Análise de Textos Históricos utilizando LLMs e Modelagem de Tópicos

Introdução

Neste trabalho, iremos investigar conexões entre tópicos de interesse de historiadores sobre a obra conhecida como "Etimologias", de Isidoro de Sevilha (c.560-636), que é uma compilação de 20 livros sobre as origens das palavras, em que o autor buscou registrar o conhecimento de escritores latinos da Antiguidade Clássica, como Varrão e Plínio o Velho. Esta obra é considerada a primeira grande enciclopédia da Idade Média, e foi copiada exaustivamente ao longo de cerca de 700 anos para ser utilizada como livro-texto base nas instituições de ensino da época.

Modelagem de tópicos

Para a modelagem dos tópicos, utilizamos o BERTopic (Grootendorst, 2022). Esta é uma técnica de modelagem de tópicos em linguagem natural que baseia-se (basicamente) em 4 passos:

1. Extração de representações semânticas vetoriais de sentenças (*text embeddings*);
2. Redução de dimensionalidade;
3. *Clustering* das projeções de baixa dimensionalidade dos *embeddings*;
4. Extração de representações textuais dos *clusters*.

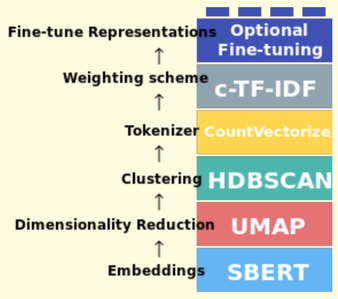


Figure 1: *Representação modular do BERTopic.*

1. Mapeamento semântico

Modelos de redes neurais artificiais denominados *word* e *sentence embedders* são capazes de extrair informação semântica de palavras e frases em linguagem natural, criando representações vetoriais.

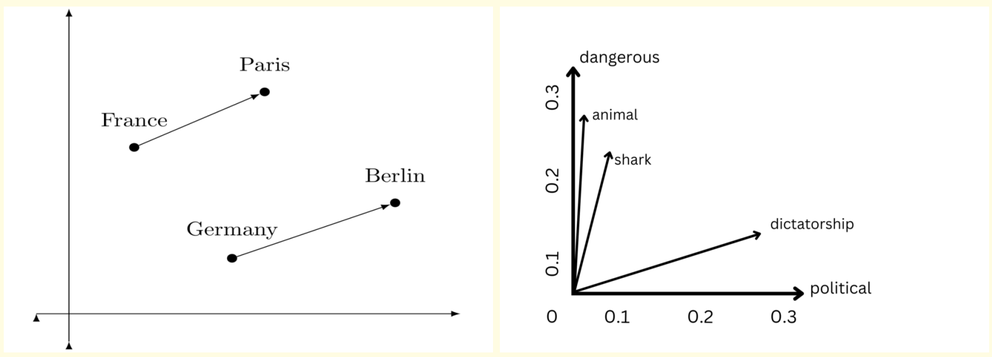
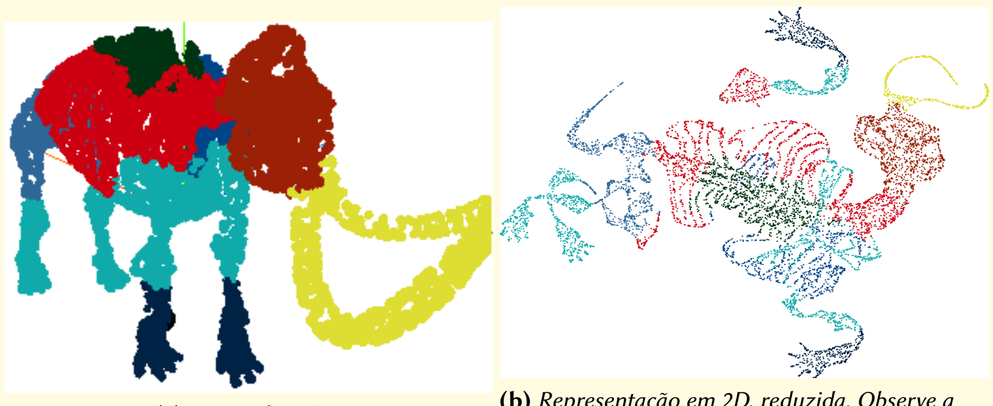


Figure 2: *Exemplos de embeddings.*

Essas representações são construídas de tal maneira que sentenças (ou palavras) com significados análogos têm suas representações vetoriais próximas em um espaço latente de *embeddings*.

