

Instituto Federal Minas Gerais - Campus Bambuí  
Departamento de Engenharia e Computação - DEC  
Curso de Engenharia de Computação - ENGCOMP

# Análise de técnicas de agrupamento de dados para notícias de futebol

Mateus Araújo Cruz

Orientador: Dr. Marcos Roberto Ribeiro

2023



# Sumário

---

1 Introdução

2 Fundamentação

3 Metodologia

4 Resultados

5 Considerações Finais

6 Referências

# Introdução - Contextualização

---

- O crescente volume de dados dificulta a localização e leitura de informações relevantes no mundo do futebol
- A segmentação de dados em grupos ajuda na análise e tomada de decisões no setor esportivo
- Existem poucos estudos abordando técnicas de agrupamento aplicadas a notícias de futebol
- A análise de técnicas de agrupamento pode auxiliar em decisões estratégicas nesse setor

# Introdução - Objetivos

## Objetivo Geral

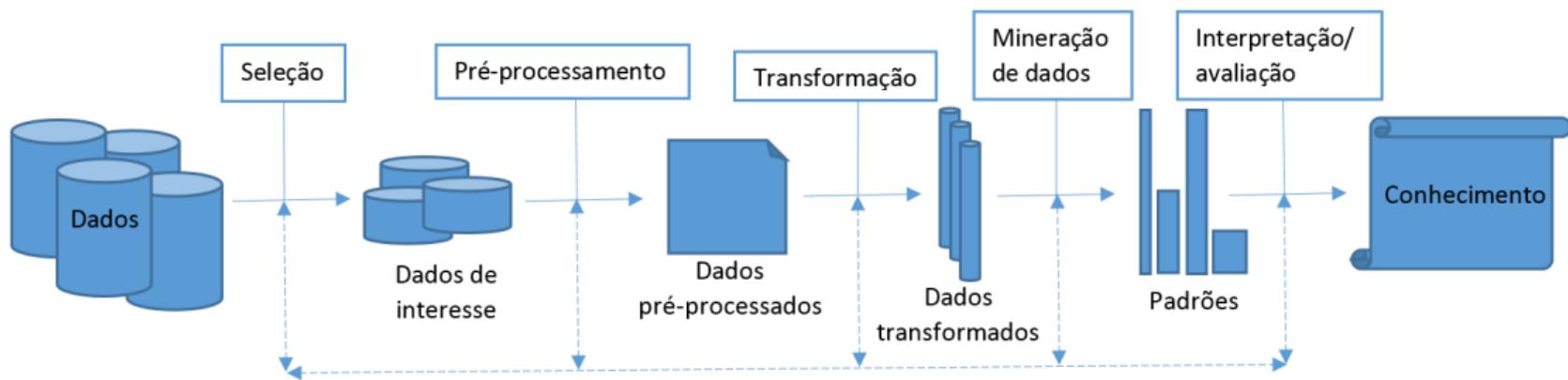
Analisar e comparar as principais técnicas de agrupamento de dados aplicadas a notícias de times de futebol

## Objetivos Específicos

- 1 Estruturar a base de dados de notícias de times de futebol
- 2 Selecionar as técnicas de agrupamento de dados mais adequadas para a base
- 3 Analisar as técnicas de agrupamento por meio de experimentos utilizando a base de dados

## Fundamentação - KDD

- Descoberta de Conhecimento em Banco de Dados, ou KDD, tem como finalidade a identificação de padrões úteis a partir de grandes conjuntos de dados (ALLAHYARI *et al.* 2017)

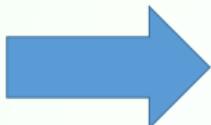


Fonte: FREITAS; MOURA; SILVA, 2015.

# Fundamentação - Pré-processamento de Texto

## ■ Tokenização

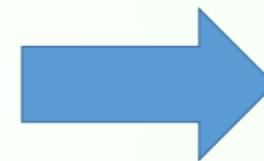
Agrupamento de dados é uma técnica que divide objetos em grupos com base em suas similaridades, permitindo a identificação de padrões e estruturas nos dados. É usado em diversas áreas, como mineração de dados e aprendizado de máquina.



Tokens
Agrupamento
de
dados
é
uma
técnica
que
divide
objetos
.....

## ■ Remoção de palavras de parada

Tokens
Agrupamento
de
dados
é
uma
técnica
que
divide
objetos
.....



Tokens
Agrupamento
dados
técnica
divide
objetos
.....

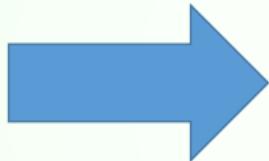
Fonte: Elaborado pelo Autor (2023).

