# Unraveling BERT: Complete Course on Topic Modeling

## Simpósio Brasileiro de Banco de Dados

Antônio Oss Boll e Letícia Maria Puttlitz

Universidade de São Paulo

15 de Outubro de 2024

# Sumário

# Slides

- GitHub

# Apresentação

**Antônio Oss Boll:**
- Graduação em Estatística pela UFRGS
- Mestrando na USP em Computação
- ML Engineer na Widelabs

**Letícia Maria Puttlitz:**
- Graduação em Estatística pela UFRGS
- Mestranda na USP em Computação
- ML Engineer na Widelabs

# Introdução

- Há uma maior dificuldade em trabalhar com dados não rotulados, uma vez que não há um resultado comparável ao predito.
- Modelagem de tópicos é uma solução comum para explorar e descobrir padrões em dados não rotulados.
- BERTopic oferece uma solução mais robusta ao usar embeddings como BERT, em relação a metodos como LDA (Latent Dirichlet Allocation).

# Introdução

Vamos abordar os seguintes tópicos:

- Transformers
- Encoder
- BERT
- BERTopic

# Transformer

O *transformer* representou uma revolução no campo de *NLP*.

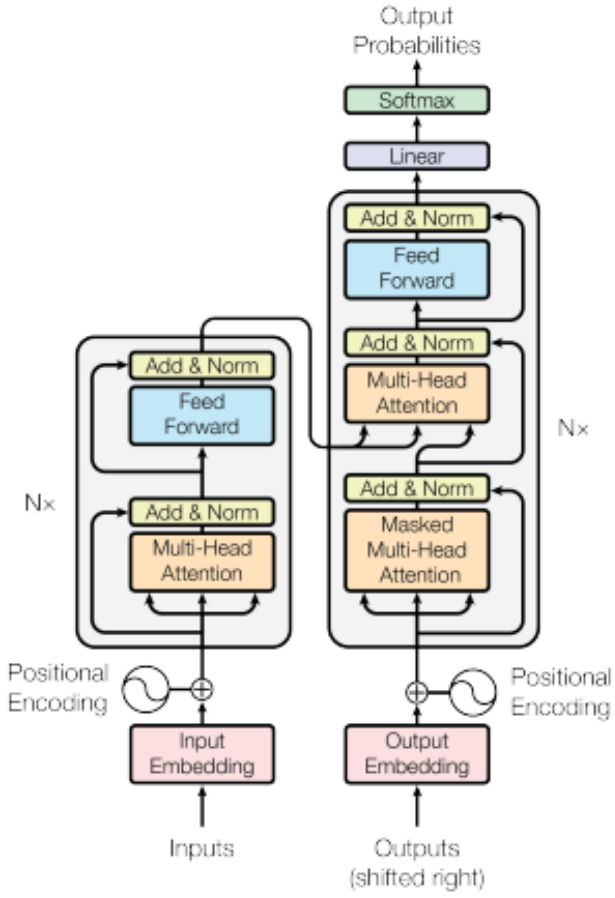



Figure: Exemplo de uma estrutura *Transformer*.[1]

[1] Fonte: Vaswani et al. [2017]

# Mecanismos de atenção

A função do mecanismo de atenção é justamente visualizar a frase inteira, permitindo traduções mais precisas e análises contextuais aprofundadas.

- **“Do you have a pen I can borrow?”**

Se fosse traduzida literalmente:

- **“Você tem uma caneta eu posso emprestado?”**

A função do mecanismo de atenção é justamente visualizar a frase inteira, permitindo traduções mais precisas e análises contextuais aprofundadas.

# Mecanismos de atenção

- **“Eu como bolo.”**

A palavra “bolo” possui um significado totalmente diferente de:

- **“Você me deu um bolo.”**.

Essa diferença é avaliada pelo mecanismo de *self-attention* ou autoatenção, que busca captar o contexto em uma frase através de uma matriz de correlação entre as palavras. Nela, o algoritmo gera pesos maiores, por exemplo, para “como” e “deu” com “bolo”, assim, conseguindo distinguir o significado de bolo para as duas sentenças.

# Encoder

O *encoder* consegue gerar uma representação vetorial de uma frase com contexto. Ele captura as relações e dependências entre partes da sentença.

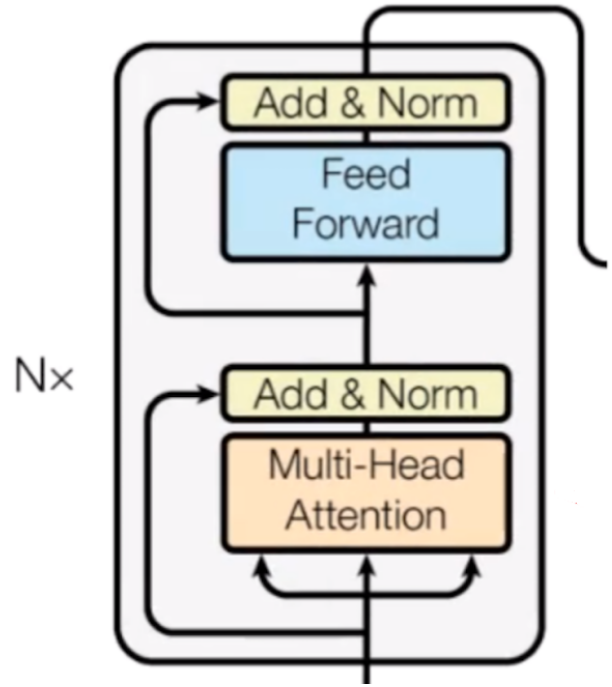



Figure: Exemplo de uma estrutura *Encoder*.[2]

[2]Fonte: Vaswani et al. [2017]

# Cálculo da Atenção

Trazendo o conceito de Query, Key e Value, vamos utilizar como exemplo a frase “Vídeo comendo chocolate na Espanha”.

- A **Query** pode ser entendida como o elemento procurado ou, em outras palavras, o foco da atenção. A *query* representa aquilo que é procurado. Ela pode ser representada pela frase completa ou focar em uma palavra central, como “chocolate”.
- A **Key** representa o que cada elemento oferece à *Query*. A *Key* seria a contribuição de cada elemento para o foco da atenção, ou seja, o que “Vídeo”, “comendo”, “chocolate”, “na” e “Espanha” oferecem para a frase completa ou para a palavra central “chocolate”.
- O **Value**, é a representação do significado de cada palavra. No exemplo, seria o que “Vídeo”, “comendo”, “chocolate”, “na” e “Espanha” realmente significam.

# Scaled Dot-Product

Esses scores representam os pesos, ou força, que cada palavra na frase tem em relação às outras. Seria a relação de cada elemento com os outros.

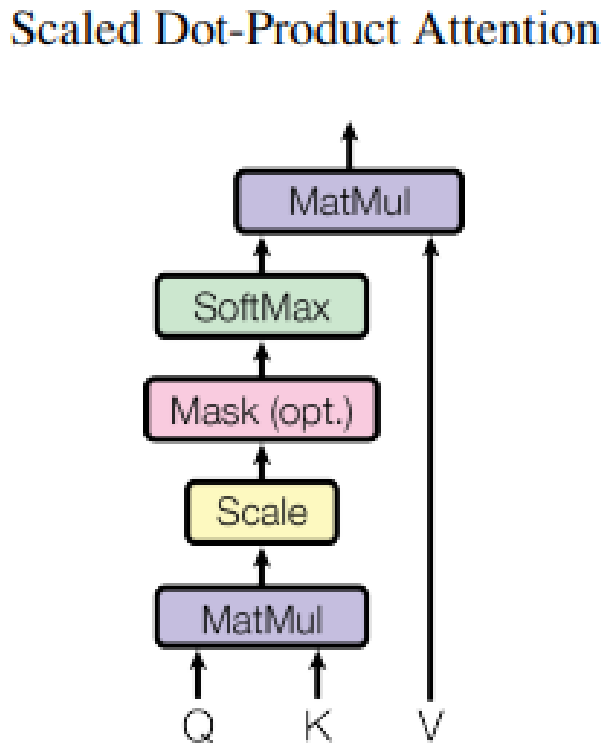



Figure: Exemplo da *Scaled Dot-Product Attention*.[3]

[3] Fonte: Vaswani et al. [2017]

# Multi-Head

São realizadas muitas vezes (diferentes cabeças) para garantir que vai englobar todos os casos.

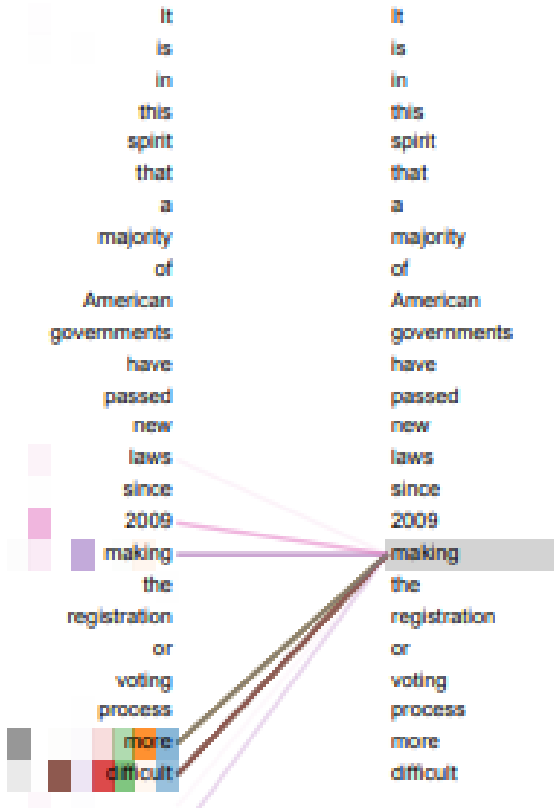



Figure: Exemplo do *Multi-Head*.[4]

[4] Fonte: Vaswani et al. [2017]

# BERT

O BERT (*Bidirectional Encoder Representations from Transformers*) possui uma estrutura de diversos *encoders* (Devlin et al. [2018]).
O BERT recebe uma frase completa e é capaz de avaliar palavras em toda a sentença para compreender o contexto. Assim, os *word embeddings* fornecidos pelo BERT são contextualizados.

# Input Representation

O BERT consegue dividir palavras, como por exemplo em “come” + “##u” e “come” + “##ndo”, permitindo a identificação de diversas sentenças ainda não aprendidas.
*O token* **[CLS]**. Ele é inserido no início da frase e é uma representação agregada da sentença inteira na forma de um *token*, ótimo para tarefas de classificação.
*O token* **[SEP]**. Ele é inserido para separar frases dentro de um conjunto de sentenças.

# Input Representation

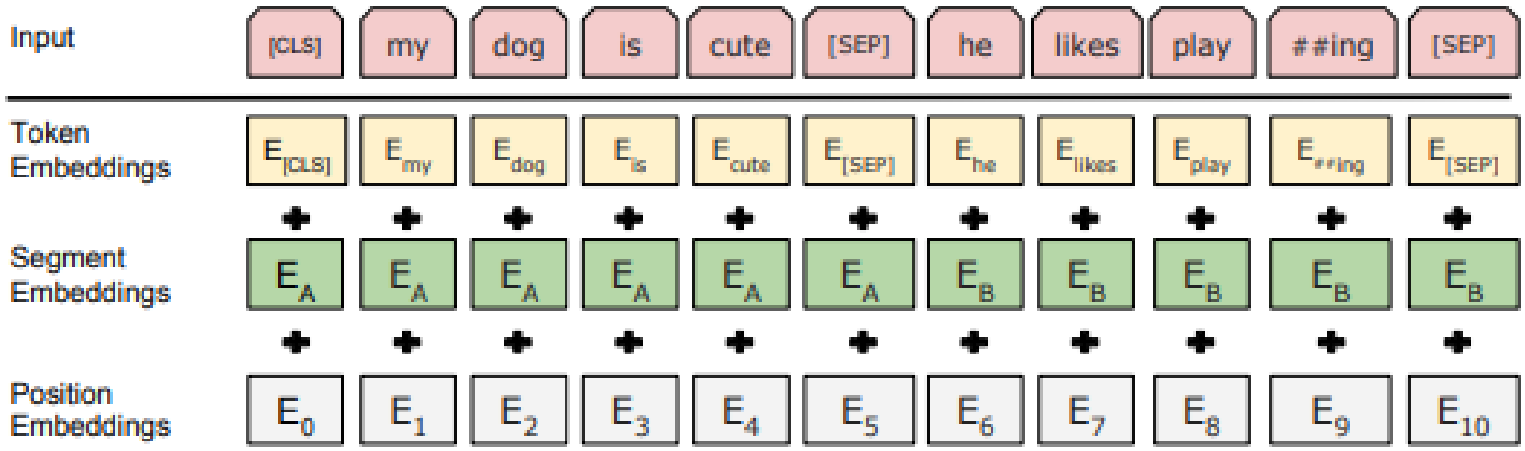



Figure: Exemplo do *Input Representation* do BERT.[5]

[5] Fonte: Devlin et al. [2018]

# Topic Modeling

Alguns modelos de topic modeling incluem:

- **Latent Dirichlet Allocation (LDA)**:
  - Algoritmo probabilístico que modela tópicos como distribuições de palavras em documentos.
- **Correlated Topic Model (CTM)**:
  - Extensão do LDA que permite a correlação entre tópicos, melhorando a modelagem em dados correlacionados.

# Modelo BERTopic

- Trabalhando sem dados rotulados, o BERTopic se torna uma alternativa aos modelos tradicionais de *topic modeling* como o LDA (Blei et al. [2003]) (Matriz de frequência) e o CTM (Blei and Lafferty [2007]).
- Diferentemente dos tradicionais, o BERTopic obtém a semântica da frase dos modelos BERT (Devlin et al. [2018]), fazendo assim, uma modelagem considerando o contexto da frase. Ele é dividido em diversas partes que serão explicadas a seguir.

# Evolução do Bertopic

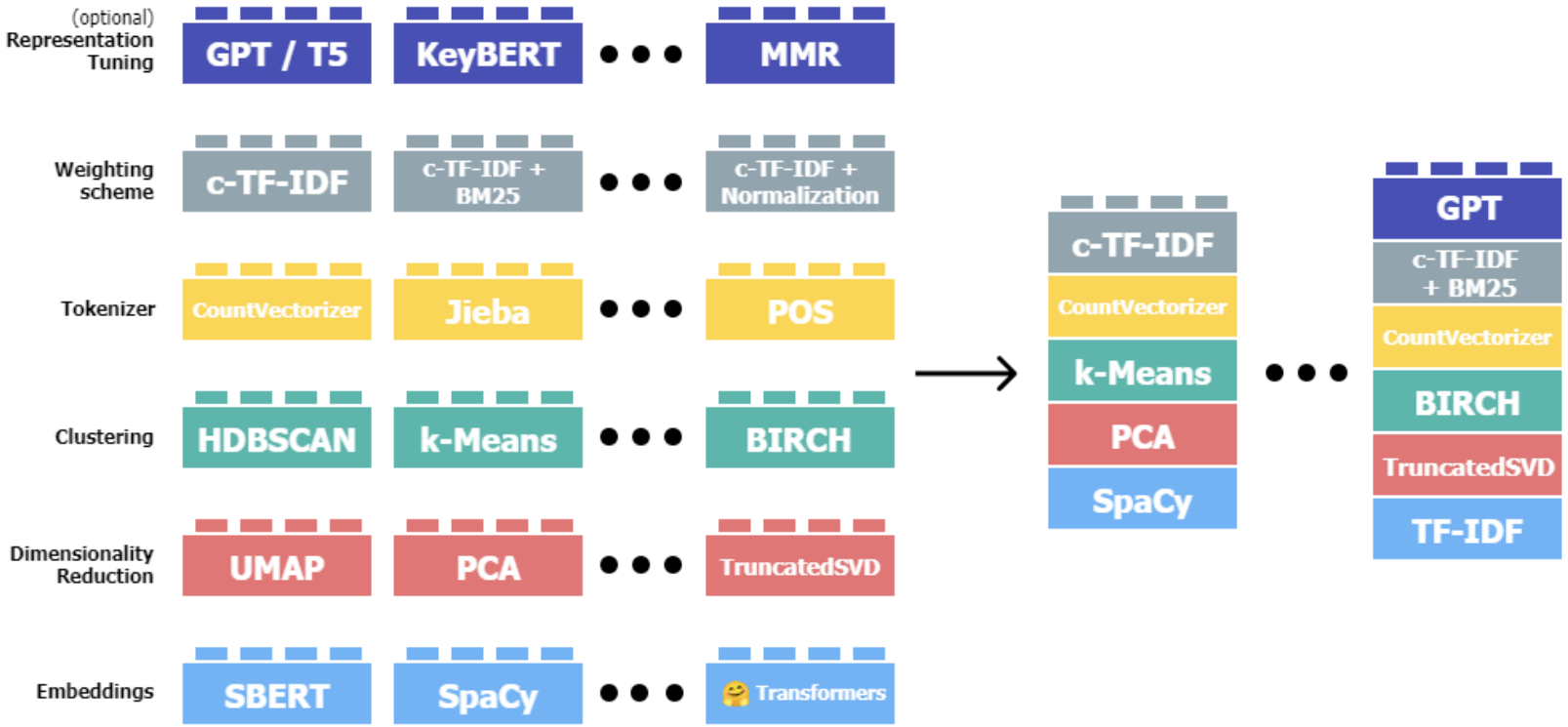



Figure: Exemplo do BERTopic.[6]

# Word Embeddings

*Word Embeddings* são a representação de palavras como vetores (Mikolov et al. [2013]).
Dessa maneira, palavras similares são posicionadas próximas umas às outras.

- Rei - Homem + Mulher = Rainha

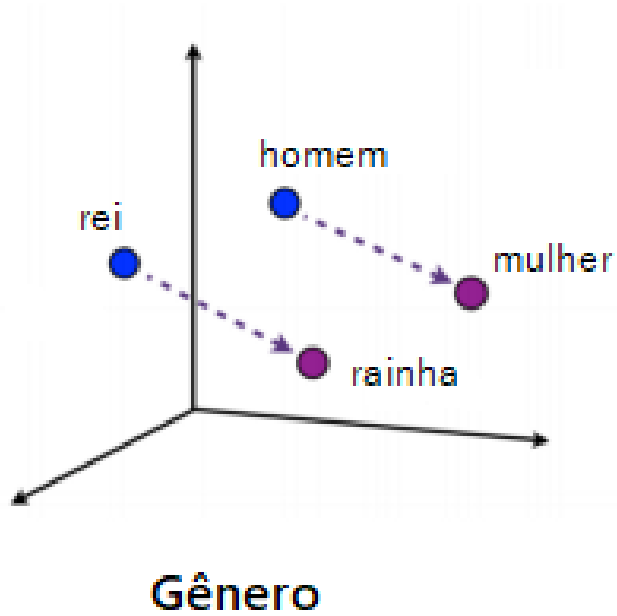



Figure: Exemplo vetorial de um Word Embedding.[7]

[7] Fonte: Artigo da Medium

# Document Embedding

Nesse inicio, os dados passam por algoritmos que os transformam em embeddings e que podem ser utilizados após em outros algoritmos.

