

Challenge - Dados

Conhecendo o Challenge

O desafio deste mês é criar um modelo de “machine learning” que prediz qual gênero de uma música somente baseado na letra dela.

O modelo será construído utilizando alguns conceitos de NLP (Processamento de Linguagem Natural) junto a conceitos estatísticos simples. Os dados serão retirados de uma base do “Kaggle” contendo uma vasta gama de músicas em português e inglês. Para esse desafio recomenda-se fortemente trabalhar com uma só dessas línguas.

Extra: É interessante também visualizar quais as palavras mais comuns de cada gênero

Requisitos Funcionais

Não será necessário que o seu modelo seja o melhor possível ou que se use algum tipo de melhoria na sua eficiência. A base usada pode ser obtida no link:

Song lyrics from 79 musical genres

Data scraped from the website vagalume.com.br

<https://www.kaggle.com/neisse/scrapped-lyrics-from-6-genres?select=artists-data.csv>



1. Dados limpos antes do treino do modelo
2. Modelo funcionando com pelo menos 90% de acertos
3. Análise da eficiência do modelo
4. Testagem com músicas de fora da base de dados usada

Etapas para a solução do challenge

- ☐ Aprender o que são e como aplicar conceitos de NLP
 - ☐ Tokenização
 - ☐ Limpeza com expressões regulares
 - ☐ Remoção de stopwords
 - ☐ Stemming
 - ☐ Lemmatization
 - ☐ Vetorização
- ☐ Aprender como funciona e como aplicar o modelo Naive-Bayes e sua forma multinomial para o uso nesse challenge