

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*.

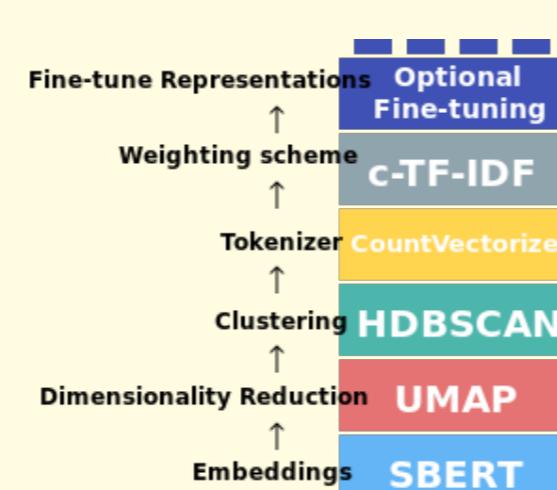


Figure 1: Representação modular do BERTopic

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.

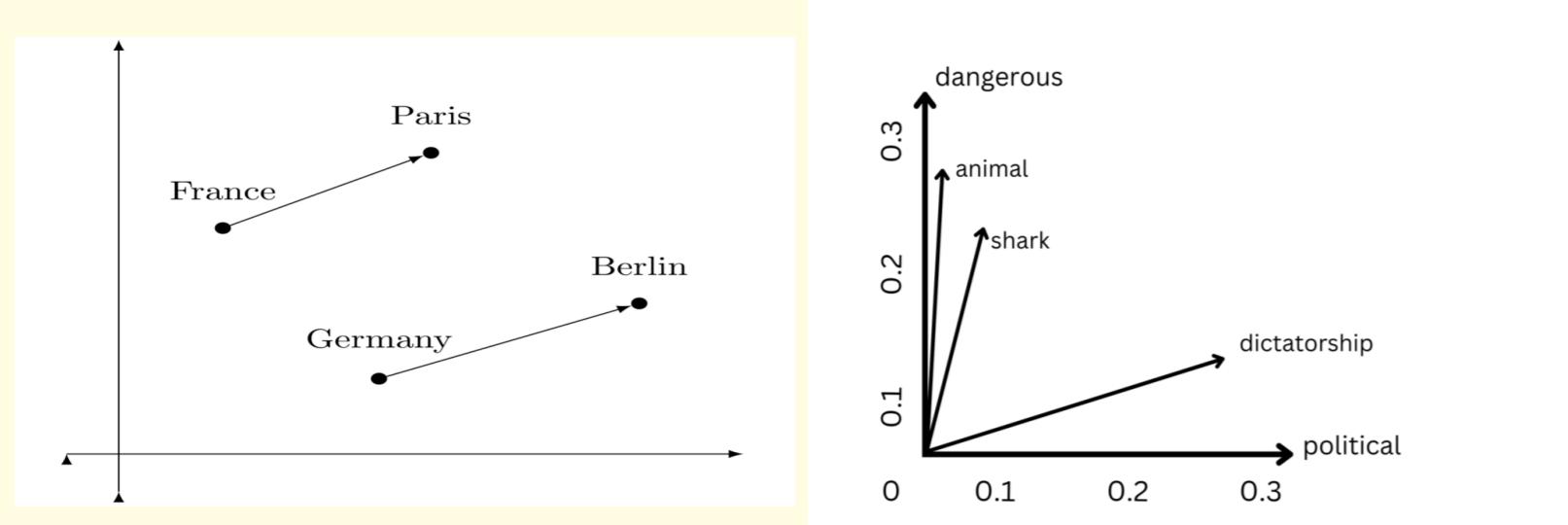


Figure 2: Exemplos de embeddings

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 ($d = 1000$, tipicamente). Devido à **maldição da dimensionalidade**, aplicamos uma técnica de redução de dimensionalidade denominada UMAP (McInnes et al., 2018), para viabilizar a etapa posterior de *clustering* (e também para visualização).

