FATEC RUBENS LARA - BAIXADA SANTISTA

SIMILARIDADE DO COSSENO
(SIMILARIDADE ENTRE OS LIVROS DE T.I)

Nome:Yamine Moura Paulino da Silva

SANTOS, SÃO PAULO
2025

similaridade do cosseno

Similaridade entre os livros de t.i

May 19, 2025

1 Introduction

O conjunto de dados Nomes e descrições de livros de TI contém 8.558 linhas, o conjunto de dados foi obtido através do Kaggle em formato Excel. O objetivo será analisar quais são os livros que tem a descrição mais similar ao livro Game Development with GameMaker Studio 2.

1.1 Preparação dos Dads

Serão utilizadas algumas bibliotecas, como Pandas para carregar o arquivo Excel; TfidfVectorizer para transformar as descrições em vetores numéricos usando TF-IDF (Term Frequency- Inverse document Frequency); Cosine Similarity para calcular a similaridade do cosseno entres os vetores gerado; e, por fim, a bibliote NLTK para importar a lista de palavra irrelevantes, como "e", "o", etc., chamadas de **stopwords**.

1.2 Trasformação dos Dados

As *stopwords* (palavras irrelevantes como 'e', 'o') são removidas do texto, e em seguida, as descrições foram transformadas em vetores numéricos utilizando o método TF-IDF (Term Frequency- Inverse document Frequency).

```
vectorizer = TfidfVectorizer(stop_words='english')
tfidf_matrix = vectorizer.fit_transform(data['description'])
```

Para calcular a similaridade do cosseno entre o livro escolhido e os outros livros, foi utilizado o método **Cosine Similarily**.

```
cosine_sim = cosine_similarity(tfidf_matrix, tfidf_matrix)
```

Foram removidas as linhas que não possuem descrições e aquelas que tem a descrições muito curtas (menos de 30 carateres).

```
data = data.dropna(subset=['description'])
data = data[data['description'].str.len() > 30].reset_index(drop=
    True)
```

O segundo passo foi escolher o livro Game Development with GameMaker Studio 2 para comparar com os demais.

```
filme_referencia = data[data['book_name'].str.contains("Game
    Development with GameMaker Studio 2", case=False)].iloc[0]
```

Para refazer a vetorização de forma avançada, foi utilizado o parâmetro **min df** para ignorar palavras que aparecem com menos de duas descrições, e o paramêtro **ngram range** para capturar tanto palavras individuais (unigramas) quanto combinações de duas palavras (bigramas).

```
vectorizer = TfidfVectorizer(stop_words=stop_words, min_df=2,
    ngram_range=(1, 2))
tfidf_matrix = vectorizer.fit_transform(data['description'])
```

Por útlimo, foi criada uma lista com os nomes dos livros e seus valores de similaredades, ignorando o livro escolhido, para não recomendar ele mesmo. Em seguida, foi exibida uma lista com os 10 livros mais similareas em ordem decrescente.

```
# Gerar lista de similaridade (excluindo ele mesmo)
similarities = [
    (data.iloc[i]['book_name'], cosine_sim[i])
    for i in range(len(data)) if i != ref_index and cosine_sim[i] >
     0
]

# Ordenar por similaridade
similarities.sort(key=lambda x: x[1], reverse=True)

# Mostrar top 10 resultados
print(f"Filmes mais similares a '{filme_referencia['book_name
    ']}':")
for title, sim in similarities[:10]:
    print(f"{title}: Similaridade = {sim:.4f}")
```

2 Resultado

Os valores variam de 0 a 1, quanto mais próximo de 0 significa que não há similaridade, quanto mais próximo de 1 significa que são idênticos.

```
Filmes mais similares a 'Game Development with GameMaker Studio 2':

GameMaker Essentials: Similaridade = 0.3714

Beginning Game Development with Amazon Lumberyard: Similaridade =
    0.3056

HTML5 Game Development from the Ground Up with Construct 2:
    Similaridade = 0.2945

C++ Game Development Cookbook: Similaridade = 0.2922

Learning Windows 8 Game Development: Similaridade = 0.2886

The Game Jam Survival Guide: Similaridade = 0.2863

```

```
Make a 2D Arcade Game in a Weekend: With Unity: Similaridade =
    0.2834

iOS Game Development By Example: Similaridade = 0.2811

Mastering Cocos2d Game Development: Similaridade = 0.2801

GameMaker: Studio For Dummies: Similaridade = 0.2759
```

GameMaker Essentials é o mais semelhante ao livro de referência, com uma similaridade de 0,37. Isso indica que a descrição do conteúdo possui muitos termos em comum com Game Development with GameMaker Studio 2. Ambos os livros abordam conceitos semelhantes relacionado ao aprendizado da linguagem GameMaker (GML) e desenvolvimento de jogos.

Beginning Game Development with Amazon Lumberyard apresenta uma similaridade de 0,30, sendo um pouco menos semelhante ao livro Game Development with GameMaker Studio 2. Os dois falam sobre desenvolvimento de jogos.

Já o livro HTML5 Game Development from the Ground Up with Construct 2 apresenta uma similaridade 0,29 com o livro Game Development with GameMaker Studio 2.

Referência


```
Kaggle - Livros de T.I https://www.kaggle.com/datasets/
    cscastilloliva90/it-books-names-and-descriptions/data
```