Fonte: Elaborado pelo Autor (2023).

# Fundamentação - Pré-processamento de Texto

## ■ Derivação

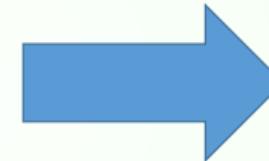
Tokens
Agrupamento
dados
técnica
divide
objetos
.....



Tokens
agrup
dad
técn
divid
objet
.....

## ■ Lematização

Tokens
Agrupamento
dados
técnica
divide
objetos
.....



Tokens
agrupamento
dado
técnico
dividir
objeto
.....

Fonte: Elaborado pelo Autor (2023).

Fonte: Elaborado pelo Autor (2023).

# Fundamentação - Pré-processamento de Texto

## ■ Representação no espaço vetorial

	Termo 1	Termo 2	Termo 3	Termo 4	Termo 5
Doc 1	0.00	0.28	0.47	0.28	0.35
Doc 2	0.63	0.00	0.31	0.00	0.53
Doc 3	0.31	0.47	0.28	0.31	0.31
Doc 4	0.45	0.34	0.00	0.47	0.00
Doc 5	0.53	0.50	0.34	0.00	0.00

Fonte: Elaborado pelo Autor (2023).

## ■ Redução de dimensionalidade

	Componente 1	Componente 2	Componente 3
Doc 1	0.77	0.07	0.62
Doc 2	0.80	0.55	0.20
Doc 3	0.97	0.13	0.19
Doc 4	0.83	0.51	0.20
Doc 5	0.94	0.16	0.16

Fonte: Elaborado pelo Autor (2023).

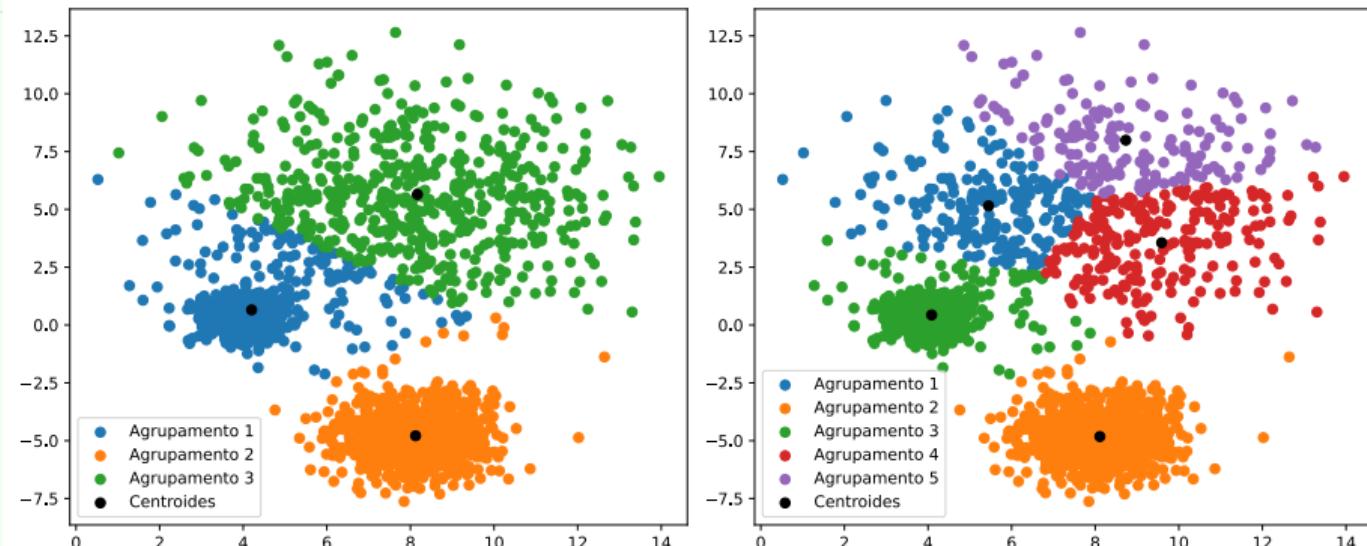
## Fundamentação - Agrupamento de Dados

---

- Técnica que consiste na separação de dados em subgrupos que contêm objetos semelhantes entre si, denominados *clusters* (RAI; SINGH, 2010)
- Estratégia de aprendizado não supervisionado (GHOSAL *et al.* 2020)
- É dividido em duas categorias: *hard clustering* e *soft clustering* (KAUSHIK, 2016)