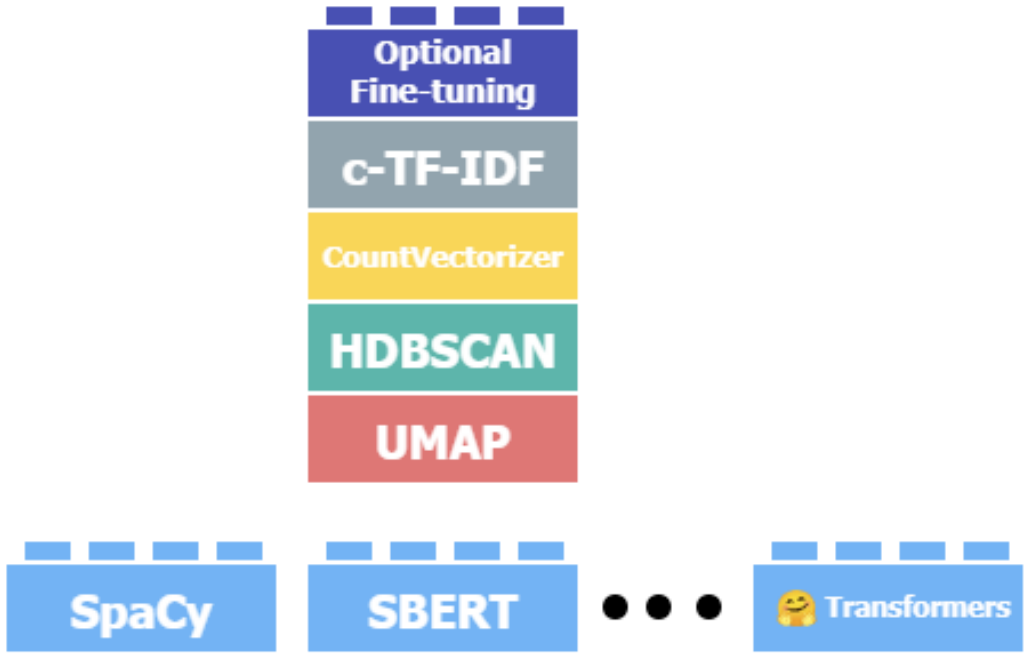



Figure: Embedding Models.[8]

# Document Embedding

Inicialmente, o BERTopic trabalha com o S-BERT (Reimers and Gurevych [2019]) para gerar *sentence embeddings*. Ele é uma variação do BERT.

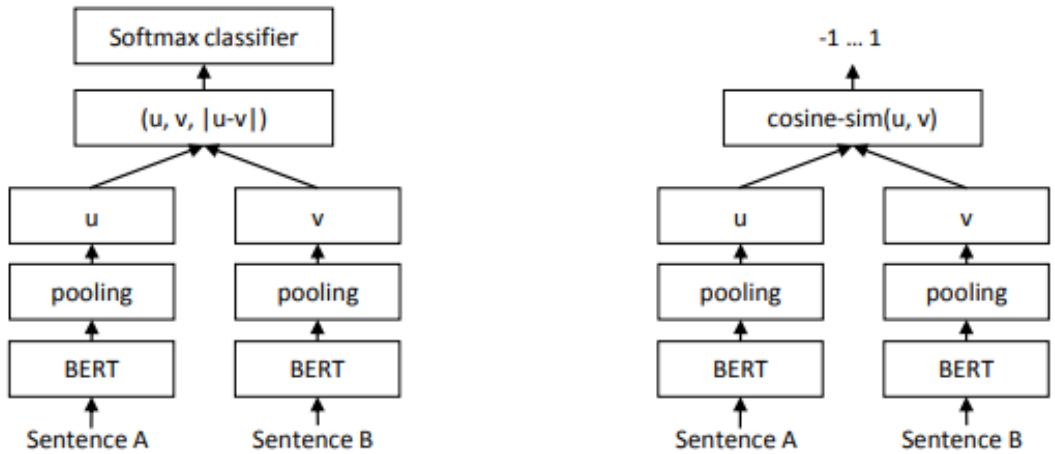



Figure: Representação do S-BERT.[9]

Os resultados do S-BERT são utilizados para criar *clusters*, o próximo passo do BERTopic.

[9] Fonte: Reimers and Gurevych [2019]

# Document Embedding

Três formas de agregação são apresentadas:
**Pooling por [CLS]**: A informação já esta resumida no *token* [CLS]. Portanto, o *pooling* é apenas a obtenção do *embedding* deste *token*;
**Pooling por Máximo**: Obtém-se o valor máximo dos *embeddings* elemento em elemento;
**Pooling por Média**: Obtém-se o valor médio dos *embeddings* elemento em elemento.
Podendo realizar Classificação, Regressão e Triplet (Aproximar três sentenças).

# Embeddings

O autor permite a utilização de diferentes algoritmos de preferência do usuário, oferecendo integração com múltiplos, como:

- Hugging Face
- Flair
- SpaCy
- Universal Sentence Encoder
- Gensim
- Scikit-Learn Embeddings
- OpenAI
- Cohere

Além disso, o algoritmo suporta:

- Embeddings Multimodais
- Backend Customizável
- TF-IDF
- Embeddings Customizáveis

# Dimensionality Reduction

O segundo passo é a redução de dimensionalidade, facilitando a clusterização e evitando a *curse of dimensionality* nos embeddings.

- A *curse of dimensionality*, termo introduzido por Bellman [1957], refere-se à dificuldade de interpretar e inferir dados em espaços de alta dimensão.

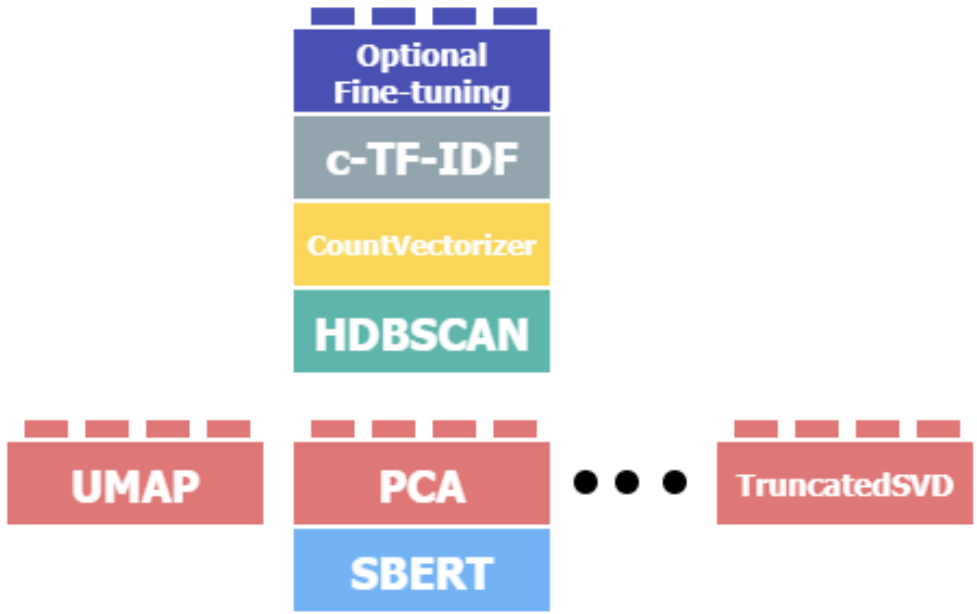



Figure: Modelos de redução de dimensionalidade.[10]

# Dimensionality Reduction UMAP

Os vetores gerados pelos *Document Embedding* possuem alta dimensionalidade. Para reduzir isso, utiliza-se o UMAP McInnes and Healy [2018], projetando dados de dimensões maiores para menores.

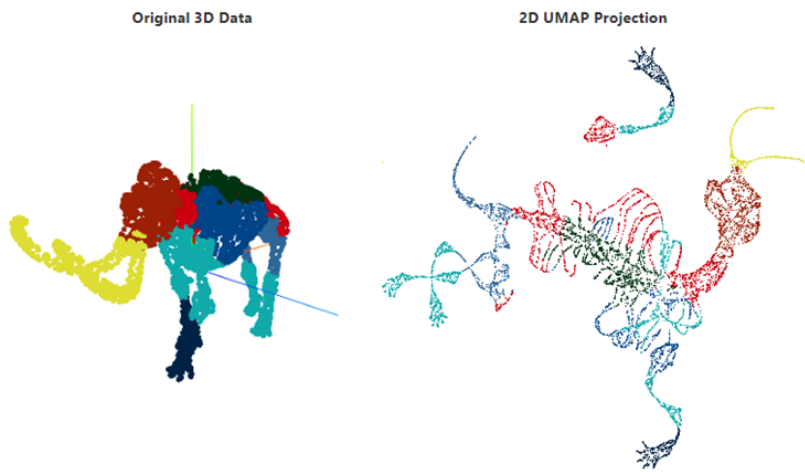



Figure: Representação do UMAP.[11]

Esse algoritmo possui como objetivo replicar os *clusters* das maiores dimensões às menores sem perder os padrões.

[11] Fonte: Khadivi et al. [2023]

# Dimensionality Reduction

Além do UMAP, é possível utilizar:

- PCA
- Truncated SVD (ou outros modelos do *scikit-learn*)
- cuML UMAP
- Ou até mesmo pular esse passo

# Document Clustering

O próximo passo é realizar a clusterização, ou seja, agrupar os dados similares representados como embeddings. Com isso, podemos gerar tópicos para dados anteriormente não rotulados.

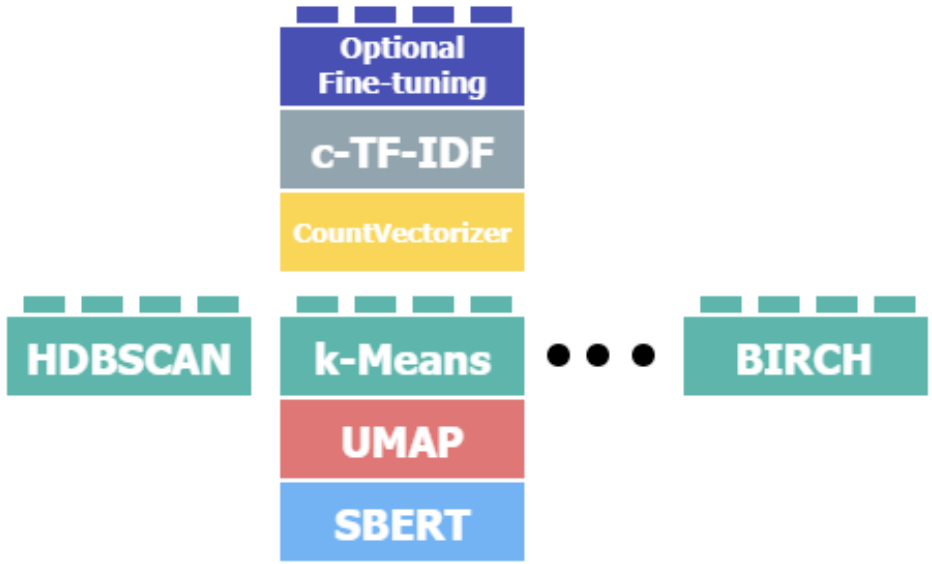



Figure: Modelos de clusters.[12]

# Document Clustering HDBSCAN

A partir do UMAP, é utilizado o HDBSCAN (Campello et al. [2013]) para *clusterizar* os *embeddings*. Ele realiza a clusterização hierárquica seguindo o padrão do DBSCAN (Ester et al. [1996])

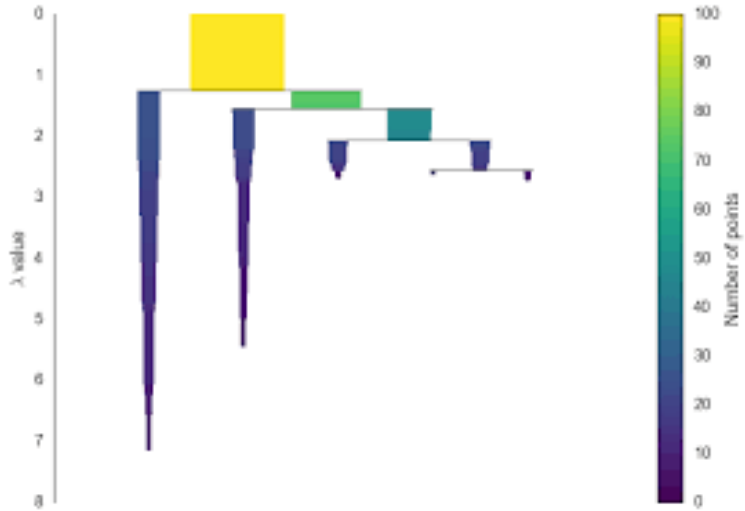



Figure: Representação do HBDSCAN.[13]

[13] Fonte: https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html

# Document Clustering

Também podemos utilizar diferentes algoritmos para clusterização:

- k-Means
- Agglomerative Clustering (ou outros do *scikit-learn*)
- cuML HDBSCAN

# Vectorizers/Tokenizer

Nessa etapa, diversos algoritmos são utilizados para ajudar nas representações de cada tópico, permitindo uma melhor compreensão do conteúdo agrupado em cada um deles.

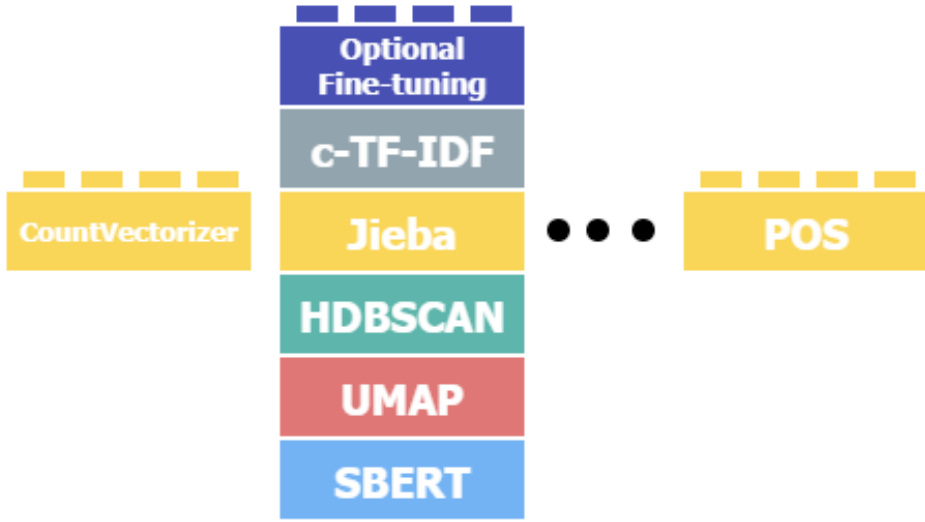



Figure: Modelos de vetorização.[14]

# CountVectorizer

Um método comum é o CountVectorizer, que permite ajustar os tópicos sem retreinar o modelo. Nele, podemos:

- Alterar os n-gramas
- Remover *stop words*
- Definir o número mínimo de aparições de uma palavra
- Definir o número máximo de nomes
- Personalizar tokenizadores

Além disso, há o OnlineCountVectorizer, que atualiza os tópicos aplicando *decay* e limpeza contínua.

# Topic Representation

Os tópicos são extraídos dos *clusters* gerados na etapa de *Document Clustering*. Para identificar as palavras-chave de cada *cluster*, utilizamos uma versão modificada do TF-IDF, chamada cTF-IDF.

- O TF-IDF calcula a importância de palavras em documentos.
- O cTF-IDF realiza os cálculos por *cluster*, gerando distribuições de palavras por tópico.

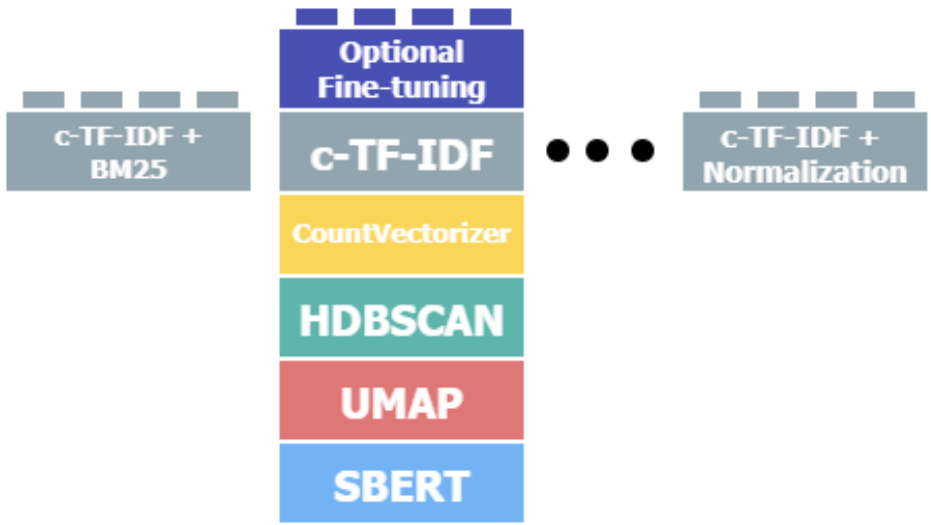



Figure: Representação dos Tópicos.[15]

# Topic Representation

Dois parâmetros importantes no **cTF-IDF**:

- **bm25_weighting**: Ativa a ponderação BM-25 baseada em classes, aumentando a robustez contra stop words em conjuntos de dados menores.
- **reduce_frequent_words**: Reduz palavras frequentemente ocorrentes que não são consideradas stop words padrão, aplicando a raiz quadrada da frequência dos termos após a ponderação. Essa mudança pode impactar significativamente a quantidade de stop words nas representações dos tópicos.

# Fine-tune Topics - Representation Models

O BERTopic oferece modelos de representação opcionais para ajustes finos nas palavras-chave de cada tópico:

- Facilita a atualização dos tópicos após o treinamento, sem necessidade de retreinamento.
- É possível refinar ainda mais as representações com diversos modelos implementados.

# Fine-tune Topics - Representation Models

KeyBERTInspired: Esse modelo ajusta a relação semântica entre palavras-chave e o conjunto de documentos em cada tópico.

