# Análise de Similaridade de Textos com TF-IDF e Ângulos Vetoriais

Autor: Arthur Davino Rizzo — FATEC — Álgebra Linear

O presente trabalho aplica conceitos de Álgebra Linear na análise de similaridade entre textos, utilizando o modelo TF-IDF (Term Frequency–Inverse Document Frequency). Essa técnica transforma documentos em vetores num espaço de alta dimensão, permitindo medir a proximidade semântica por meio do ângulo entre esses vetores.

## Tema do Trabalho

O tema escolhido foi a comparação de textos jornalísticos sobre esportes e tecnologia. O objetivo é identificar como o vocabulário e o conteúdo influenciam a similaridade entre textos desses dois domínios. Essa análise demonstra como a Álgebra Linear é utilizada em Processamento de Linguagem Natural (PLN).

## Descrição do Dataset

O conjunto de dados (dataset) é composto por 20 textos curtos em português, criados para simular notícias e comentários sobre futebol e tecnologia esportiva. Dez textos tratam de partidas, jogadores e desempenho em campo, enquanto os outros dez abordam o uso de tecnologia, inteligência artificial e análise de dados no esporte.

## Metodologia

Cada texto foi convertido em um vetor TF-IDF usando a biblioteca scikit-learn. Em seguida, foi calculada a similaridade do cosseno entre todos os pares de textos, e o ângulo entre vetores foi obtido pela fórmula: $\theta = \arccos(\text{sim}(A,B)) \times 180/\pi$. Valores de ângulo menores indicam textos mais semelhantes.

## Resultados

A análise mostrou que os textos sobre o mesmo tema apresentam ângulos menores, ou seja, maior similaridade. Os textos sobre futebol e sobre tecnologia foram agrupados corretamente pelo modelo, evidenciando a coerência temática e lexical.

## Conclusão

O modelo TF-IDF, aliado à Álgebra Linear, demonstrou ser uma ferramenta eficaz para identificar padrões de similaridade textual. Essa metodologia tem aplicações práticas em mecanismos de busca, recomendação de conteúdo e análise de notícias esportivas.