CURSO: MINERÍA DE DATOS

MAESTRÍA EN EXPLOTACIÓN DE DATOS Y DESCUBRIMIENTO DE CONOCIMIENTO

TRABAJO PRÁCTICO ENTREGABLE I

Preprocesamiento de datos y Bases de datos NoSQL

INTRODUCCIÓN

En este primer Trabajo Práctico entregable del curso, se integrarán los conocimientos relacionados con el preprocesamiento de datos, integración, construcción de variables y gestión de datos mediante una Base de Datos NO-SQL orientada a documentos.

Para la exploración de estos temas, se utilizará el IDE R-Studio del lenguaje de programación R y la Base de Datos MongoDB.

Los datos fueron generados con un script R[1] que descarga *charts* desde SpotifyCharts[2] y luego para cada artista en los charts obtiene algunos datos desde la API de Spotify[3].

OBJETIVO GENERAL

El objetivo general de este trabajo es realizar un *análisis exploratorio* del dataset y el posterior *preprocesamiento*, de acuerdo a las técnicas vistas en clase para entender las relaciones existentes entre algunas variables del dataset.

Si bien el trabajo tiene un carácter netamente exploratorio y se definen las consignas de manera abierta, se evaluará la aplicación de las técnicas vistas en clase así como también el carácter innovador de la solución propuesta.

ACERCA DE LOS DATOS

El conjunto de datos con el que se realizará el trabajo es un backup de MongoDB y está en formato JSON. En total son tres colecciones de MongoDB que contienen: charts, artistas y obra de los artistas.

[1] download-spotify_template.R
[2] https://spotifycharts.com/
[3] https://developer.spotify.com/

Las colecciones que almacenan los datos son:

- **Artist:** Incluye información de los artistas que obtuvieron alguna posición en el ranking durante el período analizado.
- **Artist_audio_features:** Guarda metadatos de las canciones que obtuvieron alguna posición en el ranking durante el período analizado.
- **Charts:** Tabla de rankeo de canciones (Top 200) durante un periodo determinado [3].

La estructura con la información de un documento de cada colección se muestra en el Anexo I. A su vez, las definiciones de los campos pueden ser consultados en la documentación de Spotify [2].
Los datasets fueron exportados con mongoexport y deben ser incorporados a la base con el comando **mongoimport** desde una consola CMD de Windows o una terminal Linux de la siguiente manera:

Archivo de artist:

```
mongoimport -h localhost -d DMUBA_SPOTIFY -c artist --file=./artist.json
```

Archivo de charts:

```
mongoimport -h localhost -d DMUBA_SPOTIFY -c charts --file=./charts-dm.json
```

Archivo de artist_audio_features:

```
mongoimport -h localhost -d DMUBA_SPOTIFY -c artist_audio_features
--file=./artist_audio_features-dm.json
```

Los datos deben ser descargados del siguiente enlace:

https://drive.google.com/drive/folders/1ym38Camd1k_cU7hF0XPi8Bbvt7LcroXd?usp=sharing

ALCANCE DEL TRABAJO

Se espera que puedan desarrollar los temas que fueron vistos durante las clases teóricas y prácticas, indagando en nuevos datos con complejidades propias de un problema del mundo real.

Los problemas principales a resolver son:

- Formular entre tres y cinco preguntas que guíen el trabajo de descubrimiento de conocimiento.
- Análisis exploratorio: Será necesario que indaguen en los datos “crudos” para poder tomar decisiones de diseño de su trabajo. Por ejemplo, decidir sobre aspectos de la integración, tareas de limpieza, identificador de presencia de ruido, etc.
- Integración de datos: La estructura jerárquica de los registros va a requerir hacer algunas transformaciones para convertir las colecciones de canciones, artistas, y ranking en una matriz de datos que sea funcional.
- Además de integrar información se podrán incorporar otras fuentes de información consideren pertinentes para enriquecer el análisis. (No es obligatorio realizar esta tarea, invierta el tiempo justo y necesario dado que puede llevar demasiado tiempo integrar datos de fuentes heterogéneas)
- Reducción de datos y dimensiones: Deberá tomar decisiones sobre qué variables duplican información, qué variables pueden dejarse afuera debido a su escaso aporte información para el objetivo de KDD.
- Limpieza de datos: Será necesario mejorar la calidad de los datos a partir de corregir inconsistencias, normalizar nombres, corregir tipos de datos, etc.
- Análisis de valores atípicos: Tratar de buscar no solo valores extremos sino también algunas observaciones que sean outliers de algún conjunto específico.
- Transformaciones: Se puede llegar a requerir re-escalado de variables o transformaciones logarítmicas. Tengan presente estas cuestiones durante los análisis.
- Generación de variables: Analice la posibilidad de combinar variables para generar otras.

- Visualizaciones: Se espera que puedan hacer un análisis en el tiempo sobre algunas variables del dataset o analizar relaciones entre variables utilizando herramientas gráficas. Es importante que tengan presente que los gráficos son una herramienta que facilita entender el problema, avalar o refutar alguna hipótesis de trabajo, es decir, es parte de la historia que quieren contar a partir de los datos. Por lo tanto, deben ser comprensibles por quien los vaya a leer. Todos los gráficos que se incorporen deben tener su correspondiente leyenda, nombres en los ejes, unidades de medidas, título, etc.