- 1) São sorteados **n** documentos representativos em cada tópico;
- 2) É calculado o c-TF-IDF para esses documentos, no qual os que melhor representam o tópico são obtidos;
- 3) São gerados *embeddings* para cada tópico;
- 4) São comparados os *embeddings* dos tópicos com as palavras;
- 5) Palavras mais próximas a cada tópico são obtidas.

# Fine-tune Topics - Representation Models

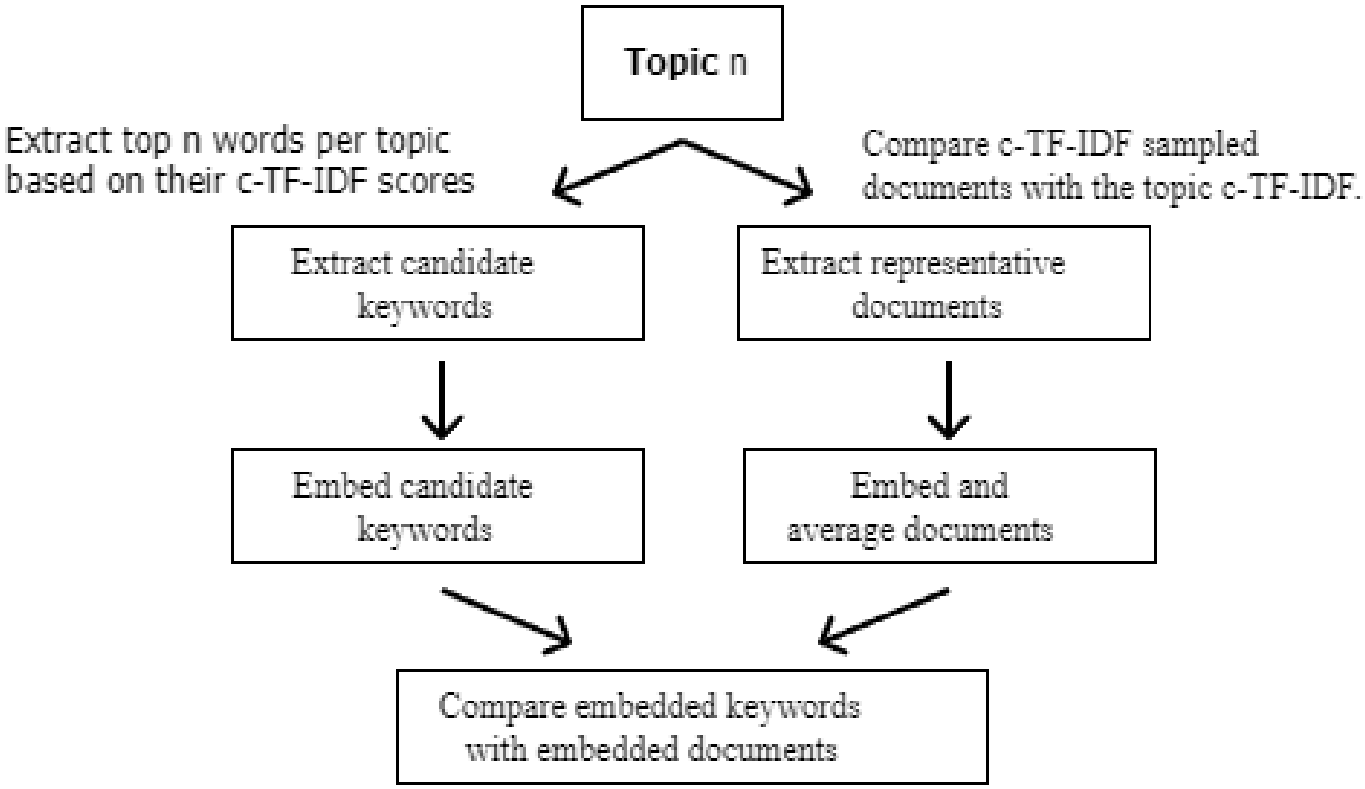



Figure: KeyBERTInspired.[16]

# Fine-tune Topics - Representation Models

PartOfSpeech (POS): utiliza informações sobre as classes gramaticais das palavras (como substantivos e adjetivos) para melhorar a extração de palavras-chave.

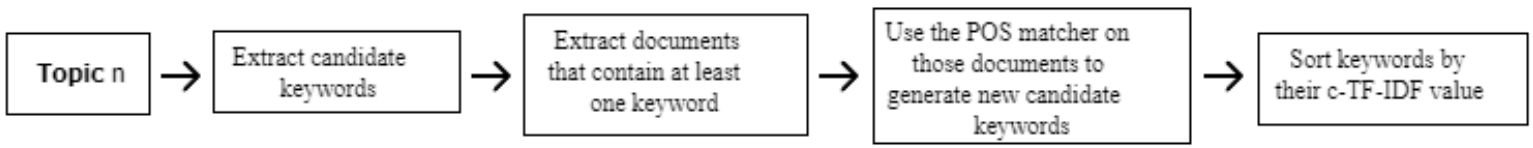



Figure: PartOfSpeech.[17]

# Fine-tune Topics - Representation Models

| Default Representation | | PartOfSpeech |
|---|---|---|
| meat \| organic \| food \| beef \| emissions \| eat \| of \| eating \| is | ⟶ | meat \| organic \| food \| beef \| emissions \| most \| health \| pesticides \| production |
| the \| explosion \| atmosphere \| eruption \| kilometers \| of \| | ⟶ | explosion \| atmosphere \| eruption \| kilometers \| eruptions \| fireball \| super |
| immune \| system \| your \| cells \| my \| and \| is \| the \| how \| of | ⟶ | immune \| system \| cells \| immunology \| adaptive \| body \| memory \| cell |
| moon \| earth \| lunar \| tides \| the \| water \| orbit \| base \| moons | ⟶ | moon \| earth \| lunar \| tides \| water \| orbit \| base \| moons \| surface \| gravity |
| eu \| european \| democratic \| vote \| parliament \| member \| union | ⟶ | democratic \| vote \| parliament \| member \| union \| states \| national \| countries |
| plastic \| plastics \| tons \| pollution \| waste \| microplastics \| polymers | ⟶ | plastic \| plastics \| tons \| pollution \| waste \| microplastics \| polymers \| bag |

Figure: PartOfSpeech Exemplo.[18]

# Fine-tune Topics - Representation Models

Maximal Marginal Relevance (MMR): Este modelo visa aumentar a diversidade de palavras-chave em um tópico.

| Default Representation | | MaximalMarginalRelevance |
|---|---|---|
| meat \| organic \| food \| beef \| emissions \| eat \| of \| eating \| is | ⟶ | meat \| organic \| beef \| emissions \| health \| pesticides \| foods \| farming \| conventional |
| the \| explosion \| atmosphere \| eruption \| kilometers \| of \| | ⟶ | explosion \| atmosphere \| eruption \| eruptions \| crust \| volcanoes \| earthquakes |
| immune \| system \| your \| cells \| my \| and \| is \| the \| how \| of | ⟶ | immune \| system \| cells \| immunology \| adaptive \| body \| memory \| antibodies |
| moon \| earth \| lunar \| tides \| the \| water \| orbit \| base \| moons | ⟶ | moon \| lunar \| tides \| moons \| surface \| gravity \| tide \| meters \| oceans \| dust |
| eu \| european \| democratic \| vote \| parliament \| member \| union | ⟶ | eu \| democratic \| vote \| parliament \| citizen \| laws \| institutions \| influence \| nations |
| plastic \| plastics \| tons \| pollution \| waste \| microplastics \| polymers | ⟶ | plastics \| tons \| pollution \| waste \| microplastics \| polymers \| ocean \| bpa \| cotton |

Figure: Maximal Marginal Relevance Exemplo.[19]

# Fine-tune Topics - Representation Models

Zero-Shot Classification: Se você já tem uma lista de tópicos candidatos, pode usar Zero-Shot Classification para atribuir automaticamente rótulos a tópicos gerados.

| Default Representation | | ZeroShotClassification |
|---|---|---|
| meat \| organic \| food \| beef \| emissions \| eat \| of \| eating \| is | ⟶ | Organic food |
| the \| explosion \| atmosphere \| eruption \| kilometers \| of \| | ⟶ | the \| explosion \| atmosphere \| eruption \| kilometers \| of |
| immune \| system \| your \| cells \| my \| and \| is \| the \| how \| of | ⟶ | Your immune system |
| moon \| earth \| lunar \| tides \| the \| water \| orbit \| base \| moons | ⟶ | moon \| earth \| lunar \| tides \| the \| water \| orbit \| base \| moons |
| eu \| european \| democratic \| vote \| parliament \| member \| union | ⟶ | eu \| european \| democratic \| vote \| parliament \| member \| union |
| plastic \| plastics \| tons \| pollution \| waste \| microplastics \| polymers | ⟶ | plastic \| plastics \| tons \| pollution \| waste \| microplastics \| polymers |

Figure: Zero-Shot Classification Exemplo.[20]

# Fine-tune Topics - Representation Models

Os Modelos Encadeados permitem a capacidade de encadear diferentes modelos de representação. Por exemplo:

- Utilize primeiro o Maximal Marginal Relevance.
- Em seguida, aplique um modelo GPT para refinar ainda mais as representações dos tópicos.

Essa abordagem proporciona maior flexibilidade e precisão nas representações dos tópicos.

# Fine-tune Topics - Representation Models

Os Modelos Customizados permitem que o usuário utilize um modelo de sua preferência. Essa flexibilidade possibilita a adaptação do modelo às necessidades específicas de cada projeto.

# Fine-tune Topics - LLM & Generative AI

Os tópicos gerados pelo **BERTopic** podem ser aprimorados com **Modelos de Linguagem** como ChatGPT, GPT-4 e soluções open-source.

- Podemos criar rótulos, resumos e descrições.

# Fine-tune Topics - LLM & Generative AI

Modelos que podem ser utilizados incluem:

- Zephyr Mistral 7B
- Llama 2
- llama.cpp
- OpenAI
- ChatGPT
- Langchain
- Cohere

| Default Representation | | Cohere |
|---|---|---|
| meat \| organic \| food \| beef \| emissions \| eat \| of \| eating \| is | ⟶ | Organic food |
| the \| explosion \| atmosphere \| eruption \| kilometers \| of \| | ⟶ | Exploding planets |
| immune \| system \| your \| cells \| my \| and \| is \| the \| how \| of | ⟶ | How your immune system works |
| moon \| earth \| lunar \| tides \| the \| water \| orbit \| base \| moons | ⟶ | How tides work |
| eu \| european \| democratic \| vote \| parliament \| member \| union | ⟶ | How democratic is the European Union? |
| plastic \| plastics \| tons \| pollution \| waste \| microplastics \| polymers | ⟶ | Plastic pollution |

Figure: Exemplo utilizando Cohere.[21]

# Fine-tune Topics - LLM & Generative AI

Tarefas que esses modelos podem realizar:

- Sumarização
- Ajuste dos tópicos
- Truncar documentos (Diminuir número de tokens)
- Selecionar documentos (4 Melhores documentos)
- Engenharia de prompt

```
prompt = """
I have topic that contains the following documents: \n[DOCUMENTS]
The topic is described by the following keywords: [KEYWORDS]

Based on the above information, can you give a short label of the topic?
"""
```

Figure: Exemplo de Prompt Engineering.[22]

# Fine-tune Topics - Multiple Representations

Em vez de gerar uma única representação de um tópico, podemos criar várias representações para o mesmo tópico.

- Utilizamos diferentes modelos de representação em paralelo.
- Exemplo: podemos usar o modelo **PartOfSpeech** e também o **KeyBERTInspired** para representar nossos tópicos.

Isso nos permite comparar os resultados e obter uma compreensão mais profunda do agrupamento.

# Fine-tune Topics - Multiple Representations

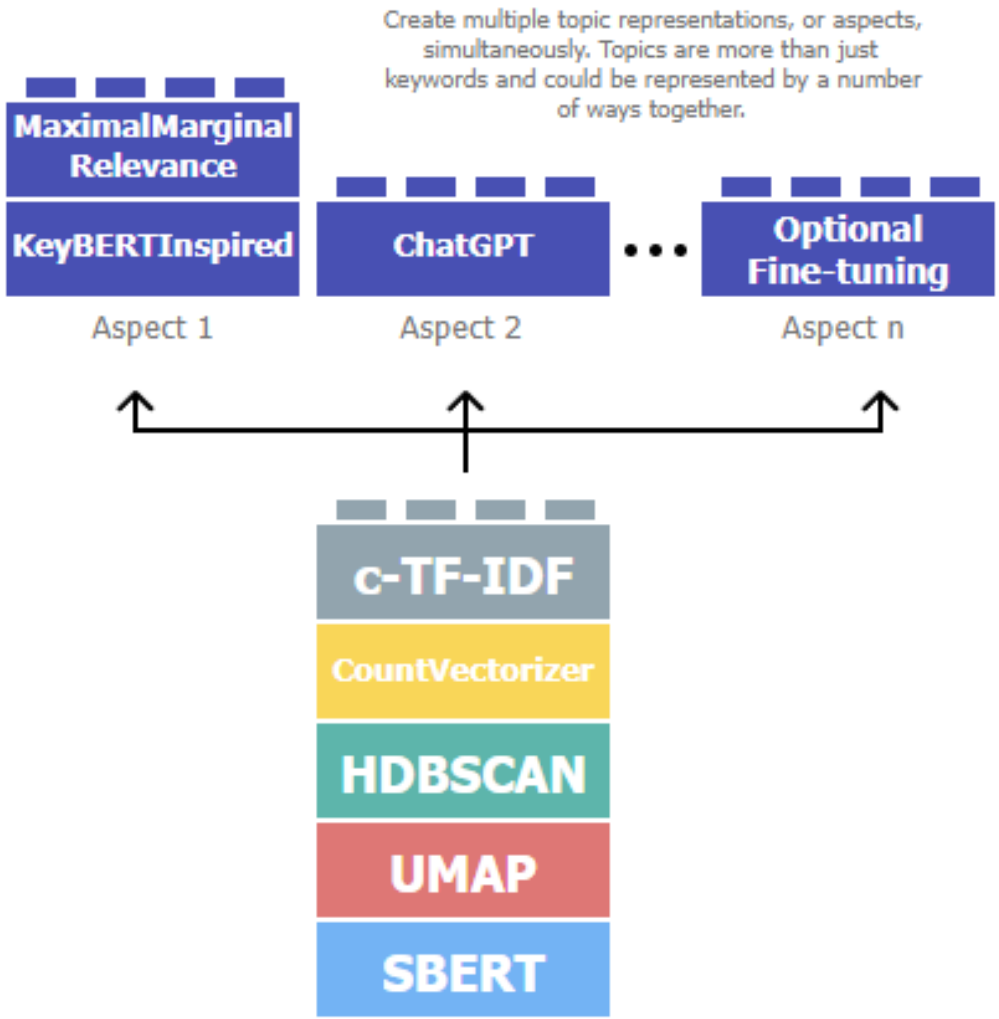



Figure: Exemplo de Multiple Representations.[23]

# Fine-tune Topics - Multiple Representations

```
topic_model.get_topic_info()
```

| | Topic | Count | Name | Representation | Aspect1 | Aspect2 | Representative_Docs |
|---|---|---|---|---|---|---|---|
| 0 | -1 | 6774 | -1_more_get_about_than | [more, get, about, than, the, information, you... | [one, use, other, more, time, new, information... | [about, than, information, your, other, system... | [\n\n*nnnnnnnng* Thank you for playing, I cann... |
| 1 | 0 | 1838 | 0_nhl_hockey_rangers_league | [nhl, hockey, rangers, league, players, baseba... | [game, team, games, season, hockey, players, y... | [nhl, rangers, espn, play, pts, series, stl, b... | [The problem with your nihilistic approach, Ro... |
| 2 | 1 | 591 | 1_encryption_crypto_nsa_encrypted | [encryption, crypto, nsa, encrypted, clipper, ... | [key, clipper, chip, encryption, keys, escrow,... | [encryption, nsa, clipper, agencies, chips, se... | [The following document summarizes the Clipper... |
| 3 | 2 | 530 | 2_hello___ | [hello, . . . . . ] | [. . . . . ] | [hello, . . . . . ... | [ites:, Hello., Hello.] |
| 4 | 3 | 474 | 3_investigation_assault_fired_fire | [investigation, assault, fired, fire, koresh, ... | [koresh, fire, compound, gas, children, agents... | [investigation, assault, fired, koresh, batf, ... | [NOTE - local tx groups trimmed out of Newsgro... |
| ... | ... | ... | ... | ... | ... | ... | ... |

Figure: Exemplo de Multiple Representations.[24]

# Variations - Dynamic Topic Modeling

A Dynamic Topic Modeling (DTM) permite analisar a evolução dos tópicos ao longo do tempo.

- No BERTopic, calcula as representações temporais dos tópicos sem precisar re-treinar o modelo para cada período.
- O modelo é inicialmente treinado sem considerar o elemento temporal.
- A representação de cada tópico ao longo do tempo é calculada usando a técnica c-TF-IDF.

