Roteiro do Projeto: Análise de Letras dos Engenheiros do Hawaii

1. Introdução e Objetivos

Este projeto atende ao desafio proposto de analisar o corpus de letras da banda Engenheiros do Hawaii, que já foi coletado e transformado em um dataset.

O objetivo principal **não é criar modelos complexos**, mas sim "explorar, brincar e refletir sobre como os algoritmos lidam com linguagem, estilo e significado" nas letras da banda.

Os experimentos selecionados para este projeto são:

- **Experimento 1:** Palavras e temas dominantes
- **Experimento 2:** Geração no estilo Engenheiros
- **Experimento 3:** Classificação intuitiva

---

2. Metodologia de Execução (Roteiro)

Cada experimento será desenvolvido em seu próprio notebook, conforme detalhado abaixo.

Experimento 1: Palavras e Temas Dominantes

Objetivo: Identificar o vocabulário central da banda e o que ele revela sobre seus temas.

Ferramentas Principais: Bibliotecas Python (Pandas, NLTK/SpaCy, Scikit-learn, Matplotlib, WordCloud).

Passo a Passo:

1. **Carregamento e Pré-processamento:**
    - Carregar o dataset de letras.
    - Limpeza dos dados (tokenização): converter para minúsculas, remover pontuação e caracteres especiais.
    - Remoção de **Stopwords:** Eliminar palavras comuns da língua portuguesa (ex: 'de', 'que', 'para', 'em') que não carregam significado temático.
2. **Análise 1: Contagem Simples (Top-10 Termos e Nuvem de Palavras)**
    - Contar a frequência de todas as palavras (tokens) restantes no corpus.
    - Gerar uma **Nuvem de Palavras (Word Cloud)** com as 100 ou 200 palavras mais frequentes para visualização.

    - Listar o "Top-10" ou "Top-20" termos mais comuns.
3. **Análise 2: TF-IDF (Term Frequency-Inverse Document Frequency)**
    - Aplicar o TF-IDF considerando cada música como um "documento".
    - O TF-IDF ajudará a identificar palavras que são importantes para uma música específica, mas não necessariamente frequentes em todas as letras.
    - Identificar os termos com maior score médio de TF-IDF no corpus.
4. **Análise e Reflexão (Resultados Qualitativos):**
    - Analisar os resultados da contagem e do TF-IDF.
    - Questões a responder: O vocabulário é mais concreto (ex: "estrada", "noite") ou abstrato (ex: "infinito", "razão")? As palavras remetem a temas recorrentes (crítica social, existencialismo, relacionamentos)?

Experimento 2: Geração no Estilo Engenheiros

Objetivo: Avaliar a capacidade de dois LLMs distintos de mimetizar o estilo lírico da banda e comparar os resultados com as letras originais.

Ferramentas Principais: Acesso a dois modelos de LLM (ex: API da OpenAI para ChatGPT e um modelo via Ollama ou Hugging Face).

Passo a Passo:

1. **Seleção dos Modelos:**
    - **Modelo 1:** ChatGPT (via API, se possível, ou interface web).
    - **Modelo 2:** Um modelo open-source (ex: Llama 3 ou Mistral) acessado via Ollama ou Hugging Face Transformers.
2. **Engenharia de Prompt (Prompt Engineering):**
    - Criar um prompt detalhado para instruir os modelos.
    - Estratégia "Few-Shot": Fornecer 2-3 exemplos de estrofes reais dos Engenheiros do Hawaii no prompt para dar contexto de estilo.
    - Instrução: "Crie uma letra de música no estilo da banda brasileira Engenheiros do Hawaii. Use ironia, referências culturais e temas filosóficos ou existenciais.”
3. **Geração:**
    - Submeter o mesmo prompt (ou prompts adaptados) aos dois modelos selecionados.
    - Coletar as letras geradas.
4. **Análise Comparativa (Resultados Qualitativos):**
    - Comparar as letras geradas (Modelo 1 vs. Modelo 2) com as letras originais da banda.
    - Questões a responder: Os modelos capturaram o "tom" da banda? Houve uso de jogos de palavras? Os temas são similares? As letras

geradas usaram o vocabulário dominante identificado no Experimento 1?

Experimento 3: Classificação Intuitiva

Objetivo: Criar um classificador simples, baseado em regras, para categorizar as letras em "melancólicas", "otimistas" ou "filosóficas".

Ferramentas Principais: Python (Pandas, NLTK/SpaCy).

Passo a Passo:

1. **Definição dos Léxicos (Dicionários):**

Criar manualmente três listas (léxicos) de palavras-chave para cada categoria:

- list_melancolicas = ['triste', 'dor', 'sozinho', 'saudade', 'noite', 'fim', ...]
- list_otimistas = ['sol', 'luz', 'novo', 'dia', 'amanhã', 'esperança', 'sorriso', ...]
- list_filosoficas = ['infinito', 'razão', 'ser', 'mundo', 'porquê', 'tempo', 'vida', ...]

2. **Desenvolvimento do Classificador (Regras Simples):**

Criar uma função em Python que:

- Recebe uma letra de música (já pré-processada, sem stopwords).
- Conta a ocorrência de palavras de cada uma das três listas.
- Atribui a classificação (tag) com base na lista que tiver a maior contagem de palavras.
- Refinamento (opcional): Normalizar a contagem pelo tamanho da música (ex: (contagem / total de palavras) * 100) para evitar que músicas longas sejam favorecidas.

3. **Aplicação e Avaliação:**
   - Aplicar o classificador em todo o dataset.
   - Analisar os resultados quantitativos: Qual a distribuição das categorias? (ex: 40% filosóficas, 35% melancólicas, 25% otimistas).
   - Avaliar qualitativamente (inspeção manual): Selecionar 5 músicas conhecidas (ex: "Infinita Highway", "Pra Ser Sincero", "O Papa é Pop") e verificar se a classificação "intuitiva" do algoritmo faz sentido para o grupo.

---

3. Entregáveis e Avaliação

Para a entrega final (até 09/12), o grupo produzirá:

1. **Repositório Git:**
    - Um notebook .ipynb para cada um dos 3 experimentos detalhados acima.
    - O código será limpo e documentado.
2. **Arquivo README.md:**
    - **Experimentos:** Descrição de quais experimentos foram feitos (os 3 listados acima).
    - **Resultados:** Prints (ex: Nuvem de Palavras, tabelas de Top-10 palavras), gráficos (ex: distribuição das classificações) e exemplos (ex: letras geradas pelas IAs).
    - **Análise Quantitativa/Qualitativa:** Apresentação de métricas (ex: contagens, scores TF-IDF, percentuais de classificação) e, o mais importante, a **interpretação (reflexão final)** do grupo sobre o que a máquina "entendeu" ou "percebeu".
    - **Exploração:** O README também incluirá as perguntas que o grupo fez durante o processo e as tentativas de entender o comportamento dos modelos.