RETO 3: Soundtrack Your Timeline

Objetivo

Poner en práctica los conceptos aprendidos en clase acerca de las estructuras de datos no lineales: árboles, los algoritmos de ordenamiento y búsquedas eficientes de información. Específicamente se desea:

a) Utilizar árboles binarios de búsqueda y árboles balanceados, en conjunto con las demás estructuras de datos del curso (listas y tablas de Hash) para solucionar los requerimientos del reto.
b) Utilizar adecuadamente el patrón Modelo, Vista Controlador.
c) Utilizar adecuadamente el ambiente de trabajo (IDE, GIT y GitHub).

Fecha Límite de Entrega

28 de abril, 11:59 p.m.

Contexto

Los sistemas de análisis musical procesan diferentes piezas musicales, para entender características y propiedades como clasificar el tipo de música que se está escuchando, comprender los estados de ánimo que representan, estimar la cantidad de voz o instrumentos presentes en una canción, entre otras. Estos análisis ayudan a seleccionar la siguiente pieza musical o a mejorar la recomendación de determinadas canciones o pistas en situaciones especiales (ej.: en el gimnasio, para estudiar, para meditar, para descansar, una fiesta, etc.).

El objetivo de este reto es procesar estos análisis de canciones para entender el comportamiento de los usuarios y para saber qué tipo de música puede ofrecérseles en un momento determinado. Para ello, vamos a utilizar el conjunto de datos provisto en Kaggle, llamado **Context-Aware Music Recommender System,** el cual tienen cerca de 11 millones de eventos asociados con información de usuarios enviada vía Twitter.

Las pistas musicales pueden ser analizadas para determinar valores numéricos que permitan caracterizar la música. A continuación, se presentan algunos de los elementos más conocidos y utilizados para la caracterización de una pieza musical.

Característica de Contenido	Descripción
Instrumentalness (Instrumentalidad)	Este valor representa la cantidad de instrumentalización en la canción. Toma valores entre 0 y 1. Entre más cerca de 1 más instrumental es la canción
Acousticness (Acústica)	Este valor describe cuan acústica es la canción. Toma valores entre 0 y 1. Un valor de 1 significa que la canción es completamente acústica o sintetizada
Liveness (Liveness)	Este valor describe la probabilidad de que la canción fue grabada con una audiencia en vivo. Un valor igual o superior a 0.8 significa que muy probablemente fue grabada en vivo. Toma valores entre 0 y 1.
Speechiness (Speechiness)	Detecta la presencia de voz en la canción. Toma valores entre 0 y 1. Un valor de 0.66 o superior indica que con alta probabilidad la pista incluye solo voz. Un valor entre 0.33 y 0.66 indica que tiene voz y música y un valor por debajo de 0.33 indica que no tiene voz
Energy (Energía)	Representa una medida porcentual de la intensidad y actividad de la pista. A mayor porcentaje, más rápida, ruidosa y sonora la pista.
Danceability (Capacidad de baile)	Describe cuan apropiada es la pista para bailar basada en la combinación de elementos como: tempo, ritmo, fuerza y regularidad. Toma valores entre 0 y 1. Cero indica que no es bailable y 1 que si lo es.
Valence (Valencia)	Indica la positividad de una pista. Toma valores entre 0 y 1. Una pista con una valencia alta es más feliz o eufórica en tanto que entre más cercana al cero es más triste y depresiva.

Tabla 1. Lista de características de contenido.[1]

