Universidade de São Paulo
Instituto de Matemática e Estatística
Bacharelado em Ciência da Computação

Análise de Textos Históricos utilizando LLMs e Modelagem de Tópicos

João Pedro Lukasavicus Silva

Monografia Final

mac 499 — Trabalho de Formatura Supervisionado

Supervisor: Mateus Espadoto

São Paulo

Agradecimentos

Tenho inúmeras pessoas a agradecer, sem as quais este trabalho provavelmente não seria realizado. Tantas que essa lista definitivamente não é exaustiva.

Primeiramente agradeço a Jesus Cristo, por sua infinita misericórdia e graça, e por me dar tudo o que eu tenho. Espero ser sempre grato pela vida que me foi dada.

Agadeço aos meus pais, Marcelo e Adriana, por todo o carinho e cuidado, todo o suporte durante todos esses anos, por fazerem o impossível para que eu tivesse uma boa vida e uma boa educação, e por me darem o que nunca receberam. Espero que seus esforços sejam produtivos, e que eu possa retribuir seu amor.

Agradeço ao meu irmão, Lucas, que em muitos momentos foi um segundo pai para mim, cuidou de mim e me fez ser quem sou. Se todos tivessem um irmão mais velho como você, o mundo seria um lugar mais justo e compassivo. Espero formar em mim um caráter como o seu.

Agradeço à minha irmã, Bia, pela riqueza de sua compania, por toda a alegria que você me proporciona e pelo seu amor. Obrigado por estar ao meu lado, e por fazer do mundo um lugar mais legal de viver. Sei que o mundo aguarda ansiosamente para ver o seu brilho, e espero poder te acompanhar nas suas jornadas.

Agradeço ao meu amor, Ana Paula, por todo o seu carinho, cuidado e dedicação. Por todo o tempo que passamos juntos, por acreditar em mim quando eu mesmo não acreditava, por me dar ânimo quando eu queria desistir, por me trazer vida e ser um refúgio em meio aos mares turbulentos da vida. Espero dividir minha vida com você, para sempre.

Agradeço à toda minha família. Vocês são o meu bem mais precioso.

Agradeço à minha terapeuta, Nathália, por me fazer enxergar um caminho, e por me ensinar a ser mais gentil comigo mesmo.

Agradeço aos meus vizinhos, Ricardo e Andrea, pelas caronas. Este trabalho não seria possível sem vocês.

Agradeço ao professor Marcel Kenji, pela orientação e pelo encorajamento quando eu estava totalmente perdido e sem esperanças.

Agradeço aos meus colegas, Carlos Marques e Carol Paixão, pelas breves conversas no IME. Vocês tornaram meus dias muito mais suportáveis.

Agradeço aos antigos companheiros de basquete do IME, Arthur Saba e Luis Hikaru, por me fazerem me sentir bem vindo e querido, apesar da minha teimosia e das minhas neuroses. Prometo treinar mais, Hikaru, e espero te encontrar novamente algum dia.

Agradeço aos meus amigos, Davi, Giovanna, Baggio, e Rodrigo, pela inestimável amizade de vocês, todas as conversas e todos os momentos que dividimos.

Agradeço ao professor Hitoshi, pela sua ajuda e paciência nos últimos anos da minha graduação.

Por fim, agradeço imensamente ao meu orientador, Mateus, por toda sua ajuda durante este trabalho, pelo seu empenho em me ajudar a me formar, pela sua enorme paciência comigo e por todas as conversas e conselhos. Muito obrigado.

Resumo

João Pedro Lukasavicus Silva. **Análise de Textos Históricos utilizando LLMs e Modelagem de Tópicos**. Monografia (Bacharelado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo.

Este trabalho investiga a aplicação de Grandes Modelos de Linguagem (LLMs) e técnicas de modelagem de tópicos na análise de textos históricos, com foco nas Etimologias de Isidoro de Sevilha. Enquanto a análise tradicional de tais *corpora* extensos é manual e custosa, este estudo propõe um *pipeline* computacional para detectar e organizar automaticamente estruturas temáticas latentes. Foram realizados experimentos utilizando *embeddings* de sentenças para avaliar a qualidade dos tópicos gerados. Os resultados indicam que uma configuração que prioriza a minimização do tamanho máximo dos clusters, em detrimento da minimização de outliers, produz grupos semanticamente mais coerentes. Além disso, a análise ao nível de sentenças mostrou-se superior no isolamento de assuntos distintos, como gramática e retórica, revelando também temas transversais dispersos pelos livros. Conclui-se que estes métodos computacionais fornecem uma ferramenta robusta de "leitura distante", capaz de replicar a categorização de especialistas e descobrir conexões semânticas na literatura medieval.

Palavras-chave: Redes neurais. LLM. Grandes modelos de linguagem. Modelagem de tópicos. Textos históricos. Isidoro de Sevilha. Etimologias.

Abstract

João Pedro Lukasavicus Silva. **Analysis of Historical Texts using LLMs and Topic Modeling**. Capstone Project Report (Bachelor). Institute of Mathematics and Statistics, University of São Paulo, São Paulo.

This work investigates the application of Large Language Models (LLMs) and topic modeling techniques to the analysis of historical texts, focusing on the Etymologies of Isidore of Seville. While traditional analysis of such extensive corpora is manual and labor-intensive, this study proposes a computational pipeline to automatically detect and organize latent thematic structures. Experiments were conducted using sentence embeddings to assess the quality of the generated topics. The results indicate that a configuration which prioritizes the minimization of maximum cluster size, rather than the minimization of outliers, yields more semantically coherent groups. Furthermore, sentence-level analysis proved superior in isolating distinct subjects, such as grammar and rhetoric, while also revealing transversal themes dispersed throughout the books. We conclude that these computational methods provide a robust "distant reading" tool, capable of replicating expert categorization and uncovering semantic connections in medieval literature.

Keywords: Neural Networks. LLM. Large Language Models. Topic Modeling. Historical Texts. Isidore of Seville. Etymologies.

Lista de abreviaturas

BERT	Bidirectional Encoder Representations from Transformers
CBOW	Continuous Bag-of-Words
cTF-IDF	Class-based Term Frequency - Inverse Document Frequency
HTML	Hyper-text Markup Language
IME	Instituto de Matemática e Estatística
LDA	Latent Dirichlet Allocation
LLM	Large Language Model
LSA	Latent Semantic Analysis (Análise Semântica Latente)
LSTM	Long Short-Term Memory
MLP	Multi-layer Perceptron
NMF	Non-negative Matrix Factorization
PCA	Principal Component Analysis
PDF	Portable Document Format
PLN	Processamento de Linguagem Natural
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network (Rede Neural Recorrente)
SBERT	Sentence-BERT
SVD	Singular Value Decomposition (Decomposição em Valores Singulares)
TF-IDF	Term Frequency - Inverse Document Frequency
t-SNE	t-Distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection
URL	Localizador Uniforme de Recursos (*Uniform Resource Locator*)
USP	Universidade de São Paulo

Lista de figuras

Lista de tabelas

Sumário

Capítulo 1

Introdução

Tradicionalmente, pesquisadores das áreas de humanidades como história, filosofia, entre outras, dependem da análise de grandes quantidades de fontes para a realização do seu trabalho. Esta análise tipicamente é realizada de forma manual, e com grande custo, tanto em termos de tempo quanto de recursos humanos, custo este que pode se tornar um fator limitador da produção científica desses pesquisadores.

A área de processamento de linguagem natural (PLN) possui métodos bem estabelecidos para análise semântica de textos, como por exemplo, modelagem de tópicos (Deerwester *et al.*, 1990; David M Blei *et al.*, 2003; Jelodar *et al.*, 2019), que visa agrupar textos que tratam de assuntos semelhantes. Com o surgimento dos grandes modelos de linguagem (LLM) baseados em transformers (Vaswani *et al.*, 2017; Devlin *et al.*, 2019), é possível observar um grande salto em termos de qualidade das ferramentas e métodos para análise de texto, o que pode ser atribuído à maior capacidade de mapeamento de conceitos em um espaço latente de atributos, que, por sua vez, possibilita um melhor agrupamento de textos em termos de semântica. Em todo caso, o uso destas técnicas requer certa familiaridade com o seu funcionamento e com os textos analisados, de modo que sejam feitos os ajustes necessários para a obtenção de bons resultados.

1.1 Objetivos

Neste trabalho pretendemos, por meio do uso de técnicas computacionais de processamento de texto, estudar os tópicos existentes na obra conhecida como *Etimologias*, de Isidoro de Sevilha (c.560-636). Esta obra, que é considerada como a primeira grande enciclopédia da Idade Média, é formada por 20 livros que tratam das origens das palavras agrupadas em diversos grandes temas. O autor, que viveu em uma época de grandes mudanças culturais, buscava registrar o conhecimento de escritores latinos da Antiguidade Clássica, como Varrão, Catão e Plínio o Velho, entre outros. A obra, que foi escrita em latim medieval, foi copiada exaustivamente ao longo de cerca de 700 anos, sendo utilizada como uma espécie de livro-texto básico nas instituições de ensino da época. Por conveniência, para este trabalho será utilizada uma tradução recente para a língua inglesa (Barney *et al.*, 2006).

1.2 Contribuições

A principal contribuição pretendida com este trabalho é conseguir mapear temas transversais discutidos na obra estudada, para além dos temas gerais propostos pelo autor. Adicionalmente, propomos um experimento de análise de similaridade semântica entre a obra estudada e possíveis fontes da antiguidade, para identificar potenciais citações sem atribuição.

1.3 Organização

O texto está organizado da seguinte forma: no Capítulo 2 tratamos do background histórico dos autores estudados, bem como das ferramentas utilizadas; no Capítulo 3 apresentamos os dados analisados, o *pipeline* de processamento criado para os dados, e a configuração dos experimentos realizados; no Capítulo 4 apresentamos e discutimos os resultados obtidos para cada experimento realizado. O Capítulo 5 conclui o texto.

Capítulo 2

Background

Neste trabalho, utilizaremos ferramentas de modelagem de tópicos e grandes modelos de linguagem com o objetivo de demonstrar sua utilidade no estudo de textos históricos. Nas seções a seguir detalhamos as ferramentas utilizadas e as obras a serem estudadas.

2.1 Redes Neurais e Grandes Modelos de Linguagem

Desde o surgimento dos primeiros computadores digitais no período logo após a Segunda Guerra Mundial (1939 - 1945), pesquisadores trabalharam para desenvolver ferramentas que pudessem executar tarefas que até então eram realizadas apenas por humanos. Cientistas como Warren McCulloch (1898 - 1969) e Walter Pitts (1923 - 1969), um médico e um lógico, estão entre os primeiros a discutirem modelos do cérebro humano, e como estes poderiam ser implementados computacionalmente. Em 1943 publicaram um artigo (McCulloch e Pitts, 1943) considerado seminal na área, aonde propuseram um modelo matemático do cérebro representado como uma rede de elementos simples interconectados. Um elemento particular recebe sinais dos elementos conectados à sua entrada, e produzem uma soma ponderada como sinal de saída, que, dependendo de um limiar pré-determinado, é enviado para elementos conectados à sua saída como um sinal ativo (valor 1) ou inativo (valor 0). Os autores demonstram no artigo que é possível calcular funções lógicas utilizando diferentes configurações de conexões entre elementos. Posteriormente, estes elementos passaram a ser conhecidos como neurônios artificiais, que serviram de base para o desenvolvimento das redes neurais artificiais, anos mais tarde.

Alan Turing (1912 - 1954), matemático e teórico da computação que ajudou a estabelecer as bases da ciência da computação como disciplina, foi um dos primeiros a se dedicar ao que veio a ser chamado posteriormente de inteligência artificial, de forma mais abrangente. Turing apresenta (Turing, 1950) um conjunto de questões e definições que seriam essenciais para o desenvolvimento da área. Em primeiro lugar, ele pergunta se "máquinas podem pensar", o que traz o primeiro problema importante, o de falta de definições claras sobre o que é "pensar" e, naquele momento, sobre o que seria a "máquina" em questão. Como definições do que é pensar, e de forma mais abrangente, do que é inteligência, representam problemas filosóficos importantes e sem resposta objetiva, Turing propôs

uma abordagem prática para isso: no lugar de "máquinas podem pensar", ele passa a perguntar se "máquinas podem agir de forma indistinguível de um humano". Em outras palavras, ele ignora o processo interno pelo qual uma ação é produzida, e se preocupa apenas com o efeito da ação por aqueles que a percebem. Para demonstrar esta ideia, Turing propôs um experimento chamado de Jogo da Imitação. Na versão mais simples do jogo, há três participantes: um humano e um computador que devem responder a perguntas de um juiz humano que só pode fazer perguntas e receber respostas escritas em papel, sem contato direto com os outros dois. Se o juiz, após uma série de perguntas, não for capaz de distinguir o humano da máquina, a máquina venceu o jogo, ou seja, conseguiu se passar por um humano. Apesar de ser uma definição prática, simples, e que serve aos propósitos do autor, esta ideia de que basta ser percebido como humano para ser considerado inteligente ou pensante é problemática para pensadores de outras áreas, que tratam de aspectos metafísicos da mente e do pensamento.

Em 1980, o filósofo John Searle (1932 -) propôs o argumento do Quarto Chinês (Searle, 1980). Neste experimento, o autor imagina uma pessoa que não fala chinês isolada em um quarto com um livro contendo instruções sobre como interpretar símbolos chineses. Se alguém colocar um texto em chinês por debaixo da porta, por exemplo, a pessoa poderia seguir as instruções do livro para produzir símbolos que representem uma resposta coerente para falantes de chinês. No entanto, esta pessoa estaria apenas seguindo regras sintáticas sem nenhuma compreensão semântica do texto que está produzindo. O autor busca com isso se contrapor a ideias funcionalistas, como a de Turing, de que a mente é apenas um sistema de processamento de informações. Ou seja, para Searle, não basta ser capaz de se comportar como se entendesse uma conversa para uma entidade ser considerada como pensante ou inteligente.

Apesar das críticas, a busca pela inteligência artificial seguiu com maior ou menor intensidade ao longo da segunda metade do século XX. Por exemplo, partindo da ideia de neurônios artificiais de McCulloch e Pitts em 1943, Frank Rosenblatt (1928 - 1971), psicólogo, desenvolveu o Perceptron (Rosenblatt, 1958) em 1957, que é considerada a primeira rede neural artificial. A ideia do Perceptron foi desenvolvida ao longo dos anos, passando por diversos altos e baixos, e com a evolução do hardware existente, cada vez mais capaz, culminou na criação do que passou a ser chamado de Deep Learning (Krizhevsky *et al.*, 2012), por volta de 2012 com o trabalho de Alex Krizhevsky (1986 -), que consiste no uso de grandes redes neurais com milhões e até bilhões de elementos aplicado a problemas de processamento de texto e imagem. Mais recentemente, por volta de 2017, a partir do trabalho de Ashish Vaswani (1986 -) e outros foram criadas novas arquiteturas de conexão entre os neurônios artificiais, batizadas de Transformers (Vaswani *et al.*, 2017), com o objetivo de permitir que mais dados pudessem ser armazenados e processados em conjunto, desta forma possibilitando fornecer mais informações de contexto para um problema computacional de processamento de texto ou imagem. A ideia dos transformers é que possibilitou o surgimento de modelos mais recentes como ChatGPT e Gemini, que são exemplares do que é considerado inteligência artificial hoje em dia.

Modelos baseados em transformers possuem grande capacidade de mapeamento semântico de dados e de geração de texto verossímil, que podem ser consideradas como suas características mais salientes. Estas capacidades são obtidas com base na observação e processamento de grande quantidade de dados de exemplo, processo que é chamado

de “treinamento” no jargão da área, realizados com o objetivo de se identificar padrões existentes nos dados para poder determinar, por exemplo, quais palavras que aparecem comumente próximas a outras em certos contextos. Concluída esta etapa de treinamento, o modelo é capaz de produzir texto em resposta a uma questão: primeiro é feito o mapeamento semântico da questão para encontrar termos e sentenças com significado parecido na memória do modelo, e segundo, com base nos termos e sentenças encontrados, o modelo faz a síntese de um texto que tenha verossimilhança. Note o uso do termo verossimilhança no lugar de corretude: como os modelos são criados de acordo com a definição de Turing, basta que forneçam respostas indistinguíveis das de um humano, e cabe a quem coloca a questão avaliar a corretude das respostas fornecidas.

2.2 Evolução das redes neurais em PLN

Hoje, modelos de redes neurais artificiais são prevalentes em diversas áreas de aplicação de machine learning, como classificação e geração de imagens, Processamento de Linguagem Natural (PLN), reconhecimento de fala, etc. A seguir, apresentaremos uma breve (e incompleta) história da evolução desses modelos em algumas sub-áreas de PLN, desde desenvolvimentos anteriores à adoção de redes neurais, até o advento dos Transformers.

Harris (1954) e Firth (1957) argumentam que o sentido de uma palavra pode ser deduzido, em parte, a partir dos contextos onde ela é comumente utilizada - "conhecerás uma palavra pela compania que ela mantém". Essa ideia, que podemos chamar de Hipótese Distribucional, direta ou indiretamente, guiou inúmeros avanços em diferentes campos na área de PLN. Deerwester *et al.* (1990), com seu modelo de Análise Semântica Latente (LSA), foi um dos pioneiros em operacionalizar essa ideia, obtendo resultados promissores na área de indexação e recuperação de informação. Neste modelo, aplica-se uma técnica de álgebra linear, denominada Decomposição em Valores Singulares (SVD), a uma matriz de frequência (ou alguma outra métrica, como TF-IDF) termo-documento, para se obter uma representação mais densa e de menor dimensionalidade desta matriz. O resultado é uma representação vetorial que consegue capturar relações semânticas entre termos e documentos, mapeando itens com significados similares a lugares comuns em um espaço latente, configurando uma forma primitiva do conceito que hoje conhecemos como *word embeddings*.

