2. Text Mining

Fecha: 6 de octubre de 2020

Docente: Enrique Rendón C.

Programa de Educación Continua y Permanente
Centro de Investigaciones para el Desarrollo - CID
Facultad de Ciencias Económicas

Temática

- Text Mining con R
- Bag of words
- Topics Modeling
- Tagging

¿Qué es Text Mining?

Podríamos decir que es el conjunto de técnicas y tecnologías usadas para analizar grandes volúmenes de materiales textuales con el fin de capturar conceptos y temas clave, descubriendo así, relaciones, tendencias ocultas, o reglas que explican el comportamiento del texto.

Text Mining usando Tidyverse

Para iniciar nuestros primeros pasos en text mining, necesitaremos primero instalar el paquete Tidyverse, este paquete incluye otros paquetes básicos fundamentales para realizar text mining, por ejemplo contiene: ggplot2, tibble, dplyr, readr, entre otros.

- Tidyverse es un framework, es el dogma conformado por una variedad de paquetes de R que comparte dicha visión.

- "comparten una filosofía de diseño subyacente, gramática y estructuras de datos“.

Text Mining usando Tidyverse

- Library(tidyverse)

```
[Workspace loaded from ~/.RData]

> library(tidyverse)
-- Attaching packages --------------------------------------- tidyverse 1.3.0 --
v ggplot2 3.3.2     v purrr   0.3.4
v tibble  3.0.3     v dplyr   1.0.2
v tidyr   1.1.2     v stringr 1.4.0
v readr   1.3.1     v forcats 0.5.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
>
```

Importar y revisar la data

- Readr: paquete que nos permitirá cargar y revisar la data.
- `Variable <- read.csv("NombreArchivo")`

- Hagamos el ejercicio de buscar el promedio del rating de un director según sus películas.

Tokenizar y limpiar

- `library(tidytext)`
- Bag of words: las palabras en un documento son independientes, se puede determinar cual es el criterio para partir el texto.
- Cada grupo separado de textos es un documento.
- Cada palabra única es un término.
- Las repeticiones de un término es un token.
- Crear una bolsa de palabras es conocido como Tokenizar.

Tokenizar y limpiar

- **unnest_tokens():** es la función para Tokenizar.
- Count(): manera sencilla de leer y contar las palabras.
- Stop Words: son las palabras comunes, como los conectores, que no agregan valor a nuestra bolsa de palabras.
- anti_join: función que nos permitirá filtrar la información usando un segundo dataframe de referencia, si se una una variable usamos la función filter()
- Tidytext incluye un dataframe que se llama stop words que nos servirá para iniciar la limpieza de las palabras que nos sirven.

Graficar el conteo de palabras

Usaremos:

- ggoplot()
- geom_col()
- coord_flip()
- ggtitle()

Pulir el conteo de palabras

¿Que hacemos cuando stop_words no es suficiente?

Creemos un data frame personalizado con las palabras que queremos excluir:

- tribble()

Ahí podemos añadir todas las palabras que queremos excluir de nuestro análisis.

Y lo unificamos creando un data frame mas completo:

- bind_rows()

Organizando diferentes graficas

Hagamos el ejercicio de organizar las palabras mas utilizadas utilizando como criterio el año:

- top_n()
- ungroup()

Luego hagamos una comparación por año de las palabras mas usadas con:

- facet_wrap()

Organizando diferentes graficas

Hagamos la grafica mas amigable para los demás con algunas modificaciones como:

- show.legend = False
- coord_flip()
- scales = “free”

¿Qué otros cambios se te ocurre para la grafica?