UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

JOÃO GABRIEL ZANDONÁ

Uma abordagem focando na utilização de modelos Out-of-the-box para recuperação de informações em atas de reuniões

Monografia apresentada como requisito parcial para a obtenção do grau de Bacharel em Ciência da Computação

Orientador: Prof. Dr. Dennis Giovani Balreira

Porto Alegre
2024

RESUMO

O fácil acesso e manejo de informações contidas em atas de reuniões é essencial para a tomada de decisões conscientes e o gerenciamento eficaz de projetos. Em sistemas de Recuperação de Informações (RI) de menor escala, soluções tecnológicas *"out-of-the-box"* são frequentemente necessárias já que, em muitos casos, o custo e o esforço envolvidos na adaptação de um sistema de RI para um domínio específico podem não ser justificáveis dada a escala do projeto ou da organização. Este trabalho apresenta uma abordagem para a recuperação de informações contidas em atas de reuniões focada na utilização de modelos *"out-of-the-box"*. A abordagem foi aplicada no contexto das reuniões do Instituto de Informática da Universidade Federal do Rio Grande do Sul (UFGRS), abrangendo reuniões do Conselho do Instituto de Informática (CONINF), Programa de Pós-Graduação (PPGC), e Colegiado do Departamento de Informática Aplicada. A aplicação da abordagem apresentou um desempenho razoável na tarefa de recuperar as informações presentes nas atas. Mesmo que utilizando um pipeline simples, a avaliação qualitativa mostrou certo um grau de precisão e revocação com o uso da abordagem, mesmo que limitado.

Palavras-chave: Processamento de Linguagem Natural. Recuperação de Informação. Ata de reuinão. Out-of-the-box.

Using LaTeX to Prepare Documents at II/UFRGS

ABSTRACT

Easy access and management of information contained in meeting minutes is essential for making informed decisions and effective project management. In smaller-scale Information Retrieval (IR) systems, *"out-of-the-box"* technological solutions are often necessary as, in many cases, the cost and effort involved in adapting an IR system to a specific domain may not be justifiable given the scale of the project or organization. This work presents an approach for recovering information contained in meeting minutes, focused on the use of models *"out-of-the-box"*. The approach was applied in the context of meetings at the Institute of Informatics of the Federal University of Rio Grande do Sul (UFGRS), covering meetings of the Council of the Institute of Informatics (CONINF), Postgraduate Program (PPGC), and Collegiate of the Department of Applied Informatics. The application of the approach showed reasonable performance in the task of retrieving the information present in the minutes. Even using a simple pipeline, the qualitative evaluation showed a certain degree of precision and recall with the use of the approach, even if limited.

Keywords: Natural Laguage Processing. Semantic information retrieval. Minute of meeting. Out-of-the-box.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

API	Application Programming Interface
BEIR	Benchmarking IR
BERT	Bidirectional Encoder Representations for Transformers
BM25	Best Match 25
CEI	Centro de Empreendimentos em Informática
CNNs	Convolutional Neural Networks
Colegiado	Colegiado do Departamento de Informática Aplicada
CONINF	Conselho do Instituto de Informática
CSV	Comma Separated Values
DCG	Discounted Cumulative Gain
GPT3	Generative Pretrained Transformer 3
GPT4	Generative Pretrained Transformer 4
INF	Instituto de informática da Universidade do Rio Grande do Sul
IDCG	Ideal Discounted Cumulative Gain
LLM	Large Language Models
MTEB	Massive Text Embedding Benchmark
NDCG	Normalized Discounted Cumulative Gain
PLN	Processamento de Linguagem Natural
PPGC	Programa de Pós Graduação em Computação
RI	Recuperção de informações
RNNs	Recurrent Neural Networks
SBERT	Sentence Bidirectional Encoder Representations for Transformers
TF-IDF	O Term Frequency-Inverse Document Frequency
UFRGS	Universidade Federal do Rio Grande do Sul

SUMÁRIO

1 INTRODUÇÃO

A gestão de informações presentes em documentos tem sido estudada há muito tempo na biblioteconomia, antes mesmo da criação dos computadores digitais. Todavia, com ajuda da criação e evolução dos computadores digitais, a gestão de informações se tornou uma tarefa muito mais robusta e escalável, uma vez que passou a ser facilmente automatizável.

