TRABALHO DE ÁLGEBRA LINEAR:
ALGORITMO DA SIMILARIDADE DE COSSENO

Angelica Fonseca Garcia

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 utilizada foi Python pelo ambiente de desenvolvimento Spyder na plataforma Anaconda.

DATASET e DESENVOLVIMENTO

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
6. `origin_2`: Origem do grão
7. `100g_USD`: Preço por 100g do grão em US Dolár
8. `rating`: Classificação do café
9. `review_date`: Data da avaliação do café

10. `desc_1`: Texto da avaliação #1
11. `desc_2`: Texto da avaliação #2
12. `desc_3`: Texto da avaliação #3

Foi escolhido o seguinte dataset pois contém o que era necessário para uma avaliação de similaridade de cosseno, podendo escolher uma coluna da linha referência, como no caso o nome do café, aplicar as técnicas de processamento do texto na coluna que possui as avaliações do café. Não foi realizado nenhuma edição no dataset, está carregado na íntegra, porém 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.

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, tokenização).
pip install googletrans==4.0.0-rc1 - para traduzir as descrições sensoriais dos cafés para o português.

A linguagem de programação utilizada para desenvolver o trabalho foi Python 3.12 no ambiente Anaconda Navigator 2.6.2.

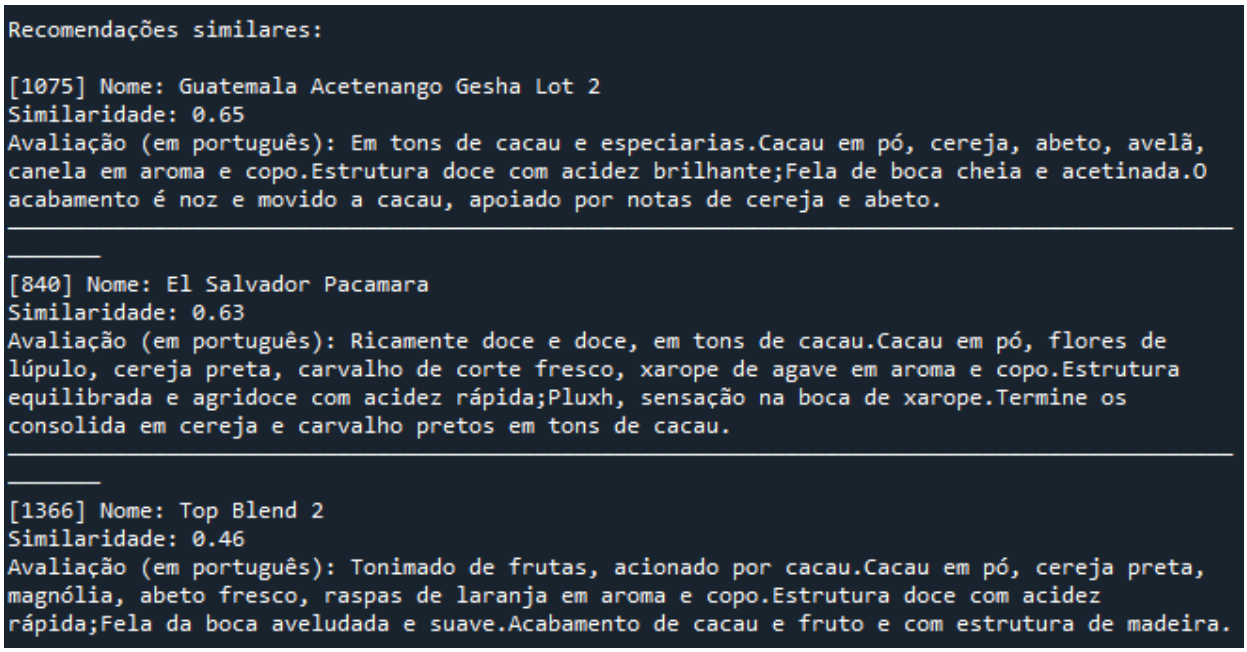
RESULTADOS

```
Digite o nome (ou parte do nome) do café que deseja buscar: catuai
Cafés encontrados:
[595] Panama Elida Natural Catuai
[1123] Honduras Catuai/Bourbon
[1426] Guatemala Finca El Principito Yellow Catuai
[1516] Elida Estate Dragonfly Catuai Lot 13
[1519] Elida Estate ASD Natural Catuai 15
Digite o número (índice) do café para ver recomendações:
```

Utilizar vetorização TF-IDF e cálculo de similaridade do cosseno para encontrar cafés com descrições semelhantes, com base em uma escolha feita pelo usuário.

```
[1426] Guatemala Finca El Principito Yellow Catuai
[1516] Elida Estate Dragonfly Catuai Lot 13
[1519] Elida Estate ASD Natural Catuai 15
Digite o número (índice) do café para ver recomendações: 1426
================================================================================
=======
Café de referência:

[1426] Nome: Guatemala Finca El Principito Yellow Catuai
Avaliação (em português):
Em tons de cacau e especiarias.Cacau em pó, cereja preta, carvalho de corte fresco, pistache,
especiarias para aroma e copo.Estrutura doce com acidez equilibrada;Fela de boca completa e
xarope.O acabamento é noz e movido a cacau, apoiado por notas de cereja e carvalho pretos.
================================================================================
=======
```



O valor da similaridade do cosseno varia de:

0.0: nenhuma similaridade entre os textos

1.0: textos idênticos (ou praticamente iguais)

exemplo: Isso significa que a descrição desse café é 62% semelhante, matematicamente falando, à descrição do café que você escolheu como referência.