2. Redução de dimensionalidade: UMAP

Os modelos de embedding tipicamente projetam sentenças em espaços de alta dimensionalidade ($d \approx 1000$, tipicamente). Devido à **maldição da dimensionalidade**, aplicamos uma técnica de redução de dimensionalidade denominada UMAP (McInnes, Healy, and Melville, 2018), para viabilizar a etapa posterior de *clustering* (e também para visualização).



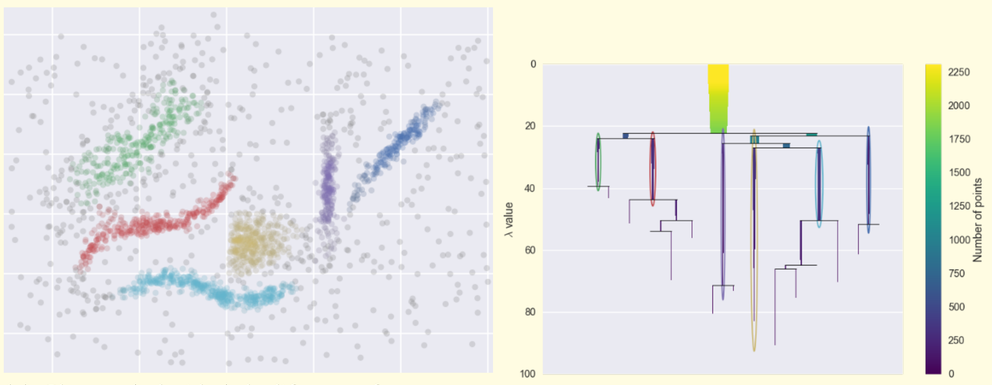
(a) *Original em 3D.*

(b) *Representação em 2D, reduzida. Observe a preservação de estruturas locais.*

Figure 3: *Exemplo de visualização do UMAP.*

3. *Clustering*: HDBSCAN

Para a etapa de aglomeração dos documentos, utilizamos o HDBSCAN (Campello et al., 2013), que é um algoritmo de *clustering* hierárquico baseado em densidade. Utilizamos essa abordagem para sermos capazes de detectar *clusters* de formatos variados, não-convexos.



(a) *Clusters (coloridos) de diferentes formas e densidades. Observe a presença de pontos classificados como ruído (cinza).*