Bengio *et al.* (2003) propôs o uso de redes neurais arificiais para modelar distribuições de probabilidade conjunta de sequências de palavras em um determinado idioma. Na arquitetura proposta, uma rede neural é treinada para executar duas tarefas simultaneamente: Associar cada palavra de um vocabulário a um vetor de $\mathbb{R}^m$ (*embedding*), e determinar a distribuição de probabilidade condicional de sequências de palavras do vocabulário (isto é, dada uma sequência de palavras, determinar a distribuição de probabilidade da próxima palavra), expressas em termos de suas representações vetoriais. Embora poderosa, essa abordagem ainda tinha uma grande complexidade computacional, inviabilizando o seu uso em grandes conjuntos de dados. Um grande avanço na área, que possibilitou a criação de modelos de embedding com um número bem maior de dimensões, e o uso de volumes muito maiores de dados no treinamento desses modelos, foi a chegada dos modelos *Continuous Skip-gram* e *Continuous Bag-of-Words* (CBOW) (Mikolov, Chen *et al.*, 2013). Estas arquiteturas dispensaram o uso de camadas ocultas, e tinham como objetivo

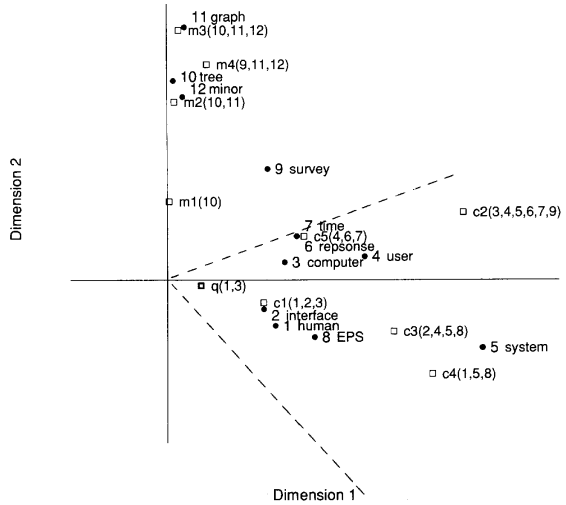


Figura 2.1: *Representação do espaço latente gerado pelo LSA.*

simplesmente prever uma palavra dadas as palavras ao seu redor (arquitetura CBOW), ou, dada uma palavra, prever as palavras ao seu redor (arquitetura Skip-gram), usando uma janela deslizante de contexto, cujo tamanho é dado como parâmetro da arquitetura. Embora, nessa arquitetura, a ordem das palavras na janela de contexto deixasse de ter importância (no caso do CBOW, usa-se a média dos vetores dessas palavras, por exemplo), os embeddings gerados conseguiam capturar bem melhor o sentido global das palavras, principalmente devido à baixa complexidade computacional do modelo, em relação ao proposto por Bengio, o que possibilitou o uso de um volume muito grande de dados em seu treinamento, e a criação de embeddings de dimensionalidades muito maiores (até 1000 dimensões, comparados com as 50-100 do modelo proposto por Bengio). Porém, vale dizer que este modelo, diferentemente do anterior, não é um modelo de linguagem probabilístico, no sentido de que o seu objetivo não é de modelar distribuições de probabilidades de sequências de palavras, mas de aprender representações semânticas eficientes. Posteriormente, usando técnicas como subamostragem e amostragem negativa, Mikolov mostrou como acelerar o processo de aprendizado e inferência dos modelos, e criar melhores representações (Mikolov, Sutskever *et al.*, 2013).

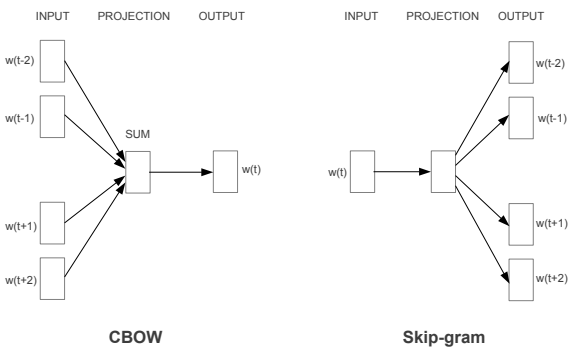


Figura 2.2: *Arquiteturas CBOW e Skip-gram.*

Na área de tradução automática de texto, um trabalho pioneiro propôs o uso de uma arquitetura do tipo *encoder-decoder* para traduzir sequências de palavras do Inglês para o Francês (Sutskever *et al.*, 2014). O modelo era composto por duas redes do tipo *Long Short-Term Memory*, ou LSTM, que é um tipo de rede neural recorrente (RNN) dotada de mecanismos de "controle de memória", para lidar com problemas de dependência de longo alcance em sequências. O modelo funcionava em duas etapas: primeiro, o *encoder*, consistindo de uma RNN mapeava sequências de palavras de tamanho variado para um vetor de dimensão fixa, denominado *estado oculto* (também chamado de *vetor de contexto*) iterativamente. Depois, o *decoder* usa o último estado oculto do *encoder* para gerar a sequência de saída. Assim, a predição de cada palavra na saída da rede é influenciada tanto por todas as palavras na entrada, quanto pelas previsões anteriores.

Uma evolução significativa dessa última abordagem foi introduzida no trabalho de Bahdanau *et al.* (2016), onde foi introduzida uma versão precursora dos mecanismos de atenção modernos. No modelo anterior, sequências de palavras eram mapeadas para um vetor de dimensão fixa, o que fazia com que a performance do modelo deteriorasse rapidamente para sequências longas (Cho *et al.*, 2014). O novo modelo proposto buscava livrar o *encoder* de ter que comprimir todo o contexto em um único vetor de tamanho fixo, gerando uma sequência de estados ocultos que o *decoder* poderia acessar. Além disso, o modelo dispunha de um mecanismo para "procurar", entre os estados ocultos, informações relevantes para a predição de uma determinada palavra dentro da sequência, e *selecionar* quais palavras deveriam influenciar a palavra atual. O resultado foi uma melhora significativa na transdução de sequências longas.

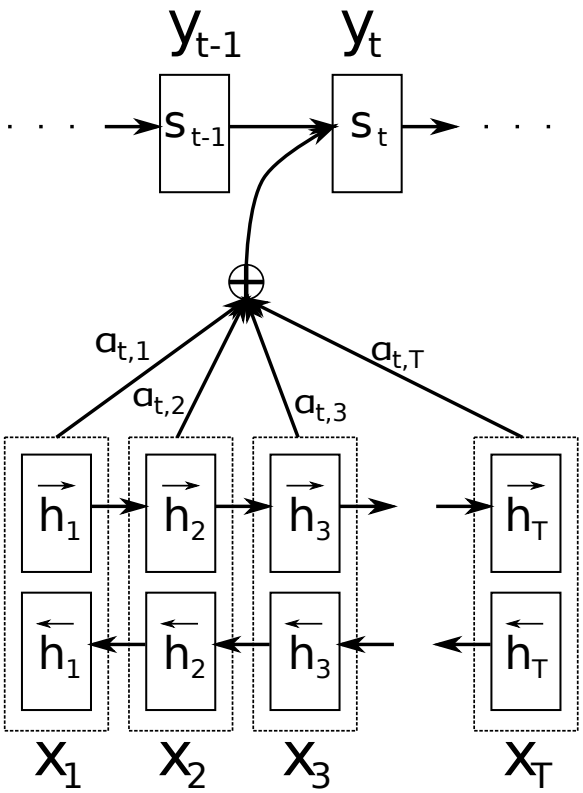


Figura 2.3: *RNNSearch.*

2.3 Transformers e desenvolvimentos subsequentes

Todos estes desenvolvimentos culminaram na criação da arquitetura batizada de Transformer (Vaswani *et al.*, 2017). Todos os modelos mencionados anteriormente dependem do processamento sequencial de tokens da entrada, devido ao emprego de redes neurais recorrentes. Na arquitetura Transformer, os tokens de entrada são processados em paralelo, o que permitiu uma enorme escalabilidade e aceleração de seus processos de treinamento e inferência. A seguir mostramos uma descrição superficial da sua arquitetura.

2.3.1 Arquitetura do modelo Transformer

Assim como os modelos de transdução já citados (Bengio *et al.*, 2003; Sutskever *et al.*, 2014; Bahdanau *et al.*, 2016), um Transformer conta com um *encoder* e um *decoder*, porém, com a diferença de não haver RNNs envolvidas.

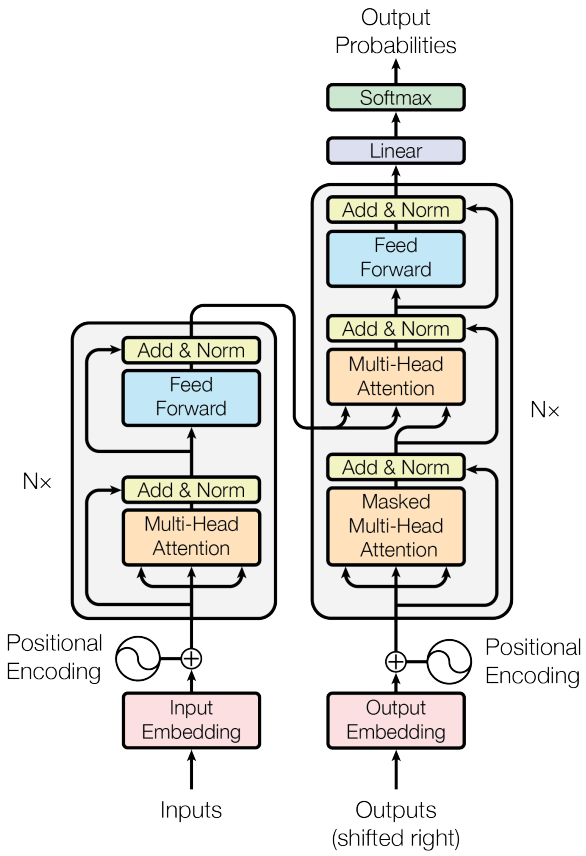


Figura 2.4: *Arquitetura de um Transformer.*

Encoder

O encoder começa com uma camada de *embedding*, juntamente com uma camada de codificação posicional, que serve para imbuir os embeddings de informação sobre suas posições, já que não estamos mais utilizando RNNs. Em cima disto, múltiplas camadas idênticas são empilhadas. Cada uma destas camadas é composta por duas sub-camadas. A

primeira implementa um mecanismo de atenção paralela (*multi-head attention*), e a segunda é uma simples rede do tipo *multi-layer perceptron* (também chamada de rede *feed-forward*).

Decoder

O decoder também é composto por várias camadas idênticas empilhadas, onde cada camada é composta por três subcamadas, duas iguais às subcamadas do encoder, e uma extra, *masked self-attention*, que consiste em uma camada de atenção modificada para evitar que uma palavra gerada pelo modelo "preste atenção" em uma palavra subsequente (isto é, seja modificada por ela).

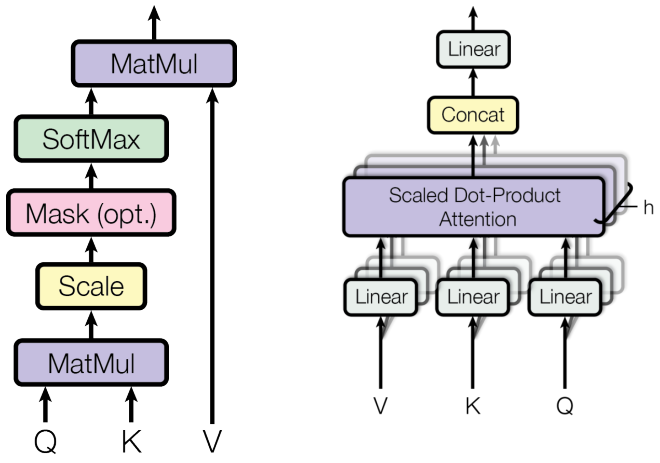
2.3.2 Atenção

Podemos entender os mecanismos de atenção presentes na arquitetura Transformer como uma maneira de palavras "informarem" umas às outras quais palavras elas podem influenciar, e de que forma. Mais especificamente, sobre cada embedding (que representa um token, na entrada ou na saída), um bloco de atenção calcula três valores, Q, K e V (*query*, *key* e *value*) usando transformações lineares simples. Para cada embedding, o seu valor de Q é comparado com o valor de K de si mesmo e todos os outros (ou dos embeddings anteriores, no caso de *masked self-attention*), usando uma função de similaridade (no caso, o produto escalar), seguida de uma normalização (*softmax*). O resultado então é usado para calcular a influência dos embeddings uns sobre os outros (efetivamente, soma-se o valor da atenção para um embedding ao valor do embedding, em conexões residuais):

$$\text{Atencao}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

onde d_k é a dimensionalidade de Q e K.

Após uma camada de atenção há uma rede *feed-forward*, com funções de ativação do tipo ReLU (*Rectified Linear Unit*).



(a) *Atenção de produto escalar.*

(b) *Multi-head attention.*

Figura 2.5: *Mecanismos de atenção nos Transformers.*

Multi-head attention

Ao invés de calcular uma única função de atenção, os valores Q, K e V são projetados linearmente h vezes em espaços de menor dimensionalidade, e vários valores de atenção são calculados em paralelo, e depois esses valores são concatenados e projetados novamente no espaço original, resultando nos valores finais de atenção. Isso permite ao modelo considerar influências entre palavras em diferentes espaços latentes, em diferentes posições.

2.3.3 BERT, SBERT e derivados

A chegada dos Transformers representou um novo paradigma para o uso de redes neurais em PLN, e em outras áreas de aplicação. Pela sua capacidade de modelagem de dependências de longo alcance, e pela sua escalabilidade, modelos baseados em Transformers logo dominaram o cenário de machine learning, impulsionados principalmente por modelos generativos como ChatGPT (Radford e Narasimhan, 2018), Gemini (Gemini Team, 2025), entre outros. Estes modelos apresentam algumas variações arquitetônicas (como uso apenas de decoders em modelos generativos) mas ainda mantêm a estrutura básica de um Transformer, com seus mecanismos de atenção. Um exemplo de modelo que conta apenas com encoders é o BERT (Devlin *et al.*, 2019). Este modelo é pré-treinado em diferentes tarefas, como predição de tokens ocultos, e predição de próxima sentença, e depois passa por um ajuste fino para tarefas específicas, como detecção de paráfrases, inferência textual, classificação de texto, etc. O seu pré-treinamento faz com que o modelo obtenha um "entendimento" básico da linguagem, que melhora consideravelmente o seu desempenho em tarefas específicas posteriores.

Apesar de alcançar bons resultados, o modelo BERT se torna computacionalmente inviável para algumas tarefas, como busca por similaridade semântica em um grande corpus, ou clustering de documentos. Para essas tarefas, é desejável um modelo capaz de gerar embeddings de frases, o que o BERT não foi projetado para fazer. Para mitigar essa limitação surgiu o SBERT (Reimers e Gurevych, 2019), que é uma modificação dos modelos BERT pré-treinados, usando redes neurais siamesas e estruturas de rede do tipo triplet para gerar boas representações semânticas de frases. Essas representações então podem ser comparadas usando distância de cosseno, por exemplo. Os embeddings usados neste trabalho são gerados por descendentes desse modelo.

2.4 Modelagem de Tópicos

Modelos de tópicos são ferramentas computacionais para descobrir automaticamente estruturas temáticas comuns em grandes quantidades de documentos, permitindo que sejam organizados e sumarizados de acordo com os assuntos presentes (David M. Blei, 2012). Modelos de tópicos são muito úteis na prospecção de documentos de interesse, em sistemas de recomendação, e em análises de séries de documentos. Existem diferentes abordagens para essa tarefa, como *Latent Dirichlet Allocation* (LDA), *Non-negative Matrix Factorization* (NMF) e modelos baseados em redes neurais. No LDA (David M Blei *et al.*, 2003), tratamos um tópico como uma distribuição de probabilidade sobre um vocabulário, e assumimos que cada documento contém uma mistura de tópicos, diferentes documentos contendo diferentes proporções dos tópicos. O modelo então supõe que os documentos

são gerados de acordo com essas distribuições, que são tratadas como variáveis ocultas, e aproxima a distribuição posterior de probabilidade, dada a distribuição observada de palavras por documento.

Modelos neurais de tópicos geralmente envolvem *embeddings* de palavras ou frases gerados por redes neurais, como o BERTopic, que utilizamos neste trabalho.

2.4.1 BERTopic

No BERTopic (Grootendorst, 2022), usamos agrupamentos de *embeddings* de trechos de documentos para representar tópicos subjacentes no texto.

Ele é constituído, basicamente, de 4 etapas:

1. Extração de embeddings
2. Redução de dimensionalidade
3. Clustering
4. Extração de representações dos tópicos

Primeiramente, extrai-se representações vetoriais das sentenças do corpus, utilizando algum modelo de *sentence embedding*. Utilizamos então um clustering desses embeddings como nosso modelo de tópicos. Entretanto, a alta dimensionalidade dos embeddings gerados acaba por prejudicar a performance de algoritmos e técnicas baseados em conceitos como proximidade, distância, ou vizinhos mais próximos, como diversos algoritmos de clustering (Radovanović *et al.*, 2010; Aggarwal *et al.*, 2001; Steinbach *et al.*, 2003). Técnicas de redução de dimensionalidade se demonstraram eficazes para mitigar esse efeito, e melhorar a qualidade de clusterings (Herrmann *et al.*, 2022; Allaoui *et al.*, 2020).

Por fim, depois de obter aglomerados de sentenças (ou trechos arbitrários de documentos), correspondendo aos tópicos detectados, precisamos extrair representações interpretáveis destes tópicos. Alguns modelos de tópicos baseados em embeddings utilizam termos próximos do centróide de um cluster como representação dos tópicos (Sia *et al.*, 2020; Angelov, 2020). No BERTopic, utilizamos os termos com maior poder discriminante entre os documentos de um cluster, calculando uma variante da métrica TF-IDF (*Term Frequency - Inverse Document Frequency*), denominada cTF-IDF (o 'c' vem de "classe"):

$$\text{cTF–IDF}_{t,c} = f_{t,c} \cdot \log\left(1 + \frac{A}{f_t}\right)$$

Onde $f_{t,c}$ é a frequência do termo t na classe (ou tópico, cluster) c, f_t é a frequência geral do termo t no corpus, e A é o número médio de termos por classe. As palavras com pontuações altas em um determinado tópico são justamente aquelas que, simultaneamente, aparecem muito nesse tópico, e pouco em outros tópicos.

