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

Modelagem de Tópicos em Textos Históricos utilizando LLMs

João Pedro Lukasavicus Silva

Monografia Final

mac 499 — Trabalho de Formatura Supervisionado

Supervisor: Mateus Espadoto

São Paulo

Esta seção é opcional e fica numa página separada;
ela pode ser usada para uma dedicatória ou epígrafe.

Agradecimentos

Do. Or do not. There is no try.
— Mestre Yoda

Texto texto. Texto opcional.

Resumo

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

Elemento obrigatório, constituído de uma sequência de frases concisas e objetivas, em forma de texto. Deve apresentar os objetivos, métodos empregados, resultados e conclusões. O resumo deve ser redigido em parágrafo único, conter no máximo 500 palavras e ser seguido dos termos representativos do conteúdo do trabalho (palavras-chave). Deve ser precedido da referência do documento. Texto texto. Texto texto.

Palavras-chave: Palavra-chave1. Palavra-chave2. Palavra-chave3.

Abstract

João Pedro Lukasavicus Silva. **Title of the document: *a subtitle***. Capstone Project Report (Bachelor). Institute of Mathematics and Statistics, University of São Paulo, São Paulo.

Elemento obrigatório, elaborado com as mesmas características do resumo em língua portuguesa. De acordo com o Regimento da Pós-Graduação da USP (Artigo 99), deve ser redigido em inglês para fins de divulgação. É uma boa ideia usar o sítio www.grammarly.com na preparação de textos em inglês. Text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text. Text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text text.

Keywords: Keyword1. Keyword2. Keyword3.

Lista de abreviaturas

CFT	Transformada contínua de Fourier (*Continuous Fourier Transform*)
DFT	Transformada discreta de Fourier (*Discrete Fourier Transform*)
EIIP	Potencial de interação elétron-íon (*Electron-Ion Interaction Potentials*)
STFT	Transformada de Fourier de tempo reduzido (*Short-Time Fourier Transform*)
ABNT	Associação Brasileira de Normas Técnicas
URL	Localizador Uniforme de Recursos (*Uniform Resource Locator*)
IME	Instituto de Matemática e Estatística
USP	Universidade de São Paulo

Lista de símbolos

ω	Frequência angular
ψ	Função de análise *wavelet*
Ψ	Transformada de Fourier de ψ

Lista de figuras

Lista de tabelas

Lista de programas

Sumário

Introdução

Capítulo 1

Background

Capítulo 2

Experimentos

2.1 Conjunto de dados

2.1.1 Textos históricos

O conjunto de dados utilizados neste trabalho consiste nas obras Etimologias - Isidoro de Sevilha, ... TODO: lista de obras

Todos os arquivos estão em formato texto puro, e foram obtidos em... TODO: Como pode-se obter esses arquivos.

2.1.2 Preparação dos dados

Primeiramente, pelo modo como cada livro está formatado, fazemos um processamento inicial para separar cada livro em capítulos, e juntar palavras com hífen. TODO: exemplos. Então, cada capítulo foi subdividido em sentenças usando a biblioteca Spacy... TODO: expandir. TODO: exemplos

Apesar de algumas tarefas em processamento de linguagem natural serem usadas em outras abordagens de modelagem de tópicos, como remoção de stop-words, lematização, etc, não fazemos esses procedimentos na fase de preparação dos dados para a abordagem utilizada neste trabalho, pois ... TODO: expandir. PERGUNTA: Devo explicar remoção de stop words e lematização?

Depois de um pré-processamento dos textos, terminando em sua divisão em sentenças, geramos embeddings para cada uma dessas sentenças. Os modelos usados para a geração desses embeddings foram: "sentence-transformers/LaBSE", "jinaai/jina-embeddings-v3", "intfloat/multilingual-e5-large-instruct", "nomic-ai/nomic-embed-text-v2-moe". Esses embeddings então foram armazenados em um banco de dados vetorial, juntamente com cada sentença, e alguns outros metadados, como nome do autor, nome da obra, e uma identificação do capítulo. Esses metadados servem para criar filtros para buscas posteriores nesse banco de dados.

A seguir, fazemos uma redução de dimensionalidade de cada um dos conjuntos de

embeddings gerados. Essa etapa é importante para o desempenho da etapa posterior do método, de clusterização, pois... TODO: expandir. Esses embeddings reduzidos são armazenados então, para sua utilização posterior.

2.2 Experimentos

Cada experimento consistiu em aplicar o BERTopic, utilizando os diferentes conjuntos de dados gerados pelos diferentes modelos de embedding, variando o número de tópicos a serem descobertos: 10, 20, 50 e 100 tópicos.

Capítulo 3

Resultados

Capítulo 4

Conclusão

Índice remissivo