

Relatório - Similaridade de Pokémon com TF-IDF e Cosseno

1. Introdução

Este relatório apresenta a aplicação de técnicas de Álgebra Linear, especificamente o cálculo de similaridade do cosseno, utilizando a vetorização TF-IDF em um conjunto de dados de Pokémon. A análise tem como objetivo mostrar como essas técnicas podem ser utilizadas para identificar a similaridade semântica entre diferentes entradas textuais, neste caso, os nomes e tipos dos Pokémon.

A similaridade do cosseno é amplamente utilizada em sistemas de recomendação, busca textual e classificação de documentos, sendo uma técnica essencial na área de Ciência de Dados e Machine Learning.

2. Objetivo

Desenvolver um algoritmo em Python que permita ao usuário comparar dois Pokémon diferentes com base na similaridade entre seus nomes e tipos, utilizando TF-IDF e a similaridade do cosseno. Este projeto tem como foco principal aplicar os conhecimentos de Álgebra Linear na prática, demonstrando como vetores e medidas de similaridade podem ser usados para análise de dados.

3. Motivo da Escolha do Dataset

O dataset de Pokémon foi escolhido por ser de fácil acesso, bem estruturado e popular entre o público, o que facilita o entendimento do problema e o engajamento com a análise. O conjunto de dados contém informações como o nome do Pokémon e seus dois tipos principais (ex: 'Fire', 'Water'), que são ideais para uma análise textual básica, pois permitem comparar características sem depender de atributos numéricos.

Relatório - Similaridade de Pokémon com TF-IDF e Cosseno

4. Algoritmo Utilizado

O algoritmo consiste em duas etapas principais:

1. Vetorização TF-IDF: Cada Pokémon é representado como um vetor numérico baseado em seus atributos textuais (nome, tipo 1 e tipo 2). O TF-IDF atribui pesos às palavras com base na sua frequência em relação ao restante do dataset, destacando termos mais relevantes e diminuindo a influência dos mais comuns.

2. Similaridade do Cosseno: Uma vez vetorizadas as descrições dos Pokémon, calcula-se a similaridade do cosseno entre dois vetores. Isso resulta em um valor entre 0 e 1, onde 1 representa vetores idênticos (alta similaridade) e 0 indica vetores ortogonais (sem similaridade).

A combinação dessas técnicas permite uma análise eficaz do quão semanticamente próximos dois Pokémon são com base em seus dados categóricos.

5. Exemplo Prático

Considerando os Pokémon 'Pikachu' (Elétrico) e 'Raichu' (Elétrico), ambos compartilham o mesmo tipo e nomes semelhantes. O algoritmo retorna uma alta similaridade entre eles, o que condiz com a realidade do universo Pokémon, já que Raichu é a evolução direta de Pikachu.

Por outro lado, ao comparar 'Pikachu' com 'Charizard' (Fogo/Voador), a similaridade é muito menor, refletindo a diferença tanto nos nomes quanto nos tipos de cada Pokémon.

Relatório - Similaridade de Pokémon com TF-IDF e Cosseno

Esses testes validam a eficácia do algoritmo em capturar padrões semânticos, mesmo com um modelo simples baseado em texto.

6. Resultados Obtidos

Foram realizados testes com diferentes pares de Pokémon. Os resultados variaram conforme a semelhança nos nomes e nos tipos. Seguem alguns exemplos de similaridades obtidas:

- Pikachu x Raichu: Similaridade aproximadamente 0.65
- Bulbasaur x Ivysaur: Similaridade aproximadamente 0.67
- Pikachu x Charizard: Similaridade aproximadamente 0.20
- Gengar x Alakazam: Similaridade aproximadamente 0.31

Estes resultados mostram que a abordagem é sensível a características comuns, validando sua aplicação em análises exploratórias e sistemas de recomendação simples.

7. Conclusão

Este projeto demonstrou a utilidade das técnicas de Álgebra Linear na análise de dados categóricos. Ao combinar TF-IDF com similaridade do cosseno, foi possível criar um sistema que compara Pokémon de forma textual, destacando relações semânticas com precisão razoável.

Apesar de ser uma abordagem introdutória, ela serve como base para aplicações mais complexas em processamento de linguagem natural (PLN), recomendação de itens e categorização automática. A escolha do dataset e a simplicidade do modelo tornam esta análise acessível e

Relatório - Similaridade de Pokémon com TF-IDF e Cosseno

pedagógica para iniciantes em Ciência de Dados.