TF-IDF TRABALHO - Enzo Ura

Enzo Ura

November 2025

1 Introdução

Sistema de Recomendação Odontológica com TF-IDF Aluno: **Enzo Shinji Sugano Ura** November 4, 2025

2 Introdução

O trabalho tem como objetivo desenvolver um sistema de recomendação de procedimentos odontológicos utilizando a técnica de **TF-IDF** (Term Frequency–Inverse Document Frequency) e a métrica de **similaridade do cosseno**.

A proposta é permitir que o usuário digite um procedimento odontológico ou uma breve descrição, e o sistema retorne os tratamentos mais semelhantes com base em um banco de dados textual. A escolha do tema visa demonstrar a aplicação prática de técnicas de mineração de texto na área da saúde.

3 Descrição do Dataset

O conjunto de dados é composto por 10 procedimentos odontológicos, com suas respectivas descrições resumidas:

- **Limpeza dental** — remoção de placa bacteriana e tártaro, polimento e aplicação de flúor.
- **Restauração dental** — tratamento para reparar dentes danificados por cáries ou fraturas.
- **Extração dental** — remoção de dentes comprometidos por cáries profundas ou infecção.
- **Clareamento dental** — procedimento estético para clarear o tom dos dentes.
- **Canal (endodontia)** — tratamento para remover a polpa dentária infectada e selar o dente.
- **Implante dentário** — colocação de pino de titânio para substituir dentes ausentes.
- **Aparelho ortodôntico** — dispositivo fixo ou móvel para corrigir o alinhamento dos dentes.
- **Profilaxia infantil** — limpeza preventiva em crianças para evitar cáries e gengivite.
- **Raspagem periodontal** — remoção profunda de tártaro abaixo da gengiva.
- **Prótese dentária** — substituição de dentes perdidos por próteses fixas ou removíveis.

4 Metodologia

O sistema foi implementado em **Python** utilizando a biblioteca `scikit-learn`. Cada descrição textual é convertida em um vetor numérico com base no modelo **TF-IDF**.

4.1 Cálculo do TF-IDF

O TF-IDF mede a relevância de uma palavra dentro de um documento e em relação ao conjunto total de documentos. A fórmula geral é:

$$TFIDF(t,d) = TF(t,d) \times IDF(t)$$

5 Análise das Semelhanças e dos Ângulos

A análise dos ângulos mostra como o modelo identifica relações semânticas entre os procedimentos. Por exemplo:

- **Limpeza dental** e **Profilaxia infantil** — ambas envolvem remoção de placa e prevenção, portanto apresentam alta similaridade (ângulo pequeno, cerca de 25° a 30°).
- **Raspagem periodontal** e **Limpeza dental** — possuem objetivos semelhantes, mas uma é mais profunda. Similaridade moderada (ângulo de 40° a 50°).
- **Clareamento dental** e **Implante dentário** — tratam de áreas distintas (estética e cirurgia), logo o ângulo entre os vetores é grande (próximo de 80°), indicando baixa similaridade.
- **Canal** e **Extração dental** — embora sejam tratamentos diferentes, ambos envolvem dentes danificados e processos de remoção de tecido, com similaridade média (ângulo de 45° a 55°).

Essas comparações mostram que o TF-IDF, mesmo sendo um método simples, consegue capturar padrões relevantes de contexto. O "ângulo" funciona como uma medida geométrica de semelhança textual: quanto mais próximas as descrições, mais "paralelos" são os vetores no espaço vetorial.

6 Interação com o Usuário

O programa é interativo: o usuário informa o nome ou uma descrição do procedimento e o sistema retorna os três mais semelhantes. Para encerrar a execução, basta digitar `sair`.

7 Código em Python

```
from␣sklearn.feature_extraction.text␣import␣TfidfVectorizer
from␣sklearn.metrics.pairwise␣import␣cosine_similarity

procedimentos␣=␣{
␣␣␣␣"Limpeza␣dental":␣"r e m o ␣␣o␣de␣placa␣bacteriana␣e␣t rtaro ,␣polimento␣e␣fl␣or.",
␣␣␣␣"R e s t a u r a ␣␣o␣dental":␣"tratamento␣para␣reparar␣dentes␣danificados.",
␣␣␣␣"E x t r a ␣␣o␣dental":␣"r e m o ␣␣o␣de␣dentes␣comprometidos␣por␣c␣r i e s␣profundas.",
␣␣␣␣"Clareamento␣dental":␣"clarear␣o␣tom␣dos␣dentes␣com␣agentes␣clareadores.",
␣␣␣␣"Canal␣(endodontia)":␣"r e m o ␣␣o␣da␣polpa␣dent r i a␣infectada␣e␣selamento.",
␣␣␣␣"Implante␣dent r i o":␣"c o l o c a ␣␣o␣de␣pino␣de␣t i t n i o␣para␣substituir␣dentes␣ausentes."
␣␣␣␣␣␣↪␣,
␣␣␣␣"Aparelho␣ortod n t i c o":␣"uso␣de␣dispositivo␣fixo␣ou␣m v e l␣para␣corrigir␣dentes.",
␣␣␣␣"Profilaxia␣infantil":␣"limpeza␣preventiva␣em␣c r i a n ␣as␣para␣evitar␣c r i e s.",
␣␣␣␣"Raspagem␣periodontal":␣"r e m o ␣␣o␣profunda␣de␣t rtaro␣abaixo␣da␣gengiva.",
␣␣␣␣"Pr t e s e␣dent r i a":␣"s u b s t i t u i ␣␣o␣de␣dentes␣perdidos␣por␣pr teses."
}

nomes␣=␣list(procedimentos.keys())
descricoes␣=␣list(procedimentos.values())

vetorizador␣=␣TfidfVectorizer()
matriz_tfidf␣=␣vetorizador.fit_transform(descricoes)

print("===␣Sistema␣de␣R e c o m e n d a ␣␣o␣Odontol g i c a␣===")
print("Digite␣um␣procedimento␣(ou␣'sair'␣para␣encerrar)")

while␣True:
␣␣␣␣entrada␣=␣input("Procedimento:␣").strip().lower()
␣␣␣␣if␣entrada␣==␣"sair":
␣␣␣␣␣␣␣␣print("Encerrando␣o␣sistema...")
␣␣␣␣␣␣␣␣break

␣␣␣␣entrada_tfidf␣=␣vetorizador.transform([entrada])
```

```
    similaridades = cosine_similarity(entrada_tfidf, matriz_tfidf)[0]
    indices_ordenados = similaridades.argsort()[::-1]

    print("\nMais semelhantes:")
    for i in indices_ordenados[:3]:
        print(f"- {nomes[i]} ({similaridades[i]:.2f})")
```

8 Conclusão

O uso do TF-IDF aliado à similaridade do cosseno mostrou-se eficaz na criação de um sistema de recomendação textual simples e funcional. Mesmo com um conjunto de dados pequeno, a técnica conseguiu identificar relações reais entre os procedimentos odontológicos com base na linguagem natural.

A interpretação dos ângulos entre vetores reforça a importância do modelo: ângulos pequenos indicam forte correlação entre termos, enquanto ângulos grandes revelam diferenças semânticas significativas.

Como aprimoramentos futuros, pode-se incluir:

- Análise semântica com embeddings (Word2Vec, BERT);
- Expansão do dataset com descrições mais detalhadas;
- Interface gráfica para facilitar a interação do usuário.