# Fundamentação - K-Means

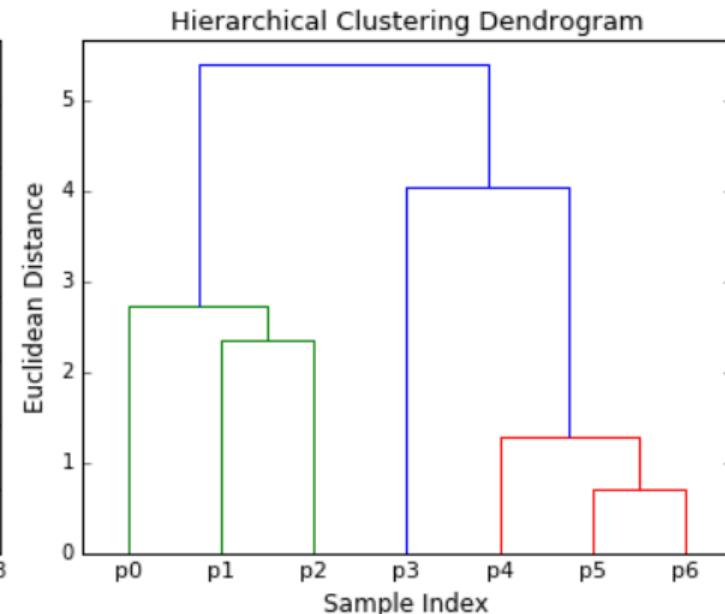
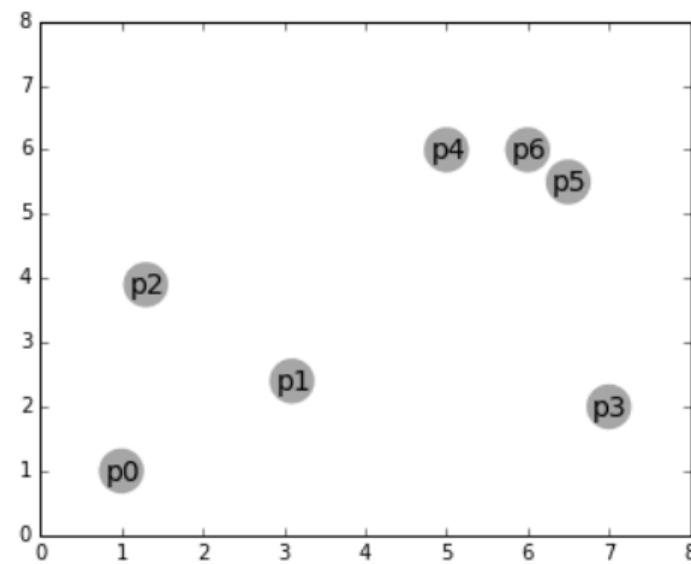
## ■ K-Means (baseado em partições)



Fonte: Elaborado pelo Autor (2023).

# Fundamentação - Agrupamento Hierárquico

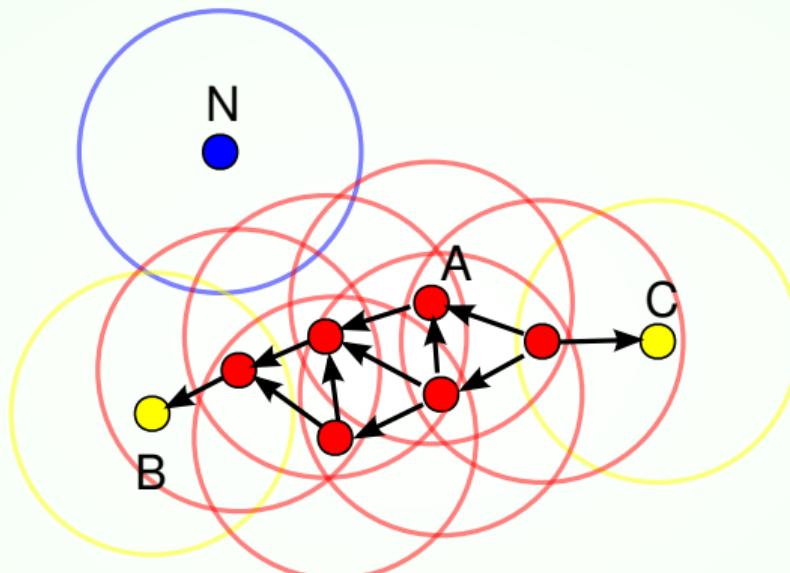
## Agrupamento Hierárquico (baseado em hierarquia)



Fonte: SCIENCE, 2023.