Tempo	Descripción
Larghissimo	very, very slowly (24 bpm and under)
Adagissimo	very slowly
Grave	very slow (25–45 bpm)
Largo	broadly (40–60 bpm)
Lento	slowly (45–60 bpm)
Larghetto	rather broadly (60–66 bpm)
Adagio	slowly with great expression (66–76 bpm)
Adagietto	slower than andante (72–76 bpm) or slightly faster than adagio (70–80 bpm)
Andante	at a walking pace (76–108 bpm)
Andantino	slightly faster than andante (although, in some cases, it can be taken to mean slightly slower than andante) (80–108 bpm)
Marcia moderato	moderately, in the manner of a march (83–85 bpm)
Moderato	at a moderate speed (108–120 bpm)
Allegretto	by the mid-19th century, moderately fast (112–120 bpm); see paragraph above for earlier usage
Allegro moderato	close to, but not quite allegro (116–120 bpm)
Allegro	fast, quick, and bright (120–156 bpm) (molto allegro is slightly faster than allegro, but always in its range; 124-156 bpm)
Vivace	lively and fast (156–176 bpm)
Vivacissimo	very fast and lively (172–176 bpm)
Allegrissimo or Allegro vivace	very fast (172–176 bpm)
Presto	very, very fast (168–200 bpm)
Prestissimo	even faster than presto (200 bpm and over)

Tabla 2. Listado del Tempo musical.[2]

[1] *Is my Spotify music boring? An analysis involving music, data, and machine learning,* URL: https://towardsdatascience.com/is-my-spotify-music-boring-an-analysis-involving-music-data-and-machine-learning-47550ae931de

[2] *Tempo*, URL: https://en.wikipedia.org/wiki/Tempo

Fuente de Datos

El conjunto de datos **Nowplaying-RS** presenta características de contenido y contexto de los eventos de escucha. Contiene 11,6 millones de eventos de escucha de música de 139.000 usuarios y 346.000 pistas recopiladas de Twitter.

A continuación, se describen las fuentes de datos para el reto. Los datos están contenidos en los archivos `usertrackhashtagtimestamp.csv`, `contextcontentfeatures.csv` y `sentiment_values.csv` disponibles en aula unificada en Bloque Neón.

Característica de Contenido	Descripción
usertrackhashtagtimestamp.csv	Contiene información básica de las etiquetas sobre un evento de escucha de una pista musical. Para cada evento de escucha se proporciona: • [redacted] • identificador del usuario (userid) • identificador de la canción (trackid) • hashtag • fecha de creación del registro (created_at)
contextcontentfeatures.csv	Contiene el análisis de los eventos enviados por los usuarios en relación con las pistas musicales. Para cada evento de escucha se proporciona: • id del evento • identificador del usuario (userid) • identificador de la canción (trackid) • identificador del artista (artistid) • características de contenido con respecto a la pista mencionada en el evento (instrumentalidad, viveza, habla, capacidad de baile, valencia, sonoridad, tempo, acústica, energía, modo, clave) • características de contexto con respecto al evento de escucha (coordenadas (como geoJSON), lugar (como geoJSON), geo (como geoJSON), idioma del tweet (tweetlanguage), fecha de creación del registro (created_at), idioma del usuario (userlang), y zona horaria (time_zone)
sentiment_values.csv	Este archivo contiene información asociada a las etiquetas utilizadas para clasificar las pistas musicales, mediante el uso de diccionarios de sentimientos. El archivo contiene la etiqueta en sí y los valores de análisis de sentimiento recopilados a través de cuatro diccionarios de sentimiento diferentes: AFINN, Opinion Lexicon, Sentistrength Lexicon y vader. Para cada uno de estos diccionarios, se enumera el mínimo, el máximo, la suma y el promedio de todos los sentimientos de los tokens de la etiqueta (si está disponible; de lo contrario, se enumeran valores vacíos).

Tabla 3. Resumen y descripción de las fuentes de datos.[3]

[3] Kaggle, Nowplayingrs, URL: https://www.kaggle.com/chelseapower/nowplayingrs.

Diccionarios de Sentimientos

Los diccionarios de sentimientos sirven para clasificar palabras, frases o textos más largos. Esta clasificación permite determinar si el texto analizado es positivo, neutral o negativo.

