Taller de programación en R: Taller 3

Facultad de Economía, Universidad de los Andes

Oct 14/2024

**Instrucciones del trabajo:**

El presente taller busca que se apliquen los conceptos vistos durante las clases Magistrales de las semanas 9-13 del curso. Adicional a esto, con los ejercicios propuestos se busca desarrollar una mejor comprensión del uso de R, de los paquetes vistos en clase y fomentar un razonamiento lógico que lleve a los estudiantes a comprender e implementar de una buena manera los tópicos tratados.

El Taller 4 debe ser diligenciado en parejas (sin excepciones) y entregado el domingo 10 de Noviembre a las 11:59 pm a través de la plataforma de Bloque Neón. La copia o plagio en este trabajo está totalmente prohibida e incurrir en esta práctica conlleva a una nota de cero (0) en el taller, así como a las sanciones correspondientes tenidas en cuenta en el reglamento de estudiantes de la universidad.

Cada pareja debe entregar:

* Un script de R donde presente el código que realizó para solucionar el mismo.
* En un archivo .pdf, docx, entre otros, un documento donde realice las interpretaciones necesarias para cada punto.

En caso de que el **punto no requiera algún desarrollo además del código**, pueden **explicar esto en el documento.** Finalmente, los scripts deben estar organizados por puntos y subpuntos.

**Punto 1 (2.0):**

* 1. Descarguen de la página del DANE [Geoportal DANE- Página de descarga datos geoestadisticos](https://geoportal.dane.gov.co/servicios/descarga-y-metadatos/datos-geoestadisticos/?cod=111) un shapefile a nivel municipal de Colombia. Cargue este shapefile a R utilizando la librería SF y elimine San Andrés y Providencia.
  2. De la pagina del CEDE descargue el Panel de Características generales de los Municipios. Mantengan las variables que contienen la palabra “pob” y las variables del PIB.
  3. Realicen un merge del shapefile de los municipios con este dataframe, revisen dos cosas. Primero, que en ambas bases los municipios estén identificados con el mismo número -revisen los 0´s al inicio-. Segundo, que para la base resultante esté la geometría de cada unidad de nivel -municipio, depto, etc.- para cada año.
  4. Realicen una visualización de un mapa a nivel municipal con la simbología de la población rural por municipio para el año 2009. Añadan todos los elementos estéticos que permitan una gráfica autocontenida y bien presentada. Hagan este mapa interactivo usando la librería ggplotly.
  5. Interpreten la gráfica y cuenten una historia acorde a la misma.

**Punto 2 (3.0):**

## Vayan a la página [All products | Books to Scrape - Sandbox](https://books.toscrape.com/) la cual tiene información sobre libros en internet y su precio.

* 1. Utilizando herramientas de web-scrapping, estructuren un dataframe donde las filas sea cada uno de estos libros y las columnas sean el respectivo título y el precio.
  2. Ejecuten un procesamiento del texto presente en los títulos. Para esto, remuevan las stop-words, pasen todo el texto a minúsculas y eliminen los caracteres especiales y los números.
  3. Guarden esta base de datos como datos\_limpios.
  4. A partir de la base datos\_limpios generen una base que se llame count la cual contiene una fila con el n-grama definido y una columna con la frecuencia de veces que aparece ese n-grama.
  5. Realicen una nube de palabras que permita entender cuáles son los n-gramas más frecuentes en los títulos de los libros. Interpreten este resultado.
  6. Por otra parte, a partir de la base de datos\_limpios, realicen una matriz de term frequency TF, la cual tiene en las filas cada uno de los libros y en las columnas cada una de las palabras presentes en el corpus, el valor de esta columna es el número de veces que aparece el n-grama.
  7. Añadan a esta base la columna de precio del libro.
  8. Elijan 10 palabras (las que quieran), que en esta matriz estarán en las columnas, y reemplacen la frecuencia por un 1 si aparece más de una vez o un 0 si no aparece.
  9. Utilizando la función group\_by() calculen para estas palabras el precio de los libros cuando estas palabras aparecen y cuando no. Es decir, cuando la variable toma el valor 1 vs 0.
  10. Presenten una visualización en forma de gráfico de barras tal que en el eje Y tenga la palabra y en el eje X el valor del libro. Añadan a la gráfica una coloración por si aparece o no aparece en el título (si la es 0 o 1). Interpreten la gráfica ¿Cuáles palabras elegidas parecen estar asociadas positivamente con el precio del libro?