# Variations - Dynamic Topic Modeling

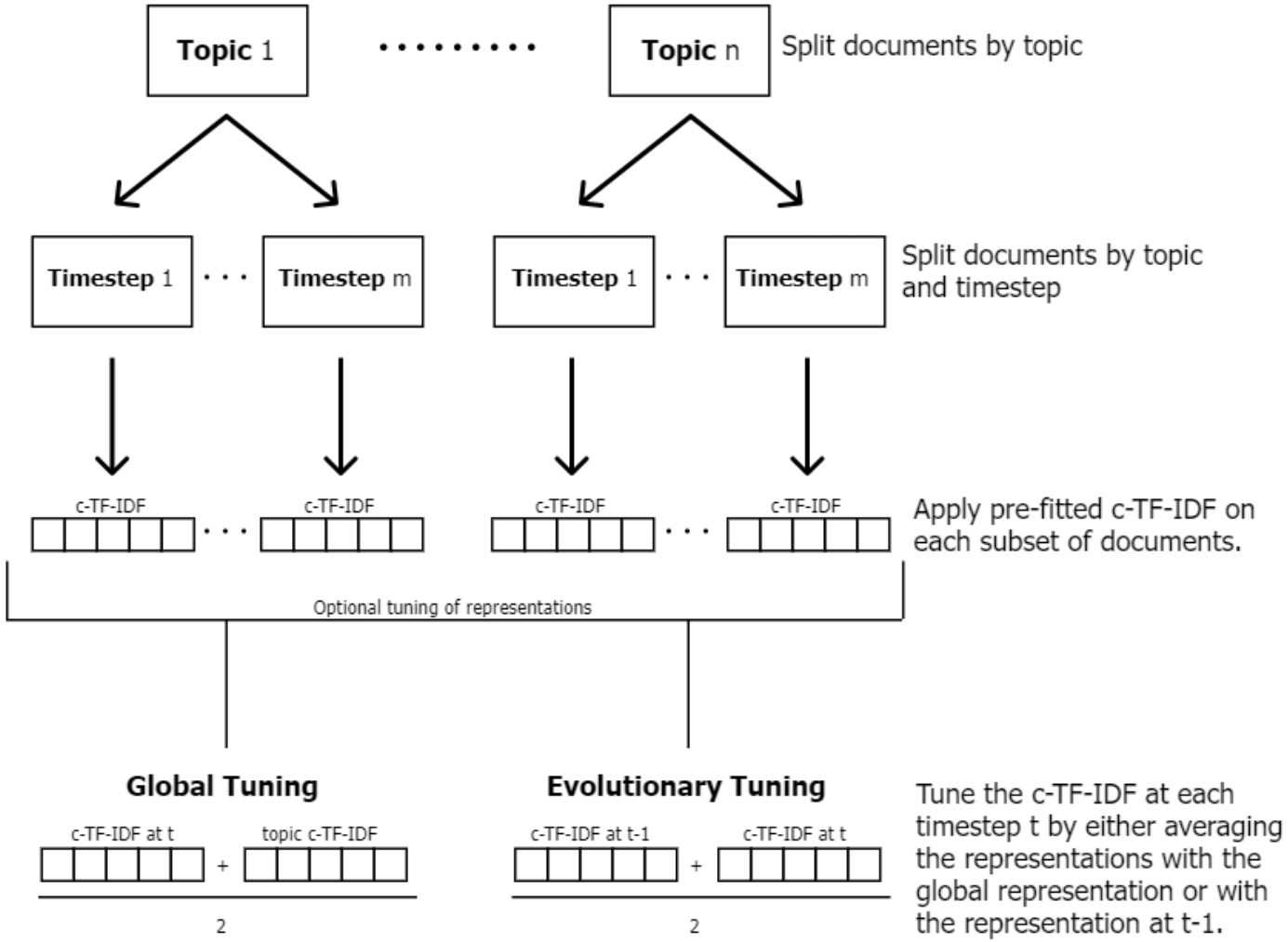



Figure: Dynamic Topic Modeling.[25]

# Variations - Hierarchical Topic Modeling

Hierarchical Topic Modeling explora a estrutura hierárquica dos tópicos identificados.

- Permite agrupar tópicos em uma hierarquia.
- Exemplo:
  - Tópicos sobre "política", "política internacional" e "eleições".
  - Podem ser representados como sub-tópicos de um tópico mais amplo: "política".

# Variations - Hierarchical Topic Modeling

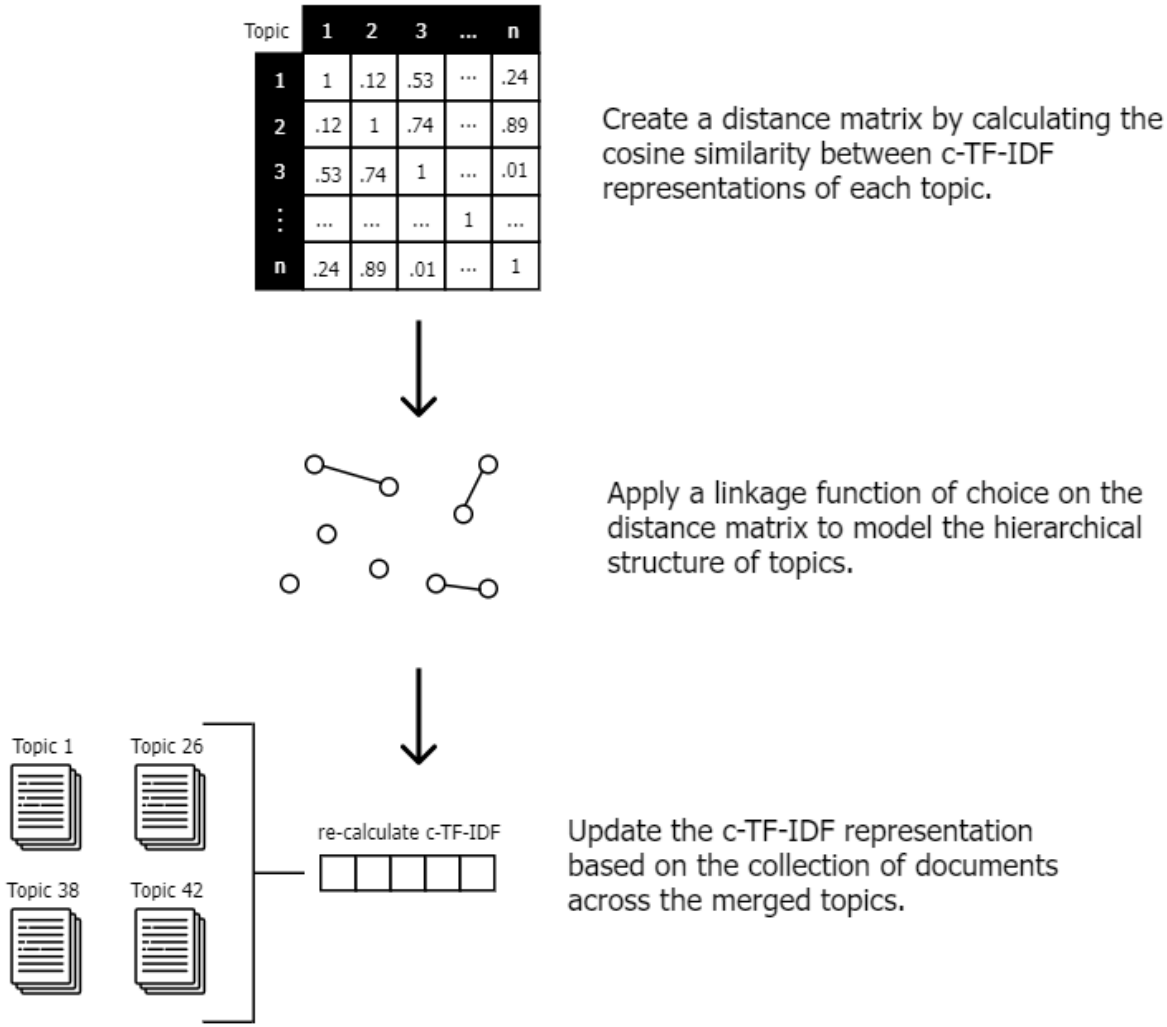



Figure: Hierarchical Topic Modeling.[26]

# Variations - Multimodal Topic Modeling

Multimodal Topic Modeling refere-se à modelagem de tópicos em dados que combinam diferentes tipos de mídias.

- Exemplos incluem a combinação de texto e imagens.
- No BERTopic, é possível usar imagens com legendas.
- Exemplo: Fotos no Instagram com descrições permitem uma análise multimodal, capturando mais contexto.

# Variations - Multimodal Topic Modeling

Text + images:

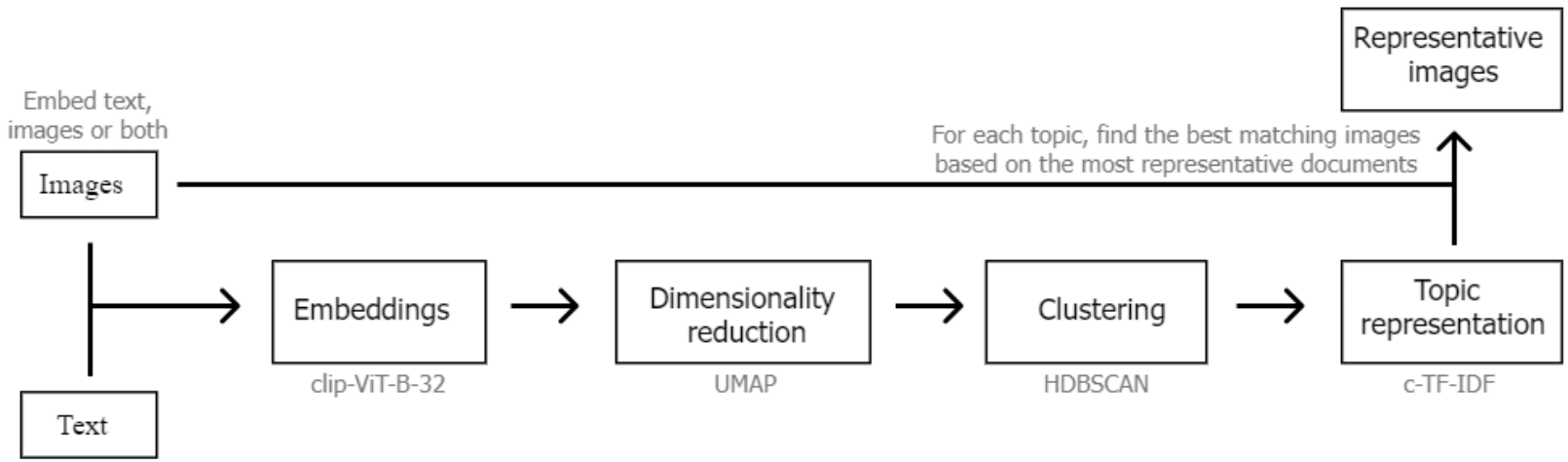



Figure: Multimodal Topic Modeling (Text + images).[27]

# Variations - Multimodal Topic Modeling

Just images:

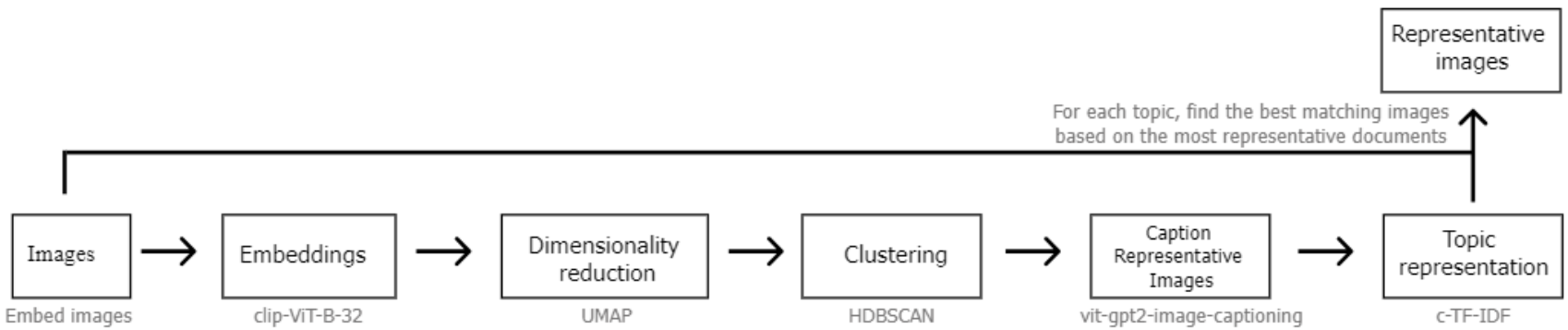



Figure: Multimodal Topic Modeling (just images).[28]

# Variations - Online Topic Modeling

Online Topic Modeling refere-se à capacidade de atualizar um modelo de tópicos continuamente conforme novos dados são inseridos.

- Diferente da modelagem tradicional, que é treinada uma vez com todos os dados disponíveis.
- Permite que o modelo ajuste-se dinamicamente.

Os principais objetivos da modelagem de tópicos online são:

- Reduzir o uso de memória.
- Atualização contínua do modelo.
- Descobrir novos tópicos.

# Variations - Online Topic Modeling

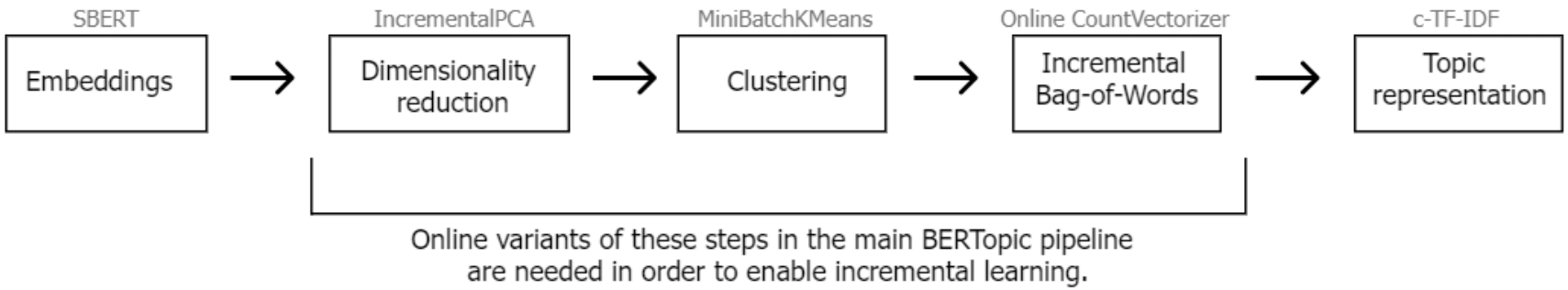



Figure: Online Topic Modeling.[29]

# Variations - Merge Multiple Fitted Models

A funcionalidade "Merge Multiple Fitted Models" no BERTopic permite combinar diferentes modelos de tópicos previamente treinados.

- Combina resultados de modelos em um único modelo.
- Preserva o conhecimento de modelos anteriores.
- Elimina a necessidade de re-treinar um único modelo desde o início.

# Variations - Semi-supervised Topic Modeling

Semi-supervised Topic Modeling no BERTopic permite guiar o processo de modelagem de tópicos.

- Cria tópicos mais alinhados com categorias ou rótulos predefinidos.
- Útil quando há conhecimento prévio sobre os dados, como rótulos ou classes.
- Garante que os tópicos gerados reflitam essas categorias.

# Variations - Semi-supervised Topic Modeling

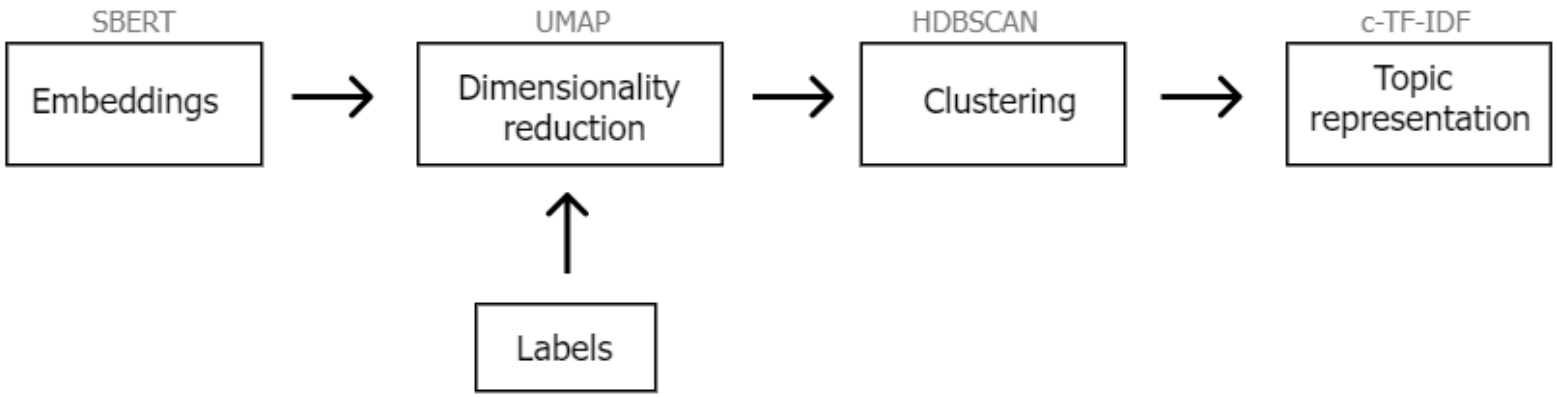



Figure: Semi-supervised Topic Modeling.[30]

# Variations - Supervised Topic Modeling

Supervised Topic Modeling no BERTopic refere-se a uma abordagem onde:

- Utiliza rótulos ou categorias pré-definidas para os documentos.
- Modela tópicos com base nesses rótulos.
- Ao contrário da modelagem não supervisionada, parte-se de informações já conhecidas.

# Variations - Supervised Topic Modeling

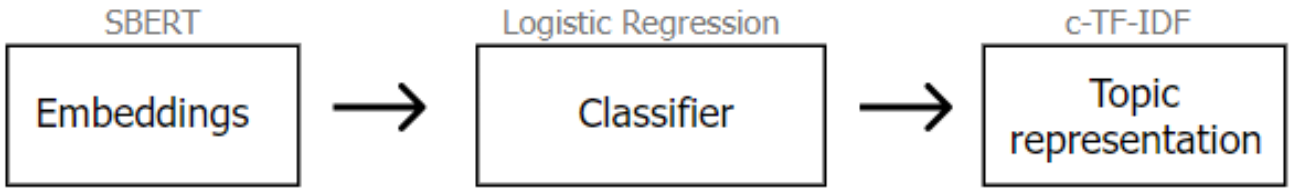



Figure: Supervised Topic Modeling.[31]

# Variations - Manual Topic Modeling

Manual Topic Modeling no BERTopic refere-se a um processo em que:

- Os tópicos já são conhecidos ou rotulados.
- Não é necessário descobrir novos tópicos por meio de clusterização ou redução de dimensionalidade.
- O objetivo é transformar rótulos preexistentes em representações de tópicos.
- Utiliza a técnica de c-TF-IDF para identificar as palavras mais representativas de cada classe de documentos.
- Por exemplo, imagine que você tem uma coleção de artigos de jornal. Esses artigos já estão classificados em diferentes seções, como esportes, política, entretenimento e economia. Agora, você quer saber quais são as palavras que melhor descrevem cada uma dessas seções.