INFORME

El informe deberá contener las siguientes secciones:

- Título, nombre y apellido de los integrantes del grupo.
- Un resumen de hasta 200 palabras.
- Enumeración de las preguntas que se abordan en el análisis.
- Descripción de cada una de las etapas del proceso de KDD que se lleva a cabo para el trabajo (no incluir conceptos teóricos sólo sus aportes).
- La extensión máxima será de **15 páginas** incluyendo texto, gráficos y anexos, no debe incluir código en el informe entregable.
- El informe deberá ser entregado en formato *pdf* (Portable Document Format) por correo electrónico al equipo docente.

CONFORMACIÓN DE GRUPOS

El trabajo deberá ser realizado en grupos de 3 personas.

FECHA DE ENTREGA

El trabajo deberá ser entregado el día 21 de Mayo de 2021 hasta las 23:59 hs.

REFERENCIAS

[1] Librería SpotifyR [link]

[2] Metadata de atributos [link]

[3] Tabla de rankeo Top 200 [link]

Anexo I

A continuación se muestra, a modo de ejemplo, la estructura con la información de un documento de cada colección:

Charts


```
{
	"_id" : ObjectId("6048b543211df45f6545ae23"),
	"Position" : 1,
	"Track_Name" : "Havana (feat. Young Thug)",
	"Artist" : "Camila Cabello",
	"Streams" : 30948101,
	"URL" : "https://open.spotify.com/track/1rfofaqEpACxVEHIZBJe6W",
	"week_start" : "2018-01-12",
	"week_end" : "2018-01-19"
}
```


Artista


```
{ "_id" : ObjectId("6054c222211df45f6548ed5e"), "Artist" : "Camila Cabello" }
```


Obra


```
{
	"_id" : ObjectId("60551111211df45f65491a2a"),
	"artist_name" : "Camila Cabello",
	"artist_id" : "4nDoRrQiYLoBzwC5BhVJzF",
	"album_id" : "3Vsbl0diFGw8HNSjG8ue9m",
	"album_type" : "album",
	"album_images" : [
		{
			"height" : 640,
			"url" : "https://i.scdn.co/image/ab67616d0000b2735f53c0dbe5190a0af0fa28f3",
			"width" : 640
		},
		{
			"height" : 300,
			"url" : "https://i.scdn.co/image/ab67616d00001e025f53c0dbe5190a0af0fa28f3",
			"width" : 300
		},
		{
			"height" : 64,
			"url" : "https://i.scdn.co/image/ab67616d000048515f53c0dbe5190a0af0fa28f3",
			"width" : 64
		}
	],
```

```
        "album_release_date" : "2019-12-06",
        "album_release_year" : 2019,
        "album_release_date_precision" : "day",
        "danceability" : 0.377,
        "energy" : 0.651,
        "key" : 4,
        "loudness" : -5.437,
        "mode" : 0,
        "speechiness" : 0.0589,
        "acousticness" : 0.0197,
        "instrumentalness" : 0.0000531,
        "liveness" : 0.174,
        "valence" : 0.0851,
        "tempo" : 129.607,
        "track_id" : "2ogKhhoMClkFXek7ZgxAhN",
        "analysis_url" : "https://api.spotify.com/v1/audio-analysis/2ogKhhoMClkFXek7ZgxAhN",
        "time_signature" : 4,
        "artists" : [
                {
                        "href" : "https://api.spotify.com/v1/artists/4nDoRrQiYLoBzwC5BhVJzF",
                        "id" : "4nDoRrQiYLoBzwC5BhVJzF",
                        "name" : "Camila Cabello",
                        "type" : "artist",
                        "uri" : "spotify:artist:4nDoRrQiYLoBzwC5BhVJzF",
                        "external_urls_spotify" :
"https://open.spotify.com/artist/4nDoRrQiYLoBzwC5BhVJzF"
                }
        ],
        "available_markets" : [
                ...
        ],
        "disc_number" : 1,
        "duration_ms" : 219742,
        "explicit" : false,
        "track_href" : "https://api.spotify.com/v1/tracks/2ogKhhoMClkFXek7ZgxAhN",
        "is_local" : false,
        "track_name" : "Shameless",
        "track_preview_url" :
"https://p.scdn.co/mp3-preview/ecbe1041a69343d4ec849876b8f3b2aa71b332a3?cid=5793b9f79e7e44cd8333d569
7cf7550a",
        "track_number" : 1,
        "type" : "track",
        "track_uri" : "spotify:track:2ogKhhoMClkFXek7ZgxAhN",
        "external_urls_spotify" : "https://open.spotify.com/track/2ogKhhoMClkFXek7ZgxAhN",
        "album_name" : "Romance",
        "key_name" : "E",
        "mode_name" : "minor",
        "key_mode" : "E minor"
}
```