Uno de los diccionarios más utilizado para este tipo de análisis es el **Valence Aware Dictionary and sEntintiment Reasoner (VADER),** este diccionario arroja tres tipos diferentes de resultados: Positivos, neutros o negativos; los cuales son representados con valores numéricos entre -1 y 1 de la siguiente manera:

- Sentimiento Positivo valores mayores o iguales a 0.5
- Sentimiento Neutro valores entre –0.5 y 0.5
- Sentimiento negativo valores inferiores o iguales a –0.5

Trabajo Propuesto

Parte 1 – Configuración Repositorio

Complete los siguientes pasos para configurar su repositorio de trabajo:

a) Cree en GitHub un repositorio llamado **Reto3-EDA202110** basado en el repositorio en el URL: https://github.com/ISIS1225DEVS/Reto3-202110-Template
b) Cree el **README** del repositorio donde aparezcan los nombres completos, correo Uniandes y códigos de los miembros del equipo de trabajo.
c) Realice el procedimiento para crear el directorio en su computador de trabajo para que relacione este directorio con el repositorio remoto que acaba de crear.
d) Descargue los datos desde la sección unificada del curso y cópielos en la carpeta **data** del repositorio local.

Parte 2 – Desarrollo

Para responder a los requerimientos presentados deberán cargar la información de los archivos entregados; recuerde que solo se permite leer una vez la información de cada archivo.

Al final de la carga de datos (desde los 3 archivos fuentes) debe reportar los siguientes datos:

- El total de registros de eventos de escucha cargados.
- El total de artistas únicos cargados (sin repetirse).
- El total de pistas de audio únicas cargadas (sin repetirse).
- Mostrar los primeros 5 y últimos 5 eventos de escucha cargados con sus características de contenido y de contexto.

Nota: Los ejemplos dados en el documento están hechos basados en el subconjunto de datos más pequeño ("-small.csv").

Parte 3 – Desarrollo de la solución a los requerimientos

Requerimiento 1 (Grupal): Caracterizar las reproducciones

Se desea conocer **cuántas reproducciones (eventos de escucha)** se tienen en el sistema de recomendación basado en **una característica de contenido** y con un **rango determinado**, El sistema debe indicar el total de canciones y el número de artistas (sin repeticiones).

Para dar respuesta a este requerimiento se debe recibir como entrada la siguiente información:

- La característica de contenido (ej.: valencia, sonoridad, etc.).
- El valor mínimo de la característica de contenido.
- El valor máximo de la característica de contenido.

Y como respuesta debe presentar en consola la siguiente información:

- El Total de los eventos de escucha o reproducciones.
- El número de artistas únicos (sin repeticiones)

EJEMPLO:
Se quieren conocer **cuántas reproducciones de piezas musicales** tienen la **instrumentalidad (Instrumentalness)** con un rango desde **0.75** hasta **1.00**.

```
++++++ Req No. 1 results... ++++++
Instrumentalness is between 0.75 and 1.0
Total of reproduction: 4179 Total of unique artists: 1769
```

Requerimiento 2 (Estudiante A): Encontrar música para festejar

Se desea encontrar **las pistas** en el sistema de recomendación que pueden utilizarse en una fiesta que se tendrá próximamente. Se desea encontrar las canciones que tengan en cuenta las variables (**Energy**, y **Danceability**). El usuario podrá indicar los valores (rangos) para esos parámetros.

Para dar respuesta a este requerimiento se debe recibir como entrada la siguiente información:

- El Valor mínimo de la característica **Energy**.
- El Valor máximo de la característica **Energy**.
- El Valor mínimo de la característica **Danceability**.
- El Valor máximo de la característica **Danceability**.

Y como respuesta debe presentar en consola la siguiente información:

- El total de pistas únicas (sin repeticiones).
- La información de 5 pistas seleccionadas aleatoriamente (incluya los valores **Energy** y **Danceability** para cada uno).

EJEMPLO:
Se quieren conocer **cuántas pistas no repetidas** tienen **Energy (Energía)** entre **0.50 y 0.75**, y **Danceability (Capacidad de Baile)** con un rango dese **0.75** a **1.00**.

```
++++++ Req No. 2 results... ++++++
Energy is between 0.5 and 0.75
Danceability is between 0.75 and 1.0
Total of unique tracks in events: 1703

--- Unique track_id ---
Track 1: a89fdcc29e25e35819e5cc357555a5df with energy of 0.597 and danceability of 0.777
Track 2: be1bba75e2767e6c06140f417473bbe4 with energy of 0.553 and danceability of 0.799
Track 3: dbe487649a65f5a56e8cd355f69d80e3 with energy of 0.587 and danceability of 0.794
Track 4: 8222ae03936c107f2c400cb0cc8cd0a5 with energy of 0.716 and danceability of 0.768
Track 5: b42c4727be43063311a98306d04df1ee with energy of 0.692 and danceability of 0.874
```