As aplicações de RI são diversas e impactam várias áreas (PIVETTA, 2024). No contexto da web, motores de busca são exemplos clássicos de aplicação da RI, permitindo aos usuários encontrar conteúdo relevante em meio a bilhões de páginas da web. Além disso, a RI tem um papel crucial no âmbito acadêmico e de pesquisa, auxiliando na descoberta de literatura científica e recursos educacionais. Em empresas, sistemas de RI são utilizados para gerenciar grandes volumes de documentos e dados, facilitando o acesso a informações corporativas (BAUDRU; ROSELLO; BERSINI,). Conforme definido em (MOREIRA, 2023) "A tarefa central da RI é casar a consulta de um usuário com os documentos que são potencialmente relevantes a ela."

A área de pesquisa de RI vem se mostrando relevante há décadas (LAY, 1985; KOBAYASHI; TAKEDA, 2000; CHOR et al., 1998), abordando os desafios inerentes à RI assim como maneiras de solucioná-los. Atualmente a pesquisa em RI continua tendo grande relevância, além disso, assim como outras áreas da linguística computacional, vêm avançando de mãos dadas com o progresso na área de PLN e dos Modelos de Linguagem de Larga Escala (ZHAO et al., 2024).

1.1 Motivação

A aplicação da RI em atas de reuniões é um exemplo específico e altamente relevante de como essa tecnologia pode ser utilizada para gerenciar e acessar informações em um contexto organizacional. As atas de reuniões são documentos essenciais para qualquer organização, pois registram discussões, decisões, ações e responsabilidades acordadas durante reuniões. Com o aumento do volume e da frequência das reuniões em ambientes corporativos, governamentais e acadêmicos, torna-se crucial ter um sistema eficiente para recuperar informações específicas desses documentos.

Em cenários de menor escala, abordagens*out-of-the-box* para sistemas de RI são frequentemente necessárias e mais práticas. Isso se deve ao fato de que, em muitos ca-

sos, o custo e o esforço envolvidos na adaptação de um sistema de RI para um domínio específico podem não ser justificáveis dada a escala do projeto ou da organização. Em tais situações, soluções pré-configuradas ou genéricas, que requerem mínimo ou nenhum ajuste, podem ser mais adequadas.

Essas soluções prontas para uso permitem que pequenas empresas ou projetos com recursos limitados implementem rapidamente sistemas de RI sem a necessidade de investimentos significativos em desenvolvimento e personalização. Embora essas soluções possam não oferecer o mesmo nível de precisão ou eficiência que um sistema personalizado, elas ainda fornecem uma funcionalidade básica de RI que pode ser suficiente para atender às necessidades de muitos cenários.

Além disso, as abordagens modernas em RI que empregam LLM e redes neurais representam avanços significativos na capacidade de processar e entender grandes volumes de texto. No entanto, é importante ressaltar que essas técnicas inovadoras ainda não tornaram obsoletas outras abordagens mais tradicionais, especialmente em cenários de fora de domínio, sendo que essas tecnologias ainda são muito usadas em conjuntos com técnicas mais novas em modelos híbridos (FINARDI et al., 2024).

As soluções tecnológicas *out-of-the-box* são projetadas para serem versáteis, e acessíveis a usuários que não são especialistas em RI. Portanto, o estudo de abordagens de RI com este foco não é apenas relevante, mas também essencial para a democratização de tais tecnologias em um mundo cada vez mais orientado por dados.

1.2 Objetivos

Este trabalho apresenta uma abordagem, para a RI, contidas em atas de reuniões. Além disso, aborda a relevância e efetividade de soluções tecnológicas de perfil *out-of-the-box* aplicadas nesse contexto.

Os objetivos deste trabalho serão buscados por meio de: (i) Estudo da literatura na área de RI (ii) Desenvolvimento e implementação de uma abordagem de RI, focando na utilização de modelos *out-of-the-box* para a gestão e recuperação eficiente de informações a partir de atas de reuniões. (iii) Aplicação e validação da abordagem no contexto do Instituto de Informática da Universidade Federal do Rio Grande do Sul.