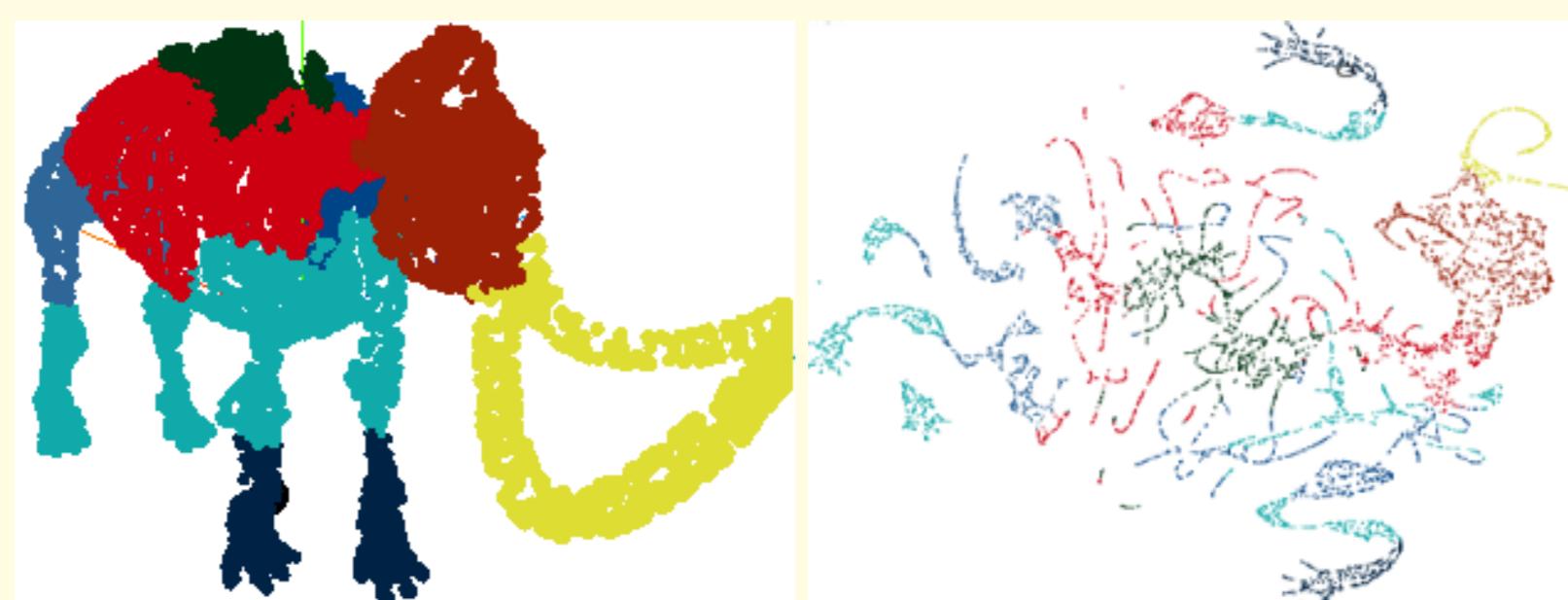


Figure 3: Exemplo de visualização do UMAP

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.