(b) *Hierarquia de clusters.*

Figure 4: *Exemplo de clusterização com HDBSCAN.*

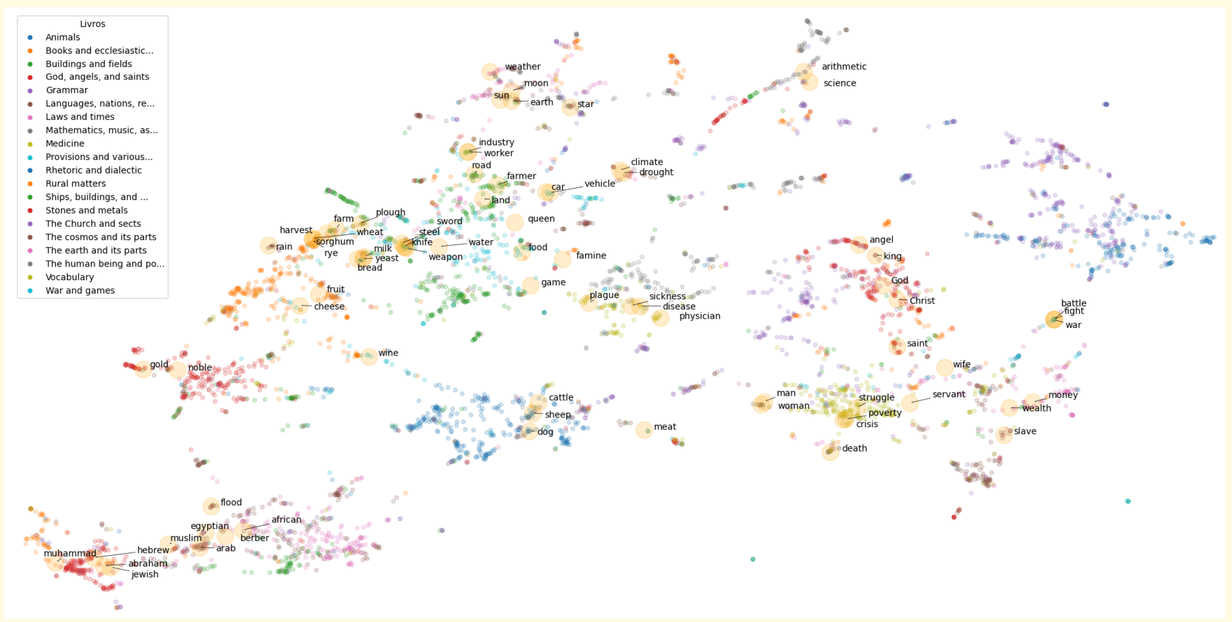
4. Representação

Ao final, calculamos uma variação da métrica TF-IDF (Salton and Buckley, 1988), denominada cTF-IDF, e selecionamos os termos com maior pontuação para representar cada cluster:

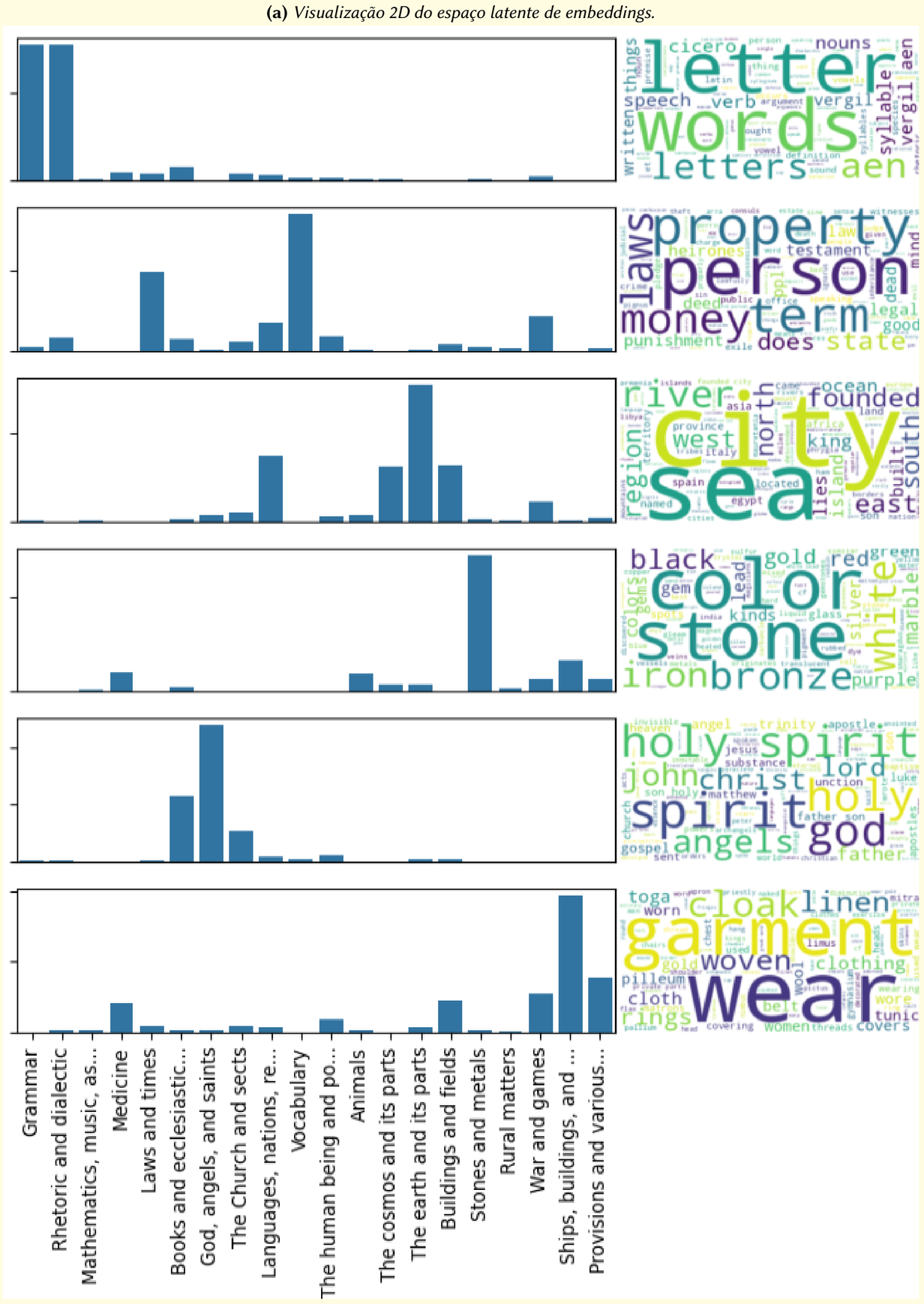
$$W_{t,c} = tf_{t,c} \cdot \log(1 + \frac{A}{tf_t})$$