Requerimiento 3 (Estudiante B): Encontrar música para estudiar

Se desea conocer **las pistas** en el sistema de recomendación que podrían ayudar a los usuarios en su periodo de estudio, para este fin se debe tener en cuenta tengan en cuenta las variables (**Instrumentalness**, y **Tempo**). Por ejemplo, para un grupo de estudio que desea trabajar de forma tranquila; en este caso se prefieren canciones **instrumentales**, sin letra o muy poca y de preferencia con un **Tempo Largo** (40–60 BPM) porque se sabe que favorece el aprendizaje.

- Para dar respuesta a este requerimiento se recibe como entrada la siguiente información:
  - El valor mínimo del rango para **Instrumentalness.**
  - El valor máximo del rango para **Instrumentalness.**
  - El valor mínimo del rango para el **Tempo.**
  - El valor máximo del rango para el **Tempo.**

Y como respuesta se espera que la consola presente la siguiente información:

- El total de pistas únicas (sin repeticiones).
- La información de 5 pistas seleccionadas aleatoriamente (incluya los valores de **Instrumentalness** y **Tempo** para cada uno).

EJEMPLO:
Para un grupo de estudio que desea trabajar de forma tranquila; en este caso se prefieren canciones **instrumentales**, sin letra o muy poca y de preferencia con un **Tempo Largo** (40–60 bpm) porque se sabe que este tipo de música favorece el aprendizaje.

```
++++++ Req No. 3 results... ++++++
Instrumentalness is between 0.6 and 0.9
Tempo is between 40.0 and 60.0
Total of unique tracks in events: 9

--- Unique track_id ---
Track 1: d5470e12b055aeb25ec11efcc474f8bd with instrumentalness of 0.893 and tempo of 53.203
Track 2: d89902d1d1677eeb3985a7a41cf6e63e with instrumentalness of 0.817 and tempo of 43.425
Track 3: 4978ab7dad5dc1d843af6b3b422a8692 with instrumentalness of 0.755 and tempo of 56.228
Track 4: 2633956017d8c9adc52ef9eb32c963f4 with instrumentalness of 0.757 and tempo of 56.441
Track 5: ac136b1a3cae741a26ca4eecaf163dcd with instrumentalness of 0.846 and tempo of 59.285
```

Requerimiento 4 (Grupal): Estudiar los géneros musicales

Dada la siguiente tabla, que relaciona el Tempo con el género musical, se desea saber cuántas canciones se tienen por cada género (definidos como un rango de Tempo como se ve en la Tabla 4); adicionalmente, cuántos artistas se tienen en cada género.

Genero	BPM Típico
Reggae	60 a 90
Down-tempo	70 a 100
Chill-out	90 a 120
Hip-hop	85 a 115
Jazz and Funk	120 a 125
Pop	100 a 130
R&B	60 a 80
Rock	110 a 140
Metal	100 a 160

Tabla 4. BPM por género musical.[4]

Nota: Es posible que una misma pista pueda pertenecer a más de un género musical; por ejemplo, **100 BPM** puede estar en **Down-Tempo, Pop, Metal, Hip-Hop** y **Chill-Out.** En este caso la pista cuenta para la suma de todos los géneros.

Para dar respuesta a este requerimiento se recibe como entrada la siguiente información:

- La lista de géneros musicales que se desea buscar. (ej.: Reggae, Hip-hop, Pop.).

En caso de desearlo, el usuario puede agregar un nuevo género musical en la búsqueda con las siguientes variables de entrada:

- Nombre único para el nuevo género musical.
- Valor mínimo del Tempo del nuevo género musical.
- Valor máximo del Tempo del nuevo género musical.

