Recomendação com Similaridade do Cosseno

Introdução:

A recomendação de textos baseada em similaridade semântica tem se tornado uma técnica essencial em diversas aplicações de Processamento de Linguagem Natural (PLN). Neste projeto, foi desenvolvido um sistema capaz de sugerir receitas culinárias semelhantes a partir de suas descrições completas, utilizando a similaridade do cosseno como métrica principal. A abordagem permite identificar relações de proximidade entre textos mesmo quando vocabulários distintos são utilizados para expressar ideias parecidas.

Objetivo:

O objetivo deste projeto foi desenvolver um sistema simples de recomendação textual utilizando a similaridade do cosseno como métrica principal. Essa técnica, fundamentada em conceitos da Álgebra Linear, é amplamente aplicada em Processamento de Linguagem Natural (PLN) para medir a semelhança entre textos, mesmo que utilizem vocabulários diferentes.

A proposta consiste em transformar as descrições das receitas em vetores numéricos por meio do modelo TF-IDF, e então aplicar a fórmula:

$$\text{Similaridade}(A, B) = A \cdot B \,/\, \|A\| \times \|B\|$$

Isso permite identificar receitas similares com base em seus textos completos, mesmo que usem expressões distintas para descrever ingredientes ou modos de preparo semelhantes.

Dados do Dataset e tratamentos:

O Dataset trata de um conjunto de receitas, e está organizado como, Linha | Nome | Tempo de Preparo | Tempo de Cozimento | Tempo Total | Porções | Rendimento | Ingredientes | Instruções | URL | Tempo Adicional.

O Dataset original tinha dados alocados em diversas colunas, por isso foi necessário tratá-lo passando os dados para a primeira coluna transformando ele assim no dataset, test_recipes_single_column.csv.xlsx.

O arquivo, test_recipes_single_column.csv.xlsx, contém uma única coluna chamada "A", onde estão reunidas as informações completas de cada receita: nome, ingredientes e instruções de preparo, separadas por caractere pipe (|).

O pré-processamento aplicado incluiu:

- Conversão de todos os textos para letras minúsculas;
- Remoção de pontuação e caracteres especiais;
- Tokenização das sentenças (divisão em palavras);
- Eliminação de stopwords da língua inglesa, que não contribuem semanticamente (como “the”, “and”, “of”).

Esse processo teve como objetivo limpar e padronizar os dados para que a vetorização com TF-IDF refletisse de forma mais fiel a essência de cada receita.

Desenvolvimento do Algoritmo:

O sistema foi implementado em Python, utilizando o ambiente do Google Colab, e envolveu as seguintes etapas:

1. Leitura do arquivo Excel com a biblioteca pandas;
2. Aplicação da função preprocess para realizar a limpeza textual com auxílio do nltk;
3. Transformação dos textos em vetores TF-IDF com TfidfVectorizer;
4. Cálculo da similaridade do cosseno entre todos os vetores;
5. Extração automática do nome da receita, que está posicionado logo após o primeiro caractere |;
6. Retorno das 5 receitas mais semelhantes, excetuando a própria entrada consultada.

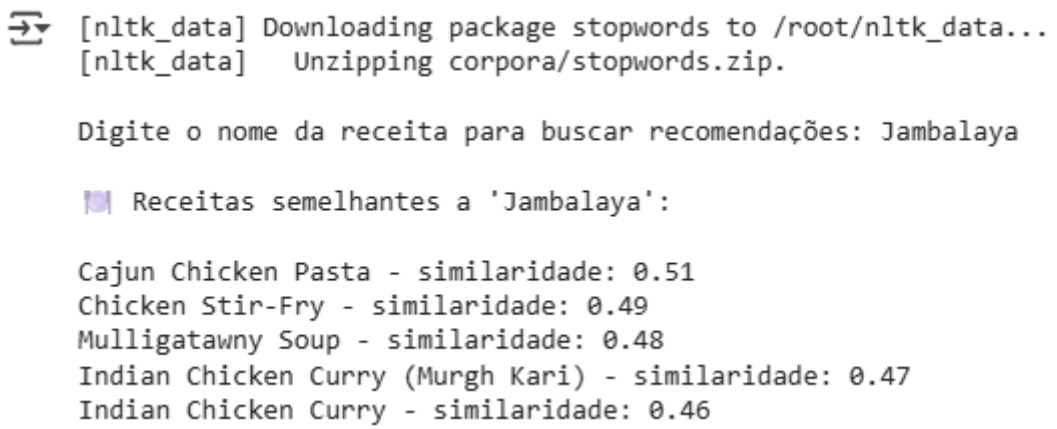
O sistema permite ao usuário digitar o nome de uma receita presente no dataset e receber como resposta outras sugestões com descrições semelhantes, considerando o texto completo e não apenas palavras coincidentes.

Resultados:

Durante os testes, o sistema demonstrou ser capaz de capturar relações semânticas entre diferentes receitas. Mesmo quando o vocabulário usado na descrição variava, o algoritmo conseguiu identificar receitas com ingredientes e métodos de preparo parecidos.

Ao digitar, por exemplo, o nome de uma receita como *“Jambalaya”*, o sistema retornava sugestões com carnes preparadas de forma semelhante, ou com molhos e acompanhamentos correlatos.

Exemplo:



Conclusão:

Este trabalho mostrou como é possível aplicar conceitos de Ciência de Dados e PLN na criação de um sistema funcional de recomendação de receitas. A similaridade do cosseno, aliada à vetorização com TF-IDF, mostrou-se eficaz na comparação de descrições culinárias, mesmo em um dataset não estruturado, além de evidenciar a importância do pré-processamento.