Projeto para processo seletivo - Cientista de Dados

Descrição do problema

Uma empresa possui um grande acervo musical, no entanto, tem muita dificuldade em diferenciar seus diferentes gêneros musicais.

Você (Cientista de Dados), poderá ajudá-los com esse problema!

Para isso, deverá desenvolver uma aplicação web (não precisa se preocupar com layout) com um *Text Area* que receberá a letra de uma música e um botão de *submit* que enviará a letra para um *Web Service* responsável pelo processamento das informações.

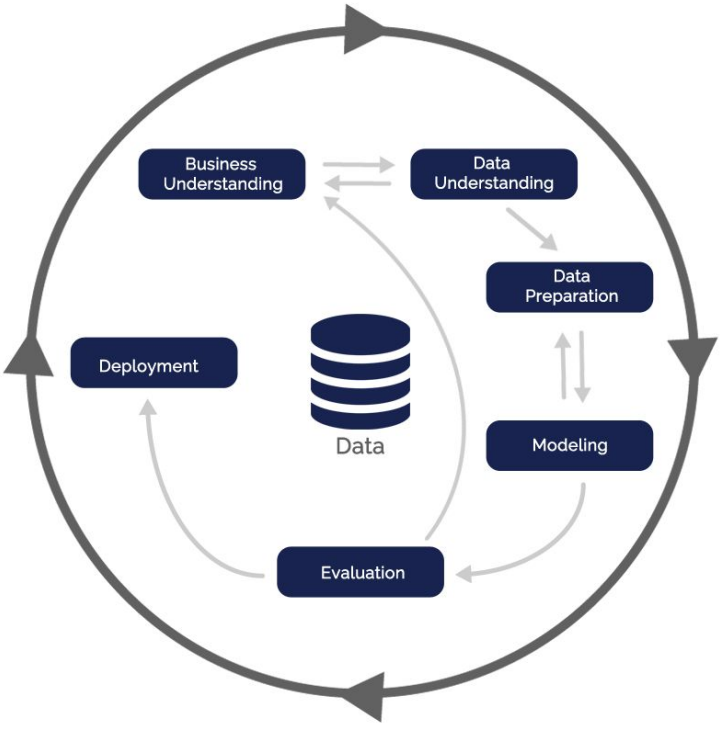
Esse serviço responderá uma página contendo o gênero musical, dentre os seguintes: **Bossa Nova, Funk, Sertanejo e Gospel.**

Conjunto de Dados

O conjunto de treinamento encontra-se no anexo **lyrics.zip.**

Critérios de Entrega

Com base na metodologia **CRISP-DM**, embasada em 6 etapas, tais como: **Entendimento do negócio, Compreensão dos dados, Preparação dos dados, Avaliação, Desenvolvimento:**



Você deverá desenvolver uma solução com os seguintes critérios e diferentes complexidades:

- O *server-side* (*Web Service*) deve ser desenvolvido em Python, utilizando estilo arquitetural REST. Poderá utilizar frameworks como (Flask, Sanic, Tornado ou qualquer outra que julgar interessante para solução).

- Realizar limpeza e pré-processamento dos dados (Justificar escolha das técnicas adotadas).

- Realizar análise exploratória do conjunto de dados de forma descritiva. Poderá utilizar abordagens como (Comportamento Gráfico, Histograma, Cálculos de momento (média, moda, mediana ou outros) e etc). Justificar escolha das técnicas adotadas.

- Aplicar técnicas de levantamento de *features.* Justificar escolha das técnicas adotadas.

- Propor modelos que possam melhor explicar e classificar o conjunto de dados em questão. Não há restrições quanto aos paradigmas dos modelos, podendo ser baseados em paradigmas (Estatísticos, Evolucionários, Conexionistas, Simbólicos, Grafos e etc). Justificar escolha das técnicas adotadas.

- Propor técnicas de particionamento de dados para avaliação dos modelos (*hold-out* e *cross-validation* com *k-folds*). O número de *k-folds* poderá ser determinado pelo candidato. Justificar escolha das técnicas adotadas.

- Propor métricas para mensurar a qualidade do modelo obtido (*score*). Poderá utilizar abordagens como (Acurácia, Matriz de Confusão, *F1 Score, Precision, Recall*). Justificar escolha das técnicas adotadas.

- Propor técnicas de engenharia e arquitetura de software para produzir a melhor solução. Poderá aplicar abordagens como *Design Patterns*, Arquitetura em Camadas e outras técnicas voltadas para soluções Orientadas a Objetos). Justificar escolha das técnicas adotadas.

- O *client-side* deve ser desenvolvido em HTML, CSS e JavaScript (apenas com jQuery, ou com algum *framework* se desejar).

Entregáveis

Você deve entregar um conjunto de artefatos, de acordo com o nível de complexidade que achar melhor. Fique à vontade para escolher o nível de complexidade que desejar!