Análisis Inteligente de Datos
Maestría en Explotación de Datos y Gestión del Conocimiento
Universidad Austral - Sede Rosario
Año 2022

Tarea N° 1

Limpieza y transformación de datos

Presentación del Problema

La primera etapa en todo análisis de datos consiste en garantizar, mediante la lectura de archivos y el ordenamiento de los datos, que la información disponible se encuentre de forma organizada y limpia, lista para transformarse en la materia prima de las etapas posteriores de análisis. El objetivo de esta tarea es poner en práctica herramientas para la recopilación de datos que, si bien provienen de diversos tipos de archivos, hacen referencia a un mismo problema.

Trabajaremos con un conjunto de datos que contiene información sobre 45 discos, junto con sus respectivas canciones, del icónico músico argentino Charly García. La información se presenta en una serie de archivos comprimidos alojados en `datos_charly.zip`.

Objetivo

Leyendo e integrando los datos provistos, deberán **generar un único conjunto de datos** con 578 observaciones (canciones) y 22 variables (mencionadas más adelante) y guardarlo en un archivo de texto separado por tabulaciones llamado `resultado.txt`.

Materiales a Entregar

Se debe entregar el archivo `resultado.txt` y una de las siguientes opciones:

a. Un único script de R con el código que lee cada uno de los archivos necesarios y produce la salida mencionada, llamado `limpieza.R`.
b. *Sugerido*. Un único script de Rmarkdown (`limpieza.Rmd`) que lee cada uno de los archivos necesarios y produce la salida mencionada y el archivo resultante de su compilación (`limpieza.html` o `limpieza.pdf`).

Aspectos a considerar

Los siguientes aspectos deben ser tenidos en cuenta y formarán parte de la evaluación:

- El código debe ser escrito de forma prolija. Se pueden observar ejemplos de la guía de estilo vigente en RStudio y Google para R en este enlace.
- El código estar bien documentado, explicando cuando no sea obvia qué acción se realiza. En el script de R o dentro de los *chunks* de Rmd se puede emplear el símbolo # para agregar comentarios. Si usan Rmd, pueden organizar con secciones y describir con texto los pasos a seguir, intercalado con código.
- El código debe poder ejecutarse sin errores. Esto implica, entre otras cosas, que deben informarse los paquetes utilizados.
- El código debe correrse teniendo como *working directory* la carpeta en la cual el script `.R` o el `.Rmd` se encuentra guardado, donde también se asume que se encuentra el archivo `datos_charly.zip`. **No deben incluirse** rutas o directorios de carpetas personales en el script final, toda ruta debe referirse a dicho *working directory*.
- El archivo `resultado.txt` debe tener la información correcta (no importa el orden de las filas o de las columnas, pero los nombres de las columnas deben coincidir con aquellos indicados en la tabla de descripción de los datos).
- Puede haber valores faltantes en columnas para las cuales algunos álbumes no presentan información.
- Se admite la entrega del resultado excluyendo los dos discos de la banda "La Máquina de Hacer Pájaros" (en ese caso, el conjunto de datos final tendrá 15 canciones menos y se deducirá algo en la calificación).

Reglas

- El trabajo es grupal, con un máximo de 3 integrantes por grupo. Deben informar la composición de los grupos en esta planilla, para poder configurar la entrega en el aula virtual.
- La fecha límite de entrega es el viernes 11/11 a las 9:00 mediante el buzón de entrega del aula virtual.
- Se promueve la discusión entre estudiantes y docentes, pero no se permite compartir código con la solución de los problemas.

Descripción de los datos

La popular aplicación Spotify registra numerosas variables para cada uno de los discos disponibles en su servicio de streaming, por ejemplo: nombre del álbum, nombre del artista, cantidad de tracks, códigos de identificación únicos, etc. Por otro lado, para cada canción registra variables que tratan de describir características musicales o sobre la *intención* de la pieza (por ejemplo, *energía*, *instrumentalidad*, *duración*, etc.). Este tipo de información está disponible en el **portal de Spotify para desarrolladores** y puede descargarse usando R mediante el paquete **Rspotify**.

El conjunto de datos con el que vamos a trabajar contiene información sobre 45 discos, junto con sus respectivas canciones (578 en total), del icónico músico argentino **Charly García**. Las bases se presentan en una serie de archivos comprimidos alojados en `datos_charly.zip`, los cuales fueron descargados mediante Rspotify y *desorganizados* con fines didácticos.

