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

Recomendação de Poemas Utilizando a Similaridade do Cosseno

Daivid Bruno Macedo Silva

Santos - SP
Maio/2025

1 Resumo

Este trabalho explora técnicas de análise e comparação de poemas, utilizando vetorização TF-IDF e similaridade do cosseno para medir proximidade entre textos e auxiliar na recomendação de poemas semelhantes no dataset.

2 Linguagem e Bibliotecas

A implementação foi realizada na linguagem Python, utilizando as seguintes bibliotecas:

- **NLTK**: Processamento de linguagem natural e remoção de *stopwords*.
- **pandas**: Manipulação e estruturação dos dados.
- **numpy**: Operações matemáticas e vetoriais.
- **scikit-learn**: Vetorização TF-IDF e cálculo da similaridade do cosseno.

3 Metodologia

O pré-processamento dos textos incluiu as seguintes etapas:

- **Remoção de stopwords**: Palavras comuns sem relevância, como "de", "para", "o", "a".
- **Normalização**: Conversão para letras minúsculas e remoção de caracteres especiais.

```
import nltk
from nltk.corpus import stopwords

nltk.download('stopwords')

stop_words = set(stopwords.words('portuguese'))

# Função para remover stopwords
def preprocess_text(text):
    words = text.split()
    filtered_words = [word for word in words if word.lower() not in stop_words]
    return ' '.join(filtered_words)

# Aplicação da remoção de stopwords
df['Content'] = df['Content'].apply(preprocess_text)
```

4 Vetorização com TF-IDF

Para representar numericamente os poemas, aplicou-se a vetorização TF-IDF, que calcula a relevância de cada palavra com base em sua frequência nos textos.

```
from sklearn.feature_extraction.text import TfidfVectorizer


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

Esse modelo atribui pesos às palavras, destacando os termos mais relevantes e reduzindo a influência de palavras muito frequentes.

5 Similaridade do Cosseno

A similaridade do cosseno é um método utilizado para medir proximidade entre textos. Calcula o ângulo entre dois vetores, sendo que ângulos menores indicam maior similaridade.

A equação da similaridade do cosseno é dada por:

$$\cos\theta = \frac{A \cdot B}{||A|| \times ||B||} \tag{1}$$

Onde:

- A e B são os vetores representando 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 valores próximos de 1 indicam maior proximidade.

```
from sklearn.metrics.pairwise import cosine_similarity


# Cálculo da similaridade entre poemas
simi = [(i, cosine_similarity(poem_vector, x[i])[0][0]) for i in range(x.shape[0])]
```

6 Resultados

O sistema identifica e retorna os dez poemas mais similares ao escolhido pelo usuário, com base na métrica de similaridade do cosseno.

```
poema_titulo = input("Digite o título do poema para buscar similaridade: ")


# Simulação de saída
print(f"Comparando com o poema: {poema_titulo}")
```

Caso o poema não exista no dataset, o sistema solicita outro título até encontrar um correspondente.