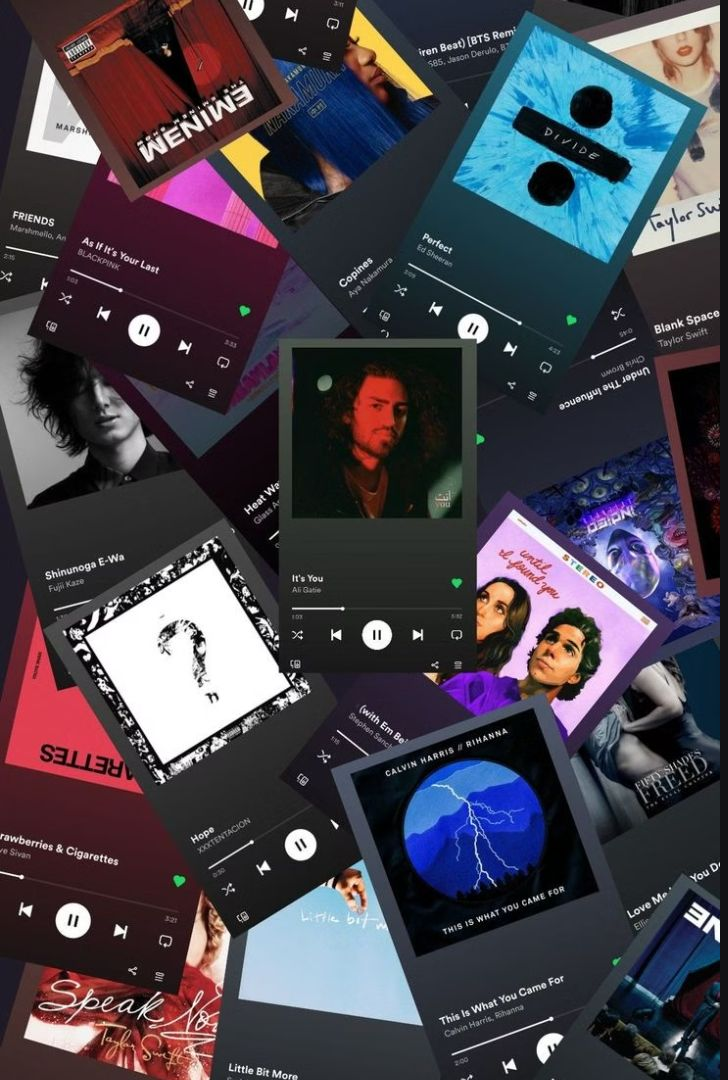


Buscador de Músicas: Recuperação de Informação com Web Scraping

Coleta, Indexação e Busca Inteligente de Letras de Músicas

Grupo: João Álvaro Cardoso de Oliveira, Júlia Vieira, João José Cardoso Ribeiro, Rafael Penido Rocha, Isabela Stefany Pereira de Faria, Henrique Fonseca Araujo

PUC Minas

Introdução

Contexto

Sistema desenvolvido para processar, indexar e buscar letras de músicas.

Técnicas de Processamento de Linguagem Natural (PLN):

- Extração de dados.
- Vetorização TF-IDF.
- Similaridade de cosseno

Objetivo

Explorar métodos para buscas eficientes e precisas.

Descrição Geral do Sistema

1

Coleta e Processamento Inicial

- Varredura e coleta de páginas HTML da internet.
- Extração de dados relevantes (título, artista, letra) e limpeza das informações.