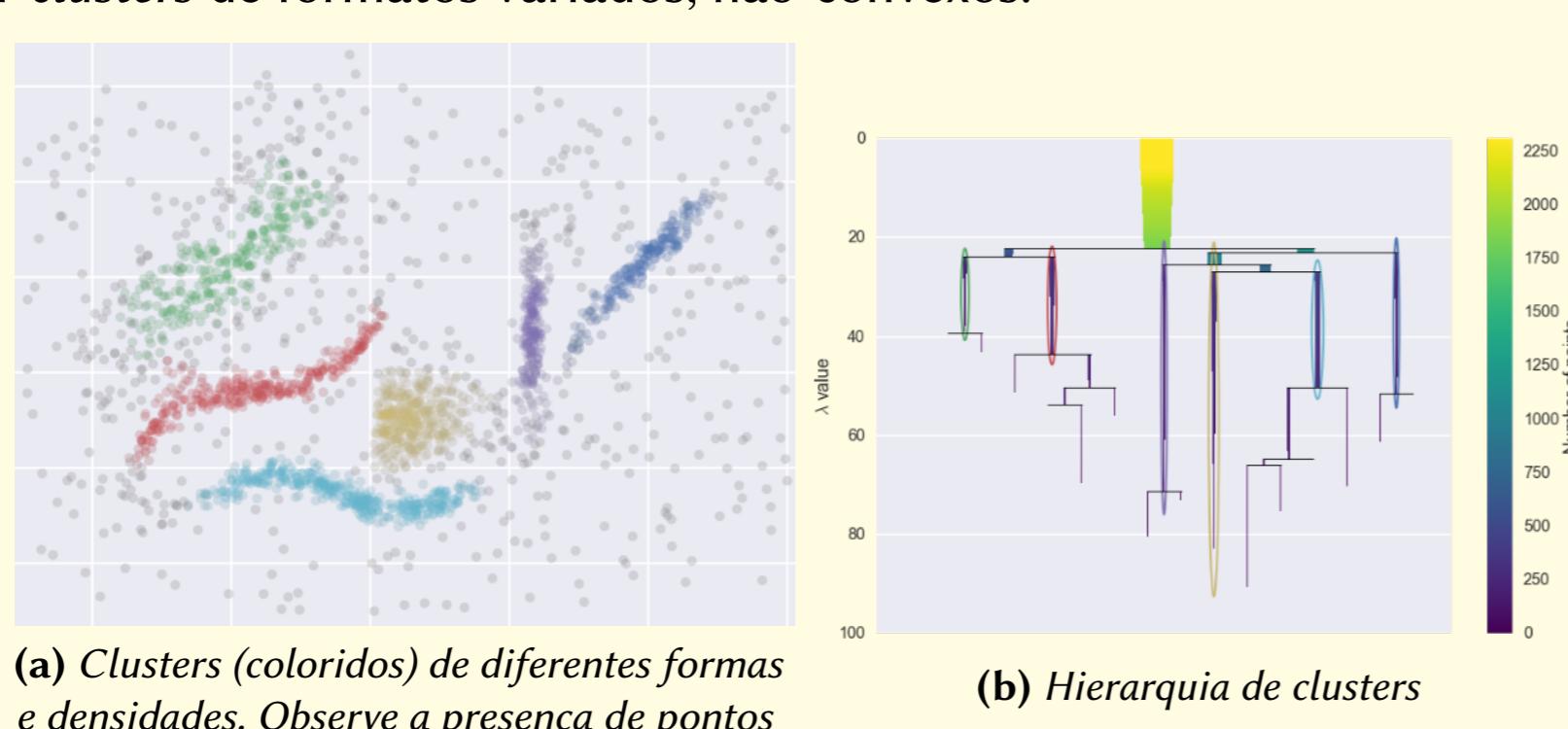
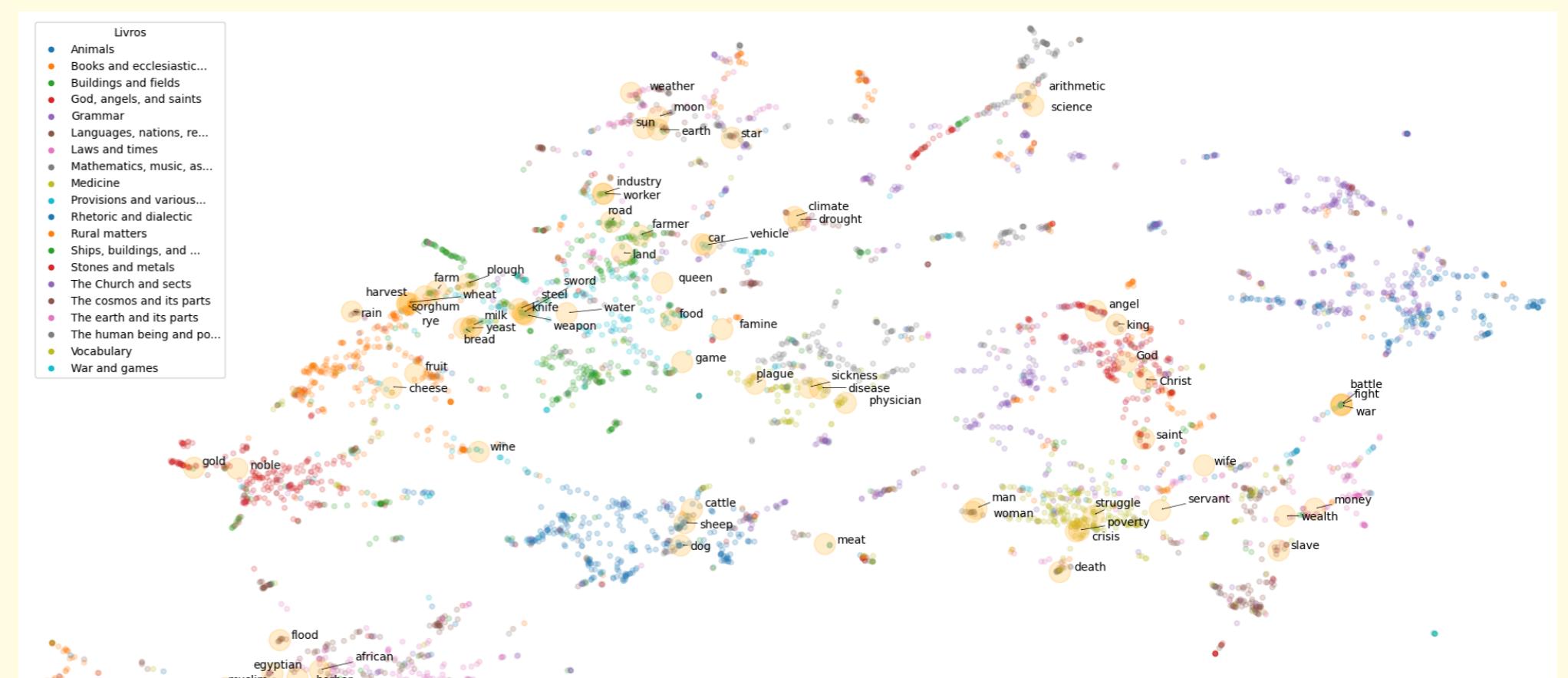
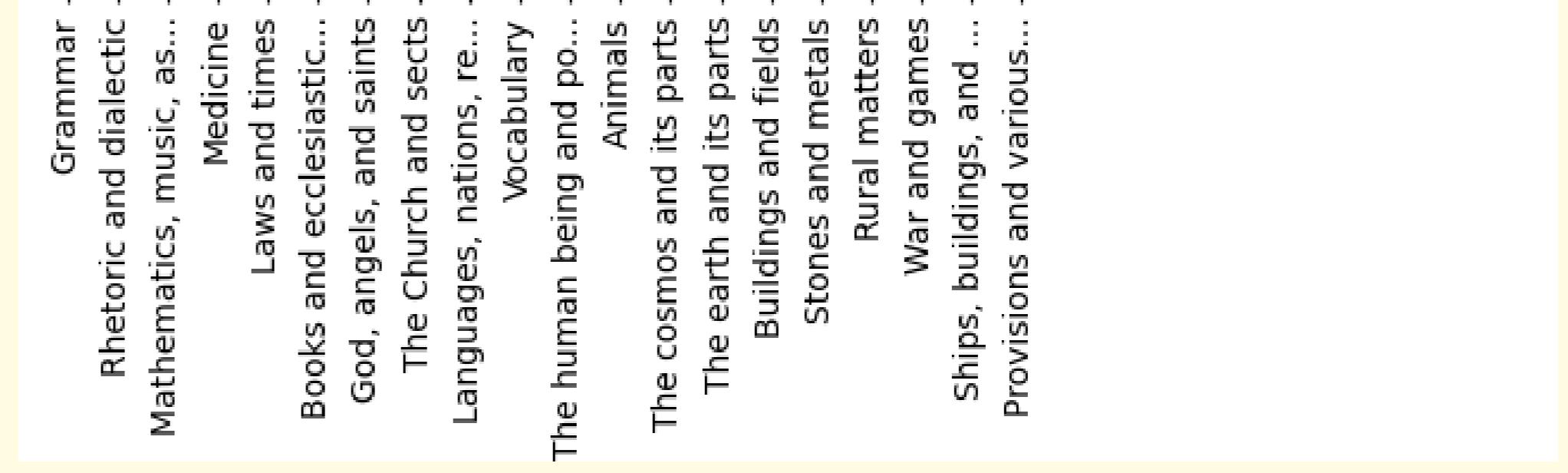
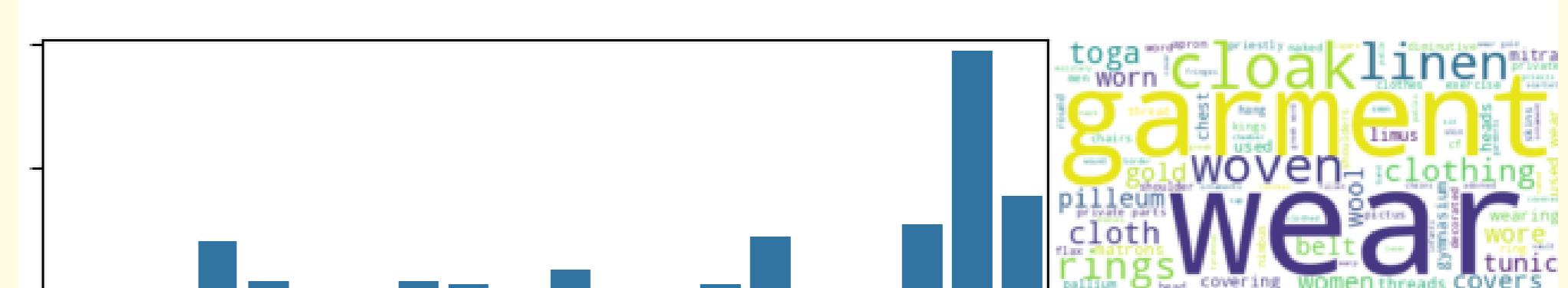