1.3 Estrutura do texto

O texto está organizado da seguinte forma: O Capítulo 2 explora conceitos básicos que são relevantes para o bom entendimento da metodologia. O Capítulo 3 explora trabalhos relacionados que mesmo com escopo diferente deste trabalho apresentam conteúdo relevante para a motivação e entendimento da abordagem explorada no quarto Capítulo. No Capítulo 4, é proposta uma abordagem de RI em atas de reunião. No Capítulo 5 essa abordagem é aplicada nas atas do Instituto de Informática da Universidade Federal do Rio Grande do Sul, e alguns experimentos são conduzidos para avaliar a performance dessa abordagem. No Capítulo 6 encontram-se as conclusões deste trabalho, assim como limitações, e trabalhos futuros.

2 CONCEITOS BÁSICOS

Este capítulo visa abordar os conceitos básicos relevantes para um bom entendimento das motivações e da metodologia utilizada neste trabalho.

2.1 Processamento de Linguagem Natural

Processamento de Linguagem Natural (PLN) é uma área de estudo e aplicação que se situa na interseção da linguística, ciência da computação e inteligência artificial. O objetivo do PLN é criar sistemas capazes de compreender, interpretar, manipular e gerar linguagem humana de maneira útil e significativa.

2.1.1 *Corpus*

No contexto de PLN e da linguística, um *corpus* (plural: corpora) é uma coleção de documentos escritos ou registros de fala usados como parte do processo de análise de linguagem. O *corpus* serve como uma fonte de dados representativa para pesquisa e desenvolvimento de sistemas de PLN, incluindo o treinamento dos modelos de linguagem. A natureza e a qualidade de um *corpus* são cruciais, pois influenciam diretamente nos resultados e na eficácia dos sistemas de PLN desenvolvidos com base nele.

2.1.2 *Dataset*

De acordo com (FREITAS, 2023) existem alguns critérios para considerar um corpus como um Dataset, são eles: tamanho, utilidade, e tipo de anotação. Quanto à terminologia, pode haver uma diferença devido ao foco da utilidade. O que a Linguística se refere como corpus anotado pode ser considerado um *Dataset* quando usado para treinar modelos de linguagem. Quando pensamos em *Dataset* para o PLN, eles são usados para avaliar e treinar ferramentas ou modelos. Os dados podem ser anotados por por humanos, ou anotados por máquinas e depois revisados por humanos. Nesse contexto, os conjuntos de dados são equivalentes a corpora padrão ouro.

2.1.3 Benchmarks

Benchmarks e datasets estão intimamente relacionados ao contexto PLN. Mais especificamente, um *benchmark* é uma ferramenta de avaliação que utiliza esses *datasets* com definições claras de tarefas e métricas para medir o desempenho de um sistema PLN. Em outras palavras, um *benchmark* pode ser visto como um "pacote"contendo três componentes principais: (i) Tarefas bem definidas, (ii) *Datasets*, (iii) Métricas.

Massive Text Embedding Benchmark(MTEB) (MUENNIGHOFF et al., 2023) é um benchmark massivo para medir o desempenho de modelos de *embedding* de texto em diversas tarefas. Esse *benchmark* inclui 58 *datasets* que cobrem um total de 112 línguas, incluindo a Língua Portuguesa.

BEIR (*Benchmarking IR*) (THAKUR et al., 2021) é um *benchmark* heterogêneo que contém diferentes tarefas de RI. Através do BEIR, é possível estudar sistematicamente as capacidades de generalização *zero-shot* de múltiplas abordagens de RI.

2.2 Modelos de linguagem de larga escala (LLM)

LLM, ou "Large Language Models"que traduzindo para o português significa Modelos de Linguagem de Larga Escala, são modelos avançados de inteligência artificial especializados no processamento e compreensão de linguagem natural. Eles são baseados em uma técnica de aprendizado de máquina chamada aprendizado profundo (*deep learning*), que permite que o modelo aprenda padrões complexos em grandes volumes de texto. Dessa maneira, esses modelos adquirem a habilidade de modelar aspectos da linguagem humana, como gramática, semântica, entre outros. Em geral, esses modelos são muito utilizados em abordagens de *transfer-learning*, por terem uma base de conhecimento extensa sobre a linguagem humana, já que eles são treinados mais de dados. Necessitando apenas de uma pequena etapa adicional de *fine-tunning* para aprender a performar conforme o objetivo específico para o qual é ajustado.