# Fundamentação - DBSCAN

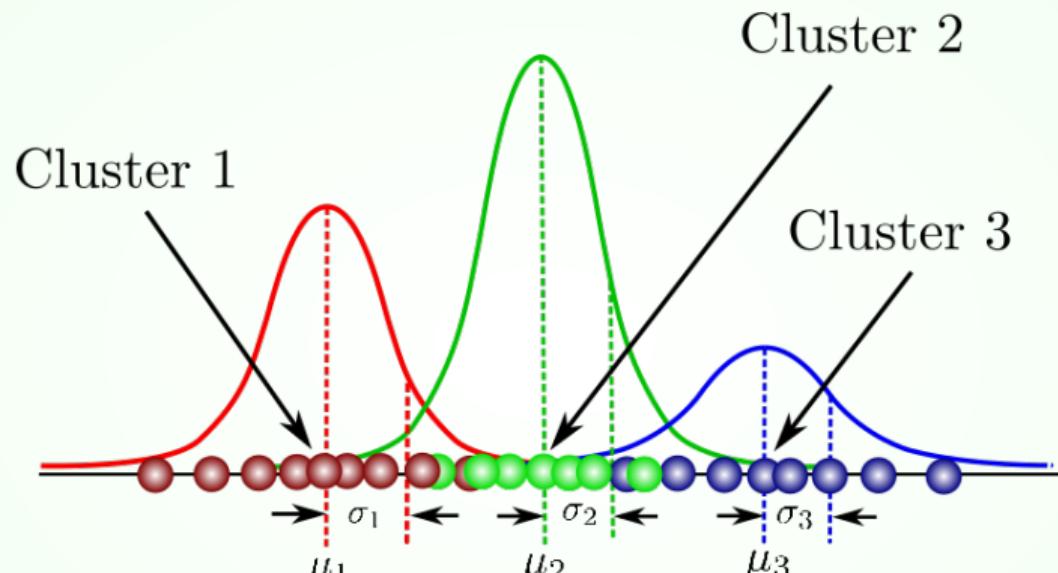
- DBSCAN (baseado em densidade)



Fonte: WIKIPEDIA, 2023.

# Fundamentação - Modelo de Mistura Gaussiana

- GMM (baseado em modelo)



Fonte: MEDIUM, 2023.

## Fundamentação - Estado da Arte

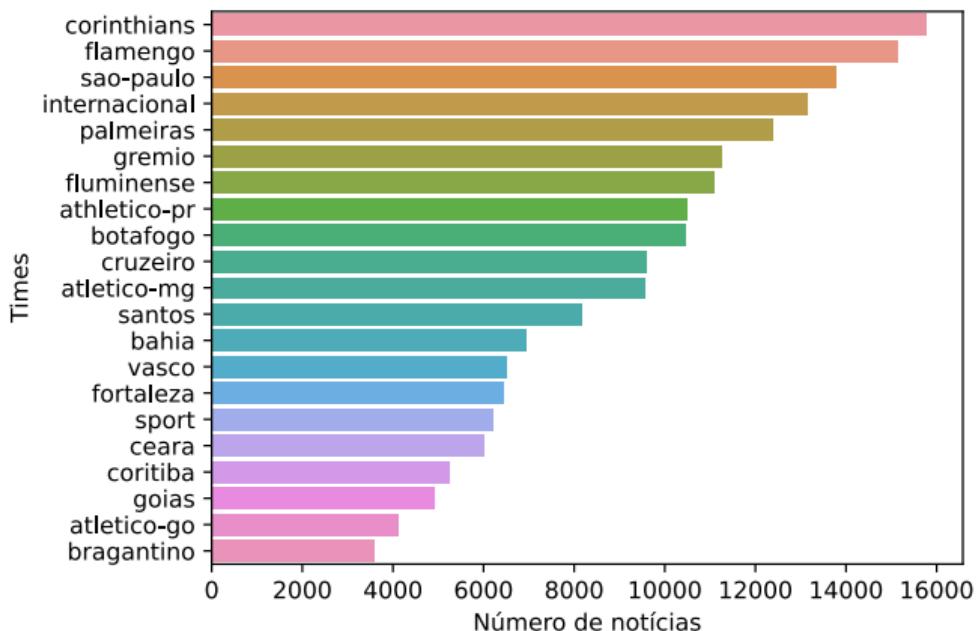
- Afonso e Duque (2014) relataram em seu trabalho os resultados de experimentos sobre agrupamento automático de texto aplicado em artigos científicos e textos de jornais em português brasileiro
- Marutho *et al.* (2018) propuseram um método de agrupamento de notícias com base em suas manchetes
- Ghosal *et al.* (2020) fazem uma revisão sobre diferentes técnicas de agrupamento e suas aplicações
- Aggarwal (2018) apresentou todo o tratamento de Recuperação de Informações e Mineração de Texto, simplificando a apresentação matemática com explicações intuitivas

