Más velocidad y menos colapsos: preprocesamiento de archivos con utilidades del sistema operativo

La creciente disponibilidad de datos masivos fácilmente puede sobrepasar nuestra paciencia, o la capacidad de procesamiento de nuestros computadores. Sin embargo, es posible reducir los tiempos de ejecución y el uso de memoria en la etapa inicial de un análisis, filtrando y procesando archivos de texto delimitado antes de importarlos a R, usando herramientas especializadas - simples pero de alta eficiencia.

Desde la versión 1.2.0 de R, es posible ejecutar comandos del sistema operativo desde R, tendiendo disponibles los resultados de salida de estos comandos para leerlos e incorporarlos al flujo de trabajo. Una de estas conexiones, creada con la función nativa ‘pipe’, permite ejecutar utilidades de terminal dentro de otras funciones y acceder a sus resultados. Para trabajar con archivos de texto delimitado, ‘awk’ es una utilidad de consola y un lenguaje de análisis semántico. Aunque data de los años setenta, hoy en día sigue siendo una opción rápida y eficiente para trabajar con archivos de texto de gran tamaño.

Este trabajo muestra como podemos especificar condiciones para descartar registros, de manera similar a la que muchos ya conocemos gracias al paquete ‘dplyr’. Por ejemplo: filtrar valores en una columna mayores o menores a algún valor particular, descartar valores NA, o seleccionar columnas de interés usando awk durante el paso de lectura de archivos con las funciones `read.csv` de R Base o `read_csv` del paquete readr. Se presentan ejemplos, con datos espaciales de biodiversidad y con estadísticas deportivas, iterando múltiples archivos, y en archivos individuales con millones de registros. El tiempo para filtrar e importar archivos tiende a reducirse a la mitad comparado con el mismo filtrado de registros dentro de R, con una importante mejoría en el uso de memoria de hasta 700%, que en computadores portátiles actuales puede ser la diferencia entre perder o no la sesión ante un colapso del sistema por falta de memoria.