2.4.2 UMAP

Como já vimos, aplicar técnicas de redução de dimensionalidade antes de executar tarefas de clustering tende a render melhores resultados. Existem diversas técnicas e

algoritmos para esse fim, com diferentes vantagens e aplicações, como PCA - *Principal Component Analysis* (Jolliffe e Cadima, 2016), t-SNE - *t-Distributed Stochastic Neighbor Embedding* (Maaten *et al.*, 2008), e UMAP - *Uniform Manifold Approximation and Projection* (McInnes *et al.*, 2018). Neste trabalho (assim como o autor da técnica de modelagem de tópicos usada), optamos por usar o UMAP, pela sua habilidade em preservar estruturas não-lineares nos dados, e manter um equilíbrio entre estruturas locais e globais, além de ter menores tempos de execução em relação a outras técnicas não-lineares.

A ideia principal do UMAP é: dada uma representação em grafo do dataset original, aprender uma representação, em um espaço de dimensionalidade reduzida, que seja o mais similar possível, em um certo sentido, ao grafo construído no espaço original. Simplificadamente, o algoritmo faz isso da seguinte maneira: primeiro, para cada ponto, ele calcula quais são seus k vizinhos mais próximos, onde o valor de k é um parâmetro do algoritmo. Então, para cada par de vizinhos, ele calcula uma certa "probabilidade assimétrica" de que esses pontos estejam conectados localmente, de acordo com a densidade local na região de cada ponto, isto é, a probabilidade de um ponto x_i estar conectado a outro ponto x_j, do ponto de vista de x_i, pode ser diferente da probabilidade de conexão de x_j e x_i, do ponto de vista de x_j. O grafo ponderado original é então construído, utilizando a probabilidade de ao menos uma conexão existir entre dois pontos, como os pesos para as arestas. O algoritmo então cria uma representação inicial dos pontos do dataset, porém em um espaço de menor dimensionalidade, e ajusta iterativamente as posições desses pontos para minimizar a diferença entre o grafo correspondente de menor dimensionalidade e o grafo original. Como toda técnica de redução de dimensionalidade, é claro que a projeção de menor dimensionalidade não é uma representação fiel do conjunto de dados. Em especial, informações sobre distâncias entre pontos considerados não-conectados são perdidas. Porém, os k pontos mais próximos de qualquer ponto no espaço original provavelmente também estarão perto na projeção de dimensionalidade reduzida, assim como pontos considerados não-conectados provavelmente estarão distantes. Assim, essa técnica serve como uma boa etapa de pré-processamento para tarefas subsequentes de clustering.

2.4.3 HDBSCAN

HDBSCAN - *Hierarchical Density-Based Spatial Clustering of Applications with Noise* é um algoritmo de clustering hierárquico baseado em densidade, capaz de extrair clusters de densidades e formas variadas, não necessariamente convexas, e ainda classificar pontos como ruído, não pertencendo a nenhuma classe (Campello *et al.*, 2013). A seguir damos uma simples descrição de seu funcionamento.

Seja $\mathbf{X} = \{x_1, \ldots, x_n\}$ um conjunto de dados, e $d(\cdot, \cdot)$ uma medida de distância. Dado um valor k, definimos uma medida de distância central $\text{core}_k(x)$, igual à maior distância entre x e um de seus k vizinhos mais próximos, para estimar a densidade ao redor de um ponto x. Definimos então outra métrica, distância de alcance mútuo, como:

$$d_{\text{mreach}-k}(x_i, x_j) = \max\{\text{core}_k(x_i), \text{core}_k(x_j), d(x_i, x_j)\}$$

Então, construímos um grafo completo, onde cada vértice corresponde a um ponto no dataset, e o peso p de uma aresta é dado por $p(x_i, x_j) = d_{\text{mreach}-k}(x_i, x_j)$. Durante a execução do algoritmo, usamos esse grafo para representar clusters: ao retirar um conjunto de arestas,

os nossos clusters são representados pelos vértices de componentes conexas do grafo.

Depois, obtemos a árvore geradora mínima deste grafo. Podemos fazer isso eficientemente utilizando o algoritmo de Prim (ou outros algoritmos, dependendo do espaço em que estamos trabalhando). Ao ordenar as arestas dessa matriz em ordem decrescente e ir retirando as arestas uma a uma, o efeito que obtemos é como um clustering divisivo: a cada aresta removida, dividimos uma componente conexa em duas. Assim podemos obter uma sequência de partições do nosso conjunto de dados, ou uma hierarquia de clusters, onde os clusters vão sendo divididos em clusters menores, usando ligação única (*single linkage*) como distância entre clusters. Podemos representar essa hierarquia usando uma árvore binária, que efetivamente é o dendrograma da clusterização. Um detalhe importante dessa construção, é que ao construir essa árvore, definimos "divisões verdadeiras" de clusters quando os clusters resultantes são maiores do que um tamanho mínimo, dado como parâmetro. Nesse caso, ambos os clusters são classificados como clusters reais. Quando um cluster resultante de uma divisão é menor que o tamanho mínimo, paramos o processo de divisão desse cluster, e classificamos os seus pontos como ruído.

Para extrair um conjunto de clusters dessa hierarquia, é definida uma métrica de estabilidade para cada cluster, que mede, informalmente, o quanto um cluster resiste ao processo de divisão. Selecionamos então os clusters cuja soma das estabilidades seja máxima, sujeito à restrição de que não podemos selecionar um cluster e algum de seus descendentes na hierarquia.

2.5 Obras e autores estudados

Neste trabalho utilizaremos edições em inglês das seguintes obras: as Etimologias (*Etymologiae*), de Isidoro de Sevilha, Sobre Agricultura (*De Agri Cultura*), de Catão, e Sobre Ciência Agrícola (*De Re Rustica*), de Varrão. A escolha destas obras não é por acaso: estudiosos da obra de Isidoro consideram que ele se baseou em diversas obras da antiguidade para escrever as Etimologias, particularmente as obras de Catão e Varrão sobre agricultura. Sendo assim, tentaremos encontrar similaridades de ideas entre os textos a partir do uso de modelos computacionais. A seguir apresentaremos detalhes sobre os autores e seu contexto.

Isidoro de Sevilha (c. 560 - 636) foi um clérigo, teólogo e pensador da alta idade média, que é considerado um dos intelectuais mais importantes do seu tempo, e cuja influência foi sentida por muitos séculos após a sua morte. Nasceu em Cartagena, que à época era parte do Reino Visigótico, estado que ocupou as regiões da Península Ibérica e atual sul da França no período seguinte ao fim do Império Romano do ocidente, e fazia parte de uma família que percentia à elite hispano-romana. Seus três irmãos ocuparam funções importantes na igreja, com destaque para seu irmão mais velho, Leandro de Sevilha (c. 534 - c. 600) foi Bispo de Sevilha, cargo que Isidoro ocuparia após a morte de Leandro. Todos os quatro irmãos são venerados como santos pela Igreja Católica.

Como Bispo de Sevilha, Isidoro exerceu grande influência no seu tempo, presidindo sínodos e concílios importantes, como os de Sevilha e Toledo, protegendo os monastérios, e ainda se envolveu na conversão dos reis Visigodos do Arianismo, uma doutrina cristã

não-trinitária, para o cristianismo Calcedoniano, que veio a se tornar a doutrina dominante na Igreja Católica.

Como intelectual, produziu diversas obras, dentre as quais se destacam as Etimologias, que são um conjunto de livros que formam uma enciclopédia etimológica, que resume e organiza o conhecimento de diversos autores da antiguidade clássica. A obra segue a tradição de enciclopedistas clássicos, como Plínio, o Velho (c. 23 - 79), com o uso de ordenação alfabética de tópicos e de uma abordagem literária para o conhecimento, baseada no pensamento analógico.

As Etimologias tratam de temas diversos como gramática, retórica, matemática, direito, a Igreja, heresias, guerra, agricultura, entre outros. A sua influência foi tão grande nos séculos seguintes que algumas das obras clássicas utilizadas como base deixaram de ser lidas e copiadas e acabaram se perdendo no tempo. Era considerado o texto base para a educação sobre o período clássico durante a maior parte da idade média.

O estudo crítico das Etimologias revela suas possíveis fontes clássicas, que na maioria das vezes não são citadas por Isidoro. De acordo com Stephen A. Barney, tradutor para o inglês (Barney *et al.*, 2006) das Etimologias, é possível identificar que o material dos livros I e II, que tratam de gramática, retórica e dialética (as disciplinas do *trivium*) provavelmente foram extraídos dos Institutos, de Cassiodoro (c. 485 - c. 585), o livro III, sobre matemática (contendo as disciplinas do *quadrivium*), provavelmente foi inspirado em Boécio (c. 480 - 524), e o livro XVII, sobre agricultura, deriva de Catão, o Velho (234 - 149 a.C.) e Varrão (116 - 27 a.C.), para citar alguns exemplos.

Marco Pórcio Catão, o Velho (234 - 149 a.C.), foi um soldado, senador e historiador romano. Nascido em uma família de plebeus, que era a classe baixa de cidadãos livres em Roma, desecndia de gerações de soldados com reputação de bravura, como seu pai e seu bisavô. Ainda na infância, com a morte de seu pai, passou a cuidar das atividades da fazenda família. Como jovem soldado, especula-se que aos 20 anos tenha participado de campanhas das Guerras Púnicas no papel de tribuno militar, uma patente de oficial do exército romano. Ao retornar do campo de batalha para a sua fazenda, e com o apoio do seu vizinho e amigo Lúcio Valério Flaco, iniciou carreira política como *questor*, cargo que possuía diversas atribuições, entre elas a de cobrança de impostos e de supervisão financeira. Daí se seguiram diversos cargos políticos importantes, como *pretor* (magistrado), *consul* (o cargo mais alto da República Romana) e *censor* (magistrado de nível superior). Em paralelo a suas atividades políticas, escreveu diversas obras, cuja maioria infelizmente foi perdida. Escreveu uma história de Roma em sete livros chamada de Origens, uma obra sobre assuntos militares, e a obra sobre agricultura que utilizaremos neste trabalho, a única preservada na íntegra. Além disso, foi famoso orador e teve cerca de 150 discursos registrados.

Marco Terêncio Varrão (116 - 27 a.C.) foi um intelectual e polímata romano, descrito por Petrarca como a "terceira grande luz" de Roma, depois de Virgílio e Cícero. Nascido em família pertencente à classe equestre de Roma, abaixo somente da classe senatorial, ocupou cargos políticos ao longo da vida, como *questor*, *pretor* e tribuno do povo. Estudou com o filólogo romano Lúcio Élio Estilo e com o filósofo platonista Antíoco de Ascalão. Foi também um líder militar sem grande prestígio durante a Guerra Civil Cesariana. Foi um escritor prolífico, que produziu cerca de 74 obras sobre temas diversos, entre as quais se destacam a Cronologia Varroniana, que lista as datas de eventos importantes de Roma,

e os nove livros das Disciplinas, organizados de acordo com os temas das artes liberais da época, e que serviram de exemplo para enciclopedistas que vieram posteriormente, como Plínio, o Velho, e o próprio Isidoro de Sevilha.

Capítulo 3

Experimentos

Neste capítulo detalhamos os dados utilizados neste trabalho juntamente com o *pipeline* de processamento empregado para a sua coleta e preparação, e a configuração experimental utilizada no estudo.

3.1 Conjunto de dados

Como conjunto de dados a ser explorado neste trabalho, temos as seguintes obras:

- Etimologias (*Etymologiae*) (Barney *et al.*, 2006), por Isidoro de Sevilha (c. 560 - 636);
- Sobre Agricultura (*De Agri Cultura*) (Cato e Varro, 1934), por Marco Pórcio Catão, o Velho (234 - 149 a.C.);
- Sobre Ciência Agrícola (*De Re Rustica*) (Cato e Varro, 1934), por Marco Terêncio Varrão (116 - 27 a.C.).

No caso das Etimologias, foi utilizada a tradução (Barney *et al.*, 2006) do latim para o inglês em formato *pdf*. Para os textos de Catão e Varrão foram utilizadas edições da *Loeb Classical Library* (Cato e Varro, 1934), atualmente em domínio público e disponíveis *online* (Thayer, 2025) em formato *html*.

3.2 Preparação dos dados

3.2.1 Limpeza

No caso das Etimologias, o texto foi extraído do arquivo *pdf* do livro e convertido para formato de texto puro, removendo-se cabeçalhos, rodapés, números de página, e outros artefatos dessa conversão que não são necessários para a análise. Devido a peculiaridades do formato *pdf* e dificuldades encontradas nas bibliotecas utilizadas para a extração do texto, parte do processo de limpeza teve de ser realizado manualmente. No caso das outras obras, disponíveis em formato *html*, o processamento foi realizado de forma totalmente automática.

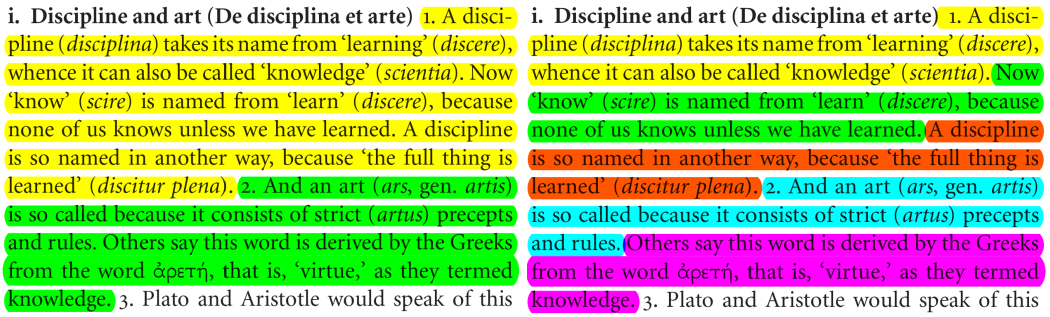
3.2.2 Pré-processamento

Primeiramente, pelo modo como cada livro está formatado, fizemos um processamento inicial para separar cada livro em capítulos, e juntar palavras com hífen. Em seguida, cada capítulo foi subdividido em sentenças usando o módulo *SentenceRecognizer* da biblioteca Spacy (Honnibal *et al.*, 2020), que possui diversos recursos para PLN.

Uma particularidade da edição escolhida das Etimologias é que cada capítulo é subdividido em seções relativamente curtas, a maioria tendo somente uma oração. Com isso, podemos fazer a modelagem dos tópicos usando dois níveis diferentes de granularidade, a depender de como definimos um "documento", unidade básica de nossa análise:

- **Seções**, já presentes no texto;
- **Sentenças**, delimitadas automaticamente;

Em nossa análise, utilizamos as duas abordagens e comparamos os resultados.



(a) *Seções.* **(b)** *Sentenças.*

Figura 3.1: *Diferentes formas de separar os textos.*

3.2.3 *Stop words, stemming* e lematização

Tarefas rotineiras de pré-processamento de linguagem natural incluem: remoção de *stop words*, que são palavras muito comuns e irrelevantes, *stemming*, e lematização, que são formas de padronizar palavras, reduzindo-as a sua forma mais básica. Apesar de outras técnicas de modelagem de tópicos adotarem essas tarefas como parte do seu processo de preparação dos dados, não utilizamos essas técnicas para este trabalho, pois os modelos de *embedding* que usamos utilizam informações contextuais de cada palavra em uma sentença, e remover palavras ou modificá-las poderia prejudicar a performance de tais modelos (Chaerul Haviana *et al.*, 2023).

3.2.4 Geração de *embeddings*

Depois da etapa de pré-processamento dos textos, geramos *embeddings* para cada documento, utilizando modelos do tipo *sentence embedders* pré-treinados. Os modelos usados para a geração desses *embeddings* foram: *sentence-transformers/LaBSE* (Feng *et al.*, 2022), *jinaai/jina-embeddings-v3* (Sturua *et al.*, 2024), *intfloat/multilingual-e5-large-instruct* (Wang *et al.*, 2024), *nomic-ai/nomic-embed-text-v2-moe* (Nussbaum e Duderstadt,

2025). Estes modelos foram desenhados e treinados para que sentenças semanticamente similares em línguas diferentes, ou traduções, estejam próximas umas das outras em um espaço latente. Podemos usar estes modelos para comparar textos em diferentes idiomas e analisar suas conexões.

3.2.5 Conjuntos de dados

Após essas etapas iniciais, definimos nossos conjuntos de dados. Cada conjunto é definido por um modelo de embedding e um nível de granularidade, já explicado.

3.3 Persistência

Os embeddings gerados, suas reduções, e diversos metadados foram armazenados em um banco de dados do *ChromaDB*, juntamente com cada documento.

Dessa forma, podemos fazer diversos tipos de busca, como busca por similaridade, busca textual, filtrar resultados baseados em metadados de cada documento, como autor, idioma, etc.

3.4 Experimentos

Como já foi explicado, a técnica de modelagem de tópicos usada (BERTopic) é composta de diferentes etapas, com várias possibilidades de escolha de algoritmos para cada etapa, e hiperparâmetros para estes algoritmos. Neste trabalho, foram usados o UMAP para redução de dimensionalidade, e o HDBSCAN para o clustering dos documentos e detecção dos tópicos.