[1]

[1] *Transfer-learning* (aprendizado por transferência), é uma técnica de aprendizado de máquina que utiliza conhecimento adquirido em uma tarefa para melhorar o aprendizado ou o desempenho em uma segunda tarefa relacionada, mas diferente. A ideia central é que, ao invés de começar o processo de aprendizado do zero para cada tarefa, podemos reutilizar modelos previamente treinados em tarefas similares como ponto de partida, ajustando-os para tarefas específicas com menos dados ou recursos computacionais.

2.2.1 Treinamento dos LLM

2.2.1.1 ***Treinamento Prévio (Pre-training)***

Durante esta fase, o modelo é exposto a uma vasta quantidade de texto. Esses textos não são específicos a uma tarefa, mas cobrem uma ampla gama de tópicos e estilos. O objetivo é que o modelo aprenda com um entendimento geral da linguagem, incluindo gramática, uso de palavras e até certos aspectos do conhecimento do mundo. O processo de treinamento ajusta os pesos internos da rede neural para que o modelo possa prever partes de um texto com base no contexto fornecido. Esse treinamento é computacionalmente intensivo e pode exigir recursos significativos.

2.2.1.2 ***Ajuste Fino (Fine-tuning)***

Após o treinamento prévio, o modelo pode ser especializado em tarefas específicas. Este processo é conhecido como "fine-tuning". Durante o fine-tuning, o modelo é treinado em um conjunto de dados mais específico, que está diretamente relacionado à tarefa que o modelo precisa executar. Por exemplo, se o LLM deve ser utilizado para responder a perguntas médicas, ele seria ajustado com um conjunto de dados de perguntas e respostas médicas. Essa fase ajusta o modelo para que ele possa aplicar seu conhecimento geral da linguagem de forma mais eficaz a uma área específica. Em geral, a fase de fine-tunning é muito menos intensiva computacionalmente do que a fase de treinamento.

2.3 In-context Learning

In-context learning é uma habilidade muito poderosa presente nos LLM, referindo-se à capacidade desses modelos de aprender e adaptar-se com base no contexto fornecido durante a fase de inferência, sem a necessidade de treinamento explícito ou ajuste fino (BROWN et al., 2020).

Em vez de serem re-treinados para tarefas específicas, os modelos de transformadores aprendem a partir de exemplos que são fornecidos diretamente no contexto de sua entrada. Por exemplo, ao fornecer exemplos, um modelo pode aprender a utilizar um novo vocabulário, como mostrado em (BROWN et al., 2020). Apesar de sua flexibilidade, o *in-context learning* tem limitações. A qualidade da saída pode variar significativamente com base nos exemplos fornecidos (ZHOU et al., 2023) e também com a capacidade ine-

rente do modelo utilizado. Modelos com uma maior quantidade de parâmetros e treinados numa quantidade maior de dados costumam apresentar capacidades maiores, esse fato é estudado em diversos trabalhos como (KAPLAN et al., 2020), e amplamente aceito na comunidade científica.

Few-Shot Learning refere-se à habilidade do modelo executar uma tarefa com poucos exemplos. Os modelos de transformadores são particularmente bons em *few-shot learning*, adaptando-se rapidamente com apenas alguns exemplos.

Zero-Shot Learning refere-se à capacidade dos modelos de linguagem de realizar tarefas sobre as quais não foram explicitamente treinados, sem a necessidade de exemplos específicos de treinamento para cada tarefa (KOJIMA et al., 2022).

2.4 Embeddings

O conceito central por trás dos *embeddings* é a transformação de palavras, frases ou até documentos inteiros em vetores de números. Estes vetores representam de forma numérica os elementos linguísticos, permitindo que os algoritmos de computador processem e analisem a linguagem de maneira eficiente.

A ideia é que estes vetores capturem não apenas a identidade da palavra, mas também aspectos de seu significado, uso e relação com outras palavras. Por exemplo, em um espaço de *embeddings* bem construídos, palavras com significados semelhantes (como inteligente e esperto) estarão representadas por vetores próximos no espaço vetorial. Isso permite que o sistema de PLN reconheça semelhanças e diferenças semânticas entre diferentes palavras, como pode ser mostrado na Figura 2.1.

