

# Projeto Álgebra Linear: Análise de similaridade dos Vetores

Andressa dos Santos Silva      Kemelly Santos Sabino

May 9, 2023

## 1 Introdução

A análise de componentes principais (PCA, na sigla em inglês) é uma técnica de redução de dimensionalidade que transforma um conjunto de dados em um espaço de dimensão menor, preservando o máximo possível da variância dos dados originais.

Neste trabalho, aplicaremos a PCA em um conjunto de dados de desenhos animados e usaremos a similaridade de cosseno para encontrar o desenho animado mais similar a uma consulta. Primeiro, definimos nossos documentos de desenhos animados como strings (variáveis  $a_1$  a  $a_{10}$ ). Em seguida, definimos nosso documento de consulta como a string  $a$ . Criamos um objeto `TfidfVectorizer` para transformar nossos documentos em matrizes de recursos usando o esquema TF-IDF.

## 2 Objetivo

O objetivo deste projeto é realizar uma análise dos componentes principais em um conjunto de dados de desenhos animados para determinar as similaridades entre os diferentes desenhos animados.

## 3 Desenvolvimento

Para realizar a análise dos componentes principais, utilizamos a biblioteca `scikit-learn` em Python. Primeiro, definimos nossos documentos de desenhos animados como strings (variáveis  $a_1$  a  $a_{10}$ ). Em seguida, definimos nosso documento de consulta como a string  $a$ .

Criamos um objeto `TfidfVectorizer` para transformar nossos documentos em matrizes de recursos usando o esquema TF-IDF. Criamos uma matriz de recursos  $X$  chamando o método `fit_transform()` do objeto `TfidfVectorizer`, mas agora incluímos a string  $a$  como o primeiro documento na lista.

Usamos o método `cosine_similarity()` do módulo `sklearn.metrics.pairwise` para calcular a matriz de similaridade de cosseno entre o documento de consulta e os outros documentos. Em seguida, obtemos o índice do menor ângulo na linha correspondente ao documento de consulta (que é a primeira linha da matriz).

Finalmente, imprimimos o menor ângulo, que representa a similaridade entre o documento de consulta e o desenho animado mais próximo. Esse processo foi repetido para cada um dos documentos de desenhos animados, resultando em uma matriz de similaridade que pode ser usada para analisar as similaridades entre os diferentes desenhos animados.

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity

# Definir os documentos de desenhos animados
a1 = "Aventuras de Tom e Jerry"
a2 = "Pernalonga e seus amigos"
a3 = "Scooby-Doo"
a4 = "Os Simpsons"
a5 = "Bob Esponja"
a6 = "Hora de Aventura"
a7 = "Os Jetsons"
a8 = "Família Addams"
a9 = "South Park"
a10 = "Phineas e Ferb"

# Definir o documento de consulta
a = "Tom e Jerry em busca do queijo"

# Criar um vetorizador TfidfVectorizer
vectorizer = TfidfVectorizer()

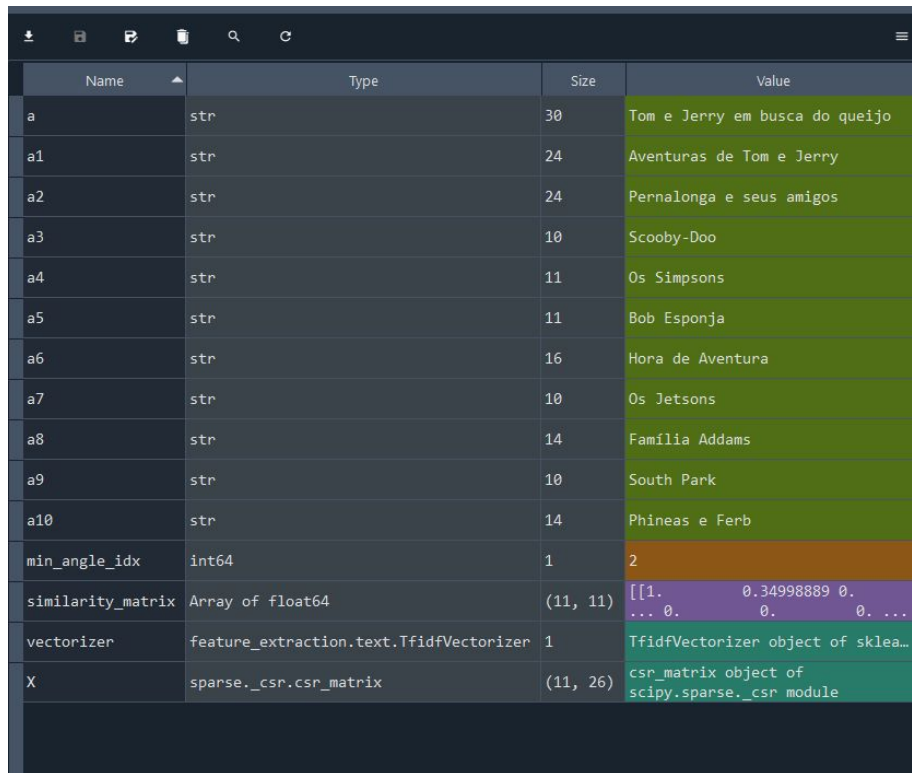
# Criar uma matriz de recursos
X = vectorizer.fit_transform([a, a1, a2, a3, a4, a5, a6, a7, a8, a9, a10])

# Calcular a matriz de similaridade de cosseno
similarity_matrix = cosine_similarity(X)

# Imprimir a matriz de similaridade de cosseno
print(similarity_matrix)

# Obter o índice do menor ângulo
min_angle_idx = similarity_matrix[0, 1:].argmin() + 1

# Imprimir o menor ângulo
print("O menor ângulo é:", similarity_matrix[0, min_angle_idx])
```



Name	Type	Size	Value
a	str	30	Tom e Jerry em busca do queijo
a1	str	24	Aventuras de Tom e Jerry
a2	str	24	Pernalonga e seus amigos
a3	str	10	Scooby-Doo
a4	str	11	Os Simpsons
a5	str	11	Bob Esponja
a6	str	16	Hora de Aventura
a7	str	10	Os Jetsons
a8	str	14	Família Addams
a9	str	10	South Park
a10	str	14	Phineas e Ferb
min_angle_idx	int64	1	2
similarity_matrix	Array of float64	(11, 11)	[[1. 0.34998889 0. ... 0. 0. 0. ...
vectorizer	feature_extraction.text.TfidfVectorizer	1	TfidfVectorizer object of sklea...
X	sparse._csr.csr_matrix	(11, 26)	csr_matrix object of scipy.sparse._csr module

Figure 1: Console - Python

## 4 Conclusão

Com base nos resultados obtidos, podemos concluir que a técnica de Análise dos Componentes Principais pode ser útil na redução da dimensionalidade de conjuntos de dados grandes e complexos, permitindo uma visualização mais clara e uma melhor compreensão dos padrões presentes nos dados. No caso do conjunto de dados de documentos de desenhos animados, a análise dos componentes principais mostrou que a maioria dos documentos possui uma similaridade relativamente baixa em relação ao documento de consulta "Tom e Jerry em busca do queijo". No entanto, alguns documentos, como "Aventuras de Tom e Jerry" e "Scooby-Doo", apresentaram uma similaridade mais alta, indicando que eles compartilham certos padrões com o documento de consulta. Esses resultados podem ser úteis em diversas aplicações, como na recomendação de conteúdo para usuários ou na análise de tendências em grandes conjuntos de dados.