Proposta - Trabalho de Formatura Supervisionado
Modelagem de Tópicos em Textos Históricos utilizando LLMs

2025

João Pedro Lukasavicus Silva (IME - USP)

Orientador: Mateus Espadoto (IME - USP)

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 (NLP) possui métodos bem estabelecidos para análise semântica de textos, como por exemplo, modelagem de tópicos [DDF+90, BNJ03, JWY+19]. No entanto, o uso destas técnicas requer certa familiaridade com o seu funcionamento, para que sejam feitos os ajustes necessários para a obtenção de bons resultados.

Com o surgimento de grandes modelos de linguagem (LLM) baseados em transformers [VSP+17, DCLT19], é possível observar um grande salto em termos de qualidade das ferramentas e métodos, 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.

2. Objetivos

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.

O texto original é em latim medieval, mas para este trabalho será utilizada a tradução para a língua inglesa das Etimologias [BLBB06], por conveniência.

Para o estudo de conexões e tópicos existentes na obra, serão utilizadas ferramentas como grandes modelos de linguagem para geração de *embeddings*, como Jina V3 [SMA+24], bibliotecas de modelagem de tópicos baseadas em transformers, como BERTopic [Gro22], UMAP [MHM18] para visualização dos *embeddings* em duas dimensões, e Graphviz [GN00] para visualização das conexões na forma de grafos.

3. Plano de trabalho e cronograma

As atividades a serem realizadas estão resumidas abaixo. Tanto a lista de atividades quanto o cronograma estão sujeitos a alterações.

1. Conversas com especialistas para entender questões de interesse
2. Estudo das ferramentas a serem utilizadas no projeto
3. Preparação do texto para processamento
4. Criação de ferramenta para modelagem de tópicos
5. Análise dos resultados e validação com especialistas
6. Apresentação do trabalho
7. Produção do texto final

Atividades	Maio	Jun.	Jul.	Ago.	Set.	Out.	Nov.	Dez.
1	■	■						
2	■	■						
3		■	■					
4			■	■	■			
5					■	■		
6						■		
7						■	■	

Tabela 1. Cronograma de execução mensal, de maio a dezembro de 2025

Referências

[BLBB06] Stephen A Barney, Wendy J Lewis, Jennifer A Beach, and Oliver Berghof. *The etymologies of Isidore of Seville*. Cambridge University Press, 2006.

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[DCLT19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and 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*, pages 4171–4186, 2019.

[DDF+90] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[GN00] Emden R Gansner and Stephen C North. An open graph visualization system and its applications to software engineering. *Software: practice and experience*, 30(11):1203–1233, 2000.

[Gro22] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[JWY+19] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia tools and applications*, 78:15169–15211, 2019.

[MHM18] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[SMA+24] Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. jina-embeddings-v3: Multilingual embeddings with task lora, 2024.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.