## Propuesta de Trabajo Fin de Máster

MASTER EN DATA SCIENCE Y BIG DATA (MDS)

#### **Título**: **Modelo de clasificación de géneros musicales basado en recuperación de información musical y características de la letra.**

#### **Estudiante**: Ludwig Gerardo Rubio Jaime

**Tutor**: Borja Foncillas García

#### **Descripción**:

Las plataformas de music streaming son uno de los claros ejemplos de adaptación de las industrias al contexto digital que debido a su volumen de usuarios y contenido se enfrentan a nuevos retos de organización y recomendación de contenido para sus usuarios.

Aún cuando la tendencia de estas plataformas para recomendar música es la de utilizar sistemas de recomendación basado en filtros colaborativos, persiste la necesidad de contar con algoritmos que mejoren la organización, así como la recomendación de música nueva para la cual aún no se cuenta con datos colaborativos.

Por tal razón, mi motivación es la de explorar alternativas que faciliten la organización de música de forma más eficiente y automatizada basada en **algoritmos de aprendizaje automático**, que además, utilicen características de la música que faciliten su escalabilidad a un sistema de recomendación músical.

Debido a que el género musical es un dato casi siempre intrínseco a los metadatos de canciones, y con el objetivo de contar con un criterio de evaluación del algoritmo, propongo generar modelos de clasificación de género basado en recuperación de información musical y características de la letra que mezclan técnicas de **MIR** con la **creación de nuevas características basadas en text-mining, análisis de sentimientos y emociones**, etc.

#### 

#### **Objetivo**:

Generar un modelo de clasificación de género musical cuyas características permitan su incorporación o escalado a sistemas de recomendación musical.

Como entregables se proponen:

- Set de datos generado

- Modelos de clasificación logrados

- Story Telling que muestre de forma clara el proyecto y permita de manera sencilla entender su aplicación

- Análisis detallado a modo de reporte acerca del proceso, resultados y conclusiones (memoria)

#### **Fuente de datos:**

|  |  |  |
| --- | --- | --- |
| **Nombre** | **Datos** | **Formato / modo de uso** |
| [Million Song Dataset](https://labrosa.ee.columbia.edu/millionsong/pages/getting-dataset) | MIR / Metadatos | Archivos MD5 / Base de datos |
| [Spotify](https://developer.spotify.com/documentation/web-api/) | MIR / Metadatos | API |
| [FMA: A Dataset For Music Analysis](https://github.com/mdeff/fma) | Audio / MIR / Metadatos | API / Archivos |
| [ISMIR](https://www.ismir.net/resources.html#datasets) | Audio / MIR / Metadatos | API / Archivos |
| [MusiXMatch](https://www.musixmatch.com) | Lyrics | API / Web Scraping |
| [AZLyrics](https://www.azlyrics.com) | Lyrics | Web Scraping |
| [Theysay](http://www.theysay.io) | Metadatos | API |

\*Algunas de las fuentes de datos que proveen metadatos serán excluidos una vez se seleccione un subconjunto como base de trabajo.

#### **Plan de trabajo:**

|  |  |  |
| --- | --- | --- |
| **Fase** | **Tiempo (Aprox)** | **Tareas** |
| Entendimiento de los datos | 1 semana | * Investigación del estado de arte * Estudio y selección de las fuentes de datos |
| Preparación de los datos | 2 semanas | * Centralización y limpieza de la información * Complementariedad de los datos   + Procesado de recuperación de información musical   + Web scraping de las letras   + Procesos de text-mining y análisis de sentimientos y emociones |
| Modelado y evaluación | 1 semana | * Generación de modelos   + Selección de variables   + Evaluación de modelos |
| Documentación | 2 semanas | * StoryTelling * Generación de memoria |

#### 