# Variations - Manual Topic Modeling

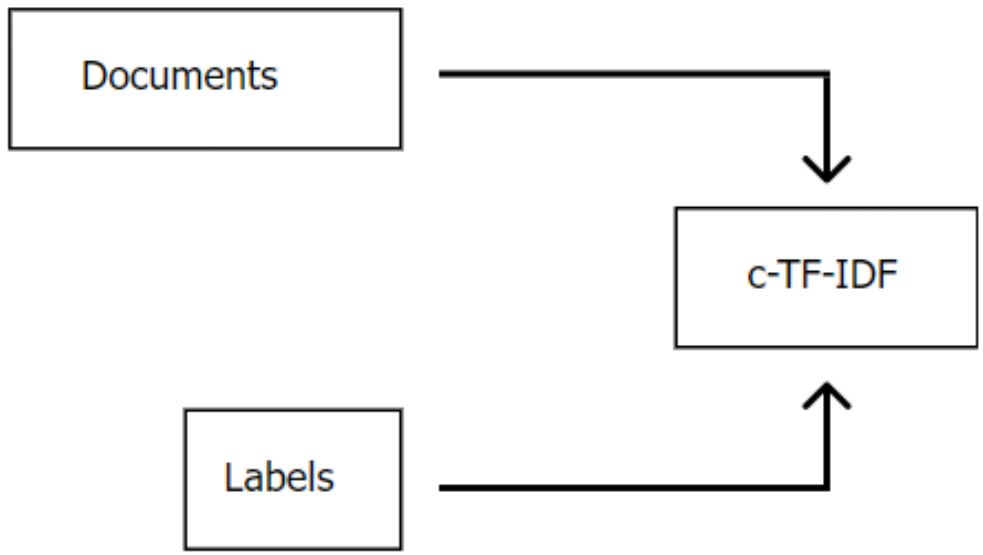



Figure: Manual Topic Modeling.[32]

# Variations - Guided Topic Modeling

Guided Topic Modeling, ou Seeded Topic Modeling, é uma técnica onde:

- O processo de modelagem de tópicos é guiado para convergir em direção a tópicos predefinidos, conhecidos como seed topics.
- Ao invés de descobrir tópicos de forma não supervisionada, o modelo é orientado por palavras-chave ou frases associadas a tópicos específicos.
- Por exemplo, imagine que você tem dados de uma livraria, onde os clientes escrevem resenhas sobre os livros que compraram. Você já tem uma ideia de quais são os tópicos de interesse, como ficção científica, autocuidado e culinária e quer descobrir o que os clientes estão falando.

# Variations - Guided Topic Modeling

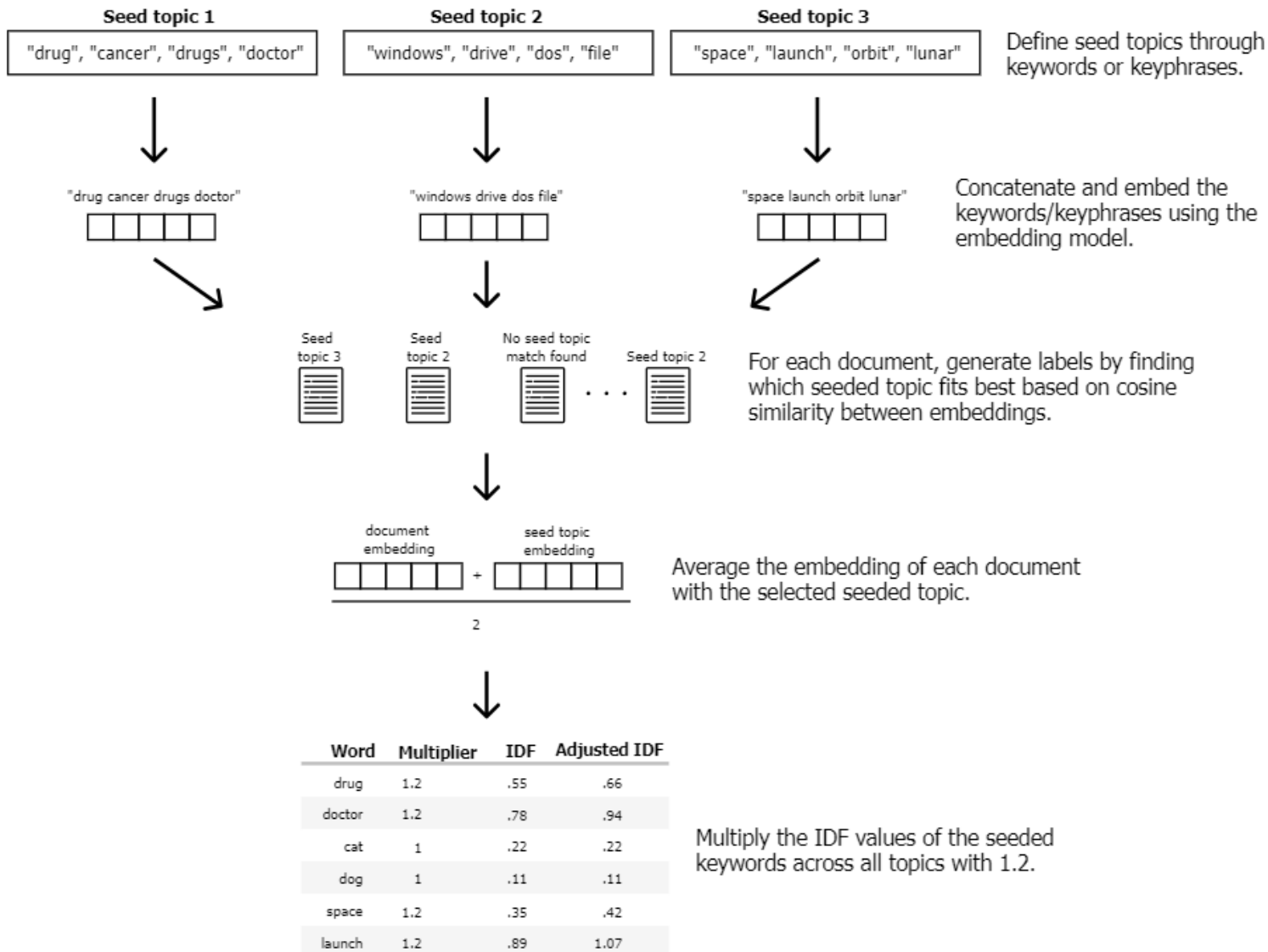



Figure: Guided Topic Modeling.[33]

# Variations - Zero-shot Topic Modeling

Zero-shot Topic Modeling é uma técnica que permite:

- Identificar tópicos em grandes quantidades de documentos com base em tópicos predefinidos, sem treinamento específico.
- Atribuir documentos a tópicos com base na similaridade semântica entre embeddings.
- Ser útil em cenários onde há uma boa noção dos tópicos presentes nos documentos.
- Oferecer flexibilidade: além de identificar tópicos predefinidos, pode criar novos tópicos para documentos que não se encaixam nos existentes.

# Variations - Zero-shot Topic Modeling

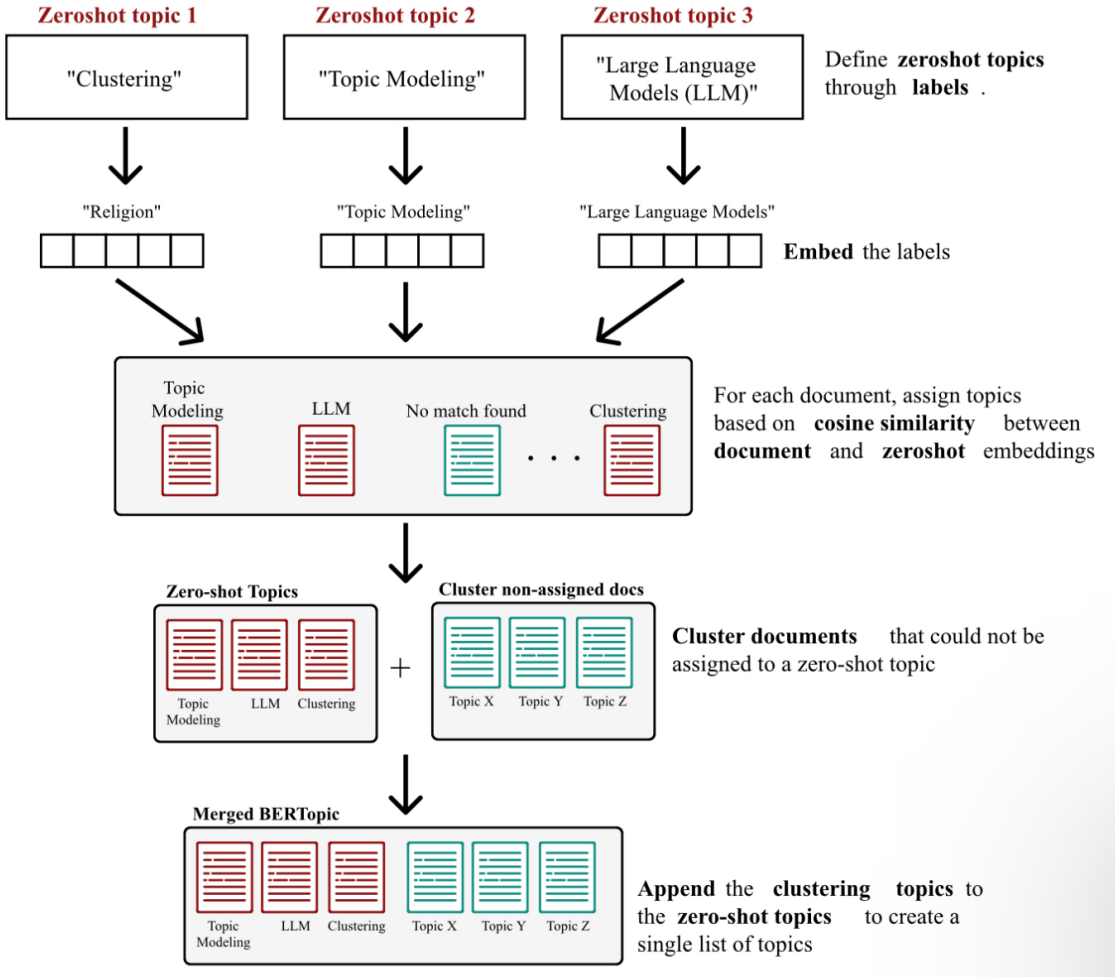



Figure: Zero-shot Topic Modeling.[34]

# Variations - Topic Distributions

Topic Distributions referem-se à maneira como um documento pode ser associado a vários tópicos, em vez de ser atribuído exclusivamente a um único tópico.

- O BERTopic permite calcular distribuições aproximadas de tópicos para cada documento.
- Essa abordagem captura a mistura de tópicos, refletindo a complexidade do conteúdo.

# Variations - Topic Distributions

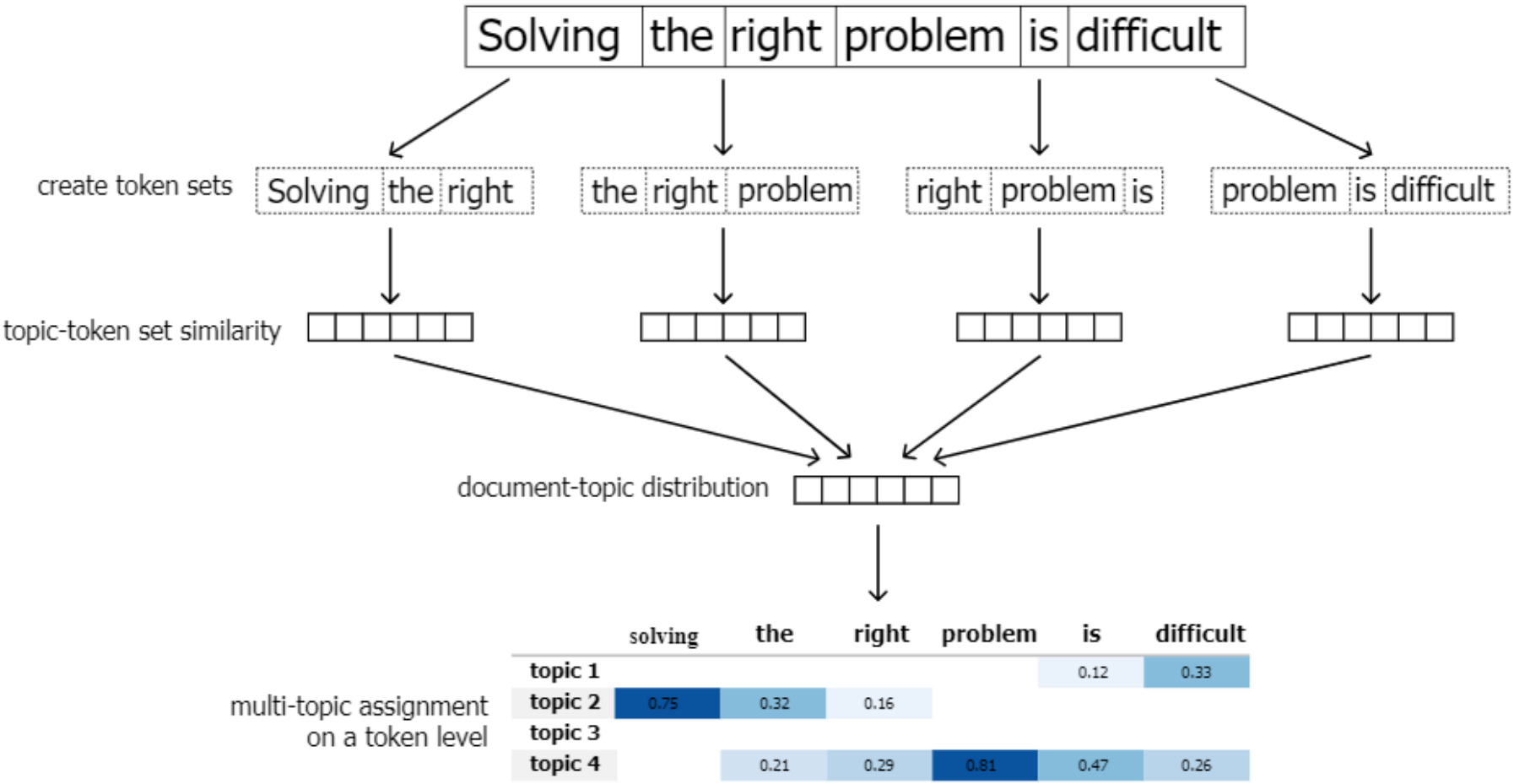



Figure: Topic Distributions.[35]

# Variations - Topics per Class

Ao invés de rodar o modelo de tópicos separadamente para cada classe, o Topics per Class permite:

- Criar um modelo global de tópicos.
- Analisar como esses tópicos são representados em diferentes subgrupos ou classes de documentos.
- Imagine que você está analisando comentários de clientes. Esses comentários estão divididos em diferentes categorias de produtos, como eletrônicos, moda, e alimentação. Você quer descobrir como os clientes falam sobre tópicos gerais (por exemplo, preço, qualidade, atendimento) em cada uma dessas categorias.

# Variations - Topics per Class

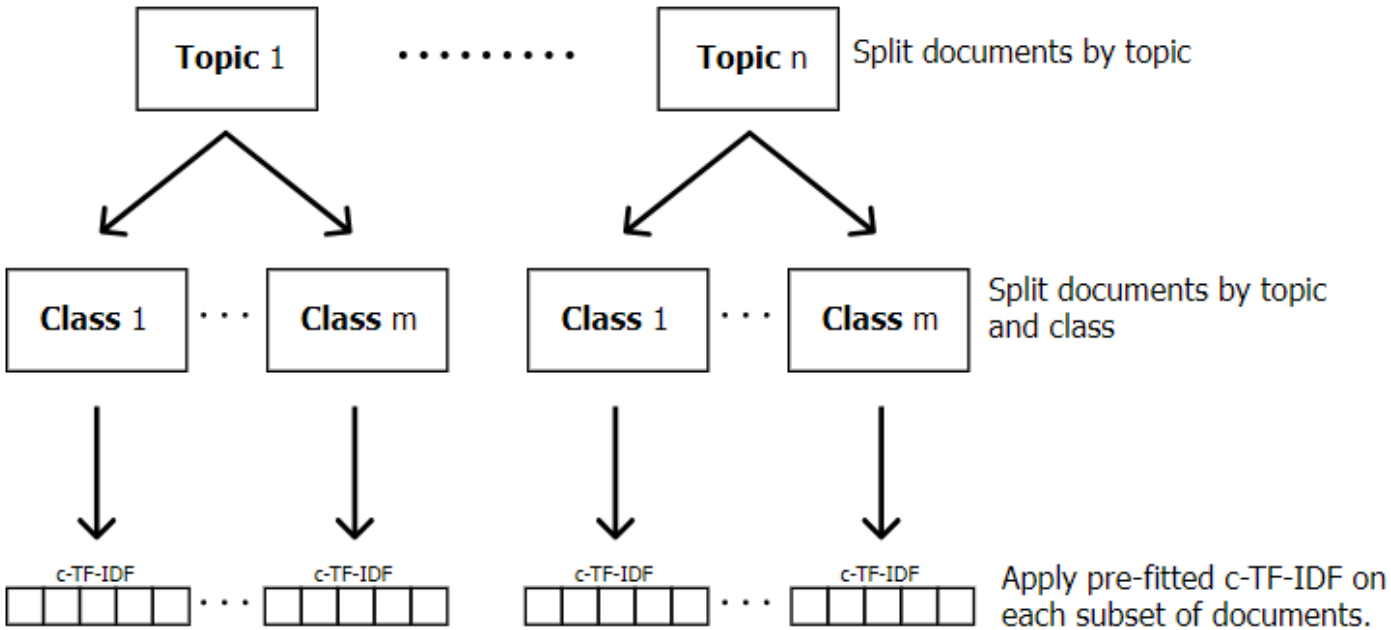



Figure: Topics per Class.[36]

# Variations - Seed Words

Seed Words são palavras-chave que você pode definir para orientar a modelagem de tópicos em dados que têm expressões específicas de um domínio.

- Essa técnica permite aumentar o peso e a importância dessas palavras na geração de tópicos.
- Influencia diretamente a forma como os tópicos são modelados.
- Imegine que você tem um conjunto de resumos de artigos e deseja realizar a modelagem de tópicos. Se você conhece o domínio dos dados (por exemplo, aprendizado por reforço), você pode saber que certas palavras, como "agente" ou "robô", são relevantes.

# Visualizações dos Tópicos

O **BERTopic** oferece diversos métodos para visualizar os tópicos:

- Gráficos de distância entre os tópicos
- Matriz de similaridade
- Tópicos ao longo do tempo
- Tópicos por classe

Essas visualizações ajudam a explorar e entender melhor os tópicos gerados.

