en cada caso.

No es necesario que envies código. Puedes presentarlos en, p.e., entradas en tu bitácora (si tienes y quieres utilizarla), en repositorios de GitHub o similares. En caso de que envíes codigo, trata de usar R Markdown y que el documento sea reproducible (i.e., que yo pueda recrear la salida).

El envío de los ficheros se hará por medio del Blackboard. La fecha límite para puntuar sobre 10 es el domingo 18 de enero de 2015. A partir de entonces, por cada semana de retraso, se descontarán dos puntos sobre los obtenidos. Es decir, la máxima nota para un ejercicio entregado el 5 de enero sería 8; para uno entregado el 13 de enero seria 6, etc.

A pesar de que hay 12 puntos en juego, las prácticas se evaluarán sobre 10, que es la máxima nota.

Clases S4 y mapas (2 puntos)

Obtén *shapefiles* de, p.e., provincias españolas (el INE los proporciona). Luego, obtén datos de algún tipo de estadística asociada a dichas entidades (población, tasa de desempleo, etc.). Con esos dos elementos, crea un `SpatialPolygonsDataFrame` y represéntalos gráficamente.

Trata de hacerlo *reproduciblemente* (¿con R MarkDown?).

JSON & XML (2 puntos)

Busca un proveedor de datos via API que te interese. Descarga unos datos, procésalos y crea una pequeña historia a su derredor.

Trata de que la obtención de los datos, etc. tenga *su dificultad*: se valorarán tanto el tratamiento de los datos como su interés o motivación.

Hazlo *reproduciblemente* (¿con R Markdown?)

Web scraping, texto y fechas (2 puntos)

Descarga información bursátil de `http://goo.gl/yD2Bwb` y crea un `data.frame` a partir de él con la misma información que en la tabla que aparece en la página. Convierte las columnas numéricas a número, etc., las fechas a fecha, etc.

Ten en cuenta que la columna hora puede tener dos tipos de información: la hora durante las horas de operación de los mercados y el día en formato dd/mm/aaaa fuera de mercado. Que tu código tenga en cuenta esa circunstancia.

Alternativamente, si encuentras algunos datos de tu interés que te interese procesar y tengan un nivel de dificulad similar al anterior, úsalos en lugar de los propuestos.

El código que envíes tiene que ser capaz de crear la tabla a la hora en que lo ejecute.

plyr, dplyr y data.table (2 puntos)

Descarga los microdatos del censo del 2011 de `http://goo.gl/guhG1M`. Puedes bajar el nacional o, si tienes problemas de memoria, alguno de los regionalizados. En esa página hay también información sobre las variables contenidas en el fichero y su formato. Puedes leerlo en R usando el paquete MicroDatosEs. Consulta la ayuda de la función `censo2010`.

El ejercicio consiste en identificar algunas variables de tu interés y construir tablas por los indicadores que creas convenniente. Por ejemplo, población por sexo y grupo de edad en cada provincia. O proporción de viudos y viudas por tramo de edad y provincia. O...

Eso sí: hazlo usando los dos paquetes `dplyr` y `data.table` (¿y `reshape2`?). Trata también con `plyr`. Añade un comentario sobre la velocidad relativa de los distintos paquetes para procesar los datos. ¿Cuál es tu favorito?

Nota: Una columna muy importante en el censo es el factor de elevación. Para contar la población de España habria que hacer `sum(factorel)` donde `factorel` es el nombre que podría recibir esa columna. Hay un ejemplo práctico de cómo usar el factor de elevación en `http://goo.gl/U6Ys8W`.

RHadoop (2 puntos)

Sube un subconjunto de datos del censo del ejercicio anterior (¿100k líneas?) a Hadoop y haz una tabulación de variables de tu interés del censo usando `mapreduce`.

Ejercicio extra (2 puntos)

Crea un paquete de R (con dos o tres funciones tontas, que hagan cualquier cosa). Súbelo a GitHub. Se valorará que el paquete pueda ser instalado usando `devtools`.