2

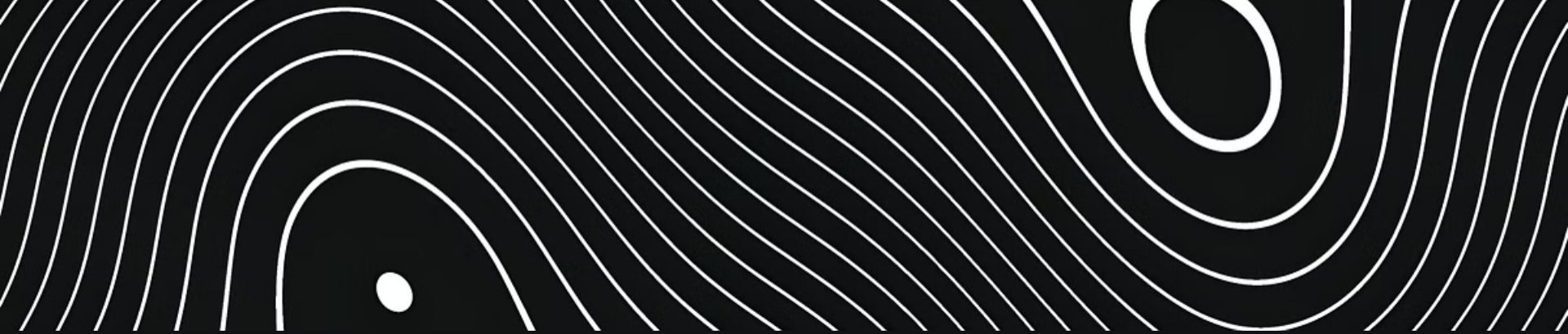
Pré-processamento dos dados

Tokenização, remoção de stopwords, stemming e vetorização dos textos utilizando TF-IDF.

3

Recuperação

Implementação de uma busca baseada em similaridade de cosseno entre a query do usuário e os documentos indexados.



Coleta e Limpeza de Dados

Técnicas

`BeautifulSoup` para facilitar a manipulação e extração dos elementos HTML.

Elementos Extraídos

Títulos (`h1`).

Artistas (`h2`).

- Letras (div com classe lyric-original tratando também as quebras de linha (
)).

Pré-Processamento

1

Detecção de Idioma

A biblioteca `langdetect` identifica a língua de cada letra, permitindo processamento específico para cada idioma.

2

Tokenização

A função `word_tokenize` separa cada letra em tokens, palavras individuais que serão processadas posteriormente.

3

Remoção de Stopwords

Palavras comuns e irrelevantes (como preposições e artigos) são removidas para melhorar a precisão da busca.

4

Stemming

O `SnowballStemmer` reduz as palavras a suas raízes, agrupando termos relacionados para uma busca mais eficiente.

Vetorização e Indexação

TF-IDF

A técnica **TF-IDF** (Term Frequency-Inverse Document Frequency) é utilizada para transformar cada letra em um vetor numérico, representando a importância de cada termo.

Vocabulário

O vocabulário do sistema possui até **30 milhões de termos**, abrangendo uma ampla variedade de palavras.

N-grams

O sistema utiliza **n-grams** (de 1 a 3 palavras) para capturar relações entre termos, melhorando a precisão da busca.

Recuperação de Informação

Similaridade de Cosseno

Método crucial para determinar a relevância entre consultas e documentos, medindo a semelhança entre vetores por meio do ângulo entre eles.

Funções de Busca

O sistema utiliza `buscar_texto` (para o documento mais similar) e `buscar_texto_multiple` (para os `n` documentos mais similares).

Estrutura do Sistema

Diretórios do sistema:

Origem

Contém arquivos HTML originais.

Destino

Contém arquivos processados.

Transformação

Armazena textos pré-processados e vetorizados.

Desempenho do Sistema

Diretório: Amostra com aproximadamente 11.000 arquivos.

Tempo de Indexação

66,84 segundos.

Consumo de Memória

453,26 MB.

Precisão Alta

Frases completas.

Precisão Baixa

Palavras isoladas.

Potenciais Melhorias

Suporte a Mais Idiomas

Expansão no pré-processamento para incluir mais idiomas. Atualmente, o sistema suporta apenas português e inglês.

Uso de Embeddings

Substituir TF-IDF por Word2Vec ou BERT. Modelos de linguagem podem capturar relações semânticas, melhorando a precisão das buscas.

Interface Web

Criar uma interface web para tornar o sistema acessível para usuários finais. Atualmente, o sistema é acessado por meio de scripts.

Indexação Incremental

Adicionar novos documentos sem reindexar o sistema inteiro. Isso reduz o tempo de processamento e permite atualizações mais frequentes.

Conclusão

Este sistema demonstrou a eficácia da combinação de técnicas clássicas de PLN, como TF-IDF e similaridade de cosseno, para recuperar informações relevantes em um conjunto de documentos. As melhorias propostas visam aumentar ainda mais a precisão e a usabilidade do sistema.

Nos próximos passos, nos concentraremos em refinar ainda mais o sistema, implementando técnicas avançadas de processamento de linguagem natural e aprendizado de máquina para melhorar a relevância e a personalização dos resultados de busca.