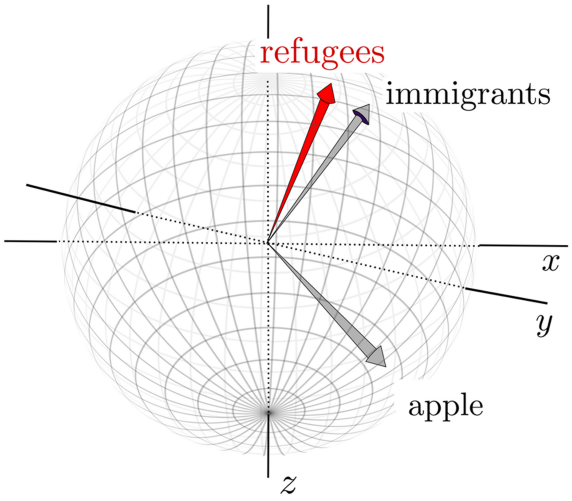
2.4.1 Embeddings Contextuais e Embeddings Não Contextuais

Embeddings podem ser classificados em dois tipos principais: contextuais e não contextuais. Esta classificação está relacionada à forma como as representações vetoriais das palavras são construídas e ao tipo de informação que elas capturam.

2.4.1.1 ***Embeddings Não Contextuais***

Neste tipo, cada palavra é representada por um único vetor, independentemente do contexto em que aparece. Isso significa que, não importa em quantos contextos diferentes

Figura 2.1 – Representação dos *embeddings* no Espaço Vetorial



Fonte: (DURRHEIM et al., 2023)

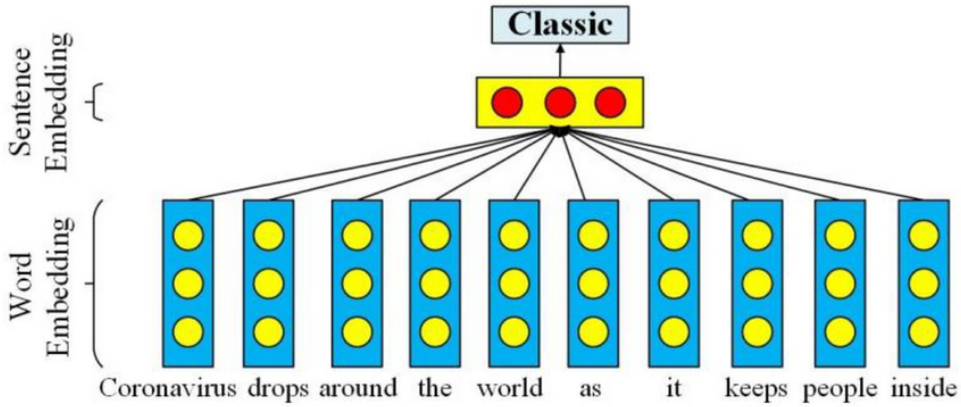
uma palavra seja usada, ela sempre terá a mesma representação vetorial. Modelos como Word2Vec (MIKOLOV et al., 2013) e GloVe (PENNINGTON; SOCHER; MANNING, 2014) são exemplos de *embeddings* não-contextuais. Eles são treinados para capturar informações gerais sobre as palavras, como similaridade semântica e relações sintáticas, com base em como as palavras co-ocorrem em grandes corpora de texto. No entanto, essa abordagem tem a limitação de não conseguir capturar os diferentes significados que uma palavra pode ter em diferentes contextos. Por exemplo, a palavra "banco"pode se referir a uma instituição financeira ou a um assento.

2.4.1.2 ***Embeddings Contextuais***

Diferentemente dos não-contextuais, os *embeddings* contextuais fornecem representações de palavras que variam conforme o contexto em que estão inseridas. Isso significa que a representação vetorial de uma palavra pode mudar dependendo das palavras que a cercam na frase. Modelos como BERT (DEVLIN et al., 2019) e GPT3 (BROWN et al., 2020) utilizam essa abordagem, que permite captar nuances de significado e uso das palavras em diferentes contextos. Por exemplo, em um modelo contextual, a palavra "banco"teria representações vetoriais diferentes quando usada em "Ele sentou no banco"e "Ele foi ao banco sacar dinheiro", refletindo os diferentes significados que assume em cada situação.

Os *embeddings* contextuais representaram um avanço no PLN, pois proporcionam uma compreensão mais refinada e precisa do uso da linguagem, permitindo que siste-

Figura 2.2 – *Embeddings* a nível de sentença e *embeddings* a nível do token (palavra)



Fonte: (SETH; SHARAFF, 2023)

mas baseados em PLN lidem melhor com ambiguidades e variações no significado das palavras.