# Visualizações dos Tópicos

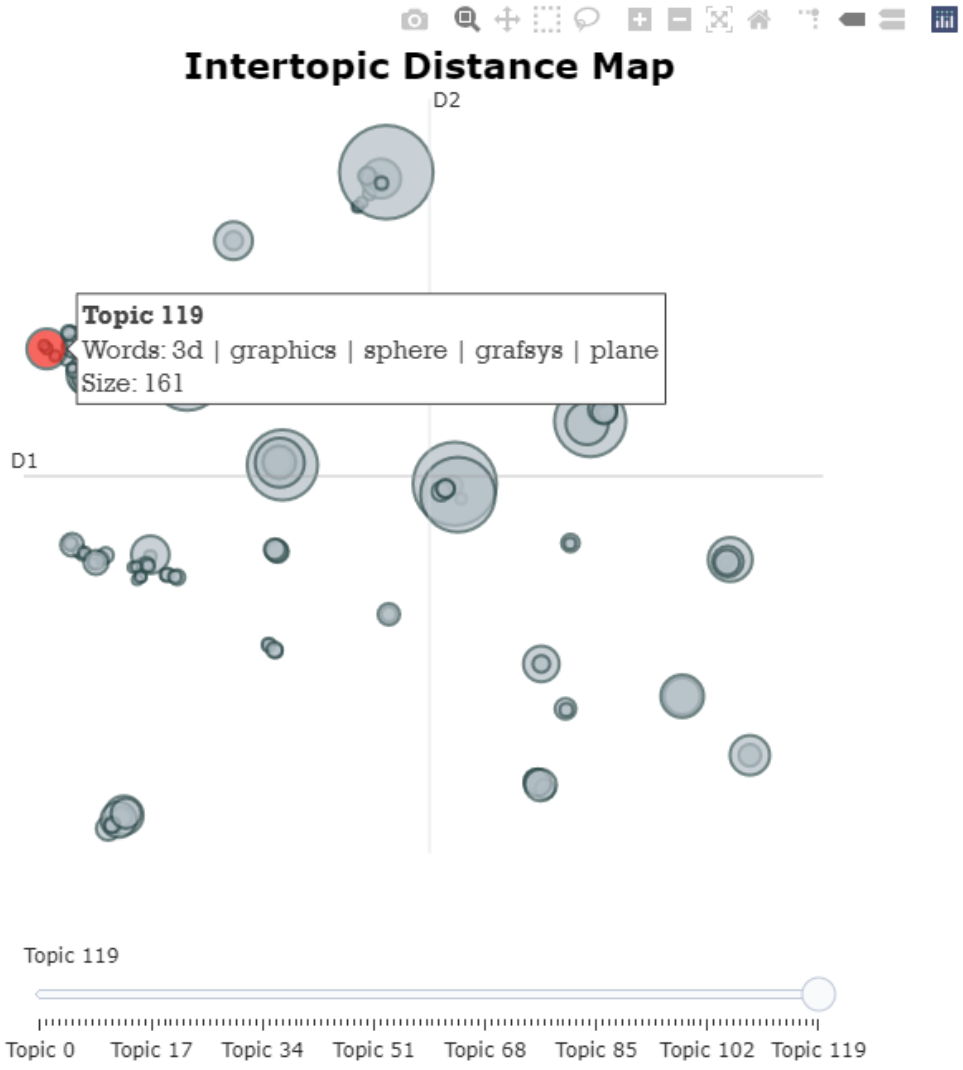



Figure: Mapa de distância entre tópicos.[37]

# Visualizações dos Tópicos

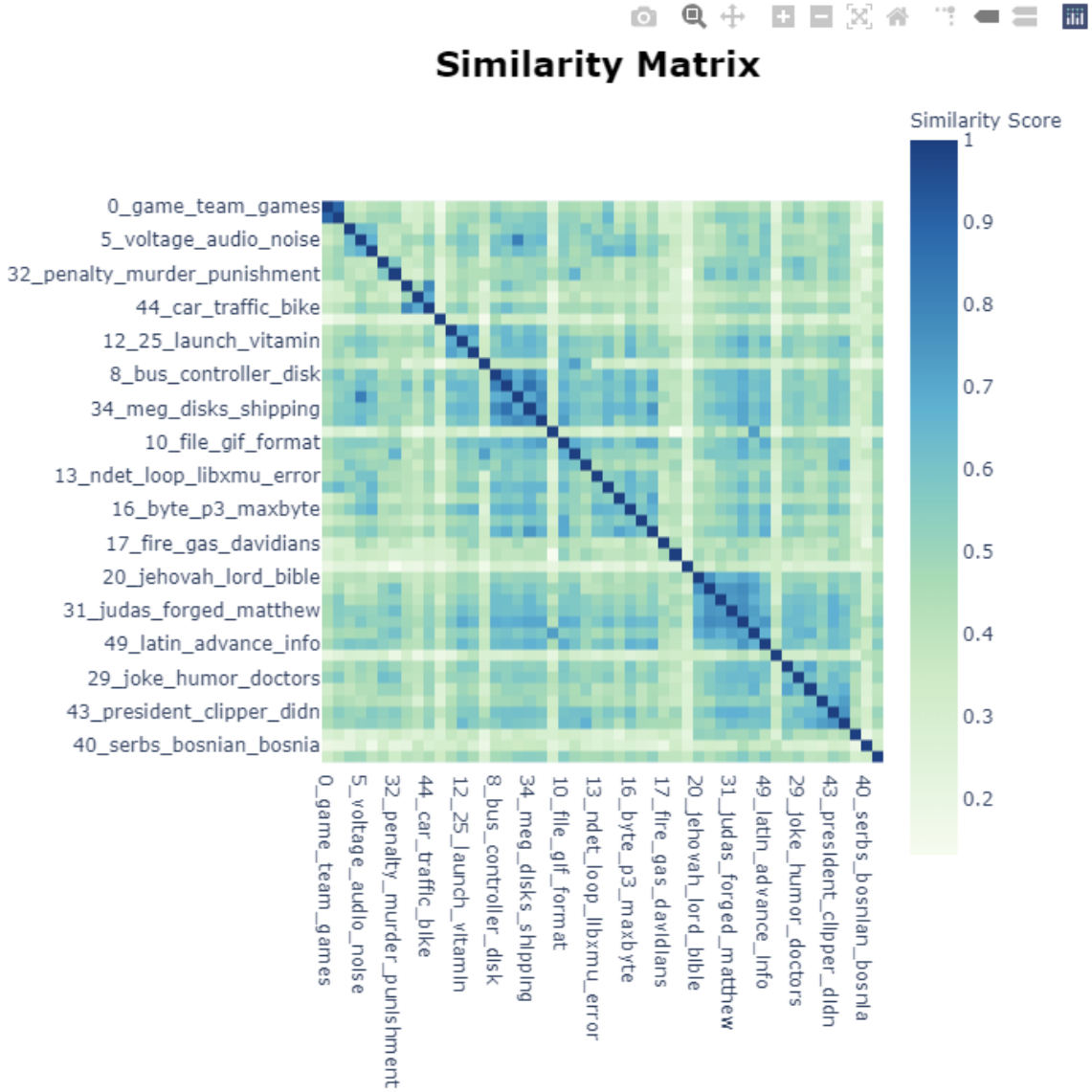



Figure: Matriz de similaridade entre tópicos.[38]

# Visualizações dos Tópicos

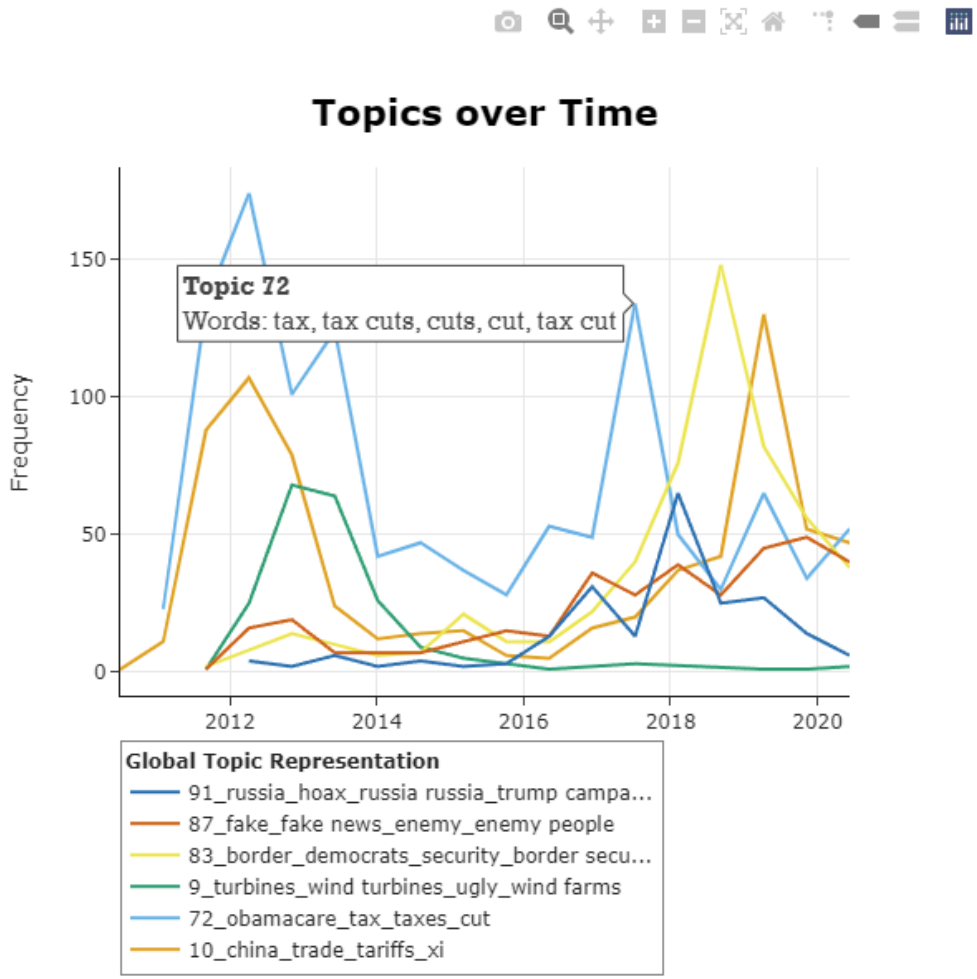



Figure: Gráfico dos tópicos ao longo do tempo.[39]

# Visualizações dos Tópicos

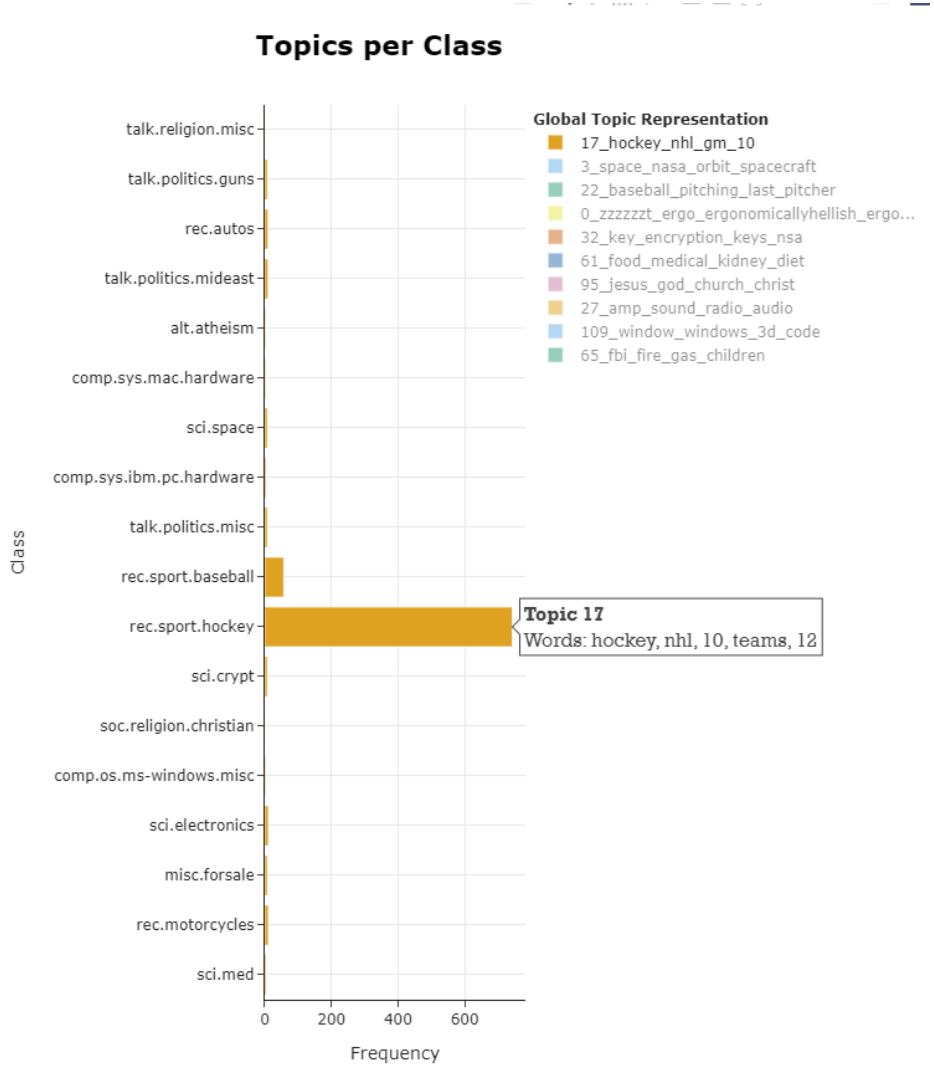



Figure: Gráfico dos tópicos por classe.[40]

# Visualizações dos Documentos

Nesta etapa, é possível observar os documentos em cada tópico com maior granularidade.

- Plotar gráficos de distribuição de probabilidade
- Visualizar a contribuição de cada token para o tópico

Essas visualizações permitem uma análise mais detalhada e precisa dos tópicos gerados.

# Visualizações dos Documentos

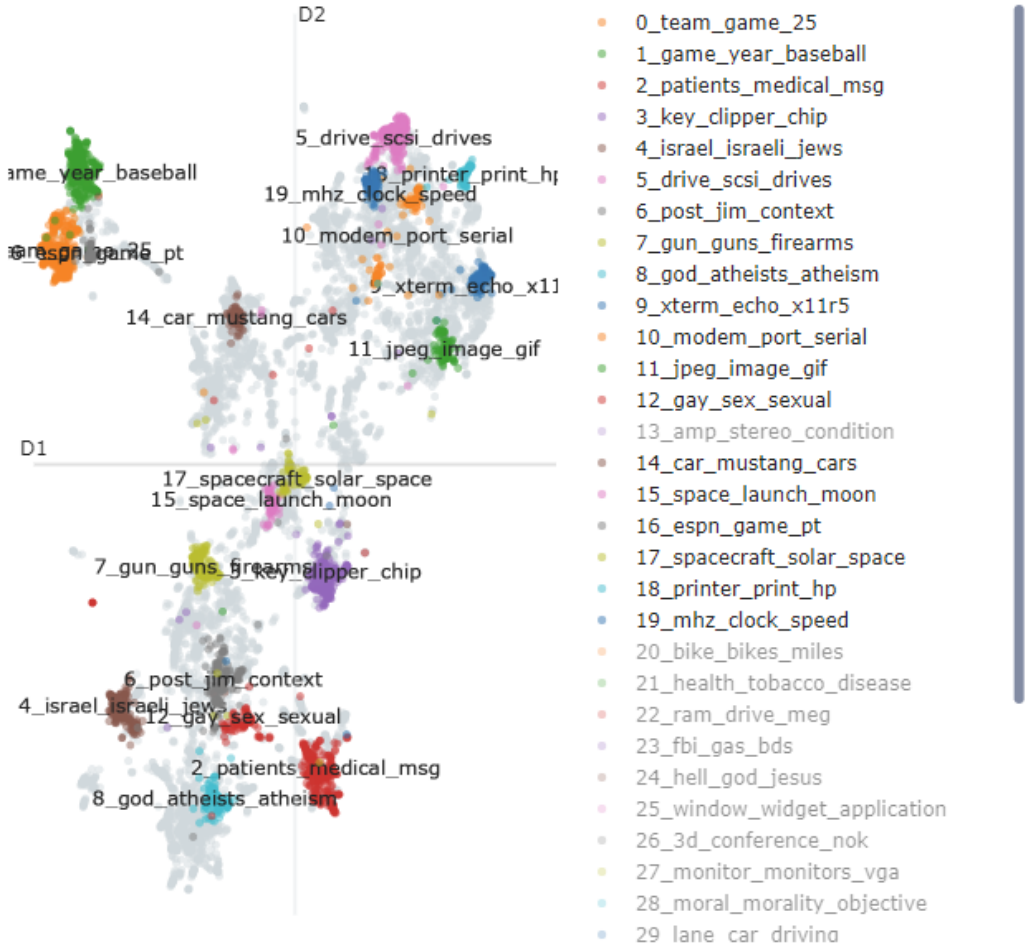



Figure: Gráfico para visualização dos documentos nos tópicos.[41]

# Visualizações dos Documentos

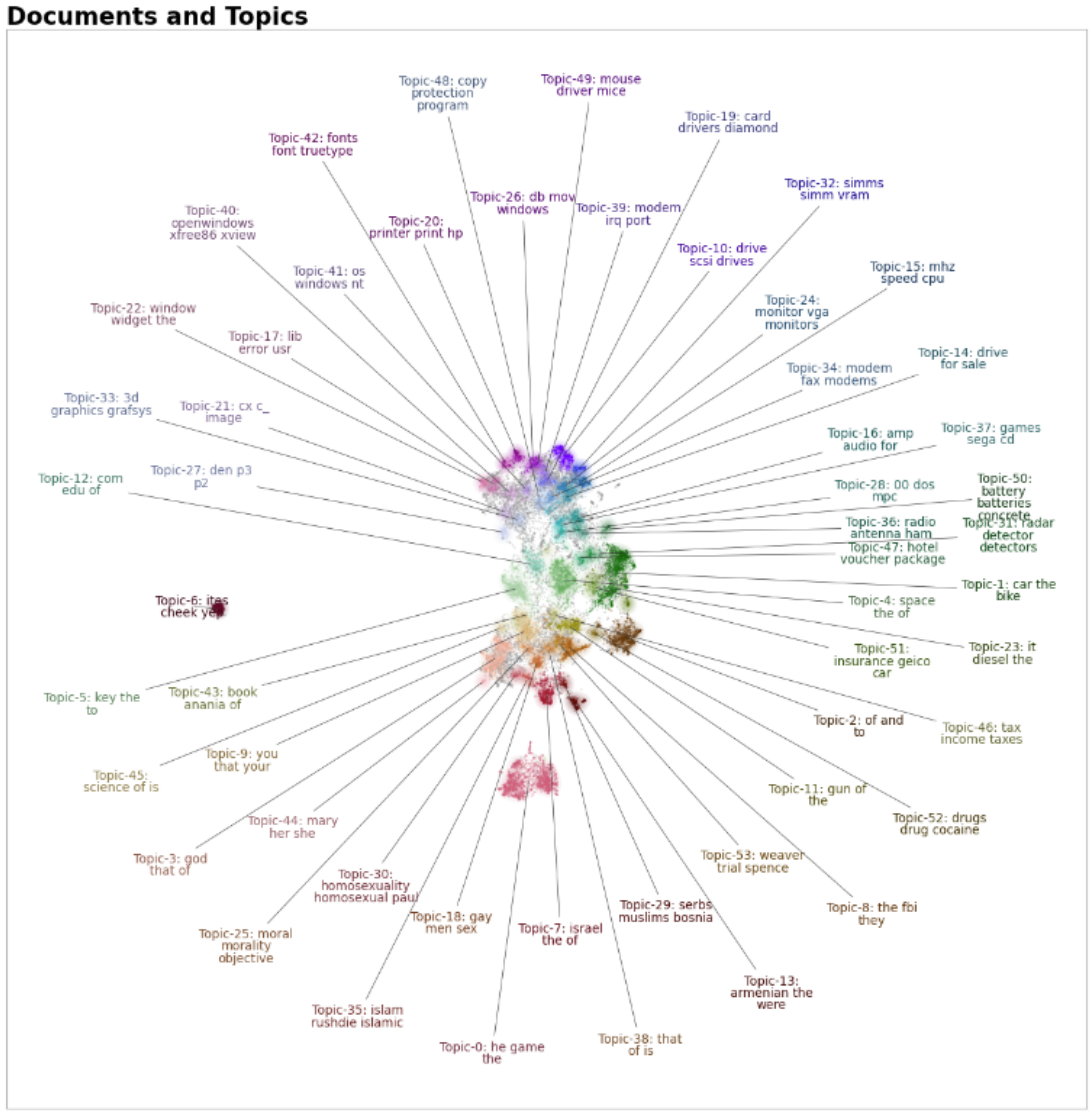



Figure: Gráfico para visualização dos documentos nos tópicos.[42]