## Metodologia - Classificação da Pesquisa

- Abordagem: quantitativa (métricas e análises estatísticas)
- Natureza: aplicada (agrupamento de notícias de futebol)
- Objetivos: exploratória (exploração das principais técnicas)
- Procedimentos: experimental (avaliação dos resultados para cada técnica)
- Metodologia específica da computação: apresentação de algo presumivelmente melhor

# Metodologia - Descrição dos Dados

- Base de dados *GE Soccer News*
- Notícias de 2015 a 2022
- Total de 191005 notícias
- 21 categorias
- Média de 9095 notícias por time



Fonte: Elaborado pelo Autor (2023).

# Metodologia - Materiais e Tecnologias

## Materiais e Tecnologias:

- Linguagem de programação Python
- Framework Natural Language Toolkit (NLTK)
- Framework Scikit-learn
- Biblioteca Pandas
- Biblioteca Matplotlib

## Metodologia de Desenvolvimento:

- Scrum proporciona uma forma flexível de trabalho em equipe (SACHDEVA, 2016)
- Kanban é uma estratégia para otimizar o fluxo de valor através de facilitação visual(VACANTI; YERET, 2021)

# Metodologia - Métodos e Procedimentos

## Etapas de análise das notícias



Fonte: Elaborado pelo Autor (2023).

## Categorias de Experimentos

- Agrupamento das notícias por time (medida de validade externa)
- Busca em grade para obtenção dos melhores parâmetros (medida de validade interna)

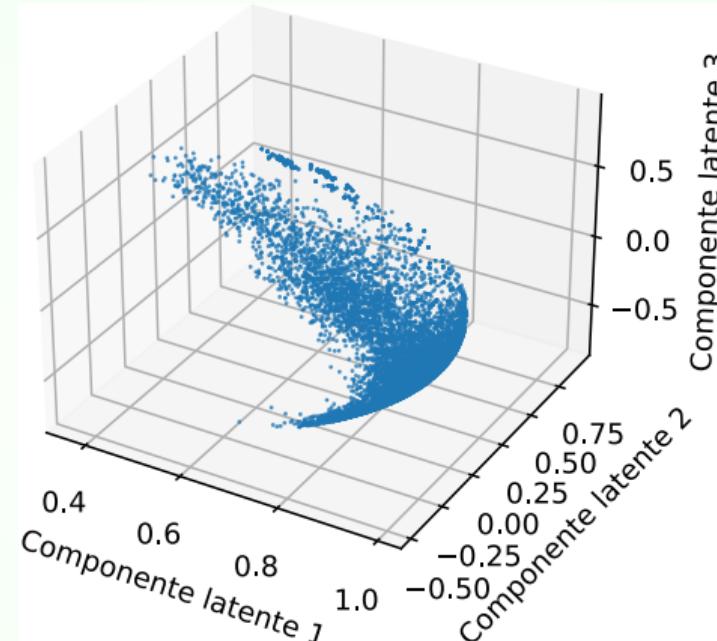
## Resultados - Dados Pré-processados

---

- Notícias do ano de 2022
- Subgrupo de 32042 notícias
- Média de 454 palavras por notícia
- Média de 169 *tokens* após o pré-processamento
- Total de 45035 termos únicos em todos os documentos

# Resultados - Redução de Dimensionalidade

- Redução de dimensionalidade para 3 componentes



Fonte: Elaborado pelo Autor (2023).

## Resultados - Redução de Dimensionalidade

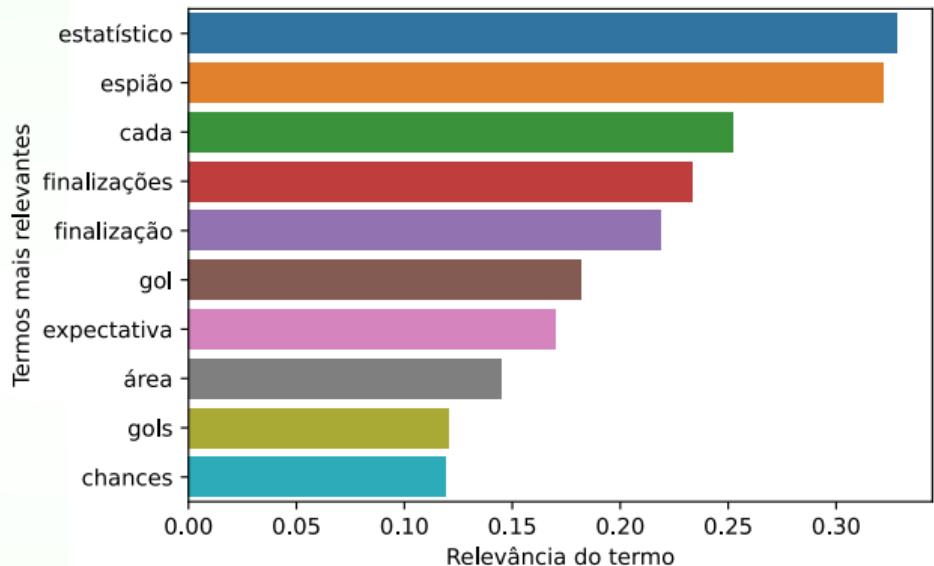
- Busca em grade para determinar o melhor número de componentes para cada algoritmo
- O número de componentes variou de 20 a 100