2.4.2 *Sentence Embeddings*

Para compreender os *sentence embeddings*, é importante primeiro entender *embeddings* a nível de token, e como essa abordagem se diferencia. Usualmente em LLM's, a representação é feita a nível de um token individual. Cada palavra ou parte de palavra (em casos onde a tokenização divide palavras mais longas em pedaços menores) é convertida em um token único. O modelo então aprende um *embedding* para cada um desses tokens. Dessa forma, uma frase é representada como uma sequência de *embeddings*. Em contraste, existem modelos como o SBERT (REIMERS; GUREVYCH, 2019), que são especificamente otimizados para gerar *embeddings* de frases inteiras. Ao invés de focar em tokens individuais, esses modelos são treinados para entender e codificar o significado global de uma frase completa, gerando um *embedding* para representar a frase inteira. A Figura 2.2 representa essa diferença.

2.5 Arquitetura Transformer

A arquitetura Transformer, introduzida no paper "Attention is all you need"(VASWANI et al., 2017), representa um avanço significativo em PLN. Essa arquitetura é distintiva principalmente pelo seu uso de mecanismos de *self-attention* como parte do bloco base de

um *transformer*, que tem como principal habilidade a geração de *embbedings* contextuais. Essa arquitetura vem mostrando desempenho muito superior em tarefas de PLN quando comparada às redes neurais recorrentes (RNNs), redes neurais convolucionais (CNNs), e suas respectivas variações. Devido à maneira como o mecanismo de *self-attention* funciona, a arquitetura *transformer* só consegue trabalhar com um conjunto de entradas de tamanho fixo. O tamanho desse conjunto é chamado de janela de contexto no âmbito dos modelos de linguagem. A janela de contexto limitada é uma das maiores limitações dessa arquitetura.

2.5.1 Self-attention

O mecanismo de *self-attention*, ou autoatenção, é uma componente fundamental da arquitetura Transformer, ele é uma das partes usadas para gerar *embeddings* contextuais. Este mecanismo permite que o modelo dê diferentes pesos de importância a diferentes partes de uma entrada, facilitando a compreensão do contexto e das relações entre as palavras em uma frase.

Cada palavra em uma sequência de entrada é representada em um vetor, por meio de *embeddings*. Ou seja, um texto é representado como sequência de *embeddings*, e serve como entrada para autoatenção, que determina a atenção que uma palavra deve dar a todas as outras palavras na sequência a fim de gerar uma representação contextual mais rica de si. Este passo é realizado para cada palavra da entrada. Dessa forma, o mecanismo gera como saída uma sequência de *embeddings* contextuais.

O mecanismo de autoatenção é considerado um dos principais fatores para o sucesso da arquitetura transformer, pois a ele é atribuída a grande capacidade de aprendizagem dos modelos inspirados nessa arquitetura. Ele permite o processamento paralelo de uma sequência, porém possui uma complexidade assintótica de $O(N^2)$, limitando a sua escalabilidade. A escalabilidade dos modelos baseados em transformer continua um problema em aberto. Existem diversas variantes do mecanismo de atenção descritos na literatura a fim de endereçar o problema da escalabilidade.

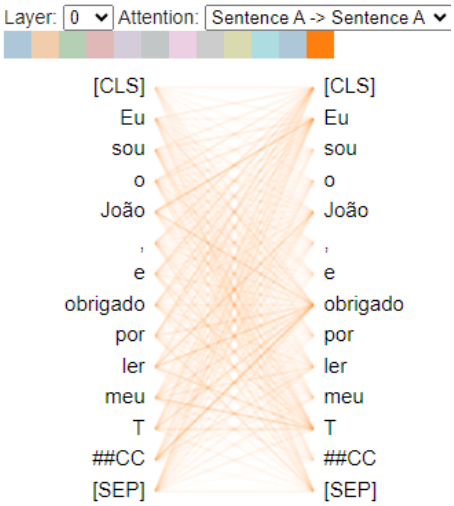


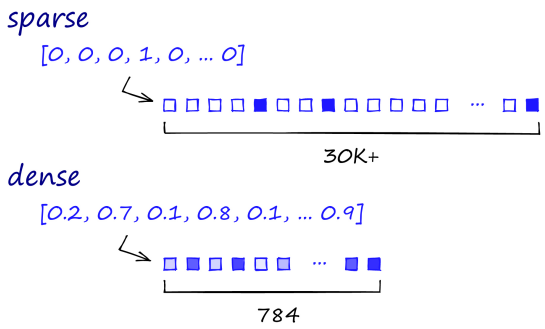
Figura 2.3 – Visualização dos escores de atenção do Bertimbau (SOUZA; NOGUEIRA; LOTUFO, 2020) com a ferramenta BertViz