# Visualizações dos Documentos

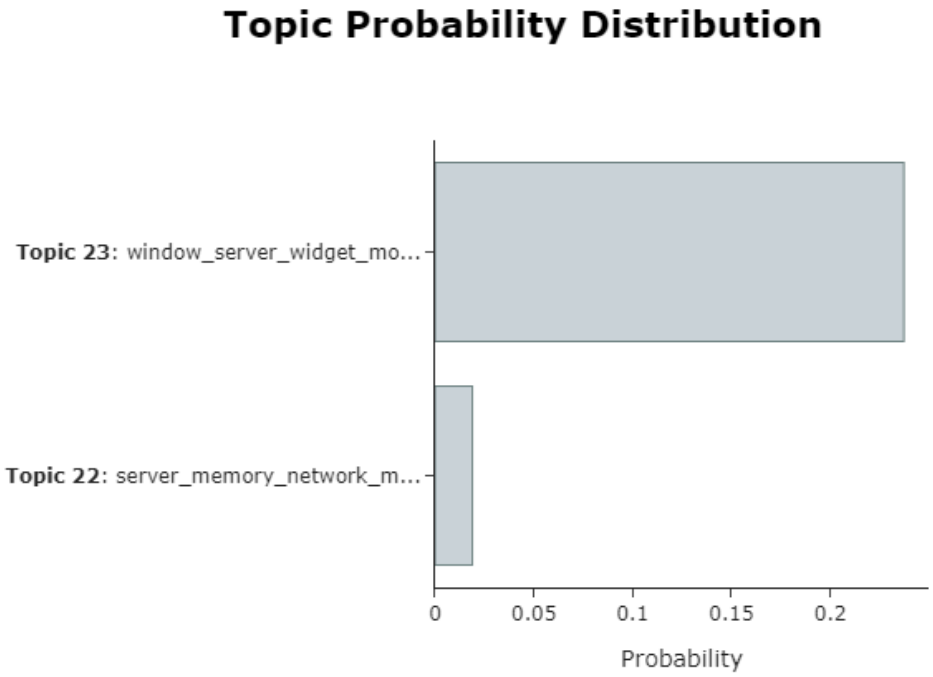
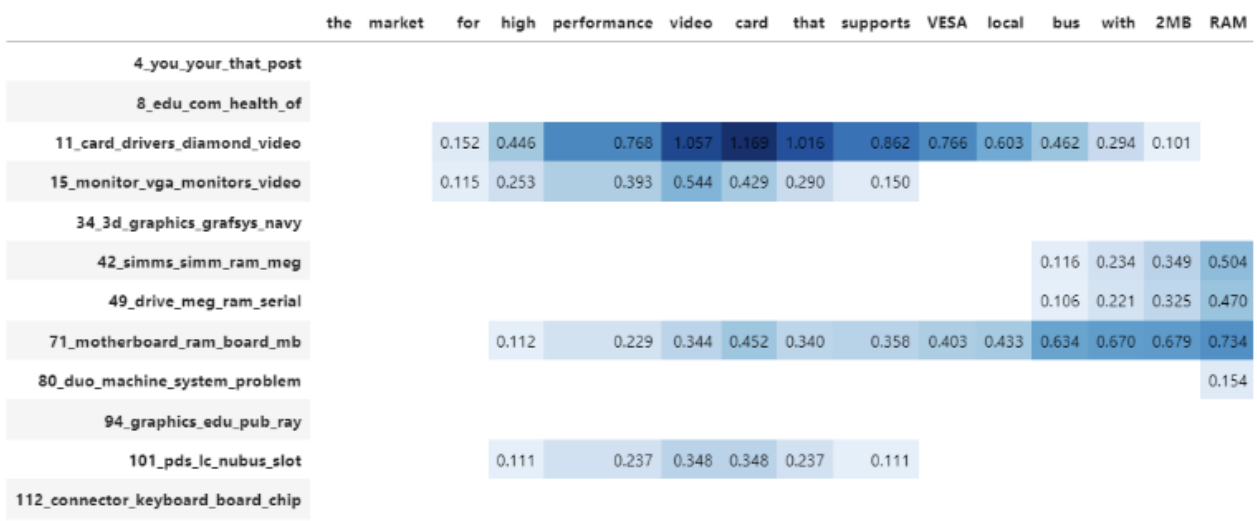



Figure: Gráfico da distribuição de probabilidade dos documentos nos tópicos.[44]

# Visualizações dos Documentos



| | the | market | for | high | performance | video | card | that | supports | VESA | local | bus | with | 2MB | RAM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4_you_your_that_post | | | | | | | | | | | | | | | |
| 8_edu_com_health_of | | | | | | | | | | | | | | | |
| 11_card_drivers_diamond_video | | | 0.152 | 0.446 | 0.768 | 1.057 | 1.169 | 1.016 | 0.862 | 0.766 | 0.603 | 0.462 | 0.294 | 0.101 | |
| 15_monitor_vga_monitors_video | | | 0.115 | 0.253 | 0.393 | 0.544 | 0.429 | 0.290 | 0.150 | | | | | | |
| 34_3d_graphics_grafsys_navy | | | | | | | | | | | | | | | |
| 42_simms_simm_ram_meg | | | | | | | | | | | | 0.116 | 0.234 | 0.349 | 0.504 |
| 49_drive_meg_ram_serial | | | | | | | | | | | | 0.106 | 0.221 | 0.325 | 0.470 |
| 71_motherboard_ram_board_mb | | | | 0.112 | 0.229 | 0.344 | 0.452 | 0.340 | 0.358 | 0.403 | 0.433 | 0.634 | 0.670 | 0.679 | 0.734 |
| 80_duo_machine_system_problem | | | | | | | | | | | | | | | 0.154 |
| 94_graphics_edu_pub_ray | | | | | | | | | | | | | | | |
| 101_pds_lc_nubus_slot | | | | 0.111 | 0.237 | 0.348 | 0.348 | 0.237 | 0.111 | | | | | | |
| 112_connector_keyboard_board_chip | | | | | | | | | | | | | | | |

Figure: Gráfico para visualização dos tokens nos tópicos.[45]

# Visualizações dos Termos

Podemos observar os termos mais frequentes dentro dos tópicos aos quais eles pertencem.
Isso permite:

- Entender melhor o agrupamento dos documentos
- Identificar padrões importantes em cada tópico

Essa análise ajuda a aprofundar a compreensão dos tópicos gerados.

# Visualizações dos Termos


Figure: Gráfico dos termos mais frequentes nos tópicos.[46]

# Visualizações das Hierarquia

Podemos observar a hierarquia entre os tópicos:

- Tópicos que derivam de outros
- Tópicos mais similares que estão juntos na hierarquia
- A hierarquia pode ser visualizada por texto ou gráficos

Além disso, podemos ver o agrupamento de tópicos gerado pelo algoritmo.

# Visualizações dos Termos

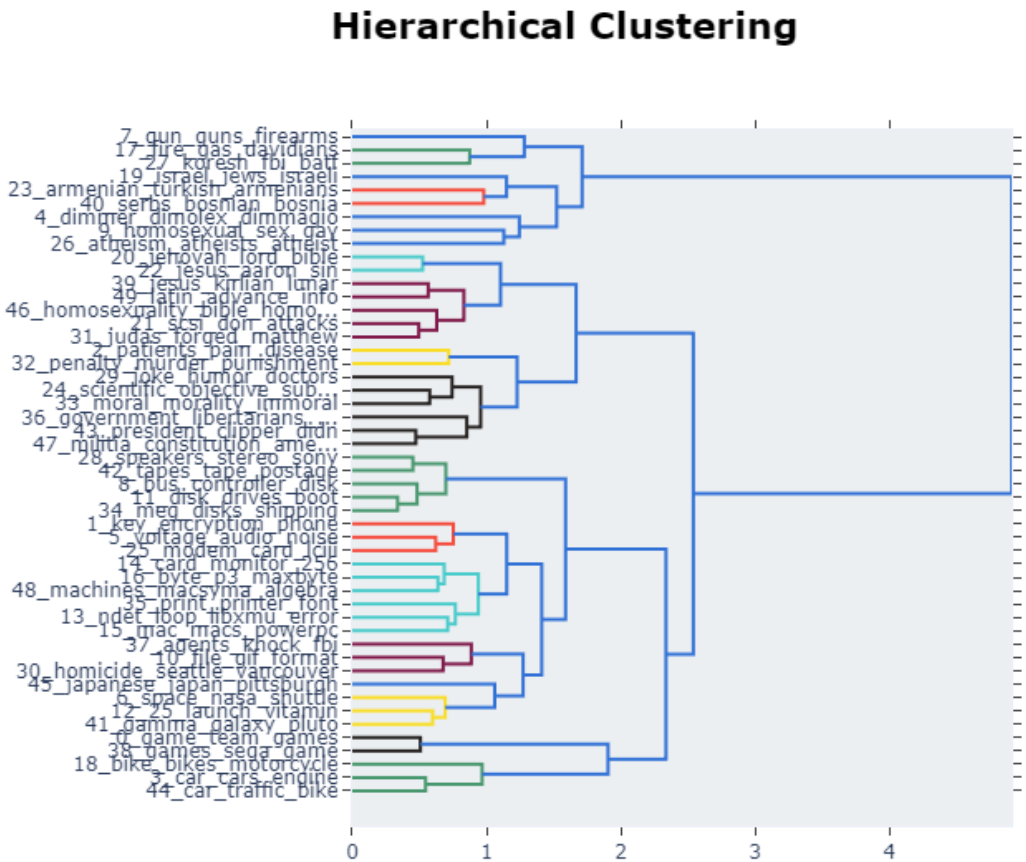



Figure: Gráfico de hierarquia nos tópicos.[47]

# Visualizações dos Termos

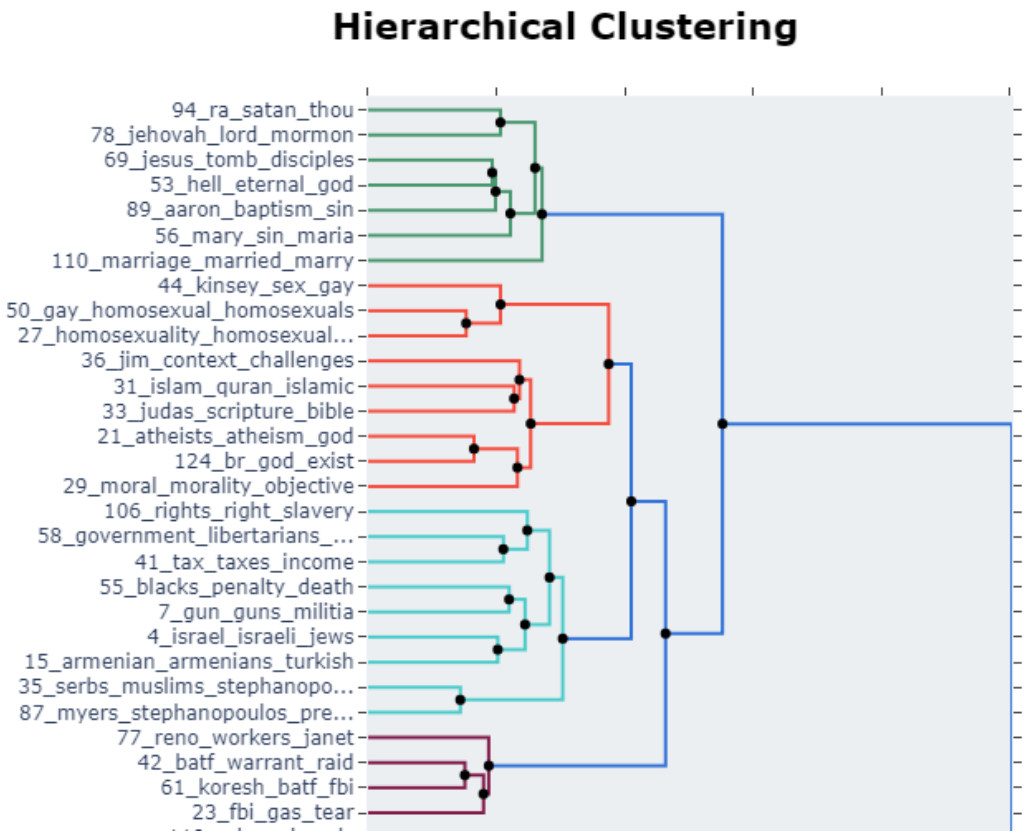



Figure: Gráfico de agrupamento por hierarquia nos tópicos.[48]

# Ajuste de Hiperparâmetros

Ao instanciar o BERTopic, há vários hiperparâmetros que você pode ajustar diretamente:

- language:
  - language = "english" (all-MiniLM-L6-v2)
  - language = "multilingual" (paraphrase-multilingual-MiniLM-L12-v2)
- top_n_words: número de palavras por tópico
- n_gram_range: intervalo de n-gramas
- min_topic_size: tamanho mínimo do tópico
- nr_topics: número de tópicos
- low_memory: se 'True', permite que o UMAP funcione em máquinas com pouca memória.
- calculate_probabilities: calcular as probabilidades de cada tópico em cada documento (computacionalmente caro)

# Ajuste de Hiperparâmetros

Ao instanciar o BERTopic, há vários hiperparâmetros que você pode ajustar diretamente:

- UMAP:
  - n_neighbors: número de amostras vizinhas usadas na aproximação do manifold
  - n_components: dimensionalidade das embeddings após a redução
  - metric: método para calcular distâncias no espaço de alta dimensionalidade (default: cosine)
- HDBSCAN:
  - min_cluster_size: tamanho mínimo de um cluster
  - min_samples: definido automaticamente como o valor de min_cluster_size e controla a quantidade de outliers gerados.
  - prediction_data: necessário para fazer previsões mais tarde.

# Integração com Outras Técnicas

Aplicar análise de sentimento em tópicos para obter insights mais profundos.

- Exemplo: "Quais são os sentimentos mais comuns em cada tópico?"

# Nuvem de Palavras e Outros Métodos de Visualização

Mostrar como usar ferramentas como wordcloud para visualização complementar.

- Visualizações ajudam a entender melhor os tópicos e suas características.

# Dicas

- Remoção de Stop Words
- Diversificar a Representação dos Tópicos: usar a técnica de Maximal Marginal Relevance (MMR) para diversificar as palavras em cada tópico, limitando palavras duplicadas.
- Pré-computar Embeddings: economiza tempo.
- Aceleração com GPU
- Instalação Leve: utilizar técnicas de embeddings baseadas em CPU, como TfidfVectorizer

# Aplicação na vida real

- **Análise de Redes Sociais**:
  - Identificação de temas de discussão.
  - Monitoramento de tendências em tempo real.
- **Recomendações de Conteúdo**:
  - Modelagem de tópicos para recomendar conteúdos baseados em padrões de consumo.
  - Personalização de feeds de notícias e artigos com base em preferências de leitura.
- **Pesquisa Acadêmica e Científica**:
  - Agrupamento de tópicos em artigos de pesquisa para facilitar revisões bibliográficas.

# Coerência de Tópicos

A coerência (Coherence) de tópicos é uma métrica que avalia a qualidade dos tópicos gerados.

- Tópicos coerentes têm palavras que frequentemente aparecem juntas.
- Avaliar a interpretação humana dos tópicos pode ser uma abordagem complementar.

# Métricas de Avaliação de Modelagem de Tópicos

- Coherence, Silhouette Score, V-Score, entre outras.

Essas métricas ajudam a comparar diferentes modelos e ajustar parâmetros.

# Use Cases

**Employee Surveys**


*"We are using BERTopic to support analysis of employee surveys. Here, we use BERTopic to compute the topics of discussion found in employee responses to open-ended survey questions. To further understand how employees feel about certain topics, we combined BERTopic with sentiment analysis to identify the sentiments associated with different topics and vice versa."*

•••

Steve Quirolgico, Ph.D.

**Principal Engineer**

U.S. Department of Homeland Security

# Exemplo

Neste trabalho, buscamos realizar o agrupamento de tópicos relacionados a diferentes produtos, utilizando as descrições contidas nas Notas Fiscais Eletrônicas (NF-e). Os dados utilizados para esta análise foram disponibilizados pela SEFAZ-RS.

- TCC

# Corpus

O corpora utilizado é composto por descrições contidas de NF-e disponibilizados pela SEFAZ-RS. Ele é dividido em:

- Corpus refinado por uma cadeia de markov (Markov [1906]) para compreender observações de leite e carne. É composto por aproximadamente 384 mil observações e passou, além da cadeia de Markov, pela remoção de acentuação e conversão de letras maiúsculas para minúsculas;
- Corpus sem refinamento composto por aproximadamente 203 mil mercadorias de NF-e da SEFAZ-RS. Ele recebeu o pré-processamento semelhante ao outro banco.

# Métrica

A métrica utilizada foi a de silhueta (Rousseeuw [1987]). Ela consiste em um valor que identifica o quão cada *embedding* é similar ao seu *cluster* comparado à outros *clusters*. Quanto maior o valor da silhueta, melhor o algoritmo realizou as clusterizações. Sua fórmula é descrita a partir da Equação (1):

$$\frac{r - s}{\max(s, r)}, \tag{1}$$

No qual **s** é a média da distância intra-*cluster* e **r** é a média do *cluster* mais próximo. O resultado final é a média de todos os pontos calculados.