La información sobre cada uno de los álbumes se encuentra en el archivo `albums.xlsx`, el cual contiene las siguientes variables:

Variable	Tipo de Registro	Descripción
id	caracter	Código de identificación del álbum.
name	caracter	Nombre del álbum.
album_type	caracter	Tipo de álbum.
artist	caracter	Artista del álbum. Aquellos discos que corresponden a su carrera solista están indicados como "Charly García"; los restantes presentan el nombre de la banda que publicó el disco, de la cual Charly era integrante.
categoria	caracter	Clasificación del álbum con respecto al tipo de lanzamiento (estudio/vivo/recopilatorio, y además solista/banda).

A su vez, la información sobre las canciones de cada uno de estos álbumes se encuentra en los restantes archivos, que difieren en su formato y modo de presentación de los datos:

Archivo	Contenido
suigeneris.csv	Datos sobre cada canción de 11 discos de la banda Sui Generis.
serugiran.xlsx	Datos sobre cada canción de 9 discos de la banda Serú Girán.
porsuigieco.txt	Datos sobre cada canción del único disco de la banda PorSuiGieco.
bbatj.sas7bdat	Datos sobre cada canción del único disco de la banda Billy Bond and The Jets, guardados como dataset de SAS.
lmdhp	Carpeta que contiene otras dos subcarpetas, una para cada uno de los discos de la banda La Máquina de Hacer Pájaros. Cada subcarpeta contiene un archivo para cada canción del álbum.
solista	Carpeta que contiene 21 archivos de texto, cada uno con datos de las canciones de 21 discos de Charly como solista.

La siguiente tabla presenta todas las variables registradas sobre las canciones. Es importante notar que algunas de ellas están ausentes para ciertos discos o que las mismas variables pueden tener nombres distintos en diferentes archivos. Estas variables son conocidas como *song features* y **una descripción ampliada de su significado puede consultarse acá**.

Variable	Tipo de Registro	Descripción
name	caracter	Nombre de la canción.
track_number	numérico	Número de orden de la canción en el disco al que pertenece.
disc_number	numérico	Número de disco (pertinente en colecciones).
album_name	caracter	Nombre del disco al que pertenece la canción.
album_id	caracter	Código de identificación del disco al que pertenece la canción.
album_artist	caracter	Artista del disco al que pertenece la canción.
id	caracter	Código de identificación de la canción.
danceability	numérico	"Bailabilidad". Medida entre 0 y 1, describe si la canción es adecuada para bailar (no = 0, sí = 1) de acuerdo a características musicales como tempo, estabilidad rítmica, fuerza del pulso, etc.
energy	numérico	Energía. Medida entre 0 y 1, describe intensidad y actividad. Por ejemplo: heavy metal tiene valor alto y Bach valor bajo.
key	numérico	Tono de la canción: 0 = Do, 1 = Do sostenido, 2 = Re, etc. Toma valores enteros entre 0 y 11. Es una variable categórica (ordinal).
loudness	numérico	Volumen promedio en decibeles.
mode	numérico	Modo (0 = menor, 1 = mayor). Es una variable dicotómica.
speechiness	numérico	Nivel de Oralidad. Medida entre 0 y 1, detecta la presencia de palabra hablada. Por ejemplo: un podcast tendría puntaje alto.
acousticness	numérico	Nivel de Acústica. Medida entre 0 y 1, detecta si la canción es acústica (valor alto significa alta chance de serlo).
instrumentalness	numérico	Instrumentalidad. Medida entre 0 y 1, detecta si la canción es instrumental. Cuanto menos sonido vocal tiene, más alto es este valor.
liveness	numérico	Vivacidad. Medida entre 0 y 1, detecta si la canción fue grabada en vivo. Por ejemplo, canciones extraídas de un recital tienen valores altos.
valence	numérico	Positividad. Medida entre 0 y 1, detecta el nivel de positividad. Canciones con valores alto suenan más positivas (alegres, eufóricas) y canciones con valores cercanos a 0 suenan más negativas (tristes, apagadas).
tempo	numérico	Tempo promedio (pulsos/minuto).
duration_ms	numérico	Duración de la canción en milisegundos.
time_signature	numérico	Compás.
uri	caracter	URI (identificador) de Spotify para la canción.
analysis_url	caracter	URL para acceder al análisis completo de la canción.