Algoritmo	Número de componentes	Percentual de acurácia
K-Means	32	75
Hierárquico	30	72
GMM	29	74

Elaborado pelo Autor, 2023.

# Resultados - Redução de Dimensionalidade

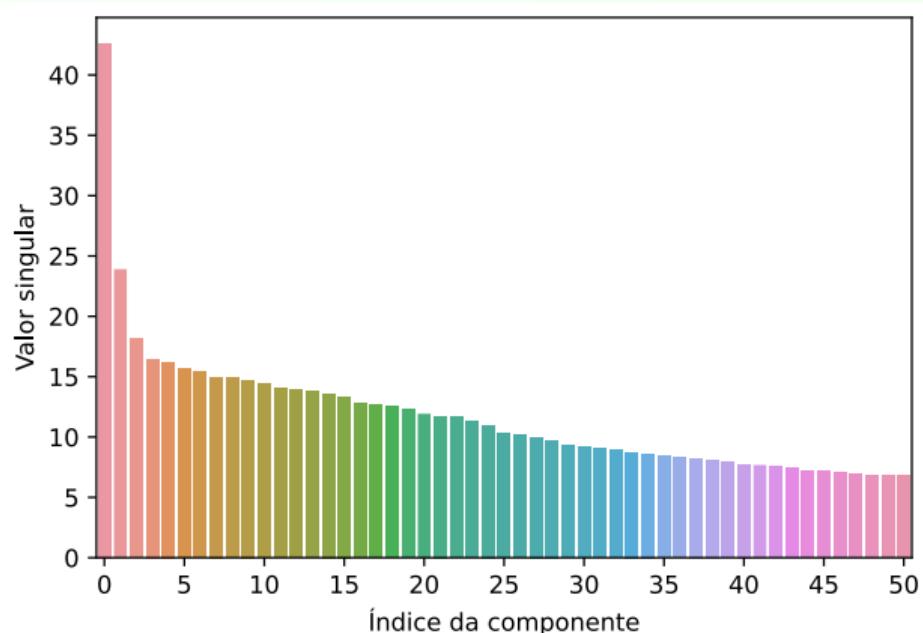
- Cada tópico representa um tema ou assunto
- Termos que mais contribuíram para a formação do tópico 1



Fonte: Elaborado pelo Autor (2023).

# Resultados - Redução de Dimensionalidade

- Matriz de valores singulares
- Representam a importância relativa dos tópicos na representação do conjunto original dos dados



Fonte: Elaborado pelo Autor (2023).

## Resultados - Agrupamento

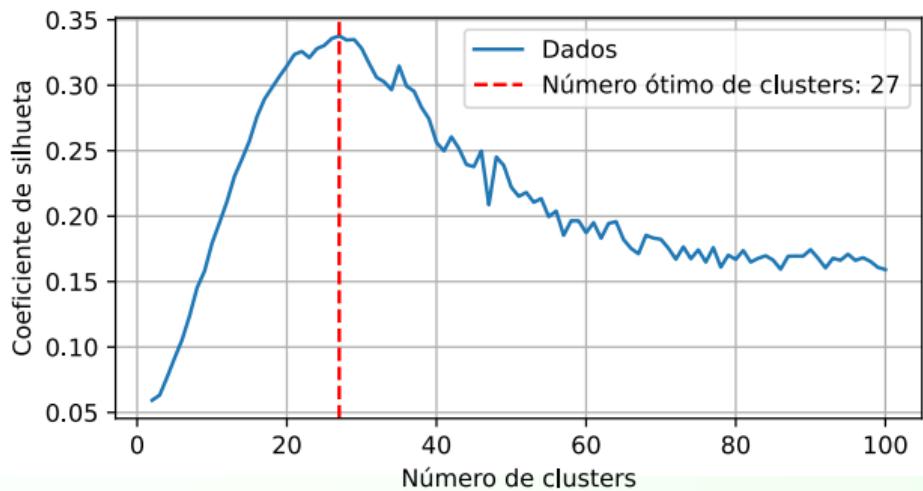
- Agrupamento usando medida de validade externa
- O nome do time foi usado para calcular o percentual de acerto
- Os dados foram divididos em 21 grupos
- O DBSCAN não foi utilizado nesse experimento