Y como respuesta se espera que la consola presente la siguiente información:

- El Total de los eventos de escucha o reproducciones.
- El Total de los eventos de escucha o reproducciones en cada género.
- El número de artistas únicos (sin repeticiones) en cada uno de los géneros musicales, y el ID de los primeros 10 artistas

EJEMPLO:
Dar el número de reproducciones y artistas únicos para los géneros de **Reggae** (de 60 a 90 BPM), **Hip-hop** (de 85 a 115 BPM) y **Pop** (de 100 a 130 BPM).

[4] Genders by ear, URL: https://www.musical-u.com/learn/rhythm-tips-for-identifying-music-genres-by-ear/

```
++++++ Req No. 4 results... ++++++
Total of reproductions: 53988

======== REGGAE ========
For Reggae the tempo is between 60.0 and 90.0 BPM
Reggae reproductions: 7545 with 2455 different artists
----- Some artists for Reggae -----
Artist 1: c2c99b9579d26fece45006b2f24b4399
Artist 2: b13ea9e1c580f759983cdde5499093e0
Artist 3: fe8ee22aaaee24fcd7fd80f39ca746c0
Artist 4: d18521e2c1c5501fe061e8641089c6fb
Artist 5: 7a3edd209fcb98ca3b7385ee8b4f9f19
Artist 6: 6f1181c6510162345ae93b2964181238
Artist 7: bf2528a296adb62d041a7519aa77f248
Artist 8: b7ec7c23c860cbf8867484eb33debf50
Artist 9: b09ddf38ed1df9fa117be5425b2cac67
Artist 10: 0c70b5ab9c28d87d3bf0878c890919f2

======== HIP-HOP ========
For Hip-hop the tempo is between 85.0 and 115.0 BPM
Hip-hop reproductions: 19978 with 4977 different artists
----- Some artists for Hip-hop -----
Artist 1: d6e00f39c099eaa24256f233ed863c97
Artist 2: 481d88c05dfb1c8709238453bbe14fee
Artist 3: d23984e186582514851dd00b64e2b921
Artist 4: a64e91d464fe9afbbc673d0f2580bdfa
Artist 5: 2ca16f767e68bb0b15e6776c84e4cf61
Artist 6: 62f5007ee28a6daf097c87d803d2fbb6
Artist 7: 7b8e61e2831286355c0508034f43a37d
Artist 8: 8c1fbfb82bad7c8958e34812839d0b9c
Artist 9: c7f9b00f61cc2799678fff79b7a99760
Artist 10: e1dc0cc7f0c6d65b72fff7c88e4eb69e

======== POP ========
For Pop the tempo is between 100.0 and 130.0 BPM
Pop reproductions: 26465 with 5891 different artists
----- Some artists for Pop -----
Artist 1: d6e00f39c099eaa24256f233ed863c97
Artist 2: 5272f90eb6120f823cb2369858401df2
Artist 3: ed73022e38d78447588e214e0d9b6a3f
Artist 4: b1d113e11165894fd12c94f2d46eb485
Artist 5: d23984e186582514851dd00b64e2b921
Artist 6: aa968850a9d255494612acd0552f8fcd
Artist 7: 62f5007ee28a6daf097c87d803d2fbb6
Artist 8: 8e972222b22628e3d170d2b5cc4272b3
Artist 9: 62f5007ee28a6daf097c87d803d2fbb6
Artist 10: 7f96a59ebec9a542be67cb83c6249664
```

Requerimiento 5 (Grupal): Indicar el género musical más escuchado en el tiempo

Dado un rango de **horas (ej.: de 10:00 am a 10:30 am)** indicar el género de música más escuchado en dicho rango teniendo en cuenta todos los días disponibles e informar el promedio para cada uno de los valores de análisis de sentimiento, en las canciones de dicho rango.

Para dar respuesta a este requerimiento se recibe como entrada la siguiente información:

- El valor mínimo de la hora del día.
- El valor máximo de la hora del día.

Y como respuesta debe presentar en consola la siguiente información:

- Género más referenciado en el rango de horas.
- Para el género más referenciado, calcular el valor promedio VADER de las pistas que contiene. tome como base para del cálculo el campo `vader_avg` de las etiquetas (Hashtag).