Uma particularidade de interesse do HDBSCAN é que ele pode classificar pontos em um dataset como não pertencentes a nenhum cluster, como "ruído". Essa característica é útil para detecção de anomalias em um dataset e para dar uma maior confiabilidade nas classificações dos pontos que pertencem de fato a um cluster, mas pode também ser prejudicial, caso o número outliers detectados seja muito grande. Experimentos preliminares com valores padrão para os hiperparâmetros dos algoritmos mostraram uma forte tendência em classificar pontos como outliers. Com uma grande quantidade de escolhas de hiperparâmetros e modelos de embedding, realizamos uma espécie de busca de grade nesse espaço de parâmetros, para obter melhores resultados.

A seguir apresentamos uma explicação dos parâmetros testados para cada algoritmo.

3.4.1 Redução de dimensionalidade

Para a etapa de redução de dimensionalidade, utilizamos o UMAP. Este algoritmo conta com os seguintes parâmetros:

A implementação usada foi a da biblioteca umap-learn.[1]

[1] https://umap-learn.readthedocs.io/en/latest/

Parâmetro	Descrição	Valores testados	Padrão do BERTopic
`n_neighbors`	Controla o equilíbrio entre estrutura local e global.	de `5` a `60`, incrementos de `5`	`15`
`n_components`	Dimensionalidade resultante.	de `5` a `40`, incrementos de `5`	`5`
`min_dist`	Controla dispersão de pontos.	`0`	`0`
`low_memory`	Menor uso de memória.	`false`	`false`
`metric`	Métrica de distância.	`cosine`	`cosine`
`random_state`	Semente determinística.	`42`	–

Tabela 3.1: *Parâmetros usados para o UMAP*

3.4.2 Clustering

Com os embeddings gerados e suas reduções já computadas, inicializamos o BERTopic de modo a pular essas etapas iniciais e começar pela etapa de clustering. O algoritmo utilizado foi o HDBSCAN, disponível na biblioteca Scikit-learn,[2] e utilizamos os seguintes parâmetros:

Parâmetro	Descrição	Valores testados	Padrão do BERTopic
`min_cluster_size`	Controla tamanho mínimo e qtd. de *clusters*.	de `10` a `40`, incrementos de `5`	`10`
`min_samples`	Controla quantidade de pontos de ruído.	de `5` a `20`, com incrementos de `5`	`10`

Tabela 3.2: *Parâmetros usados para o HDBSCAN*

3.4.3 Representação

Como já explicado, dados clusters de documentos como tópicos, o BERTopic calcula o score cTF-IDF para cada palavra em um cluster, elege as palavras mais representativas de um tópico baseado nesses scores e cria uma representação sucinta usando essas palavras. Para melhorar a representação dos tópicos gerados, fizemos alguns ajustes nestas etapas finais do método:

Vetorização

Para calcular o cTF-IDF, primeiramente precisamos calcular a frequência de cada palavra, por documento e por tópico. Para isso usamos a classe `CountVectorizer`, do próprio BERTopic. Nessa etapa podemos fazer uso da remoção de stop words, pois os embeddings já foram gerados, para que eles não acabem entrando nas representações dos tópicos. Além disso, usamos também bigramas como termos candidatos para representação.

cTF-IDF

Ao calcular as pontuações cTF-IDF das palavras (e bigramas) no texto, utilizamos duas modificações da métrica original. Fazemos isso passando parâmetros específicos ao instanciar a classe `ClassTfidfTransformer`, do BERTopic: `ClassTfidfTransformer(True, True)`.

[2] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.HDBSCAN.html

Recobrando a fórmula original:

$$\text{cTF–IDF}_{t,c} = f_{t,c} \cdot \log\left(1 + \frac{A}{f_t}\right)$$

A versão modificada é dada então pela fórmula:

$$\text{cTF–IDF}_{t,c} = \sqrt{f_{t,c}} \cdot \log\left(1 + \frac{A - f_t + 0.5}{f_t + 0.5}\right)$$

3.4.4 Busca por similaridade semântica

Uma outra possível aplicação de embeddings semânticos é, dado um conjunto de documentos e um exemplo, encontrar os documentos no conjunto que tenham um conteúdo semântico mais próximo possível do exemplo, baseado nas distâncias entre o embedding do exemplo e cada documento do conjunto: o documento mais parecido com o exemplo será o documento cujo embedding estiver mais próximo da projeção do exemplo no espaço latente. Podemos usar isso para fazer cruzamentos entre diferentes autores, e descobrir citações diretas ou indiretas, por exemplo.

Neste trabalho, também fazemos um mini experimento de similaridade semântica: Buscamos, entre as sentenças presentes em Etimologias, as sentenças mais parecidas com dois trechos do livro Sobre Ciência Agrícola *(De Re Rustica)*, de Varrão (em inglês):

> **varrao_en.1.48.2.S2**: The beard is called arista from the fact that it is the first part to dry (arescere).
>
> **varrao_en.1.64.1.S0**: Amurca, which is a watery fluid, after it is pressed from the olives is stored along with the dregs in an earthenware vessel.

Os resultados dos experimentos são apresentados no próximo capítulo.

Capítulo 4

Resultados

A seguir relataremos os resultados dos experimentos realizados.

Ao executar o algoritmo do BERTopic com diferentes parâmetros, obtivemos diferentes resultados. Em geral, o número de pontos classificados como ruído, ou outliers, foi alto. Porém, ao inspecionar as configurações que minimizaram essa quantidade, descobrimos que algumas geravam agrupamentos muito concentrados em poucos clusters, misturando temas diferentes. A solução adotada foi escolher as configurações que geraram clusters mais homogêneos, minimizando o tamanho máximo de um cluster. Essa abordagem também acabou gerando um número reduzido de outliers, como pode-se ver nas figuras 4.1 e 4.2.

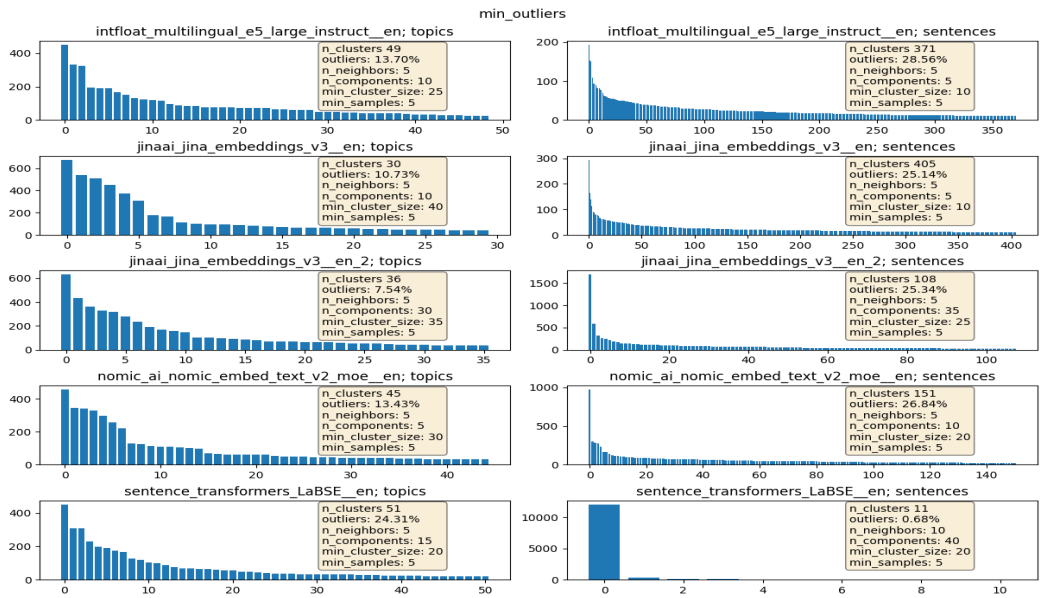


Figura 4.1: *Contagem de documentos por tópico - minimização de outliers.*

A segmentação do texto em seções gerou tópicos mais coerentes do que a segmentação em sentenças, na maioria dos casos. Porém, algumas seções de maior comprimento contribuíram para clusters impuros, com temas mistos. Em contrapartida, as configurações que segmentaram o texto em sentenças criaram mais clusters, com maior concentração das

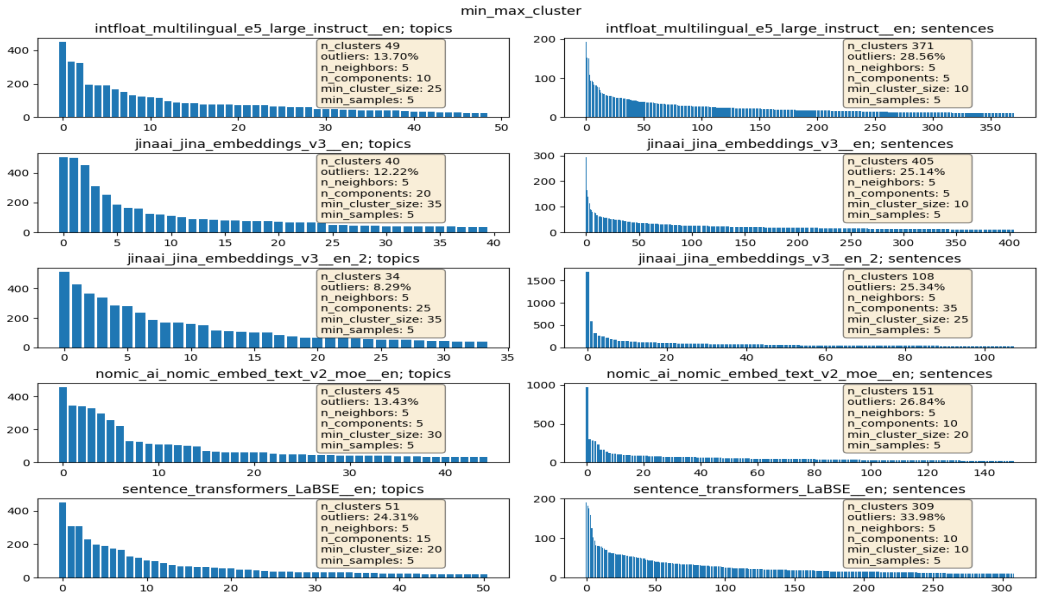


Figura 4.2: *Contagem de documentos por tópico - minimização de cluster máximo.*

palavras em clusters dominantes. A quantidade de ruído também foi significativa e consistentemente maior nas modelagens usando sentenças, o que pode-se atribuir a dificuldades na extração do texto dos PDFs e na deteção de sentenças, no pré-processamento. Porém, essa abordagem foi capaz de detectar uma quantidade maior de temas, além de temas mais distribuídos na obra. Isso nos mostra que ambas abordagens não são mutuamente exclusivas, mas complementares, e podem ser usadas em conjunto a depender do objetivo da análise.

No geral, o modelo de embedding que gerou melhores resultados foi o ***jinaai/jina-embeddings-v3*** (Sturua *et al.*, 2024), com um ajuste fino para tarefas de *clustering*.

A seguir apresentamos as principais diferenças resultantes da segmentação dos textos em seções e sentenças.

4.1 Segmentação por seções

Na figura 4.3, vemos uma visualização do espaço latente de *embeddings*, projetada para um plano 2D pelo UMAP, com cada ponto correspondendo a uma seção de um livro, e colorido de acordo com o livro ao qual a seção pertence. Os círculos em amarelo transparente são projeções de palavras nesse espaço. Podemos ver que as projeções tendem a agrupar palavras relacionadas, como *jewish, Abraham, hebrew, Muhammad, muslim, Christ, God, angel, saint*, relacionadas a religiões abraâmicas, *rye, sorghum, yeast, wheat, harvest, plough, farm, bread, milk*, relacionadas a agricultura, *dog, sheep, cattle*, relacionadas a animais, *earth, start, moon, sun, weather, climate*, relacionados a corpos celestes e fenômenos atmosféricos, e *war, fight, battle*, projetados muito próximos uns aos outros, pelo seu alto grau de similaridade semântica.

Vemos também grupos passagens de um mesmo livro, o que é esperado, mas também a presença de sub-agrupamentos de passagens dentro de um livro. Note, por exemplo,

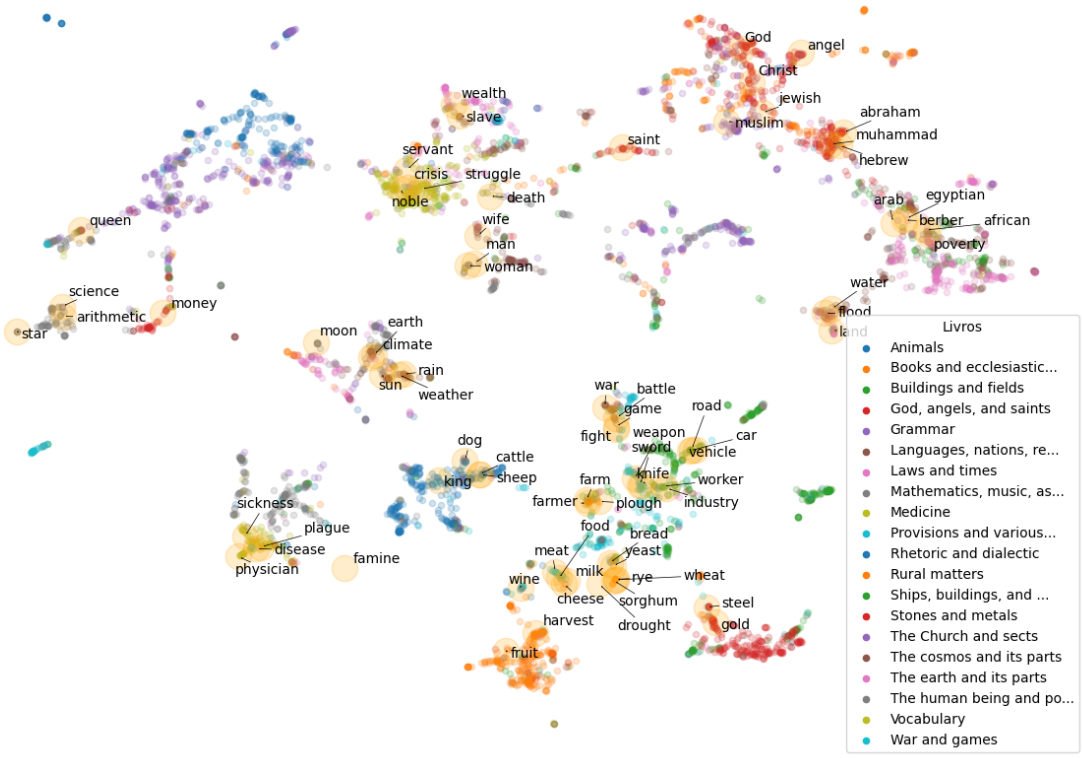


Figura 4.3: *Visualização 2D do espaço de embeddings - seções.*

a presença de dois aglomerados de passagens do livro *Stones and metals* - um no canto inferior direito, próximo de *steel* e *gold*, e um na parte central-esquerda da imagem, perto da palavra *money*. Note também a presença de um aglomerado de passagens do livro *Rural matters* ao redor da palavra *fruit*, e um de passagens do livro *Animals*, próximo às palavras *dog*, *sheep* e *cattle*.

Ao observar a distribuição dos tópicos pelos livros, na figura 4.4, podemos observar a presença tanto de temas concentrados, como o primeiro, presente majoritariamente nos livros *Grammar* e *Rhetoric and dialectic* (contendo as disciplinas do *Trivium*), o sexto, praticamente exclusivo ao livro *Stones and metals*, com passagens sobre materiais, metais, e cores, quanto temas mais transversais, distribuídos em diferentes livros, como o terceiro, distribuído principalmente entre os livros *Mathematics, music, astronomy*, *The cosmos and its parts*, *Laws and times* e *The Church and sects*, tratando de corpos celestes, estações do ano, a passagem do tempo e fenômenos atmosféricos.

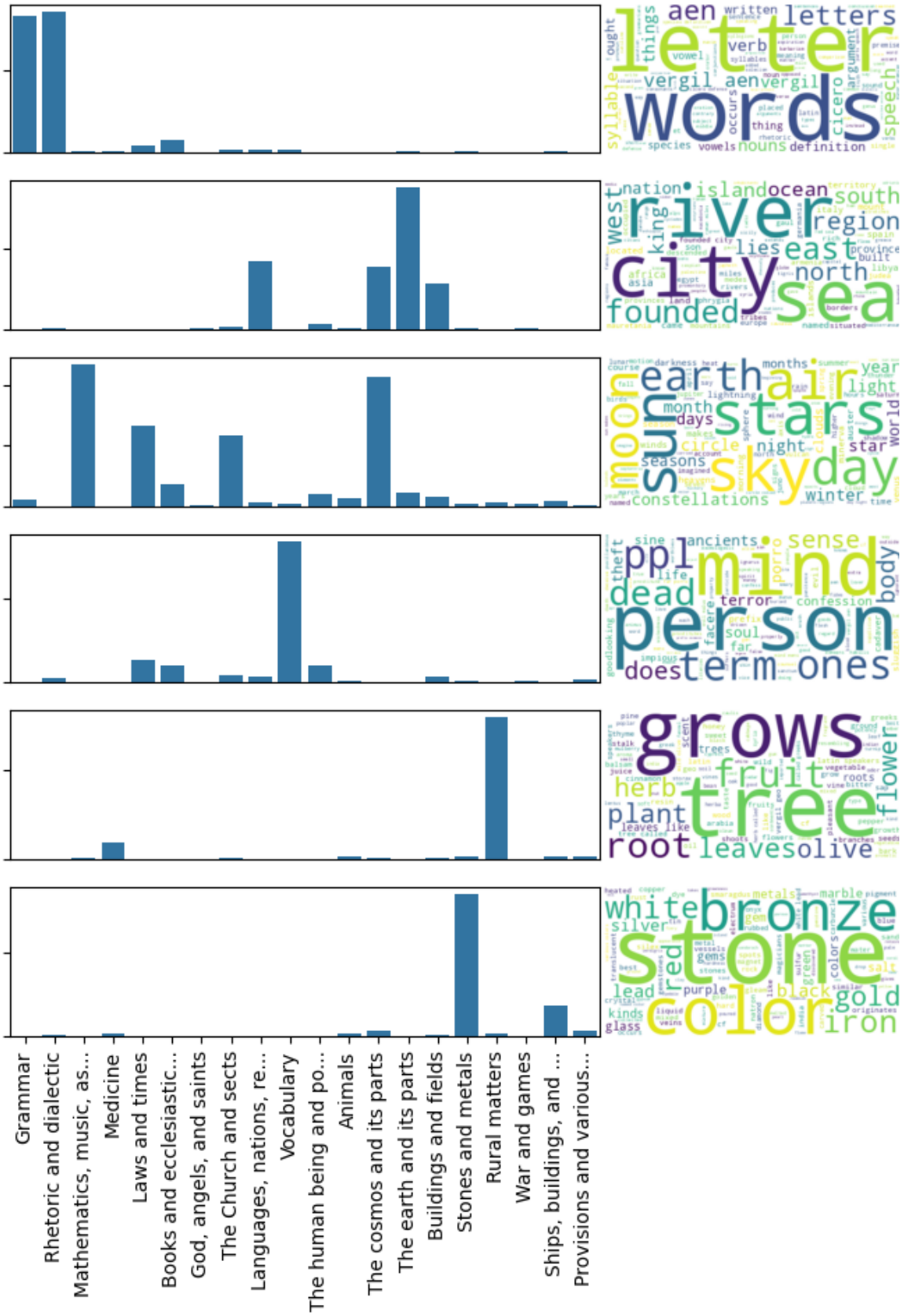


Figura 4.4: *Composição de livros por tópico - seções.*

4.1.1 Documentos por tópico

A seguir apresentamos alguns exemplos das passagens presentes em cada tópico, mostrando no máximo uma passagem por capítulo, e no máximo 5 capítulos de cada livro. A cada livro, mostramos a porcentagem de passagens daquele livro atribuídas ao tópico, menos as passagens classificadas como "ruído".

