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

Daivid Bruno Macedo Silva

RECOMENDAÇÃO DE POEMAS UTILIZANDO A SIMILARIDADE DO COSSENO

Santos - SP
Maio/2025

RESUMO

Este trabalho utiliza técnica de análise e comparação o de textos, no caso poemas, através da vetorização TF-IDF e na similaridade do cosseno, que é um método muito eficaz para medir a proximidade entre os textos. O objetivo é mostrar a aplicação o dessas técnicas matemáticas para encontrar similaridade entre poemas contidos no dataset.

LINGUAGEM E BIBLIOTECAS

A linguagem utilizada foi Python e se fez necessário utilizar as seguintes bibliotecas: NLTK, pandas, numpy e scikit-learn, essas bibliotecas são necessárias pois houve necessidade de processamento de texto, manipulação de dados entre outros.

METODOLOGIA

Após instalar todas as bibliotecas uteis e importa-las, foi necessário fazer um pré-processamento dos textos. O processo inclui:

- Remoção de *stopwords*, que são palavras comuns que na o agregam significado para a análise, como por exemplo: “de”, “para”, “o”, “a”.

```
8   # Baixar stopwords em português
9   nltk.download('stopwords')
10  stop_words = set(stopwords.words('portuguese'))
11
12  # Função para remover stopwords
13  def preprocess_text(text):
14      words = text.split()
15      filtered_words = [word for word in words if word.lower() not in stop_words]
16      return ' '.join(filtered_words)
17
```

```
25  # Aplicar a remoção de stopwords
26  df['Content'] = df['Content'].apply(preprocess_text)
27
```

- Normalizaça o: conversa o para letras minu sculas e remoça o de caracteres especiais.

VETORIZAR COM TF-IDF

Para representar numericamente os poemas contidos no dataset, foi aplicado o TF-IDF, que calcula a relevância de cada palavra com base na frequência que ela aparece no total dos textos.

```
28    # Vetorização dos textos usando TF-IDF
29    tfvec = TfidfVectorizer(max_features=10000)
30    x = tfvec.fit_transform(df['Content'])
```

Esse modelo atribui pesos para as palavras, de maneira a destacar termos importantes e reduzir a influência de palavras muito frequentes.

SIMILARIDADE DOS COSSENOS

A similaridade do cosseno e um método muito utilizado para medir proximidade entre textos. Calcula o ângulo entre dois vetores e quanto menor for esse ângulo, maior a similaridade.

```
37    # Calcular a similaridade com todos os poemas
38    simi = [(i, cosine_similarity(poem_vector, x[i])[0][0]) for i in range(x.shape[0])]
```

A equação da similaridade do cosseno é dada por:

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

- A e B são os vetores que representam os textos.
- $\|A\|$ e $\|B\|$ são as normas desses vetores.
- $A \cdot B$ é o produto interno entre os vetores.

A similaridade varia entre 0 e 1, onde os valores próximos de 1 (valores multiplicados por 100 para entrega em porcentagem) indicam textos que são mais semelhantes.

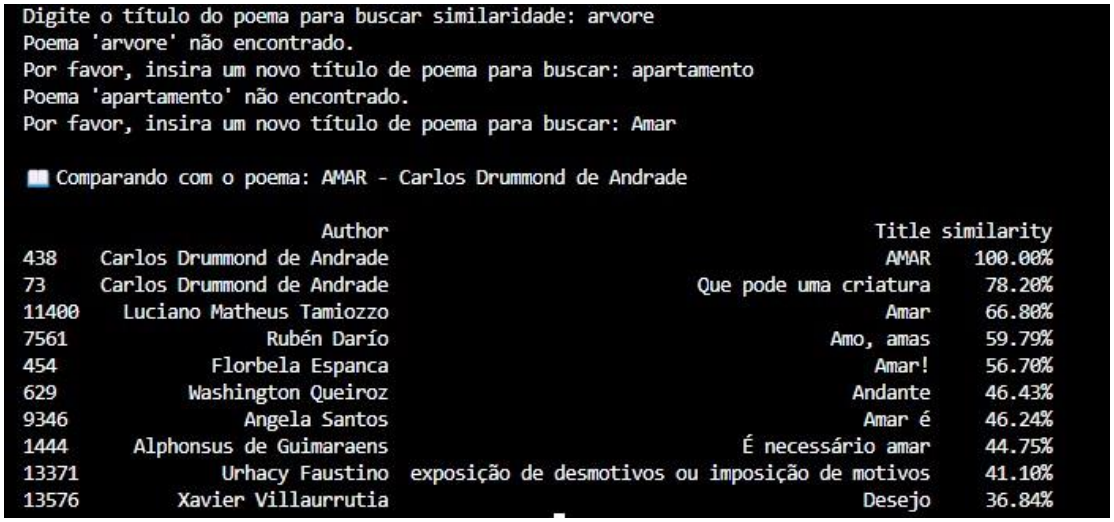
```
Digite o título do poema para buscar similaridade: retrato

📖 Comparando com o poema: Retrato - Cecília Meireles

                    Author                                  Title similarity
0          Cecília Meireles                                Retrato    100.00%
9741          Orlando Neves                                  1954     34.76%
```

No exemplo da figura mostra que o poema 1954 é o que tem maior similaridade com o poema Retrato de acordo com a similaridade do cosseno.

RESULTADOS

O modelo permitiu identificar poemas com alta similaridade em relação ao texto de referência. O sistema retorna os 10 poemas mais semelhantes, de acordo com sua proximidade percentual.

Ao rodar o código, aparecerá um campo para digitar um título de poema. Caso o poema não exista no dataset, aparecerá uma mensagem para inserir outro poema até aparecer um que esteja na base de dados.



Quando se digita um título de poema existente, o programa calcula a similaridade do poema em si (na o do título) e retorna os dez mais próximos, imprimindo na tela o nome do autor, o título e o grau de similaridade.