CENTRO ESTADUAL DE EDUCAÇÃO TECNOLÓGICA PAULA SOUZA
Faculdade de Tecnologia Baixada Santista Rubens Lara

Curso Superior de Tecnologia em Ciência de Dados

Álgebra Linear – Algoritmo de Similaridade de Cosseno

Luís Felipe Ruas do Nascimento

Santos 2025

Relatório Técnico – Aplicação do TF-IDF e Similaridade do Cosseno

Resumo Técnico – Projeto de Ciência de Dados

1 Resumo:

Este relatório apresenta a aplicação do algoritmo *Term Frequency–Inverse Document Frequency* (TF-IDF) em um conjunto de dados relacionado ao futebol, com o objetivo de calcular similaridade entre descrições textuais de jogadores. O método foi implementado em Python com suporte das bibliotecas `NLTK` e `Scikit-learn`, realizando pré-processamento, vetorização e cálculo de similaridade pelo cosseno.

2 Introdução

A análise de similaridade textual é amplamente utilizada em sistemas de recomendação, mineração de texto e recuperação de informação. O algoritmo TF-IDF é um dos métodos mais eficientes para converter documentos em vetores numéricos, permitindo a aplicação de métricas matemáticas como a similaridade do cosseno.

Neste estudo, utiliza-se um dataset contendo informações e descrições de jogadores de futebol, com a finalidade de identificar semelhanças entre perfis textuais.

3 Descrição do Dataset

O conjunto de dados utilizado possui uma coluna principal com descrições textuais de jogadores, contendo atributos como posição, características técnicas e estilo de jogo. O arquivo foi importado em formato CSV e manipulado por meio da biblioteca `pandas`.

4 Metodologia

4.1 Pré-processamento dos Dados

O texto foi padronizado por meio das seguintes etapas:

- Conversão para letras minúsculas;
- Remoção de pontuação utilizando expressões regulares;
- Tokenização;
- Remoção de *stopwords* em português (`nltk.corpus.stopwords`);

- Reconstrução do texto limpo em formato string.

4.2 Aplicação do TF-IDF

Após o pré-processamento, utilizou-se a classe `TfidfVectorizer` para transformar cada descrição em um vetor numérico baseado na relevância das palavras no corpus.

4.3 Similaridade do Cosseno

A similaridade entre os vetores TF-IDF foi calculada através da métrica de cosseno, gerando uma matriz de similaridade que indica o grau de proximidade semântica entre jogadores.

$$\cos\text{sim}(u, v) = \frac{u \cdot v}{\|u\| \, \|v\|} \quad \text{onde} \quad u \cdot v = \sum_{i=1}^{n} u_i v_i$$

5 Resultados

A matriz de similaridade permitiu identificar jogadores com perfis próximos a partir de suas descrições. O sistema demonstrou que textos contendo termos técnicos comuns, como "meia ofensivo"ou "velocidade e drible", apresentaram alta similaridade numérica.

Os valores gerados variaram entre 0 e 1, onde 1 indica total equivalência textual e 0 ausência completa de similaridade. Em relação a angulação, quanto mais perto de 0° mais similaridade terá e quanto mais perto de 90° menos similaridade terá (ortogonal)

6 Conclusão

O uso do TF-IDF combinado com a similaridade do cosseno mostrou-se eficiente para análise automática de proximidade descritiva entre jogadores de futebol. O método pode ser expandido para aplicações como recomendação de atletas, agrupamento de perfis e análises comparativas em scouting esportivo.