2.6 Recuperação de Informação por Palavras-Chave (*Keyword-Search*)

Tradicionalmente, as técnicas utilizadas em RI utilizam o casamento de palavras-chave entre a consulta do usuário e os documentos presentes nas bases de dados. Estas técnicas partem da premissa que documentos que possuem mais palavras iguais a da consulta são mais relevantes do que documentos com menos palavras iguais. Assim, o casamento entre palavras-chave é usada para estimar a relevância entre a consulta e o documento. Essa abordagem é também conhecida como **Recuperação de Informação Esparsa**, uma vez que um texto pode ser representado por um vetor esparso de dimensionalidade igual à do vocabulário como representado no vetor mais acima na Figura 2.4. Dentro da RI esparsa são utilizadas funções de ranqueamento como BM25 (AMATI, 2009), TF-IDF (EL-KHAIR, 2009). Estas funções são úteis aos algoritmos de RI, uma vez que fornecem um ranking dos documentos mais relevantes, além de possuírem um baixo custo computacional.

Embora a premissa utilizada nessas funções de ranqueamento seja uma boa heurística para determinar relevância, ela não é verdadeira para todos os casos, sendo especialmente sensível ao vocabulário utilizado nas consultas e documentos. Os sinônimos e termos polissêmicos podem causar confusão no casamento de palavras, induzindo falsos positivos e falsos negativos, além da limitação inerente à representação Bag-of-Words.

Figura 2.4 – Vetores Densos e Vetores Esparsos



Fonte: (BRIGGS, 2024)

2.7 Recuperação de Informação Semântica

Diferente das abordagens que se utilizam do casamento de termos, a recuperação de informação semântica estima a relevância entre uma consulta e um documento com base na similaridade semântica entre ambos. Essa abordagem se aproveitou dos recentes avanços na área de PLN desde a criação da arquitetura de Transformers em (VASWANI et al., 2017). Uma das formas de representar o conteúdo semântico de uma sentença é por meio de **vetores densos** situados em um espaço vetorial (*embeddings*). Dessa forma, partindo da premissa de que documentos semanticamente similares a uma consulta são relevantes para ela, a distância dos vetores no espaço de *embeddings* pode ser interpretada com uma função de relevância para a busca. Usualmente são utilizados *embeddings* a nível de sentença para comparar similaridade semântica.

2.7.1 Sentence-BERT, Sentence Embeddings, e os impactos na Busca Semântica

Um dos trabalhos que implementou *embedding* a nível de sentenças (SBERT) foi (REIMERS; GUREVYCH, 2019), no qual foi utilizado o modelo pré-treinado em (DEVLIN et al., 2019) como ponto de partida numa abordagem de aprendizagem por transferência de conhecimento (*transfer learning*) utilizando uma estrutura de redes siamesas. O modelo de *sentence-embeddings* como apresentado em (REIMERS; GUREVYCH, 2019), permitiu que modelos baseados no BERT fossem usados em grande escala em tarefas como comparação de similaridade semântica, recuperação de informação e clusterização a nível de sentenças. Anteriormente, não era possível devido ao grande custo computacional associado.

Uma das principais vantagens do SBERT é sua eficiência em comparação com o BERT para tarefas de comparação de frases. O BERT padrão não é otimizado para comparar frases de forma eficiente, pois requer que cada par de frases seja passado pelo modelo para calcular sua semelhança. O SBERT, ao contrário, gera *embeddings* que podem ser comparados diretamente usando medidas de distância, como a distância euclidiana ou a similaridade do cosseno, sem a necessidade de processamento adicional.