# Resultados Corpus refinado "carne e leite"

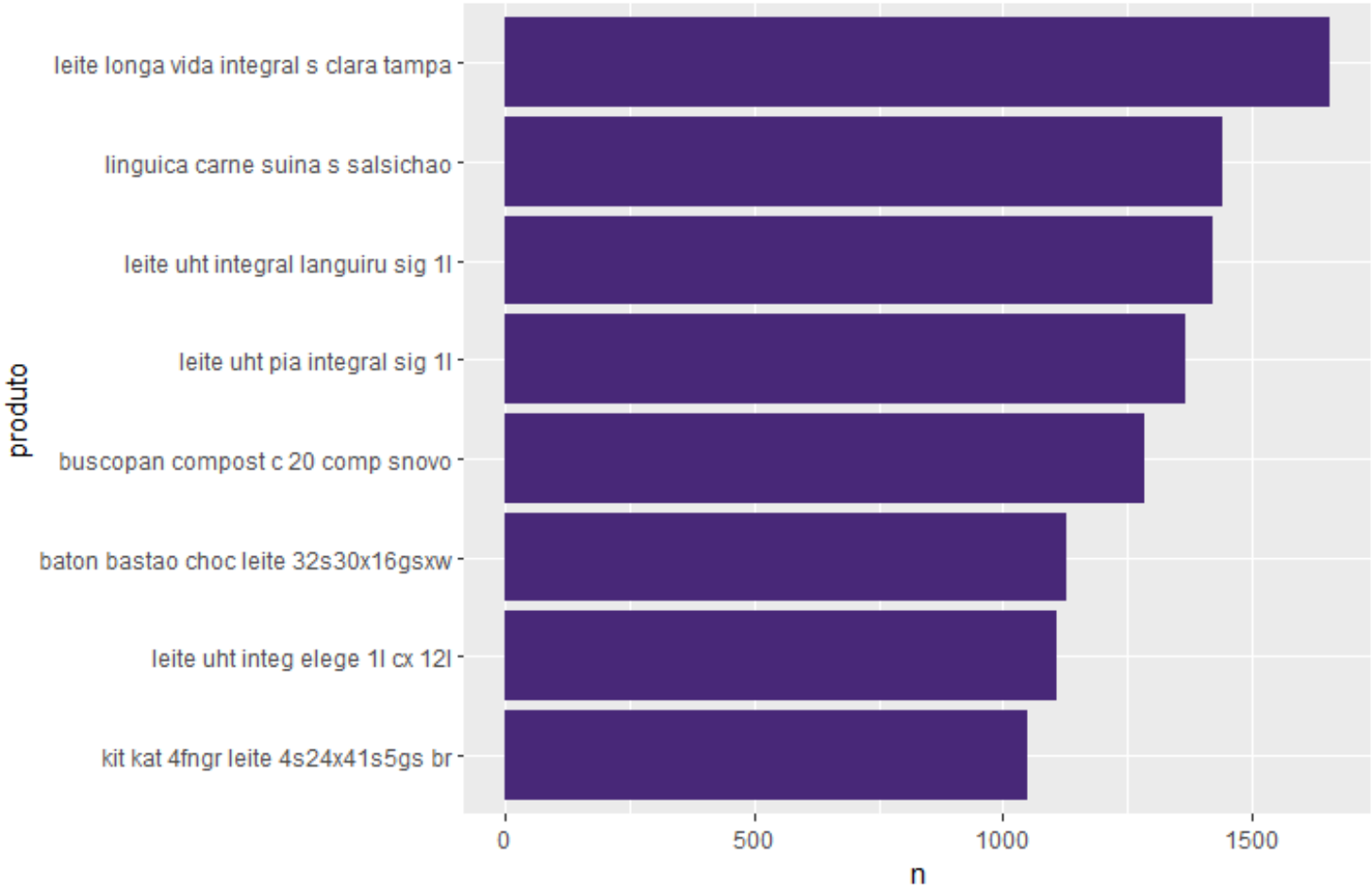
| Modelo | Modelo de Embedding | Palavras Chave | Tamanho mínimo | Número de tópicos | Modelo de Representação | Tempo | Silhueta | Banco utilizado (%) |
|---|---|---|---|---|---|---|---|---|
| 3 | "all" | 5 | 500 | X (77) | X | 1d 4h 36 Min | 0,70 | 100 |
| 12 | "p-m" | 5 | 500 | 15 (13) | KeyBERT Inspired | 8 Min | 0,67 | 20 |
| 13 | "p-m" | 5 | 500 | 10 (10) | Mistral Zephyr 7B | 8h 37 Min | 0,01 | 100 |
| 15 | "p-m" | 10 | X | X (5470) | BART | 1h 17 Min | 0,93 | 50 |
| **16** | **"p-m-MiniLM-L12-v2"** | **10** | **1000** | **X (7)** | **BART** | **9 Min** | **0,67** | **20** |

# Melhor configuração: Características básicas

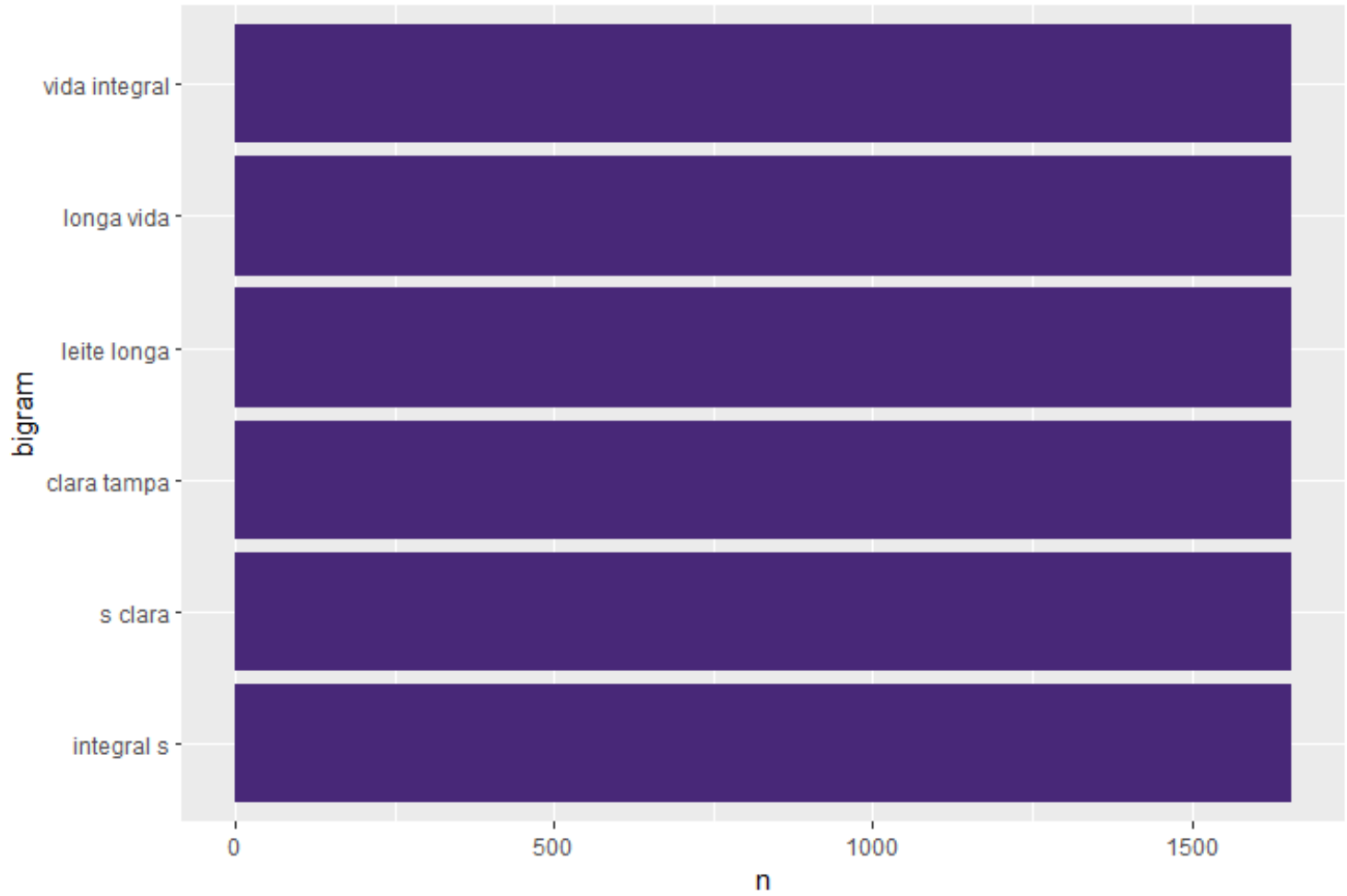
```
-1              -1_kg_leite_frango_carne
0                           0_leite___
1         1_compost_buscopan_snovo_comp
2                           2_leite___
3                           3_leite___
4                          4_frango___
5                           5_leite___
6                           6_leite___
7                           7_carne___
8                           8_leite___
9                           9_leite___
10                         10_leite___
11          11_pao_pullman_integral_500g
12    12_salsichao_linguica_suina_carne
13                         13_carne___
14                         14_leite___
15                         15_leite___
16                         16_carne___
17        17_osso_3303speito_suino_bovi
18                  18_kg_po_sabao_tixan
19                      19_tipo_po_sx_ss
Name: Name, dtype: object
```

```
-1     317843
0        1706
1        1284
2        1127
3        1492
4        3999
5        1247
6        1680
7        1933
8        1072
9        2248
10       1169
11       2104
12       1443
13       1007
14       1197
15       3909
16       2578
17       1151
18       2333
19      31732
Name: Count, dtype: int64
```

# Melhor configuração: Palavras mais frequentes



# Melhor configuração: Bigrams mais frequentes



# Melhor configuração - “Leite”

```
Topic                                                              0
Count                                                           1706
Name                                                       0_leite___
Representation                                  [leite, , , , , , , , , , ]
Representative_Docs    [leite uht pia integral sig 1l, leite uht pia ...
Name: 1, dtype: object
Topico 0 - 0_leite___ - Top 5 de cada topico:
['leite', '', '', '', '', '', '', '', '', '']
Topic                                                              3
Count                                                           1492
Name                                                       3_leite___
Representation                                  [leite, , , , , , , , , , ]
Representative_Docs    [leite uht integral languiru sig 1l, leite uht...
Name: 4, dtype: object
Topico 3 - 3_leite___ - Top 5 de cada topico:
['leite', '', '', '', '', '', '', '', '', '']
 Topic                                                             6
 Count                                                          1680
 Name                                                       6_leite___
 Representation                                 [leite, , , , , , , , , , ]
 Representative_Docs    [leite longa vida integral s clara tampa, leit...
 Name: 7, dtype: object
 Topico 6 - 6_leite___ - Top 5 de cada topico:
 ['leite', '', '', '', '', '', '', '', '', '']
 Topic                                                             10
 Count                                                          1169
 Name                                                      10_leite___
 Representation                                 [leite, , , , , , , , , , ]
 Representative_Docs    [leite uht integ elege 1l cx 12l, leite uht in...
 Name: 11, dtype: object
 Topico 10 - 10_leite___ - Top 5 de cada topico:
 ['leite', '', '', '', '', '', '', '', '', '']
```

# Melhor configuração - “Leite”

```
Topic                                                                  2
Count                                                               1127
Name                                                          2_leite___
Representation                                  [leite, , , , , , , , , ]
Representative_Docs    [baton bastao choc leite 32s30x16gsxw, baton b...
Name: 3, dtype: object
Topico 2 - 2_leite___ - Top 5 de cada topico:
['leite', '', '', '', '', '', '', '', '', '']
```

```
Topic                                                                  5
Count                                                               1247
Name                                                          5_leite___
Representation                                  [leite, , , , , , , , , ]
Representative_Docs    [leite magnesia phil 350ml, leite magnesia phi...
Name: 6, dtype: object
Topico 5 - 5_leite___ - Top 5 de cada topico:
['leite', '', '', '', '', '', '', '', '', '']
```

```
Topic                                                                  8
Count                                                               1072
Name                                                          8_leite___
Representation                                  [leite, , , , , , , , , ]
Representative_Docs    [kit kat 4fngr leite 4s24x41s5gs br, kit kat 4...
Name: 9, dtype: object
Topico 8 - 8_leite___ - Top 5 de cada topico:
['leite', '', '', '', '', '', '', '', '', '']
```

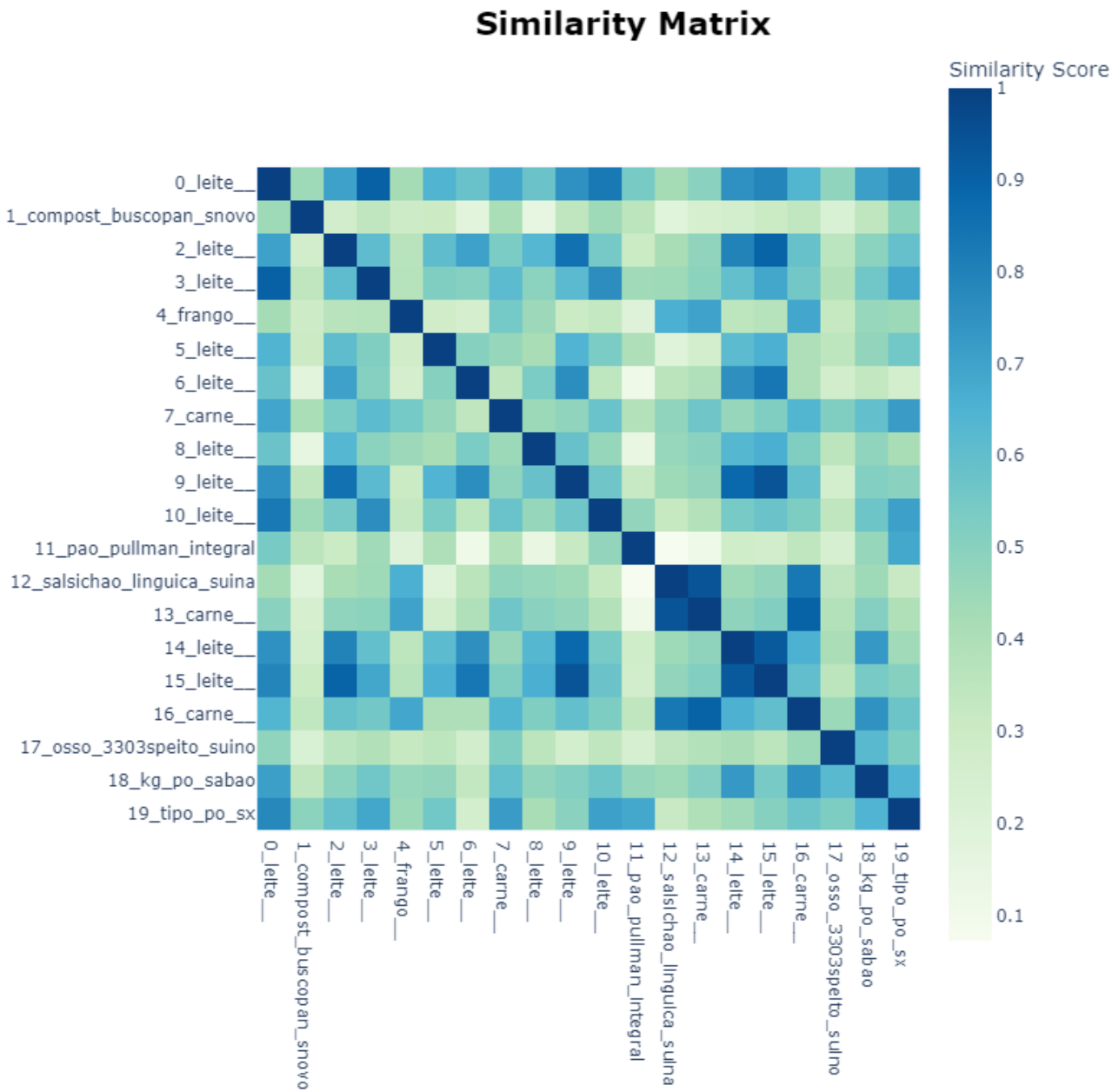
```
Topic                                                                  9
Count                                                               2248
Name                                                          9_leite___
Representation                                  [leite, , , , , , , , , ]
Representative_Docs    [danoninho leite fermentado 450g 1x450gr, dano...
Name: 10, dtype: object
Topico 9 - 9_leite___ - Top 5 de cada topico:
['leite', '', '', '', '', '', '', '', '', '']
```

# Melhor configuração - “Leite”

```
Topic                                                                14
Count                                                              1197
Name                                                      14_leite___
Representation                                  [leite, , , , , , , , , , ]
Representative_Docs    [leite po 1 kg, leite po 1 kg, leite po 1 kg]
Name: 15, dtype: object
Topico 14 - 14_leite___ - Top 5 de cada topico:
['leite', '', '', '', '', '', '', '', '', '']


Topic                                                                15
Count                                                              3909
Name                                                      15_leite___
Representation                                  [leite, , , , , , , , , , ]
Representative_Docs    [cuca doce leite, cuca doce leite, cuca doce l...
Name: 16, dtype: object
Topico 15 - 15_leite___ - Top 5 de cada topico:
['leite', '', '', '', '', '', '', '', '', '']
```

# Melhor configuração - Matriz de similaridade



# Conclusão

- Obter agora características de tópico como o preço para observar, por exemplo, fraudes.
- Comparar preços praticados no RS com àqueles do Sistema de Registro de Preços (SRP).

# Referências I


R. E. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, 1957.

D. M. Blei and J. D. Lafferty. Correlated topic models. *Advances in Neural Information Processing Systems*, 20:147–154, 2007.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

R. J. Campello, D. Moulavi, and J. Sander. Density-based clustering by means of core samples. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 160–168. SIAM, 2013.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231. AAAI Press, 1996.

M. Khadivi, S. Akbarpour, M. R. Feizi Derakhshi, and B. Anari. Persian topic detection based on human word association and graph embedding, 02 2023.

A. Markov. Rasprostranenie zakona bol'shih chisel na velichiny, zavisyaschie drug ot druga. *Izvestiya Fiziko-matematicheskogo obschestva pri Kazanskom universitete*, 15:135–156, 1906.

# Referências II


L. McInnes and J. Healy. Umap: Uniform manifold approximation and projection for dimension reduction. 02 2018.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.

P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. ISSN 0377-0427. doi: https://doi.org/10.1016/0377-0427(87)90125-7. URL `https://www.sciencedirect.com/science/article/pii/0377042787901257`.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Obrigado!

**Antônio Oss Boll:**

- Linkedin

**Letícia Maria Puttlitz:**

- Linkedin

- GitHub