Nota 1: Si el valor del campo `vader_avg` no existe, no tenga en cuenta esa etiqueta (Hashtag).

Nota 2: Recuerde que una reproducción puede tener varias etiquetas asociadas.

EJEMPLO:
Cuál es el género más escuchado **entre las 7:15:00 a.m. y las 9:45:00 a.m**. con sus **valores VADER** promedio para cada **etiqueta (Hashtag)** asociada.

```
++++++ Req No. 5 results... ++++++
There is a total of 16629 reproductions between 07:15:00 and 09:45:00
====================== GENRES SORTED REPRODUCTIONS ======================
TOP 1: Metal with 4136 reps
TOP 2: Rock with 2668 reps
TOP 3: Pop with 2664 reps
TOP 4: Chill-out with 2274 reps
TOP 5: Hip-hop with 2001 reps
TOP 6: Down-tempo with 1388 reps
TOP 7: Reggae with 716 reps
TOP 8: Jazz and funk with 490 reps
TOP 9: R&b with 292 reps
...

The TOP GENRE is Metal with 4136 reproductions...

========================= Metal SENTIMENT ANALYSIS =========================
Metal has 1395 unique tracks...
The first TOP 10 tracks are...

TOP 1 track: 3d02f9fcad37e6bb227682761039498c with 5 hashtags and VADER = 0.6
TOP 2 track: 5758909ef03fc3a2efaa57408ad43f22 with 5 hashtags and VADER = 0.6
TOP 3 track: 173803b09a0688817f58d7236e1ab9d4 with 5 hashtags and VADER = 0.6
TOP 4 track: 9d7125c57ba071920ca6ad711338909e with 5 hashtags and VADER = 0.6
TOP 5 track: 7188f922ba73eb2509c02d15ec62ff0e with 5 hashtags and VADER = 0.6
TOP 6 track: 76a18f3ba9d4c32bff693ea1614e7605 with 4 hashtags and VADER = 0.6
TOP 7 track: c2da30eb3450e8a3e5bfa16e8fa527da with 4 hashtags and VADER = 0.6
TOP 8 track: 462b1fc2bc97a2c8de40946cba29fd21 with 4 hashtags and VADER = 0.7
TOP 9 track: 0ec56289c0cc2ebc3cb1ce1a03e3355e with 4 hashtags and VADER = 0.6
TOP 10 track: 0fc1136e32ea77a75aa2cbc211552618 with 3 hashtags and VADER = 0.6
```

Parte 4 – Análisis de resultados

Cree un archivo en formato **PDF** para la entrega y guárdelo en la carpeta **Docs** del repositorio, el documento debe contener las siguientes secciones:

- Análisis de complejidad en **notación O** para cada uno de los requerimientos.
- Análisis de tiempo de ejecución y uso de memoria para cada uno de los requerimientos implementados.
- Análisis de tiempo de ejecución y consumo de memoria para cada uno de los requerimientos.

Recuerde indicar quien implemento los requerimientos individuales

TIPS:

- Tomen como guía las herramientas y los análisis realizados en los laboratorios 4, 5, y 7.
- Tomen los tres subconjuntos más grandes que puedan trabajar en sus computadores, ej.: si el máximo que pueden cargar son los archivos al 80.0 %, utilicen los del 80.0 %, 50.0 % y 30.0 %.
- Ejecute las pruebas de los requerimientos siempre con los mismos parámetros de entrada.
- Utilicen solo la configuración optima de TADs para su implementación ej.: para Map PROBING un factor de carga por defecto de 0.50, entre otros.

Entrega

1. Para hacer la entrega del taller usted debe agregar **los usuarios de los monitores y su profesor** a su repositorio **Github.**
2. Si No da acceso a su repositorio a los monitores y al profesor, el taller **NO** podrá ser calificado.
3. Recuerden que cualquier documento solicitado durante las actividades debe incluirse en el repositorio GIT y que solo se calificara hasta el último **COMMIT** realizado dentro de las fechas límites.