

Tarea 3. Diseño de experimentos para análisis de texto

Helena Patricia Carrillo Soto

Facultad de Ciencias Físico-Matemáticas, UANL

Maestría Ciencia de Datos: Procesamiento y Clasificación de Datos

Resumen

Se analizaron varias canciones de diez artistas diferentes. De dicho análisis se obtienen y comparan estadísticas descriptivas de texto así como se intenta clasificar cada una de las canciones con su respectivo artista.

Keywords: Artistas; Análisis de texto; Canciones

1. Introducción

Análisis a realizar En este artículo se realizará un análisis de diversas canciones por diez diferentes artistas. Se busca encontrar si es posible, y cual es la mejor forma de hacerlo, entrenar un modelo para clasificar las canciones de cada artista basándose en sus letras.

Canciones elegidas En su mayoría se eligieron canciones de artistas pop o de hip-hop los listados son músicos ampliamente conocidos. Se tienen artistas que son conocidos por escribir su música y artistas que abiertamente se apoyan de terceros para la escritura de sus canciones.

2. Descripción de los datos

Los artistas a analizar son los siguientes:

- **Billie Eilish**. Conocida por escribir sus canciones junto con su hermano y productor Finneas O'Connell.
- **Charlie Puth**. Artista altamente involucrado en la escritura y producción de sus canciones.
- **Coldplay**. La mayoría de sus canciones son escritas por su vocalista Chris Martin.
- **Drake**. Es de conocimiento público que utiliza varios escritores para la creación de sus canciones.
- **Ed Sheeran**. Reconocido por escribir no solo sus canciones sino también por haber escrito múltiples canciones para otros artistas.
- **Eminem**. Conocido por escribir la mayoría de sus canciones y ha escrito para otros artistas.
- **Justin Bieber**. Usa diferentes escritores y productores para la creación de la mayoría de sus canciones.
- **Khalid**. Conocido por escribir y producir la mayoría de sus canciones.
- **Rihanna**. Utiliza diferentes compositores en la escritura de sus canciones.
- **Taylor Swift**. Mundialmente conocida por escribir la gran mayoría sus canciones.

3. Metodología

Para realizar el análisis de las canciones primero se obtuvieron varias canciones de los artistas de la base de datos en Kaggle que se puede encontrar en la siguiente liga: <https://www.kaggle.com/deepshah16/song-lyrics-dataset>.

Posteriormente se separaron en frases o versos y se obtuvieron algunos estadísticos como la cantidad de palabras, la cantidad de caracteres y la densidad de palabras promedio de cada artista. Después se quitaron las palabras vacías y se lematizaron los textos para obtener las palabras más características de cada verso. Para esto se utiliza la librería `nltk`.

Luego se analizaron usando tres tipos de vectorizaciones y dos diferentes divisiones de entrenamiento y prueba.

Los tipos de vectorizadores que se usaron son los siguientes:

- **CountVectorizer** usando el analizador de palabras. Convierte un texto en una matriz de conteo de características.
- **TfidfVectorizer** usando el analizador de palabras. Convierte un texto en una matriz de características TF-IDF el cuál se utiliza para medir que términos son más relevantes.
- **TfidfVectorizer** usando el analizador de caracteres.

Así mismo se usaron dos divisiones de entrenamiento y prueba: 80-20 y 70-30.

4. Resultados

Análisis Estadístico Los primeros estadísticos que se obtienen son el número de palabras, caracteres y la densidad de palabras. Los resultados se pueden observar en el cuadro junto con los resultados de cada uno de los artistas [1](#).

De los resultados de dicho cuadro observamos que...

Cuadro 1: Número de palabras, caracteres y densidad de palabras por artista.

Artist	word_count	char_count	word_density
Billie Eilish	31.298	152.480	4.960
Charlie Puth	34.486	166.524	4.888
Coldplay	26.788	130.653	5.0375
Drake	64.903	317.682	5.009
Ed Sheeran	44.631	214.032	4.802
Eminem	76.860	385.649	5.153
Justin Bieber	38.872	187.159	4.967
Khalid	45.503	224.854	5.071
Rihanna	40.336	192.679	4.998
Taylor Swift	33.402	164.103	4.984