Tópico 1: *words, letter, letters, aen, vergil aen, speech, vergil, cicero, nouns, syllable*

Livro: 2 - Rhetoric and dialectic (85.78%)

isi_en.2.I.1.Sall *1. Rhetoric is the art of speaking well in civil cases, [and eloquence (eloquentia) is fluency (copia)] for the purpose of persuading people toward the just and good. Rhetoric is named from the Greek term p?top(c)S?tv, that is, fluency of speech, for p?otç in Greek means "speech,"p?tYp means "orator."*

isi_en.2.II.2.Sall *2. For while one has a treatise on rhetoric in hand, the sequence of its content as it were clings to the memory, but when it is set aside all recollection of it soon slips away. Accomplished knowledge of this discipline makes one an orator.*

isi_en.2.III.1.Sall *1. An orator therefore is a good man, skilled in speaking. A man's goodness is based on his nature, his behavior, his training in the arts. One skilled in speaking is grounded in artful eloquence, which consists of five parts: invention, arrangement, style, memory, pronunciation (inventio, dispositio, elocutio, memoria, pronuntiatio), and of the goal of this office, which is to persuade of something.*

Livro: 1 - Grammar (82.79%)

isi_en.1.II.1.Sall *1. There are seven disciplines of the liberal arts. The first is grammar, that is, skill in speaking. The second is rhetoric, which, on account of the brilliance and fluency of its eloquence, is considered most necessary in public proceedings. The third is dialectic, otherwise known as logic, which separates the true from the false by very subtle argumentation.*

isi_en.1.III.1.Sall *1. The common letters of the alphabet are the primary elements of the art of grammar, and are used by scribes and accountants. The teaching of these letters is, as it were, the infancy of grammar, whence Varro also calls this discipline 'literacy' (litteratio). Indeed, letters are tokens of things, the signs of words, and they have so much force that the utterances of those who are absent speak to us without a voice, [for they present words through the eyes, not through the ears].*

isi_en.1.IV.6.Sall *6. Now they are vowels, and now semivowels, and now medials (i.e. glides). They are vowels because they make syllables when they are positioned alone or when they are joined to consonants. They are considered consonants in that they sometimes have a vowel set down after them in the same syllable, as Ianus, vates, and they are considered as consonants.*

Livro: 6 - Books and ecclesiastical offices (7.43%)

isi_en.6.IX.2.Sall *2. Hence it was said among scribes, "You shall not strike wax with iron."Afterwards it was established that they would write on wax tablets with bones, as Atta indicates in his Satura, saying (12): Let us turn the plowshare and plow in the wax with a point of bone. The Greek term graphium is scriptorium in Latin, for ypa??? is "writing."*

isi_en.6.VIII.7.Sall *7. A panegyric (panegyricum) is an extravagant and immoderate form of discourse in praise of kings; in its composition people fawn on them with many lies. This wickedness had its origin among the Greeks, whose practised glibness in speaking has with its ease and incredible fluency stirred up many clouds of lies.*

isi_en.6.XIV.2.Sall *2. The scribe (scriba) got his name from writing (scribere), expressing his function by the character of his title.*

Livro: 5 - Laws and times (4.35%)

isi_en.5.XXV.32.Sall *32. Cessio is a concession (concessio) of one's own property, such as this: "I cede by right of affinity,"for we say 'cede' (cedere) as if it were 'concede' (concedere), that is, those things that are our own; for we 'restore' the property of another, we do not 'cede' it. In fact, technically speaking, someone is said to cede when he gives in to another in spite of the truth, as Cicero (Defense of Ligarius 7.22): "He ceded,"he says, "to the authority of a very distinguished man, or rather, he obeyed."*

isi_en.5.XXVI.9.Sall *9. Deceit (falsitas) is so called from saying (fari, ppl. fatus) something other than the truth.*

isi_en.5.XXVII.27.Sall *27. Report does not possess a trustworthy name, because it is especially untruthful, either adding many things to the truth, or distorting the truth. It lasts just as long as it is not put to the test, but whenever you put it to the test, it ceases to be, and after that is called fact, not report.*

Livro: 8 - The Church and sects (1.98%)

isi_en.8.IX.29.Sall *29. The salisatores are so called because whenever any part of their limbs leaps (salire), they proclaim that this means something fortunate or something unfortunate for them thereafter.*

isi_en.8.VII.5.Sall *5. Tragedians (tragoedus) are so called, because at first the prize for singers was a goat, which the Greeks call tp?yoç. Hence also Horace (Art of Poetry 220): Who with a tragic song vied for a paltry goat. Now the tragedians following thereafter attained great honor, excelling in the plots of their stories, composed in the image of truth.*

No primeiro tópico, observamos uma grande coesão, com pouca variabilidade temática, o que é esperado. Porém, apesar de estar bem concentrado nos livros *Grammar* e *Rhetoric and dialectic*, aparecem passagens de outros livros aparentemente não relacionados, mas realmente tratando de temas similares. Nas passagens apresentadas, o autor trata de assuntos ligados à retórica e à dialética.

Tópico 2: *city, river, sea, founded, east, region, west, north, south, island*

Livro: 14 - The earth and its parts (87.17%)

isi_en.14.II.3.Sall *3. Whence it is clear that two of them, Europe and Africa, occupy half of the globe, Asia the other half by itself. But the former pair are divided into two regions, because from the Ocean the Mediterranean enters in between them and separates them. Wherefore, if you divide the globe into two parts, the east and the west, Asia will be in one, Europe and Africa in the other.*

isi_en.14.III.46.Sall *46. Lycia is so called because in the east it borders on Cilicia, for Cilicia borders on it in the east and it has the sea to the west and the south; in the north lies Caria. There lies Mount Chimera, which exhales fire in nightly surges, like Etna in Sicily and Vesuvius in Campania.*

isi_en.14.IV.30.Sall *30. Furthermore there are two Spains: Inner Spain, whose area extends in the north from the Pyrenees to Cartagena; and Outer Spain, which in the south extends from Celtiberia to the straits of Cadiz. Inner (citerior) and Outer (ulterior) are so called as if it were citra (on this side) and ultra (beyond); but citra is formed as if the term were 'around the earth' (circa terras), and ultra either because it is the last (ultimus), or because after it there is not 'any' (ulla), that is, any other, land.*

Livro: 9 - Languages, nations, reigns, the military, citizens, family relationships (41.58%)

isi_en.9.II.2.Sall *2. Now, of the nations into which the earth is divided, fifteen are from Japheth, thirty-one from Ham, and twenty-seven from Shem, which adds up to seventythree - or rather, as a proper accounting shows, seventytwo. And there are an equal number of languages, which arose across the lands and, as they increased, filled the provinces and islands.*

isi_en.9.III.2.Sall *2. Every nation has had its own reign in its own times - like the Assyrians, the Medes, the Persians, the Egyptians, the Greeks - and fate has so rolled over their allotments of time that each successive one would dissolve the former. Among all the reigns on earth, however, two reigns are held to be*

glorious above the rest - first the Assyrians, then the Romans - as they are constituted differently from one another in location as much as time.

isi_en.9.IV.28.Sall *28. Burghers (burgarius) are so called from 'fortified villages' (burgus), because in common speech people call the many dwelling-places established along the frontiers burgi. Hence also the nation of Burgundians got their name: formerly, when Germania was subdued, the Romans scattered them among their camps, and so they took their name from these places.*

Livro: 13 - The cosmos and its parts (38.17%)

isi_en.13.XIII.9.Sall *9. There is a lake in the country of the Troglodytes; three times a day it becomes bitter, and then, just as often, sweet again. The Siloan spring at the foot of Mount Zion has no continuous flow of water, but bubbles forth at certain hours and days. In Judea a certain river used to go dry every Sabbath.*

isi_en.13.XIX.4.Sall *4. People say that a lighted lamp floats on top, but when its light is extinguished, it sinks. This is also called the Salt Sea, or Lake Asphalti, that is, 'of bitumen,' and it is in Judea between Jericho and Zoara. In length it stretches 780 stades (i.e. about ninety miles) to Zoara in Arabia and its width is 150 stades, up to the neighborhood of Sodom.*

isi_en.13.XV.1.Sall *1. Greek and Latin speakers so name the 'Ocean' (oceanus) because it goes around the globe (orbis) in the manner of a circle (circulus), [or from its speed, because it runs quickly (ocius)]. Again, because it gleams with a deep blue color like the sky: oceanus as if the word were mU?v?oç ("blue"). This is what encircles the edges of the land, advancing and receding with alternate tides, for when the winds blow over the deep, the Ocean either disgorges the seas or swallows them back.*

Livro: 15 - Buildings and fields (28.57%)

isi_en.15.I.53.Sall *53. But Ascanius, after he had left his kingdom to his stepmother Lavinia, built Alba Longa. It was called Alba, 'White,' because of the color of a sow, and Longa because the town is elongated, in keeping with the great extent of the hill on which it is sited. From the name of this city the kings of the Albans took their names.*

isi_en.15.II.31.Sall *31. The Capitolium of Rome is so called because it was the highest head (caput) of the Roman city and its religion. Others say that when Tarquinius Priscus was uncovering the foundations of the Capitolium in Rome, he found on the site of the foundation the head (caput) of a human marked with Etruscan writing, and hence he named it the Capitolium.*

Livro: 11 - The human being and portents (3.14%)

isi_en.11.III.20.Sall *20. The Artabatitans of Ethiopia are said to walk on all fours, like cattle; none passes the age of forty.*

Aqui o tema parece estar ligado à localização e à formação de diferentes povos e nações, com diferentes focos de acordo com o livro de onde a passagem foi retirada. Este tema parece um pouco mais distribuído do que o primeiro, mas aparece predominantemente ainda nos livros *The earth and its parts* e *Languages, nations, reigns, the military, citizens, family relationships.*

Tópico 3: *sun, sky, stars, air, day, moon, earth, constellations, star, month*

Livro: 3 - Mathematics, music, astronomy (58.46%)

isi_en.3.L.2.Sall *2. Wandering farther to the south it makes winter, so that the earth grows fertile with wintry moisture and frost. When it approaches closer to the north, it brings summer back, so that crops grow firm in ripeness, and what was unripened in damp weather mellows in its warmth.*

isi_en.3.LI.2.Sall *2. When the sun runs across the south, it is the closer to the earth; but when it is near the north, it is raised higher in the sky. [Thus God made diverse locations and seasons for the sun's course, so that it does not consume everything with its daily heat by always tarrying in the same place. But,*

as Clement said, "The sun takes diverse paths, by means of which the temperature of the air is meted out according to the pattern of the seasons, and the order of its changes and permutations is preserved. Thus when the sun ascends to the higher reaches, it tempers the spring air; when it reaches its zenith, it kindles the summer heat; dropping again it brings back the temperance of autumn. But when it goes back to the lowest orbit, it bequeaths to us from the icy framework of the sky the rigor of winter cold."]

isi_en.3.LIII.2.Sall *2. Others maintain on the contrary that the moon does not have its own light, but is illuminated by the rays of the sun, and for this reason undergoes an eclipse when the earth's shadow comes between it and the sun. [For the sun is located higher than the moon. Hence it happens that when the moon is beneath the sun, the upper part of the moon is lighted, but the lower part, which is facing the earth, is dark.]*

Livro: 13 - The cosmos and its parts (53.44%)

isi_en.13.I.3.Sall *3. There are four zones in the world, that is, four regions: the East and the West, the North and the South.*

isi_en.13.II.4.Sall *4. In number, take for example eight divided into four, and four into two, and then two into one. But one is an atom, because it is indivisible. Thus also with letters (i.e. speech-sounds), for speech is divided into words, words into syllables, syllables into letters. But a letter, the smallest part, is an atom and cannot be divided. Therefore an atom is whatever cannot be divided, like a point in geometry, for tómoç means "division"in Greek, and ?tomoç means "non-division."*

isi_en.13.III.3.Sall *3. For this reason, all the elements are present in all, but each one has taken its name from whichever element is more abundant in it. The elements are assigned by Divine Providence to the appropriate living beings, for the Creator himself has filled heaven (i.e. the fiery realm) with angels, air with birds, water with fish, and earth with humans and the rest of the living things.*

Livro: 5 - Laws and times (33.15%)

isi_en.5.XXIX.1.Sall *1. Intervals of time are divided into moments, hours, days, months, years, lustrums, centuries, and ages. A moment (momentum) is the least and shortest bit of time, so called from the movement (motus) of the stars.*

isi_en.5.XXVII.38.Sall *38. For that reason the Romans forbade water and fire to certain condemned people - because air and water are free to all and given to everyone - so that the condemned might not enjoy what is given by nature to everyone.*

isi_en.5.XXX.16.Sall *16. Evening (suprema) is the last part of the day, when the sun turns its course toward its setting - so called because it 'still exists' (superesse) up to the final part of the day.*

Livro: 8 - The Church and sects (28.97%)

isi_en.8.IX.17.Sall *17. Haruspices are so named as if the expression were 'observers (inspector) of the hours (hora)'; they watch over the days and hours for doing business and other works, and they attend to what a person ought to watch out for at any particular time. They also examine the entrails of animals and predict the future from them.*

isi_en.8.VI.21.Sall *21. Whence also Varro says that fire is the soul of the world; just as fire governs all things in the world, so the soul governs all things in us. As he says most vainly, "When it is in us, we exist; when it leaves us, we perish."Thus also when fire departs from the world through lightning, the world perishes.*

isi_en.8.XI.45.Sall *45. Mercury (Mercurius) is translated as "speech,"for Mercury is said to be named as if the word were mediuscurrens ("go-between"), because speech is the gobetween for people. In Greek he is called 'Epμ?ç, because 'speech' or 'interpretation,' which pertains especially to speech, is called spμ?v?(c)a.*

Livro: 6 - Books and ecclesiastical offices (8.91%)

isi_en.6.II.3.Sall *3. The book of Genesis is so called because the beginning of the world and the begetting (generatio) of living creatures are contained in it.*

isi_en.6.XIX.2.Sall *2. The office of Vespers takes place at the beginning of night, and is named for the evening star Vesper, which rises when night falls.*

isi_en.6.XVII.3.Sall *3. It is calleda cycle (cyclum) because it is set out in the form of a wheel, and arranged as if it were in a circle (circulum) it comprises the order of the years without variation and without any artifice.*

Aqui vemos passagens tratando principalmente da relação entre corpos celestes e a passagem do tempo e estações do ano, bem como manifestaçoes atmosféricas e elementos da natureza. Temos aqui uma variabilidade temática um pouco maior, e maior transversalidade também, com uma maior distribuição deste tópico por diferentes livros.

