Taller de programación en R: Taller 3

Facultad de Economía, Universidad de los Andes

Oct 14/2024

**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.**

Este punto no requiere explicación, el procedimiento está explícito en el código.

* 1. **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.**

Este punto no requiere explicación, el procedimiento está explícito en el código.

* 1. **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**.

Este punto no requiere explicación, el procedimiento está explícito en el código.

* 1. **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.**
  2. **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.**

Este punto no requiere explicación, el procedimiento está explícito en el código.

* 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.**

Este punto no requiere explicación, el procedimiento está explícito en el código.

* 1. **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.**

Este punto no requiere explicación, el procedimiento está explícito en el código.

* 1. **Guarden esta base de datos como datos\_limpios.**

Este punto no requiere explicación, el procedimiento está explícito en el código.

* 1. **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.**

Este punto no requiere explicación, el procedimiento está explícito en el código.

* 1. **Realicen una nube de palabras que permita entender cuáles son los n-gramas más frecuentes en los títulos de los libros. Interpreten este resultado.**



* 1. **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.**

Este punto no requiere explicación, el procedimiento está explícito en el código.

* 1. Añadan a esta base la columna de precio del libro.

Este punto no requiere explicación, el procedimiento está explícito en el código

* 1. 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.

Para seleccionar las 10 palabras primero seleccionamos las a las palabras que tienen una frecuencia mayor a 10 y luego escogimos 10 palabras aleatorias dentro de esta muestra.

* 1. 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.

Este punto no requiere explicación, el procedimiento está explícito en el código.

* 1. 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?

Gráfico, Gráfico de barras

Descripción generada automáticamente

En el gráfico las palabras que parecen estar asociadas positivamente con el precio de los libros son: “girl”, “black” y “vol”, ya que en promedio el precio de los libros que contienen esas palabras es mayor que el precio de los libros que no la contienen.