TRABALHO DE ÁLGEBRA LINEAR:
ALGORITMO DA SIMILARIDADE DE COSSENO

Angelica Fonseca Garcia

1. INTRODUÇÃO

Este é um trabalho desenvolvido para a disciplina de Álgebra Linear, no curso de Ciência de Dados do 2º semestre de 2025 da Fatec Rubens Lara. Tem como objetivo desenvolver o algoritmo de processamento de linguagem natural por meio da técnica TF-IDF (Term Frequency-Inverse Document Frequency) utilizando a Similaridade do Cosseno conforme a análise de textos do dataset com a linguagem de programação para sugerir produtos similares conforme sua descrição. A linguagem de programação inicial utilizada foi Python pelo ambiente de desenvolvimento Spyder na plataforma Anaconda, finalizada no Google Colab.

2. DESENVOLVIMENTO

2.1 DATASET

O dataset utilizado foi obtido na plataforma Kaggle com o nome de Coffee Reviews Dataset e contém diversas informações sobre cafés avaliados, incluindo:

1. `name`: Nome do grão ou do blend (blend é a mistura de grãos)
2. `roaster`: Nome do torrador
3. `roast`: Tipo de torra (clara, média clara, média, média escura e escura)
4. `loc_country`: Localização do torrador
5. `origin_1`: Origem do grão 1
6. `origin_2`: Origem do grão 2
7. `100g_USD`: Preço por 100g do grão em US Dólar
8. `rating`: Classificação do café
9. `review_date`: Data da avaliação do café
10. `desc_1`: Texto da avaliação1
11. `desc_2`: Texto da avaliação2
12. `desc_3`: Texto da avaliação3

Foi escolhido o este dataset em CSV, que contém os dados necessários para uma avaliação de similaridade de cosseno, podendo escolher uma coluna da linha referência, como no caso o nome do café e aplicar as técnicas de processamento do texto na coluna que possui as avaliações do café. Não foi realizado tratamento no dataset, e as colunas utilizadas para o trabalho foram:

`name:` Nome do grão ou do blend (blend é a mistura de grãos)
Coluna de referência para orientação à coluna da análise do texto de avaliação sensorial do café.

`desc_1:` Texto da avaliação #1
Coluna contendo texto de avaliação sensorial do café que se aplica o algoritmo de PLN.

2.2 IMPORTAÇÃO DE BIBLIOTECAS
Para iniciar o desenvolvimento do algoritmo foi certificado as instalações das bibliotecas no Anaconda Prompt:

- pip install pandas - para manipulação e análise dos dados.
- pip install scikit-learn - Para vetorização dos textos (TF-IDF) e cálculo da similaridade do cosseno.
- pip install nltk - para processamento dos textos (remoção de stopwords).
- pip install googletrans==4.0.0-rc1 - para traduzir as descrições sensoriais dos cafés para o português.

posteriormente no Spyder foi utilizado as seguintes ferramentas de nltk:
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')
e para Colab foi necessário: nltk.download('punkt_tab')

2.3 CARREGAMENTO E PRÉ-PROCESSAMENTO

Após a instalação é solicitado o download dos recursos nltk que realizará tratamentos nas frases e palavras como:
punkt que ensina o NLTK a saber onde termina uma frase ou palavra em um texto.
stopwords onde são removidas dos textos uma lista de palavras durante o pré-processamento (como “the”, “and”, “em”, “de”, etc.).

wordnet (opcional): reduzir uma palavra à sua forma base (por exemplo, "running" vira "run").

Utilização das ferramentas de nltk *stopwords.words('english')) e lemmatizer = WordNetLemmatizer().* Com a função de pré-processamento *preprocess_text* para tratar os textos e funções. *TfidfVectorizer(), fit_transform() e cosine_similarity() p*ara aplicar as colunas do dataset, convertendo os textos em vetores com TF-IDF e calcular a similaridade do cosseno entre todos os cafés.

2.4 FUNÇÃO DE RECOMENDAÇÃO E BUSCA

A função ‘recomendar cafés’ utiliza a matriz de similaridade do cosseno para identificar os cafés mais parecidos com base na descrição sensorial do café. O café de referência é exibido junto com uma lista de recomendações similares.
Para melhor visualização do texto comparativo operou a função ‘traduzir_texto’ usa a biblioteca ‘googletrans’para traduzir automaticamente as descrições do inglês para o português.
A função ‘buscar_por_nome’ permite que o usuário digite parte do nome ou variedade de um café, e receba uma listagem em índice para a escolha do café referência, após a escolha do índice é apresentado 3 recomendações por nível de similaridade onde o das descrições das avaliações sensoriais.

3. RESULTADOS

Faça o upload do arquivo `coffee_analysis.csv` que se encontra no repositório do Github

Para executar no Google Colab entre no link:
https://colab.research.google.com/drive/14f5_XfTknL5NV0dPyX4LL-5Au8d9gzSY#scrollTo=kbtDmzkV77tn

Digite parte do nome ou variedade do café que deseja pesquisar.
Devido se tratar de cafés especiais, cito algumas principais variedades de cafés para inspiração:
a. Bourbon
b. Catuai
c. Acaia
d. Icatu

Será entregue ao usuário nomes de café que contém o nome escolhido, escolha o café pelo índice e aperte Enter.

Serão apresentadas 3 recomendações baseado no valor em porcentagem de similaridade ao café escolhido.