Tópico 4: *person, mind, term, ones, ppl, dead, does, sense, body, ancients*

Livro: 10 - Vocabulary (83.64%)

isi_en.10.A.11.Sall *11. Indecisive (anceps), wavering this way and that and doubting whether to choose this or that, and distressed (anxius) about which way to lean. Abominable (atrox), because one has loathsome (taeter) conduct. Abstemious (abstemius), from temetum, that is, 'wine,' as if abstaining (abstinere) from wine. [Neighboring (affinis) . . .] Weaned (ablactatus), because one is 'withdrawn from milk' (a lacte ablatus).*

isi_en.10.B.31.Sall *31. Baburrus, "stupid, inept."Biothanatus (i.e. a martyr who dies a violent death), because he is 'twice dead,' for death is 9?vatoç in Greek.*

isi_en.10.C.40.Sall *40. Constant (constans) is so called because one 'stands firm' (stare, present participle stans) in every situation, and cannot deviate in any direction. Trusting (confidens), one full of faith (fiducia) in all matters. Whence Caecilius (fr. 246): If you summon Confidence, confide (confidere) everything to her.*

Livro: 5 - Laws and times (13.59%)

isi_en.5.XXVI.1.Sall *1. Crime (crimen) has its name from lacking (carere) - like theft, deceit, and other actions that do not kill, but cause disgrace.*

isi_en.5.XXVII.1.Sall *1. Harm (malum) is defined in two ways: one definition being what a person does; the other, what he suffers. What he does is wrongdoing (peccatum), what he suffers is punishment. And harm is at its full extent when it is both past and also impending, so that it includes both grief and dread.*

Livro: 11 - The human being and portents (10.47%)

isi_en.11.I.1.Sall *1. Nature (natura) is so called because itcauses something to be born (nasci, ppl. natus), for it has the power of engendering and creating. Some people say that this is God, by whom all things have been created and exist.*

isi_en.11.II.31.Sall *31. Death (mors) is so called, because it is bitter (amarus), or by derivation from Mars, who is the author of death; [or else, death is derived from the bite (morsus) of the first human, because when he bit the fruit of the forbidden tree, he incurred death].*

isi_en.11.III.27.Sall *27. They claim also that in the same India is a race of women who conceive when they are five years old and do not live beyond eight.*

Livro: 6 - Books and ecclesiastical offices (9.90%)

isi_en.6.XIX.35.Sall *35. A holocaust (holocaustum) is a sacrifice in which all that is offered is consumed by fire, for when the ancients would perform their greatest sacrifices, they would consume the whole sacrificial victim in the flame of the rites, and those were holocausts, for o2oç in Greek means "whole,"mauotç means "burning,"and holocaust, "wholly burnt."*

Livro: 8 - The Church and sects (4.37%)

isi_en.8.II.7.Sall *7. It is greater than the other two, because he who loves also believes and hopes. But he who does not love, although he may do many good things, labors in vain. Moreover every carnal love (dilectio carnalis) is customarily called not love (dilectio) but 'desire' (amor). We usually use the term dilectio only with regard to better things.*

isi_en.8.III.6.Sall *6. Superstition (superstitio) is so called because it is a superfluous or superimposed (superinstituere) observance. Others say it is from the aged, because those who have lived (superstites) for many years are senile with age and go astray in some superstition through not being aware of which ancient practices they are observing or which they are adding in through ignorance of the old ones.*

isi_en.8.IV.11.Sall *11. The Hemerobaptistae [who wash their bodies and home and domestic utensils daily,] [so called because they wash their clothes and body daily (cf. ¡µspa, "day,"and ßapt(c)S?tv, "to wash")].*

Tópico 5: *tree, grows, fruit, root, leaves, plant, flower, herb, olive, leaves like*

Livro: 17 - Rural matters (79.26%)

isi_en.17.III.15.Sall *15. We speak improperly of the 'ear' (spica) of ripe fruit, for properly the ear exists when the beards, still thin like spear-tips (spiculum), project through the husk of the stalk, that is the swelling tip.*

isi_en.17.IV.6.Sall *6. 'French bean' (faselum; cf. ???o?2oç) and chickpea (cicer; cf. mptóç) are Greek names. But faselum . . .*

isi_en.17.IX.105.Sall *105.A fern (filix) is so called from the singleness of its leaf (folium, cf. filum, "a single strand"), for from one stalk a cubit high grows one divided leaf, with an intricate structure like a feather's. Oats (avena) . . . Darnel (lolium) .*

Livro: 4 - Medicine (9.38%)

isi_en.4.IX.9.Sall *9. Catapotia, because a little is drunk (potare) or swallowed down. Diamoron got its name from the juice of the mulberry (morum), from which it is made; likewise diacodion, because it is made from the poppy-head (codia; cf. mÛ6?ta, "poppyhead"), that is, from the poppy; and similarly diaspermaton, because it is made from seeds (cf. opspµa, "seed").*

isi_en.4.X.3.Sall *3. A dinamidia describes the power of herbs, that is, their force and capability. In herbal medicine, potency itself is called 6ávaµtç, whence also the books where herbal remedies are inscribed are called dinamidia.A 'botanical treatise' (butanicum, i.e. botanicum, cf. ßot?v?, "herb") about plants is so called because plants are described in it.*

isi_en.4.XII.2.Sall *2. It is called incense (thymiama) in the Greek language, because it is scented, for a flower that bears a scent is called thyme (thymum). With regard to this, Vergil (Geo. 4.169): And (the honey) is redolent with thyme.*

Livro: 16 - Stones and metals (1.59%)

isi_en.16.II.8.Sall *8. The Greek term aphronitrum is 'foam of natron,' spuma nitri in Latin. Concerning this a certain poet says (Martial, Epigrams 14.58): Are you a bumpkin? You don't know what my Greek name is. I am called 'foam of natron.' Are you Greek? It is aphronitrum. It is gathered in Asia where it distills in caves; from there it is dried in the sun. It is thought to be best if it has as little weight and is as easily crumbled as possible, and is almost purple in color.*

isi_en.16.IV.14.Sall *14. Memphitis is named from a place in Egypt (i.e. Memphis); it has the nature of a gem. When ground and mixed with vinegar and smeared on those parts of the body that have to be burned or cut it makes them so numb that they do not feel pain.*

isi_en.16.VII.14.Sall *14. Myrrhites is so named because it has the color of myrrh. When it is compressed until it becomes warm it exudes the sweet smell of nard. Aromatitis is found in Arabia and Egypt; it has the color and scent of myrrh, whence it takes its name (cf. aroma, "spice").*

Livro: 12 - Animals (1.37%)

isi_en.12.VI.37.Sall *37. The squatus is named because it has 'sharp scales' (squamis acutus). For this reason wood is polished with its skin.*

isi_en.12.VII.23.Sall *23. The cinnamolgus is also a bird of Arabia, called thus because in tall trees it constructs nests out of cinnamon (cinnamum) shrubs, and since humans are unable to climb up there due to the height and fragility of the branches, they go after the nests using lead-weighted missiles. Thus they dislodge these cinnamon nests and sell them at very high prices, for merchants value cinnamon more than other spices.*

isi_en.12.VIII.8.Sall *8. Butterflies (papilio) are small flying creatures that are very abundant when mallows bloom, and they cause small worms to be generated from their own dung.*

Livro: 20 - Provisions and various implements (1.22%)

isi_en.20.II.36.Sall *36. Honey (mel) is froma Greek term (i.e. μs2t), which is shown to have its name from bees, for a bee in Greek is called μs2tooa. Formerly honey was from dew and was found in the leaves of reeds. Hence Vergil (Geo. 4.1): Forthwith, the celestial gifts of honey from the air. Indeed in India and Arabia honey is still found attached to branches, clinging like grains of salt. Nevertheless, all honey is sweet; Sardinian honey is bitter because of wormwood, with whose abundance the bees of that region are fed. The honeycomb (favus) is so called because it is eaten rather than drunk, for the Greeks say ??ay?±v for "eat."*

isi_en.20.VII.3.Sall *3. A pyx (pyxis) is a little container for ointment made of boxwood, for what we call 'boxwood' the Greeks call pá(oç.*

Aqui vemos uma concentração muito grande de passagens sobre agricultura no livro *Rural matters*, como esperado. Note também a presença de passagens falando do uso de ervas com fins terapêuticos, no livro *Medicine*.

Tópico 6: *stone, color, bronze, white, iron, gold, red, black, lead, silver*

Livro: 16 - Stones and metals (84.92%)

isi_en.16.I.10.Sall *10. There are four kinds of sulfur. The 'living' kind, which is dug up, is translucent and green; physicians use it alone of all the kinds of sulfur. The second kind, which people call 'lump sulfur' (i.e. fuller's earth), is used only by fullers. The third kind is 'liquid'; it is used for fumigating wool because it gives whiteness and softness. The fourth kind is particularly suitable for preparing lamp-wicks. The power of sulfur is so great that it cures the 'comitial sickness' (see IV.vii.7) through its vapor when it is set on the fire and burns. When a person puts sulfur in a goblet of wine and carries it around with hot coals beneath he glows with the eerie pallor of a corpse from the reflection of the blaze.*

isi_en.16.II.10.Sall *10. But it is now produced elsewhere in caves, because, having collected as a liquid, it drips down there and solidifies into 'grape clusters.' It also occurs in hollow trenches, from whose sides the hanging drops coalesce; it is also made, like salt, under the most blazing sun. Its power is so concentrated that, when sprinkled into the mouths of lions and bears, they are unable to bite because of its astringent force.*

isi_en.16.III.11.Sall *11. Sand (arena, i.e. harena) is named from 'dryness' (ariditas), not from 'cementing' (adhaerere) building materials, as some people would have it. The test of its quality is if it grates when squeezed in one's hand, or if it leaves behind no stain when sprinkled on white cloth.*

Livro: 19 - Ships, buildings, and clothing (18.00%)

isi_en.19.X.3.Sall *3. Stones that are suitable for building: white stone, Tiburtine, columbinus, river stone, porous, red, and the others.*

isi_en.19.XVII.5.Sall *5. Syrian (Syricum) is a pigment with a red color, with which the chapter-heads in books are written. It is also known as Phoenician, so called because it is collected in Syria on the shores of the Red Sea, where the Phoenicians live.*

isi_en.19.XXVIII.1.Sall *1. Dying (tinctura) is so named because cloth is 'soaked in color' (tinguere), tinted to another appearance, and colored for the sake of beauty. What we call red or vermilion (vermiculus), the Greeks call mómmoç; it is a small grub (vermiculus) from the foliage of the forest.*

Livro: 13 - The cosmos and its parts (3.05%)

isi_en.13.X.1.Sall *1. The celestial rainbow (arcus) is named for its likeness to the curve of a bow (also arcus). Iris is its proper name. It is called iris as if the word were aeris, that is, something that descends to earth through the air (aer). It takes its light from the sun, whenever hollow clouds receive the sun's rays from the opposite side and make the shape of a bow. This circumstance gives it various colors, because the thinned water, bright air, and misty clouds, when illuminated, create various colors.*

isi_en.13.XVII.2.Sall *2. The Red Sea is so named because it is colored with reddish waves; however, it does not possess this quality by its nature, but its currents are tainted and stained by the neighboring shores because all the land surrounding that sea is red and close to the color of blood. From there a very intense vermilion may be separated out, as well as other pigments with which the coloring of paintings is varied.*

isi_en.13.XX.5.Sall *5.A drop (gutta) is that which stands hanging, stilla (i.e. another word for 'drop') is that which falls. Hence the word stillicidium (i.e. drippings from eaves), as if it were 'falling drop' (stilla cadens). Stiria (lit. "frozen drop, icicle,"here simply "drop") is a Greek word, that is 'drop' (gutta); from it the diminutive that we call stilla is formed. As long as it stands or hangs suspended from roofs or trees, it is a gutta, as if 'glutinous' (glutinosus), but when it has fallen it is a stilla.*

Livro: 20 - Provisions and various implements (3.05%)

isi_en.20.II.31.Sall *31. Galatica is named for its milky color, for the Greeks call milk y?2a. Meatballs (sphera) are named with this Greek word for their roundness, for whatever is round in shape is called o??a±pa in Greek.*

isi_en.20.IV.8.Sall *8. Chrysendetus vessels are those inlaid with gold; the term is Greek (cf. ypUoóç, "gold"; sv6?±v "bind on"). Vessels in bas-relief (anaglypha) are those carved on top, for the Greek ?vY means "above,"y2U???, "carving,"that is, carved above.*

isi_en.20.VII.2.Sall *2. An alabastrum isa vessel for ointments, and is named from the kind of stone of which it is made, which they call alabaster (alabastrites), which keeps ointments unspoiled.*

Livro: 4 - Medicine (1.56%)

isi_en.4.VI.16.Sall *16. A carbuncle (carbunculus) is so called because at first it glows red, like fire, and then turns black, like an extinguished coal (carbo).*

isi_en.4.VII.32.Sall *32. A cauculus is a stone that occurs in the bladder, and it took its name from that (i.e. calculus, "pebble"). It is formed from phlegmatic matter.*

Aqui podemos ver uma sutil diferença de foco: tanto as passagens do livro *Stones and metals* quanto as de *Ships, buildings, and clothing* tratam de diferentes minerais, porém as provenientes do primeiro tratam mais da natureza desses minerais, enquanto as do segundo de parecem tratar de minerais como matéria-prima para construção, fabricação, etc.

4.2 Segmentação por sentenças

Agora investigamos os resultados da modelagem de tópicos usando a segmentação dos textos em sentenças.

Na figura 4.5 temos uma representação 2D do espaço dos embeddings. Em comparação à figura 4.3, podemos ver que, apesar dos pontos aparecerem menos concentrados, as principais ilhas temáticas se mantém, mesmo que em posições diferentes.

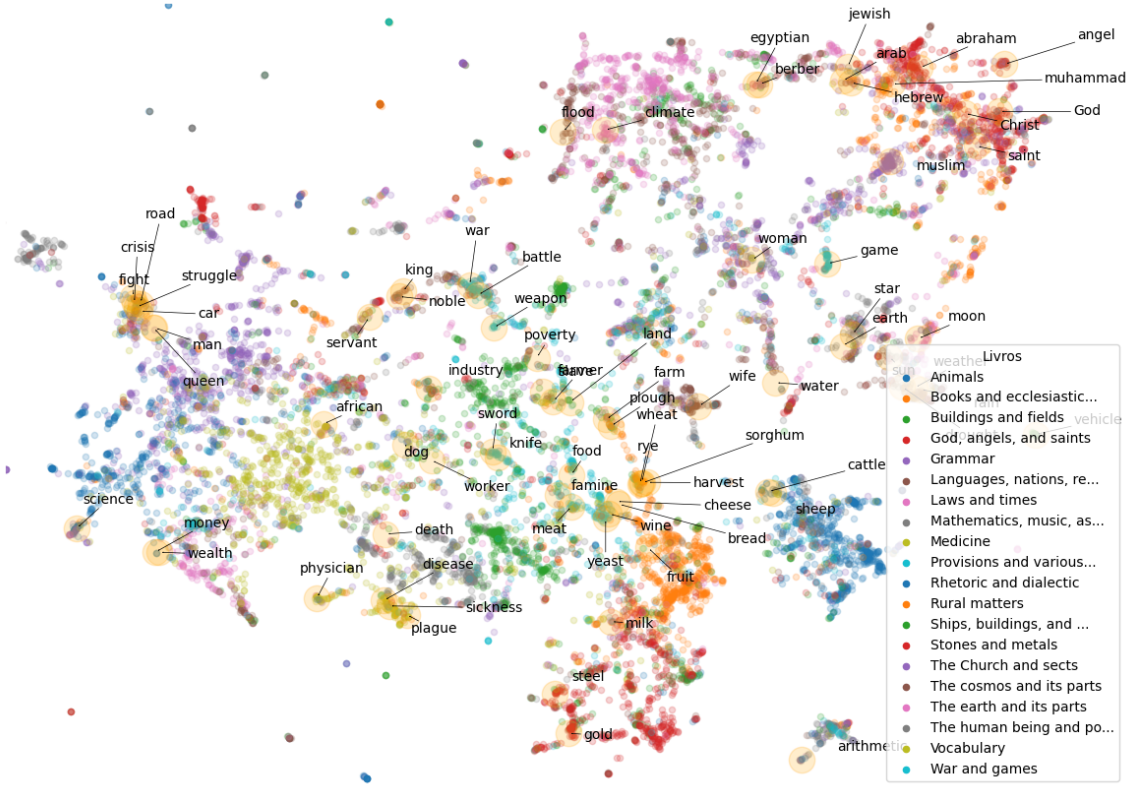


Figura 4.5: *Visualização 2D do espaço de embeddings - sentenças.*

4.2.1 Documentos por tópico

Tópico 1: *tree, stone, color, grows, white, bronze, black, fruit, leaves, root*

Livro: 17 - Rural matters (84.50%)

isi_en.17.II.7.S0 *7.A grain-field (seges) is [*

isi_en.17.III.10.S4 *The second is distichon barley, because it has two rows; many call this Galaticum.*

isi_en.17.IV.11.S0 *11. 'Bitter vetch' (ervum) derives its name from Greek, for they call it opoßoç, because while it is dangerous for some livestock it nevertheless makes bulls (cf. ßouç, "bull") plump.*

isi_en.17.IX.106.S2 *We call a handful of hay a 'sheaf' (manipulum), and it is called manipulum because it "fills the hand"(manum implere).*

isi_en.17.V.3.S0 *The labrusca is a wild vine that grows in marginal land, whence it is called labrusca, from the margins (labrum) and extremities of the land.*

Livro: 16 - Stones and metals (72.24%)

isi_en.16.I.10.S2 *The second kind, which people call 'lump sulfur' (i.e. fuller's earth), is used only by fullers.*

isi_en.16.II.3.S6 *In some places it is so hard that people make walls and houses out of masses of salt, as in Arabia.*

isi_en.16.III.9.S0 *Gypsum (gypsum) is related to limestone; it is a Greek term (i.e. yáyoç).*

isi_en.16.IV.26.S0 *The stone melanites is so called because it exudes a sweet honeyed (melleus) fluid.*

isi_en.16.IX.5.S0 *Amethystizontas is so named because the sparkle on its surface ranges towards the violet color of amethyst.*

Livro: 20 - Provisions and various implements (40.89%)

isi_en.20.II.23.S0 *Fried (frixus) is so termed from the sound food makes when it is seared in oil.*

isi_en.20.III.10.S1 *Honey-wine (mulsum) is wine mixed with honey, for it is a drink made from water and honey, which the Greeks call μ?2(c)mpatov.*

isi_en.20.IV.3.S0 *3. Ceramic dishes are said to have been first invented on the island of Samos, made from white clay and hardened by fire, hence 'Samian dishes.'*

isi_en.20.V.4.S2 *An amystis (cf. ?*

isi_en.20.VI.2.S2 *Later they passed into use for wine, but keeping the Greek term with which they had their origin.*

Livro: 19 - Ships, buildings, and clothing (23.99%)

isi_en.19.II.9.S2 *): The shining mast-heads (carchesium) of the tall mast glitter.*

isi_en.19.IV.10.S0 *The sounding-lead (catapirates) is a line with a lead weight, with which the depth of the sea is tested.*

isi_en.19.VI.7.S4 *Once extinguished it lasts so incorruptibly that the people who fix boundaries spread charcoals below the surface and place stones on top, so as to prove the boundary to a litigant however many generations later, and they recognize a stone fixed in this way to be a boundary.*

isi_en.19.VII.1.S0 *The anvil (incus) is that tool on which iron is beaten out.*

isi_en.19.X.12.S4 *Green silex is itself stubbornly resistant to fire, but there is nowhere where it is abundant, and it is found only as a stone and not as a rocky outcrop.*

Livro: 4 - Medicine (15.67%)

isi_en.4.IX.8.S1 *The remedy hiera is so called as if it were 'holy,' (cf. ¬?póç, "holy").*

isi_en.4.VI.16.S0 *A carbuncle (carbunculus) is so called because at first it glows red, like fire, and then turns black, like an extinguished coal (carbo).*

isi_en.4.VII.25.S0 *Paralysis (paralesis) is so called from damage to the body caused by excessive chilling, occurring either in the entire body, or in one part.*

isi_en.4.VIII.3.S1 *Latin speakers call erysipelas (erisipela) 'sacred fire' - speaking in antiphrasis, as it should be cursed - inasmuch as the skin grows flame-red on its surface.*

isi_en.4.X.3.S0 *A dinamidia describes the power of herbs, that is, their force and capability.*

Este tópico se assemelha um pouco aos tópicos 4.1.1 e 4.1.1 encontrados na modelagem utilizando segmentação por seções, distribuindo-se primariamente nos livros *Rural matters* e *Stones and metals*, mas também englobando temas relacionados a mantimentos e ervas medicinais.

