Lo que aprendí sobre R y reproducibilidad durante mi doctorado

Anónimo

Palabras clave: RMarkdown - Reproducibilidad - Trabajos científicos

En el ámbito académico R es usado para análisis estadístico y visualización de datos entre otras tantas actividades, pero además permite generar trabajos científicos y otros documentos académicos, publicarlos y compartirlos en distintos formatos. Al mismo tiempo R y su ecosistema de paquetes provee las herramientas necesarias para hacer que esos trabajos sean reproducirles.

El objetivo de esta presentación es compartir el conjunto de herramientas y paquetes de R que utilicé a lo largo de mi doctorado para hacer análisis y visualizaciones de datos, pero en particular para compartir resultados y escribir trabajos científicos y la tesis final. Con esta presentación espero dar a conocer estas herramientas y mostrar como utilizarlas juntas para hacer un trabajo científico reproducible y abierto. Pero estas herramientas y procesos son extensibles a otros ámbitos, por lo que puede ser de utilidad para otros usuarios y usuarias de R.

Los primeros años

Los primeros años de doctorado están plagados de pruebas y errores, explorando las herramientas disponibles para trabajar y definiendo cómo queremos hacerlo. En esta primera etapa organicé todo mi trabajo en un **proyecto de R** que incluía reportes con análisis de datos, visualizaciones y notas sobre las ideas que iba desarrollando. Estos reportes utilizaban el formato **RMarkdown** que me permitía incluir código y texto en un mismo archivo. También eran prácticos para mostrar a otras personas, ya sea localmente desde mi computadora o desde un repositorio en **GitHub**. En este proyecto también incluía **scripts** de R con algunas funciones o código que generaba y usaba para procesar o analizar datos. Si bien esta estrategia funcionó bien al principio, con el paso del tiempo empezó a mostrar desventajas. Por ejemplo, cargar todas las funciones de un script en cada reporte no era muy práctico y recordar como se usaba cada función sin contar con una documentación apropiada era casi imposible.

Mi primer trabajo científico

Comenzar a escribir mi primer trabajo científico me dio la oportunidad de comenzar un proyecto desde cero, aplicando lo que aprendí. En primer lugar inicié el nuevo proyecto como un **paquete de R**, si bien es posible que las funciones que escribí para hacer mi análisis de datos no sean útiles para muchas otras personas, tener el código organizado de esa manera, incluyendo documentación y ejemplos, me permitió trabajar mejor los años subsiguientes. Para todo esto utilicé el paquete **rrtools** que permite generar proyectos con la estructura de carpetas necesarias para manejar un paquete de R y al mismo tiempo un trabajo científico. Por supuesto, utilicé RMarkdown para escribir el trabajo pero además aproveché el paquete **rticles** que incluye plantillas de muchas revistas científicas para darle el formato necesario. Pensando en la reproducibilidad de mi trabajo comencé a utilizar **renv** para mantener un registro de los paquetes de R que se utilizaban en el proyecto y sus versiones. Esto fue imprescindible al cambiar de computadora y recuperar el ambiente de trabajo. Utilicé otras herramientas, no tan relacionadas con R, por ejemplo **Zenodo** para alojar los datos que utilicé (que se pueden descargar programáticamente con R) y **GitHub** para trabajar con control de versiones.

La tesis

Podríamos pensar que una tesis de doctorado es como la suma de todo el trabajo previo, pero suele ser mucho más. Para escribir mi tesis de doctorado aproveché todo lo que venía haciendo pero al mismo tiempo

incorporé algunas herramientas nuevas. Gracias al trabajo previo, tenía un paquete con las funciones principales que necesitaba, un conjunto de scripts de R para hacer el pre-procesamiento de mis datos y muchas notas y un trabajo científico en formato RMarkdown. Gracias a la versatilidad de RMarkdown, transferir el trabajo previo a un nuevo proyecto para generar un producto distinto no requirió mucho esfuerzo. En ese tiempo descubrí el paquete **thesisdown**, que usa **bookdown** y plantillas para generar la tesis en pdf, página web y otros formatos. Y, si bien usar esto requirió trabajar sobre la plantilla de LaTeX y de html para adaptarlas al formato requerido por la universidad, poder generar la tesis en estos formatos aprovechando las maravillas de RMarkdown fue una gran ventaja. La infraestructura de estos paquetes permiten convertir un conjunto de archivos RMarkdown que incluyen todo el texto y código para generar la tesis final tanto en el formato oficial de la universidad en pdf como en una web abierta a todo el mundo. Finalmente, todas estas herramientas permiten incluir el código que genera los resultados y disponibilizar los datos en un repositorio público; pasos necesarios hacia la reproducibilidad y la ciencia abierta.

Conclusiones

A lo largo de todo este proceso aprendí, en primer lugar, que casi nunca es necesario volver a inventar la rueda; existen una infinidad de paquetes y herramientas disponibles para usar. En segundo lugar, que vale la pena compartir de manera abierta lo que hacemos para que otras personas puedan aprovechar nuestro trabajo y experiencia. Y finalmente, que la comunidad de R juega un rol importantísimo a la hora de descubrir nuevos paquetes y herramientas, entender como otras personas trabajan y como esa experiencia puede ayudarnos.

Aprender a usar nuevas herramientas al mismo tiempo que estudiamos, hacemos un doctorado o continuamos con un trabajo no es simple. Muchas veces seguir utilizando aquello que conocemos y que *más o menos* resuelve el problema implica dedicar menos tiempo y esfuerzo a la tarea en ese momento. Sin embargo incorporar nuevas herramientas puede significar ahorrar tiempo en el futuro, mejorar el resultado final o tal vez hacer nuestro trabajo más reproducible o abierto y, en mi experiencia, creo que vale la pena el esfuerzo.