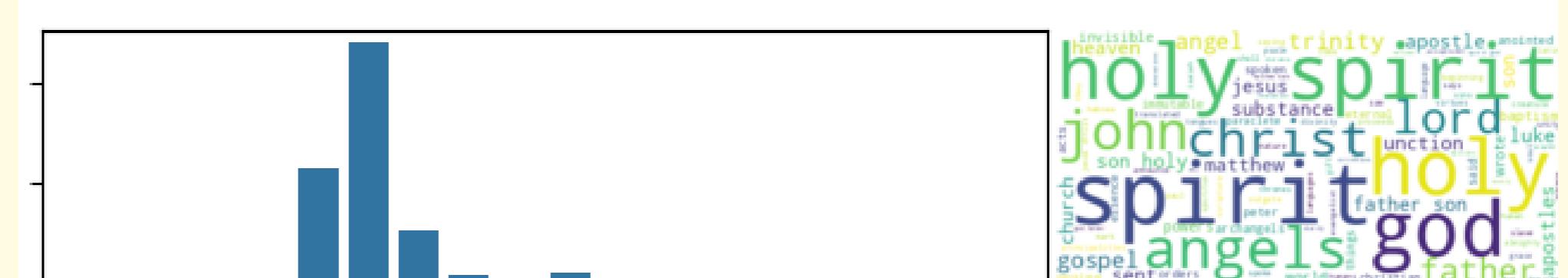
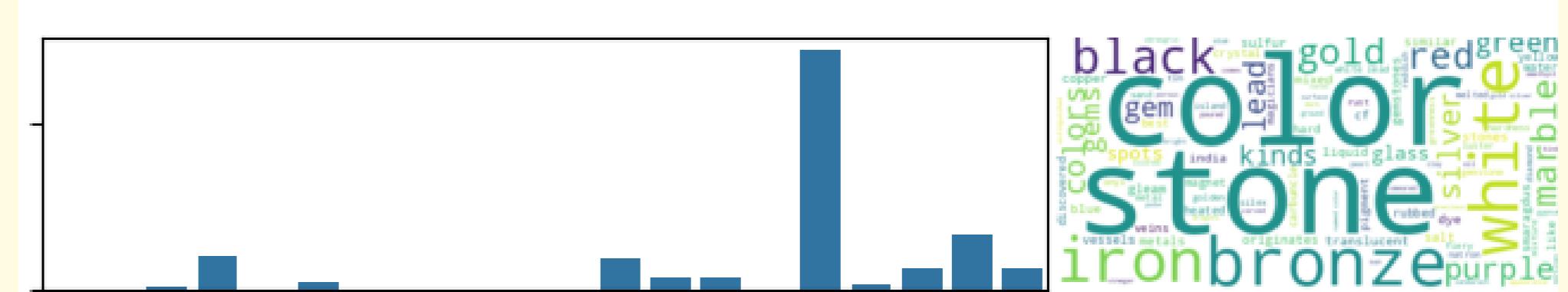
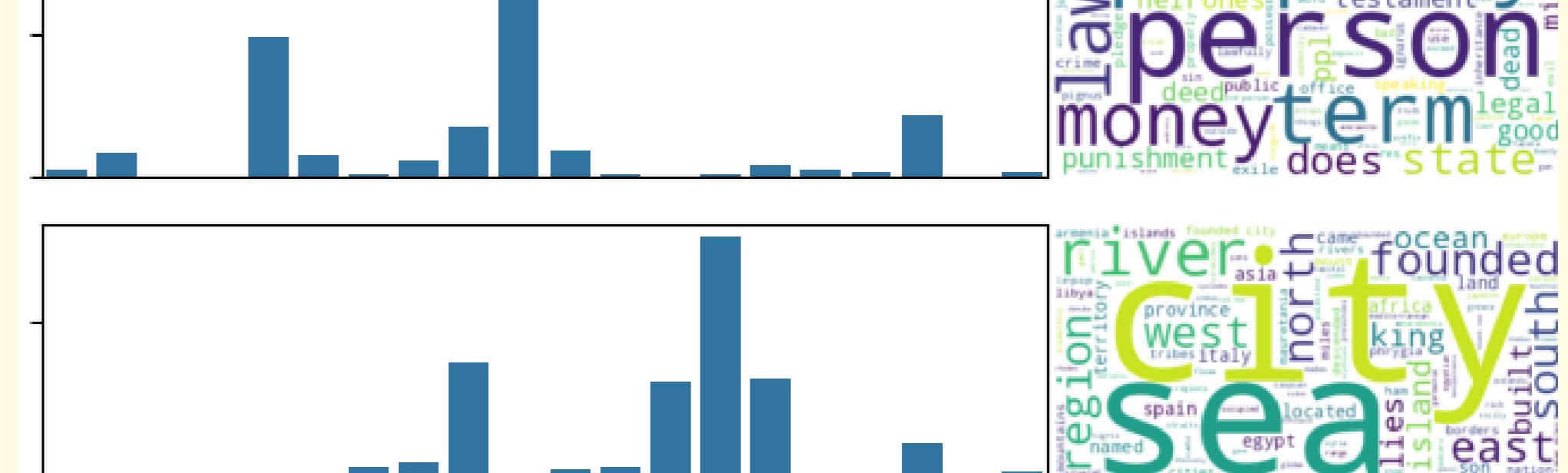
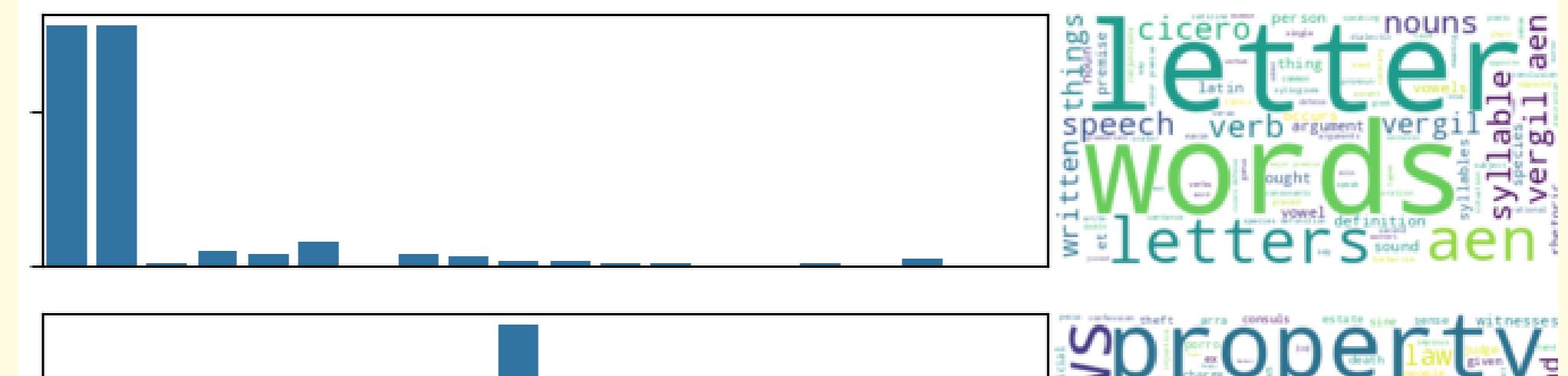


Figure 4: Exemplo de clusterização com HDBSCAN

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, Maarten (2022). "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". *arXiv preprint arXiv:2203.05794*.
- McInnes, Leland et al. (2018). "Umap: Uniform manifold approximation and projection for dimension reduction". *arXiv preprint arXiv:1802.03426*.