Tópico 2: *bird, birds, animal, venom, poison, owl, offspring, animals, fly, lions*

Livro: 12 - Animals (69.86%)

isi_en.12.I.12.S0 *Although the Greeks name the lamb (agnus) from ?yvóç ("holy") as if it were sacred, Latin speakers think that it has this name because it recognizes (agnoscere) its mother before other animals, to*

the extent that even if it has strayed within a large herd, it immediately recognizes the voice of its parent by its bleat.

isi_en.12.II.1.S1 *They are called beasts (bestia) from the force (vis) with which they attack.*

isi_en.12.III.9.S1 *It has great shrewdness, for it provides for the future and prepares during the summer what it consumes in the winter; during the harvest it selects the wheat and does not touch the barley.*

isi_en.12.IV.4.S1 *The Greeks call it 6p?mYv, whence the term is borrowed into Latin so that we say draco.*

isi_en.12.V.7.S0 *. Slugs (limax) are mud vermin, so named because they are generated either in mud (limus) or from mud; hence they are always regarded as filthy and unclean.*

Livro: 11 - The human being and portents (7.87%)

isi_en.11.I.39.S2 *They meet together to renew the power of the gaze with their frequent motion.*

isi_en.11.II.25.S0 *The 'elder' (senior) is still fairly vigorous.*

isi_en.11.III.26.S0 *In India there are said to be a race called Mampóßtot, who are twelve feet tall.*

Livro: 20 - Provisions and various implements (5.50%)

isi_en.20.II.23.S3 *Rancid (rancidus) is named after its defect, because it makes meat harsh (raucus).*

isi_en.20.III.2.S2 *The ancients called wine venom (venenum), but after poison from a lethal sap was discovered they called the one thing wine, the other venom.*

isi_en.20.XII.4.S1 *): Mothers in soft carriages (pilentum).*

isi_en.20.XIV.8.S0 *A hoe (sarculus)- of these there are the single-bladed and the two-pronged.*

isi_en.20.XV.4.S0 *The 'wolf' (lupus) or 'little dog' (canicula) is an iron grapple that takes such names because if anything falls in a well it snatches them and draws them out.*

Livro: 8 - The Church and sects (3.69%)

isi_en.8.IX.10.S0 *Hence also Lucan (Civil War 6.457): The mind, polluted by no poison of swallowed venom, yet perishes under a spell.*

isi_en.8.V.69.S4 *Some walk barefoot, others do not eat with other people.*

isi_en.8.XI.63.S2 *They furnish her with tame lions below to show that no kind of beast is so wild that it cannot be subjugated and ruled by her.*

Livro: 18 - War and games (2.42%)

isi_en.18.III.3.S0 *The standard of dragons originated in the killing of the serpent Python by Apollo.*

isi_en.18.VII.4.S1 *Indeed, "Theyintercept (excipere) boars, lie in wait for lions, and penetrate bears, if only one's hand is steady"(cf.*

isi_en.18.X.2.S1 *It is wound up with a thong of sinew and hurls either spears or stones with great force.*

isi_en.18.XII.5.S2 *): A cetra covers their left arms.*

Tópico 3: *garment, hair, wear, cloak, woven, cloth, linen, rings, tunic, silk*

Livro: 19 - Ships, buildings, and clothing (33.75%)

isi_en.19.I.27.S1 *It is also known as the litoraria, and as the caudica, made from a single hollowed piece of wood (cf.*

isi_en.19.V.3.S1 *It is also called the 'seine' (verriculum) because verrere means "drag."*

isi_en.19.VII.1.S4 *tussus) on it, that is, stretched out.*

isi_en.19.XVII.6.S0 *Now, 'silk' (sericum) is one thing, and 'Syrian' (Syricum) is another, for silk is a fiber that the Chinese (Seres; East Asians generally) export, while Syrian is a pigment that the Syrian Phoenicians gather at the shores of the Red Sea.*

isi_en.19.XVIII.3.S0 *A string (linea) is named from its material, for it is made from flax (linum).*

Livro: 11 - The human being and portents (11.24%)

isi_en.11.I.26.S0 *The crown (vertex) is the place where the hairs of the head concentrate, and where the hair growth spirals (vertere, "turn") - whence it is named.*

Livro: 20 - Provisions and various implements (5.50%)

isi_en.20.III.14.S1 *It is called passum from 'suffering' (pati, ppl.*

isi_en.20.IX.5.S0 *Bag (saccus) is so called from the word 'blanket' (sagum), because it is made by sewing one up, as if the word were sagus.*

isi_en.20.V.5.S1 *An ampulla (ampulla) is as if the word were a 'large bubble' (ampla bulla), for it has the spherical shape of bubbles that are made from foaming water and thus are inflated by the wind.*

isi_en.20.VII.1.S1 *A scortia is a vessel for oil, so called because it is made of leather (corium).*

isi_en.20.XII.4.S4 *Unless they were chaste, matrons could not use these; nor, likewise, could they wear fillets.*

Livro: 18 - War and games (5.31%)

isi_en.18.LXIX.1.S0 *A ball (pila) is properly so called because it is stuffed with hair (pilus).*

isi_en.18.XII.4.S0 *A peltum (i.e. pelta) is a very short buckler shaped like a half moon.*

isi_en.18.XIII.2.S0 *2. 'Scale-armor' (squama) is an iron cuirass made from iron or bronze plates linked together in the manner of fish scales (squama), and named for their glittering likeness to fish scales.*

isi_en.18.XIV.1.S1 *The cassis was named by the Etruscans, and I think they named that helmet cassis from the word 'head' (caput).*

Livro: 5 - Laws and times (2.41%)

isi_en.5.XXVII.7.S0 *. Foot-shackles (compes) are named because they 'restrain the feet' (continere + pes).*

Tópico 4: *exile, ignarus, extra, comparative, comparison, learned, terror, doctus, outside, ignorant ignarus*

Livro: 10 - Vocabulary (33.00%)

isi_en.10.A.8.S3 *Ambitious (ambitiosus), because one solicits (ambire) honors.*

isi_en.10.B.31.S0 *Baburrus, "stupid, inept."*

isi_en.10.C.61.S1 *Similarly a lunatic (lunaticus) because [*

isi_en.10.D.77.S3 *Indolent (desidiosus), "sluggish, lazy,"so called from 'settling down' (desidere), that is, from sitting too much.*

isi_en.10.E.83.S3 *Destitute (expers), because 'without a share (pars),' for such a one lacks a share.*

Livro: 1 - Grammar (3.29%)

isi_en.1.I.1.S2 *A discipline is so named in another way, because 'the full thing is learned' (discitur plena).*

isi_en.1.VI.2.S2 *The adverb is taken from the noun, as in 'a learned one, learnedly' (doctus, docte).*

isi_en.1.VII.1.S1 *Indeed, unless you know its name (nomen), the knowledge of a thing perishes.*

isi_en.1.XXVII.15.S0 *There is a question about how maxumus or maximus ("greatest"), and any similar pairs, ought to be written.*

isi_en.1.XXVIII.2.S0 *If any one of these is lacking, it is no longer analogy, that is, similarity, but rather anomaly, that is, outside the rule, such as lepus ("hare") and lupus ("wolf").*

Livro: 5 - Laws and times (2.07%)

isi_en.5.XXVII.30.S0 *Proscription (proscriptio) is a condemnation of exile at a distance, as if it were a 'writing afar' (porro scriptio).*

Livro: 2 - Rhetoric and dialectic (1.84%)

isi_en.2.IX.15.S1 *is] an enemy."*

isi_en.2.V.7.S1 *Comparison (comparatio) occurs when some deed of another person is argued to be proper and useful, because as that deed happened, so this deed at issue is said to have been committed.*

isi_en.2.XXI.9.S1 *Rutilius Lupus, Schemata Lexeos 1.4): "While you call yourself wise instead of cunning, brave instead of reckless, thrifty instead of stingy."*

isi_en.2.XXIV.6.S2 *Temperance (temperantia), how passion and the desire for things may be reined in.*

Livro: 8 - The Church and sects (0.79%)

isi_en.8.V.6.S0 *The Gnostics (Gnosticus) wish to call themselves thus because of the superiority of their knowledge (cf. yv?otç, "knowledge").*

isi_en.8.XI.88.S0 *They name Genius (Genius) thus because he possesses the force, as it were, of generating (gignere, ppl.*

Tópico 5: *law, laws, punishment, consuls, judge, justice, judicial, ius, charge, unjust*

Livro: 5 - Laws and times (25.17%)

isi_en.5.I.3.S0 *Numa Pompilius, who succeeded Romulus to the throne, first established laws for the Romans.*

isi_en.5.II.1.S0 *All laws are either divine or human.*

isi_en.5.III.1.S0 *Jurisprudence is a general term, and a law is an aspect of jurisprudence.*

isi_en.5.IV.2.S1 *Now this, or whatever is similar to it, is never unjust, but is held to be natural and fair.*

isi_en.5.IX.1.S2 *It concerns such things as legal inheritances, cretio (i.e. formal acceptance of an inheritance), guardianship, usucapio (i.e. acquisition of ownership by use): these laws are found among no other group of people, but are particular to the Romans and established for them alone.*

Livro: 9 - Languages, nations, reigns, the military, citizens, family relationships (12.76%)

isi_en.9.I.7.S2 *Then Mixed, which emerged in the Roman state after the wide expansion of the Empire, along with new customs and peoples, corrupted the integrity of speech with solecisms and barbarisms.*

isi_en.9.III.6.S1 *Because the Romans would not put up with the haughty domination of kings, they made a pair of consuls serve as the governing power year by year - for the arrogance of kings was not like the benevolence of a consul, but the haughtiness of a master.*

isi_en.9.IV.24.S0 *Curiales are the same as decurions, and they are called curiales because they 'have charge of ' (procurare) and carry out civic duties.*

Livro: 2 - Rhetoric and dialectic (9.47%)

isi_en.2.IV.2.S0 *. Judicial, in which a decision for punishment or reward is rendered according to the deed of that person.*

isi_en.2.V.2.S2 *Under purpose, judicial (iudicialis) and 'related to affairs' (negotialis).*

isi_en.2.VIII.2.S0 *The doubtful, in which either the judgment is doubtful, or a case is of partly decent and partly wicked matters, so that it arouses both benevolence and offense.*

isi_en.2.X.6.S0 *A law will be decent, just, enforceable, natural, in keeping with the custom of the country, appropriate to the place and time, needful, useful, and also clear - so that it does not hold anything that can deceive through obscurity - and for no private benefit, but for the common profit (communis utilitas) of the citizens.*

isi_en.2.XII.3.S3 *In words, when someone is said to have used words that are ugly and not appropriate to someone's authority, as if someone were to defame Cato the Censor himself as having incited young people to wickedness and lechery.*

Livro: 18 - War and games (7.25%)

isi_en.18.XV.5.S0 *A lawsuit consists either of argumentation or of evidence.*

Livro: 10 - Vocabulary (6.40%)

isi_en.10.A.7.S1 *Fair (aequus), meaning "naturally just,"from 'equity' (aequitas), that is, after the idea of what is equal (aequus) - whence likewise 'equity' is so called after a certain equalness (aequalitate).*

isi_en.10.C.61.S3 *Confounded (confusus), so called from one's confession (confessio) of a wicked deed; hence also 'confounding' (confusio).*

isi_en.10.D.80.S0 *Condemned (damnatus) and condemnable (damnabilis): of these the former has already been sentenced, the latter can be sentenced.*

isi_en.10.F.107.S0 *Criminal (facinorosus), so called from the commission of a particular deed, for he does (facere) what harms (nocere) another.*

isi_en.10.I.149.S1 *Infitiator, "one who denies,"because he does not confess (fateri) but strives against the truth witha lie.*

Aqui temos passagens relacionadas a temas jurídicos. Como esperado, há uma concentração maior de passagens no livro *Laws and times*, mas o tema está presente em vários outros livros.

Tópico 6: *disease, bile, intestines, accompanied, blood, lungs, pain, illness, fever, phlegm*

Livro: 4 - Medicine (49.31%)

isi_en.4.IX.11.S3 *Enema (enema) in Greek is called 'loosening' (relaxatio) in Latin.*

isi_en.4.V.1.S0 *Health is integrity of the body and a balance of its nature with respect to its heat and moisture, which is its blood - hence health (sanitas) is so called, as if it were the condition of the blood (sanguis).*

isi_en.4.VI.1.S0 *An ò(?±a is an acute illness that either passes quickly or kills rather quickly, such as pleurisis, or phrenesis, for in Greek ò(áç means acute and swift.*

isi_en.4.VII.1.S0 *Chronic disease (chronia) is an extended illness that lasts for a long time, like gout or consumption, for ypóvoç in Greek means "time."*

isi_en.4.VIII.2.S0 *2. Parotids (parotida) are areas of hardness or accretions that emerge in the vicinity of the ears, caused by fever or something else, whence they are called papYt(c)6?ç, for Yta is the Greek word for ears.*

Livro: 11 - The human being and portents (17.60%)

isi_en.11.I.57.S0 *In the Gallic language toles (cf. classical Latin toles, "goiter") - what in the diminutive are commonly called tonsils (tusilla, i.e. tonsilla) - is the name for the part in the throat that often swells up (turgescere).*

isi_en.11.II.30.S3 *Indeed, there are two things whereby the forces of the body are diminished, old age and disease.*

Livro: 20 - Provisions and various implements (8.93%)

isi_en.20.I.3.S5 *Eating and drinking, as (cf.*

isi_en.20.II.2.S1 *Youths take nourishment in order to grow, the elderly to endure, for the flesh cannot subsist unless it is strengthened with nourishment.*

isi_en.20.IV.13.S0 *Spoons (coclear) are so called because they were first used for snails (coclea).*

isi_en.20.VIII.1.S0 *Any vessel intended for cooking (coquere) is called coculum.*

Livro: 6 - Books and ecclesiastical offices (3.04%)

isi_en.6.XIX.65.S0 *Fasting (ieiunium) is parsimony of sustenance and abstinence from food, and its name is given to it from a certain portion of the intestines, always thin and empty, which is commonly called the jejunum (ieiunum).*

Livro: 10 - Vocabulary (1.31%)

isi_en.10.C.61.S0 *Epileptic (caducus, "falling sickness"), so called from falling down (cadere).*

isi_en.10.D.70.S0 *70. 'Thickly smeared' (delibutus), anointed with oil as is the custom for athletes or youths in the wrestling arena.*

isi_en.10.F.99.S1 *and] blood arouses beauty.*

isi_en.10.L.161.S0 *Panic-stricken (lymphaticus), because one fears water (cf. lympha, "water"), one whom the Greeks call ú6po??óßoç ("hydrophobic").*

isi_en.10.M.176.S1 *Evil (malus) is named after black bile; the Greeks call black µs2aç.*

Aqui podemos observar temas relacionados à saúde e ao bem-estar, e uma boa concentração das passagens no livro *Medicine*. Podemos notar a diferença no conteúdo semântico das passagens deste tópico e dos tópicos 4.1.1 e 4.2.1, que também contém passagens desse livro, porém com um foco em ervas medicinais.