Onde $tf_{t,c}$ é a frequência do termo t no cluster c, tf_t é a frequência geral do termo t no corpus, e A é o número médio de termos por cluster.

Resultados



(a) *Visualização 2D do espaço latente de embeddings.*



(b) *Distribuição de tópicos por livro.*

Bibliografia

- Campello, Ricardo J. G. B. et al. (2013). "Density-Based Clustering Based on Hierarchical Density Estimates". In: *Advances in Knowledge Discovery and Data Mining*. Ed. by Jian Pei et al. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 160–172. ISBN: 978-3-642-37456-2.
- Grootendorst (2022). "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". *arXiv preprint arXiMaartenv:2203.05794*.
- Grootendorst, Maarten (2024). *BERTopic Algorithm*. Página oficial da documentação. URL: https://maartengr.github.io/BERTopic/algorithm/algorithm.html (visited on 10/28/2025).
- McInnes, Leland, John Healy, and Steve Astels (2016). *HDBSCAN - How it works*. Documentação oficial do pacote HDBSCAN para Python. URL: https://hdbscan.readthedocs.io/en/latest/advanced_hdbscan.html (visited on 10/28/2025).
- McInnes, Leland, John Healy, and James Melville (2018). "Umap: Uniform manifold approximation and projection for dimension reduction". *arXiv preprint arXiv:1802.03426*.
- PAIR Code, Google Research (2019). *Understanding UMAP*. URL: https://pair-code.github.io/understanding-umap/ (visited on 10/28/2025).
- Salton, Gerard and Christopher Buckley (1988). "Term-weighting approaches in automatic text retrieval". *Information Processing and Management* 24.5, pp. 513–523. ISSN: 0306-4573. DOI: https://doi.org/10.1016/0306-4573(88)90021-0. URL: https://www.sciencedirect.com/science/article/pii/0306457388900210.
- Wikimedia Commons (2007). *Distributional semantics*. Licença: Creative Commons Attribution-Share Alike 4.0 International. URL: https://commons.wikimedia.org/wiki/File:Distributional_semantics.png (visited on 10/28/2025).
- – (2020). *Word embedding illustration*. Licença: Creative Commons Attribution-Share Alike 4.0 International. URL: https://commons.wikimedia.org/wiki/File:Word_embedding_illustration.svg (visited on 10/28/2025).

Autor: João Pedro Lukasavicus Silva joao.lukasavicus.silva@usp.br
Orientador: Mateus Espadoto mespadot@ime.usp.br
Departamento de Ciência da Computação — Instituto de Matemática e Estatística, Universidade de São Paulo