Experimentos	K-Means	Hierárquico	GMM
Usando apenas <i>tokenização</i>	75	72	74
Usando <i>tokenização</i> e derivação	75	71	74
Usando <i>tokenização</i> e lematização	74	69	68

Elaborado pelo Autor, 2023.

# Resultados - Busca em Grade

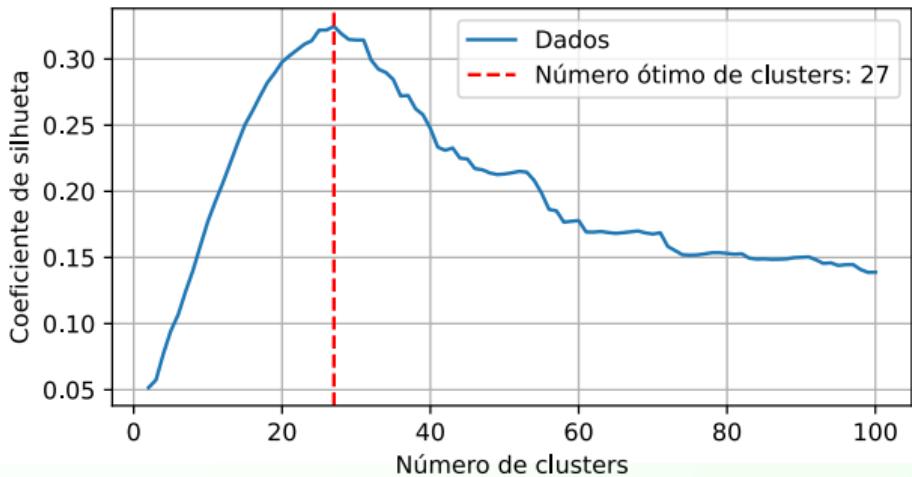
- Busca em grade para K-Means
- Melhor número de *clusters*: 27
- Melhor valor de coeficiente de silhueta: 0.337



Fonte: Elaborado pelo Autor (2023).

# Resultados - Busca em Grade

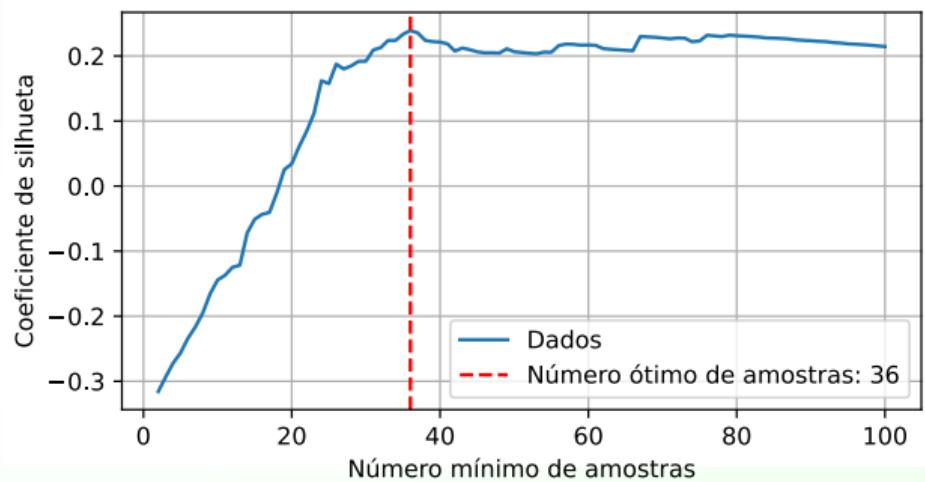
- Busca em grade para Agrupamento Hierárquico Aglomerativo
- Melhor número de *clusters*: 27
- Melhor valor de coeficiente de silhueta: 0.324



Fonte: Elaborado pelo Autor (2023).