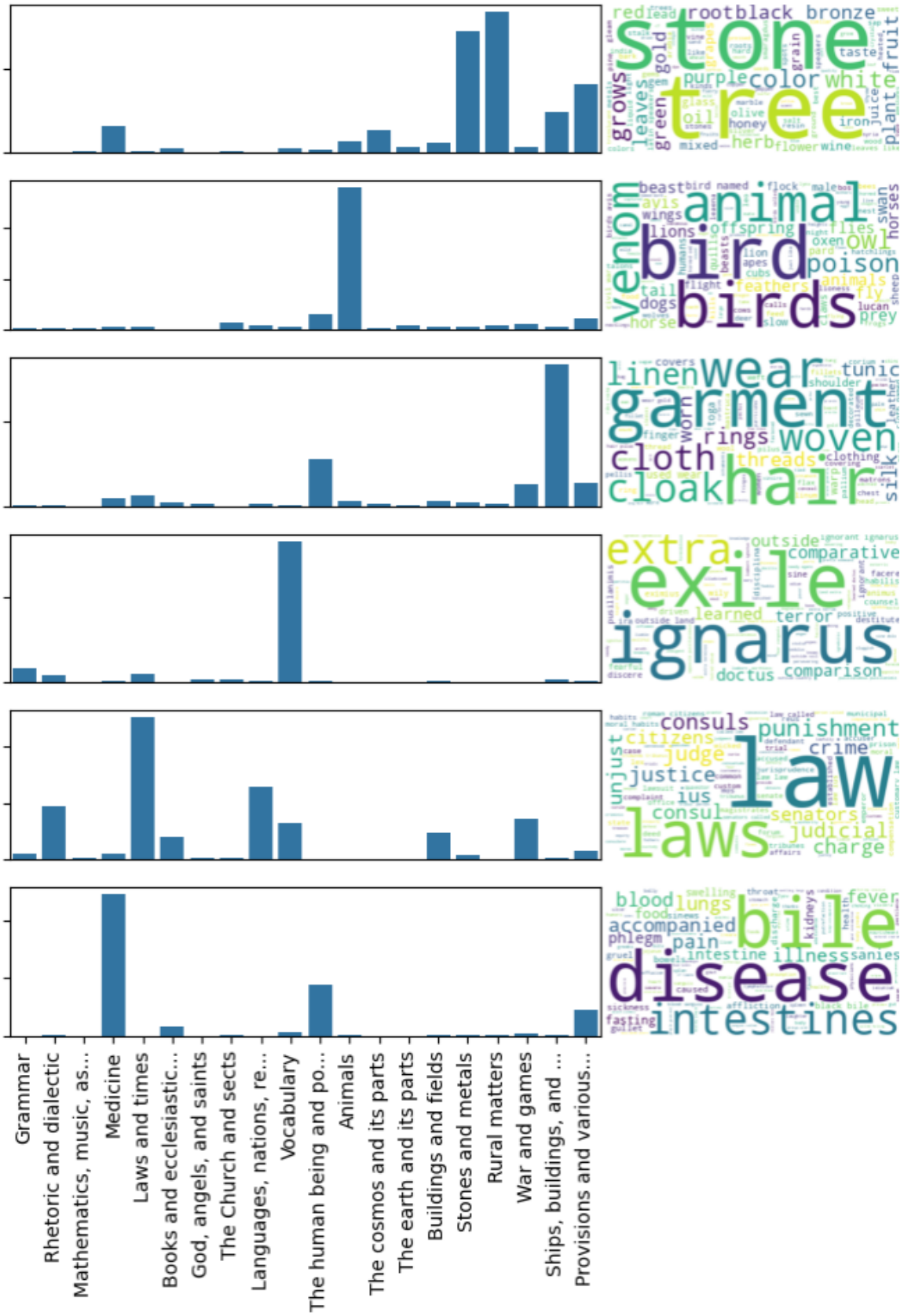


Figura 4.6: *Composição de livros por tópico - sentenças.*

4.3 Similaridade semântica

A seguir apresentamos os resultados do mini experimento de similaridade semântica, onde procuramos pelas sentenças mais próximas das sentenças de exemplo.

Exemplo	**Sentença**	**Localização**	**Similaridade de cosseno**
varrao_en.1.48.2.S2: *The beard is called arista from the fact that it is the first part to dry (arescere).*	*The beard (arista) is so called because it is the first to dry up (arescere).*	Livro: Rural matters; capítulo: III; seção: 16; sentença: 0	0.97997
	The ancients named the beard (barba) that which pertains to men (vir) and not to women.	Livro: The human being and portents; capítulo: I; seção: 45; sentença: 2	0.79411
	They are so called also for this reason, that from the cheeks the beard begins to grow (gignere, ppl.	Livro: The human being and portents; capítulo: I; seção: 43; sentença: 1	0.71945
varrao_en.1.64.1.S0: *Amurca, which is a watery fluid, after it is pressed from the olives is stored along with the dregs in an earthenware vessel.*	*The amurca of olive oil is the watery part, so called from 'emerging' (emergere), that is, because it sinks (mergere) below the oil, and it is the oil's dregs.*	Livro: Rural matters; capítulo: VII; seção: 69; sentença: 0	0.86643
	But what is pressed from white olives is called 'Spanish oil,' and òµ???mtov in Greek.	Livro: Rural matters; capítulo: VII; seção: 68; sentença: 1	0.74530
	Pickled olives (colymbas) are so called . . .	ivro: Rural matters; capítulo: VII; seção: 67; sentença: 3	0.73650

Tabela 4.1: *Resultados da busca por similaridade*

Capítulo 5

Conclusão

Neste trabalho exploramos a intersecção entre Computação e História, investigando o uso de ferramentas de Processamento de Linguagem Natual como Grandes Modelos de Linguagem (LLMs) e técnicas de modelagem de tópicos para a análise de textos históricos, especificamente as *Etimologias* de Isidoro de Sevilha e textos relacionados. O objetivo principal foi o de mapear temas transversais e estruturas semânticas latentes que são custosos e difíceis de serem realizados por meio de análise manual.

Os experimentos realizados demonstraram que a utilização do *pipeline* proposto, baseado em *embeddings* gerados por LLMs, é eficaz para organizar e sumarizar grandes *corpora* de textos históricos. A análise dos resultados da modelagem de tópicos revelou um *trade-off* importante entre a minimização de outliers e a pureza dos clusters. Observamos que configurações que forçam uma redução drástica de ruído tendem a criar agrupamentos heterogêneos, mesclando temas distintos. A abordagem mais eficaz consistiu em minimizar o tamanho máximo dos clusters, o que resultou em tópicos mais coerentes e uma representação mais fiel da estrutura das obras estudadas. Além disso, a comparação entre diferentes granularidades mostrou que a segmentação por sentenças permitiu uma concentração temática superior à segmentação por seções, isolando conceitos específicos com maior precisão.

Do ponto de vista historiográfico e literário, a visualização do espaço de *embeddings* confirmou a existência de temas altamente concentrados, como Gramática, Retórica e Dialética, alinhados à estrutura educacional do *trivium* e *quadrivium*. Mais importante ainda, foi possível identificar temas transversais que permeiam diferentes livros da obra, validando a capacidade do modelo de capturar conexões semânticas para além da organização explícita proposta por Isidoro. As nuvens de palavras geradas (e.g., termos agrícolas, conceitos teológicos e materiais) evidenciam a capacidade do modelo de recuperar o vocabulário distintivo de cada tópico.

Em suma, este estudo conclui que, embora o uso de LLMs e modelagem de tópicos não substituam a leitura crítica do especialista, ele oferece uma ferramenta poderosa de "leitura distante"e prospecção. As técnicas empregadas conseguiram não apenas replicar a categorização temática humana, mas também sugerir novas conexões semânticas. Trabalhos futuros podem se beneficiar do refinamento dos parâmetros de clustering para reduzir

ainda mais a taxa de outliers sem sacrificar a coerência, bem como aprofundar a análise comparativa de similaridade direta entre as sentenças de Isidoro e os textos de Catão e Varrão para identificar citações não atribuídas com maior precisão.

Referências

[Aggarwal *et al.* 2001] Charu C. Aggarwal, Alexander Hinneburg e Daniel A. Keim. "On the surprising behavior of distance metrics in high dimensional spaces". In: *Proceedings of the 8th International Conference on Database Theory*. ICDT '01. Berlin, Heidelberg: Springer-Verlag, 2001, pp. 420–434. isbn: 3540414568 (citado na pg. 11).

[Allaoui *et al.* 2020] Mebarka Allaoui, Mohammed Lamine Kherfi e Abdelhakim Cheriet. "Considerably improving clustering algorithms using umap dimensionality reduction technique: a comparative study". In: jul. de 2020, pp. 317–325. isbn: 978-3-030-51934-6. doi: 10.1007/978-3-030-51935-3_34 (citado na pg. 11).

[Angelov 2020] Dimo Angelov. *Top2Vec: Distributed Representations of Topics*. 2020. arXiv: 2008.09470 [cs.CL]. url: https://arxiv.org/abs/2008.09470 (citado na pg. 11).

[Bahdanau *et al.* 2016] Dzmitry Bahdanau, Kyunghyun Cho e Yoshua Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate*. 2016. arXiv: 1409.0473 [cs.CL]. url: https://arxiv.org/abs/1409.0473 (citado nas pgs. 7, 8).

[Barney *et al.* 2006] Stephen A Barney, Wendy J Lewis, Jennifer A Beach e Oliver Berghof. *The etymologies of Isidore of Seville*. Cambridge University Press, 2006 (citado nas pgs. 1, 14, 17).

[Bengio *et al.* 2003] Yoshua Bengio, Réjean Ducharme, Pascal Vincent e Christian Janvin. "A neural probabilistic language model". In: *Journal of machine learning research*. 2003. url: https://api.semanticscholar.org/CorpusID:221275765 (citado nas pgs. 5, 8).

[David M Blei *et al.* 2003] David M Blei, Andrew Y Ng e Michael I Jordan. "Latent dirichlet allocation". *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022 (citado nas pgs. 1, 10).

[David M. Blei 2012] David M. Blei. "Probabilistic topic models". *Commun. ACM* 55.4 (abr. de 2012), pp. 77–84. issn: 0001-0782. doi: 10.1145/2133806.2133826. url: https://doi.org/10.1145/2133806.2133826 (citado na pg. 10).

[Campello *et al.* 2013] Ricardo J. G. B. Campello, Davoud Moulavi e Joerg Sander. "Density-based clustering based on hierarchical density estimates". In: *Advances in Knowledge Discovery and Data Mining*. Ed. por Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda e Guandong Xu. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 160–172. isbn: 978-3-642-37456-2 (citado na pg. 12).

[Cato e Varro 1934] Marcus Porcius Cato e Marcus Terentius Varro. *On Agriculture; On Agriculture*. Ed. por William Davis Hooper e Harrison Boyd Ash. Vol. 283. Loeb Classical Library. Latin text with facing English translation; revised edition. Cambridge, MA: Harvard University Press, 1934. isbn: 978-0674993136. doi: 10.4159/DLCL.cato-agriculture.1934 (citado na pg. 17).

[Chaerul Haviana *et al.* 2023] Sam Farisa Chaerul Haviana, Sri Mulyono e Badie'Ah. "The effects of stopwords, stemming, and lemmatization on pre-trained language models for text classification: a technical study". In: *2023 10th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*. 2023, pp. 521–527. doi: 10.1109/EECSI59885.2023.10295797 (citado na pg. 18).

[Cho *et al.* 2014] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau e Yoshua Bengio. *On the Properties of Neural Machine Translation: Encoder-Decoder Approaches*. 2014. arXiv: 1409.1259 [cs.CL]. url: https://arxiv.org/abs/1409.1259 (citado na pg. 7).

[Deerwester *et al.* 1990] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer e Richard Harshman. "Indexing by latent semantic analysis". *Journal of the American society for information science* 41.6 (1990), pp. 391–407 (citado nas pgs. 1, 5).

[Devlin *et al.* 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee e Kristina Toutanova. "Bert: pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics*. 2019, pp. 4171–4186 (citado nas pgs. 1, 10).

[Feng *et al.* 2022] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan e Wei Wang. *Language-agnostic BERT Sentence Embedding*. 2022. arXiv: 2007.01852 [cs.CL]. url: https://arxiv.org/abs/2007.01852 (citado na pg. 18).

[Firth 1957] J. R. Firth. "A synopsis of linguistic theory 1930-55." 1952-59 (1957), pp. 1–32 (citado na pg. 5).

[Gemini Team 2025] Gemini Team. *Gemini: A Family of Highly Capable Multimodal Models*. 2025. arXiv: 2312.11805 [cs.CL]. url: https://arxiv.org/abs/2312.11805 (citado na pg. 10).

[Grootendorst 2022] Grootendorst. "Bertopic: neural topic modeling with a class-based tf-idf procedure". *arXiv preprint arXiMaartenv:2203.05794* (2022) (citado na pg. 11).

[Harris 1954] Zellig S. Harris. "Distributional structure". *WORD* 10.2-3 (1954), pp. 146–162. DOI: 10.1080/00437956.1954.11659520. eprint: https://doi.org/10.1080/00437956.1954.11659520. URL: https://doi.org/10.1080/00437956.1954.11659520 (citado na pg. 5).

[Herrmann *et al.* 2022] Moritz Herrmann, Daniyal Kazempour, Fabian Scheipl e Peer Kröger. *Enhancing cluster analysis via topological manifold learning*. 2022. arXiv: 2207.00510 [cs.LG]. URL: https://arxiv.org/abs/2207.00510 (citado na pg. 11).

[Honnibal *et al.* 2020] Matthew Honnibal, Ines Montani, Sofie Van Landeghem e Adam Boyd. "spaCy: industrial-strength natural language processing in python". *arXiv preprint arXiv:2002.06201* (2020). URL: https://arxiv.org/abs/2002.06201 (citado na pg. 18).

[Jelodar *et al.* 2019] Hamed Jelodar *et al.* "Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey". *Multimedia tools and applications* 78 (2019), pp. 15169–15211 (citado na pg. 1).

[Jolliffe e Cadima 2016] Ian T. Jolliffe e Jorge Cadima. "Principal component analysis: a review and recent developments". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065 (abr. de 2016), p. 20150202. ISSN: 1364-503X. DOI: 10.1098/rsta.2015.0202. eprint: https://royalsocietypublishing.org/rsta/article-pdf/doi/10.1098/rsta.2015.0202/1381479/rsta.2015.0202.pdf. URL: https://doi.org/10.1098/rsta.2015.0202 (citado na pg. 12).

[Krizhevsky *et al.* 2012] Alex Krizhevsky, Ilya Sutskever e Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks". In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'12. Lake Tahoe, Nevada: Curran Associates Inc., 2012, pp. 1097–1105 (citado na pg. 4).

[Maaten *et al.* 2008] Laurens van der Maaten, Geoffrey Hinton e Yoesoep Rachmad. "Viualizing data using t-sne". *Journal of Machine Learning Research* 9 (nov. de 2008), pp. 2579–2605 (citado na pg. 12).

[McCulloch e Pitts 1943] Warren S. McCulloch e Walter Pitts. "A logical calculus of the ideas immanent in nervous activity". *Bulletin of Mathematical Biophysics* 5 (1943), pp. 115–133 (citado na pg. 3).

[McInnes *et al.* 2018] Leland McInnes, John Healy e James Melville. "Umap: uniform manifold approximation and projection for dimension reduction". *arXiv preprint arXiv:1802.03426* (2018) (citado na pg. 12).

[Mikolov, Chen *et al.* 2013] Tomas Mikolov, Kai Chen, Greg Corrado e Jeffrey Dean. *Efficient Estimation of Word Representations in Vector Space*. 2013. arXiv: 1301.3781 [cs.CL]. URL: https://arxiv.org/abs/1301.3781 (citado na pg. 5).

[Mikolov, Sutskever *et al.* 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado e Jeffrey Dean. *Distributed Representations of Words and Phrases and their Compositionality*. 2013. arXiv: 1310.4546 [cs.CL]. url: https://arxiv.org/abs/1310.4546 (citado na pg. 6).

[Nussbaum e Duderstadt 2025] Zach Nussbaum e Brandon Duderstadt. *Training Sparse Mixture Of Experts Text Embedding Models*. 2025. arXiv: 2502.07972 [cs.CL]. url: https://arxiv.org/abs/2502.07972 (citado na pg. 18).

[Radford e Narasimhan 2018] Alec Radford e Karthik Narasimhan. "Improving language understanding by generative pre-training". In: 2018. url: https://api.semanticscholar.org/CorpusID:49313245 (citado na pg. 10).

[Radovanović *et al.* 2010] Miloš Radovanović, Alexandros Nanopoulos e Mirjana Ivanović. "Hubs in space: popular nearest neighbors in high-dimensional data". *J. Mach. Learn. Res.* 11 (dez. de 2010), pp. 2487–2531. issn: 1532-4435 (citado na pg. 11).

[Reimers e Gurevych 2019] Nils Reimers e Iryna Gurevych. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. 2019. arXiv: 1908.10084 [cs.CL]. url: https://arxiv.org/abs/1908.10084 (citado na pg. 10).

[Rosenblatt 1958] F. Rosenblatt. *The Perceptron: A Theory of Statistical Separability in Cognitive Systems (Project Para)*. Cornell Aeronautical Laboratory, Inc. Cornell Aeronautical Laboratory, 1958. url: https://books.google.com.br/books?id=7Q4-AQAAIAAJ (citado na pg. 4).

[Searle 1980] John R. Searle. "Minds, brains, and programs". *Behavioral and Brain Sciences* 3.3 (1980), pp. 417–457 (citado na pg. 4).

[Sia *et al.* 2020] Suzanna Sia, Ayush Dalmia e Sabrina J. Mielke. *Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too!* 2020. arXiv: 2004.14914 [cs.CL]. url: https://arxiv.org/abs/2004.14914 (citado na pg. 11).

[Steinbach *et al.* 2003] Michael Steinbach, Levent Ertöz e Vipin Kumar. "The challenges of clustering high dimensional data". *Univ. Minnesota Supercomp. Inst. Res. Rep.* 213 (jan. de 2003). doi: 10.1007/978-3-662-08968-2_16 (citado na pg. 11).

[Sturua *et al.* 2024] Saba Sturua *et al. jina-embeddings-v3: Multilingual Embeddings With Task LoRA*. 2024. arXiv: 2409.10173 [cs.CL]. url: https://arxiv.org/abs/2409.10173 (citado nas pgs. 18, 24).

[Sutskever *et al.* 2014] Ilya Sutskever, Oriol Vinyals e Quoc V. Le. *Sequence to Sequence Learning with Neural Networks*. 2014. arXiv: 1409.3215 [cs.CL]. url: https://arxiv.org/abs/1409.3215 (citado nas pgs. 7, 8).

[Thayer 2025] William P. Thayer. *LacusCurtius: A Gateway to Ancient Rome*. penelope.uchicago.edu/Thayer/E/Roman/home.html. Bill Thayer's Web Site, hosted by the University of Chicago. Last updated: 20 October 2025 (according to the main site). 2025. (Acesso em 08/11/2025) (citado na pg. 17).

[Turing 1950] Alan Mathison Turing. "Computing machinery and intelligence". *Mind* 49 (1950), pp. 433–460 (citado na pg. 3).

[Vaswani *et al.* 2017] Ashish Vaswani *et al.* "Attention is all you need". *Advances in neural information processing systems* 30 (2017) (citado nas pgs. 1, 4, 8).

[Wang *et al.* 2024] Liang Wang *et al.* "Multilingual e5 text embeddings: a technical report". *arXiv preprint arXiv:2402.05672* (2024) (citado na pg. 18).