# Resultados - Busca em Grade

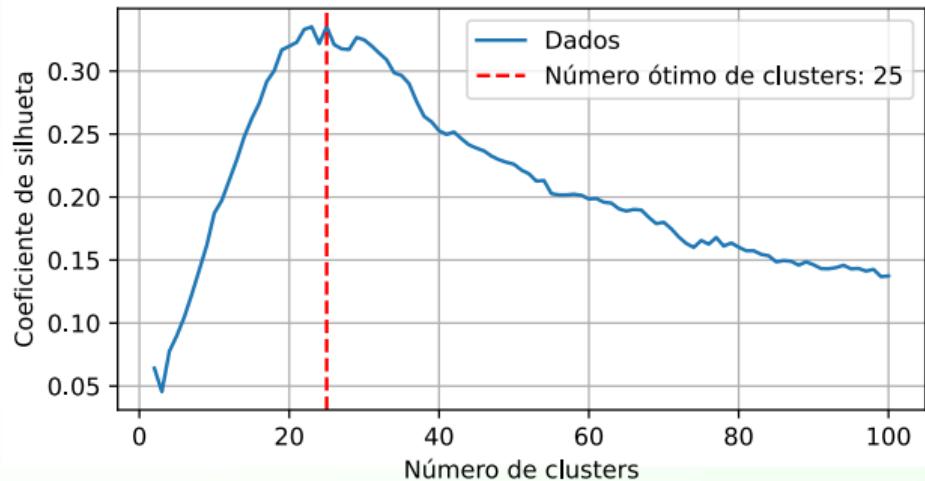
- Busca em grade para DBSCAN
- Melhor número de *clusters*: 32
- Melhor valor de coeficiente de silhueta: 0.239



Fonte: Elaborado pelo Autor (2023).

# Resultados - Busca em Grade

- Busca em grade para Modelo de Mistura Gaussiana
- Melhor número de *clusters*: 25
- Melhor valor de coeficiente de silhueta: 0.335



Fonte: Elaborado pelo Autor (2023).

# Resultados - Busca em Grade

## ■ Resumo da busca em grade

Algoritmo	Número de <i>clusters</i>	Coeficiente de Silhueta
K-Means	27	0.337
Hierárquico	27	0.324
DBSCAN	32	0.239
GMM	25	0.335

Elaborado pelo Autor, 2023.

# Considerações Finais

## Considerações Gerais

- Dificuldade e importância do pré-processamento de texto
- K-Means e GMM obtiveram os melhores índices de acurácia
- As técnicas utilizadas conseguiram capturar os padrões nas notícias de futebol

## Trabalhos Futuros

- Ampliar o período de tempo selecionado para a análise das notícias
- Aplicação de uma abordagem dinâmica na análise de notícias

## Referências I

---

- AFONSO, A. R.; DUQUE, C. G. Automated text clustering of newspaper and scientific texts in brazilian portuguese: analysis and comparison of methods. *Journal of Information Systems and Technology Management*, v. 11, n. 2, p. 415–436, 2014.
- AGGARWAL, C. C. **Machine Learning for Text: An Introduction**. Cham: Springer International Publishing, 2018.
- ALLAHYARI, M. *et al.* A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. In.
- FREITAS, N.; MOURA, C.; SILVA, M. Sistema multiagente para mineração de imagens de satélite. *XVII Simpósio Brasileiro de Sensoriamento Remoto*, p. 7351–7358, 2015.
- GHOSAL, A. *et al.* A Short Review on Different Clustering Techniques and Their Applications. Springer Singapore, Singapore, p. 69–83, 2020.

## Referências II

---

- KAUSHIK, S. **Clustering Introduction, Different Methods and Applications.** 2016. Disponível em: <https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering>. Acesso em: 8 abr. 2023.
- MARUTHO, D. *et al.* The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News. p. 533–538, 2018.
- MEDIUM. **Agglomerative Clustering vs. K-Means Clustering.** 2023. Disponível em: <https://medium.com/@namanbhandari/agglomerative-clustering-vs-k-means-clustering-766a90b37dc0>. Acesso em: 22 nov. 2023.
- RAI, P.; SINGH, S. A Survey of Clustering Techniques. **International Journal of Computer Applications**, v. 7, n. 12, 2010.
- SACHDEVA, S. Scrum methodology. **International Journal Of Engineering And Computer Science**, v. 5, n. 6, p. 16792–16799, 2016.

## Referências III

---

- SCIENCE, T. D. **Gaussian Mixture Models Explained.** 2023. Disponível em:  
<https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95>. Acesso em: 22 nov. 2023.
- VACANTI, D.; YERET, Y. **Kanban Guide for Scrum Teams.** 2021. Disponível em:  
<https://www.scrum.org/resources/kanban-guide-scrum-teams>. Acesso em: 23 mai. 2023.
- WIKIPEDIA. **DBSCAN.** 2023. Disponível em: <https://en.wikipedia.org/wiki/DBSCAN>. Acesso em: 18 ago. 2023.