2.8 *Term Frequency-Inverse Document Frequency*

O *Term Frequency-Inverse Document Frequency* (TF-IDF) é um método estatístico utilizado para avaliar a importância de uma palavra em um documento quando comparado a um corpus. Este método é amplamente empregado em tarefas de processamento de linguagem natural, como recuperação de informações e mineração de texto, sendo usado para identificar a relevância de palavras em documentos específicos. No contexto de RI, esse método é usado para indexar documentos de um corpus. Cada documento pode ser representado por um vetor de dimensionalidade igual à do dicionário de termos, em que cada elemento do vetor contém o escore TF-IDF daquele termo presente no documento. Essa representação gera um vetor esparso que captura termos importantes no documento. Quando um usuário realiza uma consulta, é calculado o vetor TF-IDF do texto da consulta. Esse vetor possibilita comparar as representações vetoriais dos documentos com o vetor TD-IDF da consulta.

Limitações: Esse método não considera a posição da palavra, já utiliza a representação *Bag Of Words* do texto, isso implica na perda da estrutura sintática e de parte do significado semântico. Palavras de alta frequência, mas com pouca relevância semântica, podem receber altas pontuações.

Cuidados na implementação: derivadas das limitações citadas no item anterior, muitas vezes o texto necessita de um pré-processamento, como remoção de *stop words*, a fim de evitar atribuir importância demasiada para termos irrelevantes. Um exemplo são preposições, que são muito frequentes no português, mas, em geral, não são tão importantes para determinar o significado semântico de um texto.

Em resumo, o TF-IDF é uma ferramenta valiosa no processamento de linguagem natural para avaliar a importância das palavras em documentos. Ela possui um bom desempenho, ainda que relativamente simples comparados com métodos atuais. Existem métodos que estendem o TF-IDF e apresentam um desempenho superior, como, por

exemplo, o *Best Match 25*, explicado a seguir.

2.9 Best Match 25

Best Match 25 (BM25) é um algoritmo amplamente utilizado para a recuperação de informações, especialmente em sistemas de busca. Ele é uma evolução dos modelos de correspondência de informações mais antigos, como o TF-IDF. O BM25 é projetado para melhorar a maneira como os sistemas de busca avaliam a relevância de documentos em relação a uma consulta de pesquisa. O BM25 se diferencia de outros modelos como TF-IDF, pois ajusta a pontuação de relevância com base no comprimento do documento. Documentos mais longos têm uma probabilidade naturalmente maior de conter mais ocorrências de um termo, em vista disso, o BM25 utiliza uma estratégia para normalizar esse efeito.

Além da sensibilidade ao comprimento dos documentos, o BM25 inclui parâmetros que podem ser ajustados para otimizar o desempenho em diferentes coleções de documentos ou tipos de consultas. Estes incluem parâmetros como $k1$ e b, que controlam a sensibilidade do algoritmo à frequência do termo e ao comprimento do documento, respectivamente.

Modelos que utilizam o BM25 ainda apresentam desempenho competitivo em alguns casos de uso, sendo superiores a modelos baseados em redes neurais em alguns casos.

2.10 Banco de Dados Vetoriais

Bancos de dados vetoriais são sistemas de gerenciamento de banco de dados projetados para armazenar, indexar e consultar dados representados como vetores em um espaço multidimensional.

Os bancos de dados vetoriais frequentemente empregam tecnologias avançadas de aprendizado de máquina para a transformação de dados e algoritmos de indexação otimizados para alta dimensão. Eles são construídos para serem altamente escaláveis e performáticos, suportando operações de alta velocidade em grandes conjuntos de dados.

No contexto da recuperação de informação, esses bancos são utilizados em conjunto com *embeddings* para indexar documentos de uma base de dados. Eles permitem

comparar os índices de maneira rápida em aplicações como similaridade semântica.

A abordagem mais simples para achar os vetores mais próximos do *embedding* de uma consulta seria usando o algoritmo dos k-vizinhos mais próximos. Infelizmente, essa abordagem não é viável em abundância de dados, devido ao custo computacional proibitivo. Dessa maneira, os bancos de dados vetoriais implementam técnicas para acelerar esse processo, como, por exemplo, buscas aproximadas em estruturas de dados eficientes.

Uma solução utilizada em várias implementações desses bancos é o algoritmo *Approximate Nearest Neighbors* (INDYK; MOTWANI, 1998), que funciona para buscar de forma eficiente dados similares em grandes bases de dados vetoriais. Ele permite operações rápidas em espaços de alta dimensão ao sacrificar um pouco da precisão em troca de uma grande melhoria na velocidade, sendo um componente chave na engenharia de sistemas de RI modernos.

2.11 Métricas de avaliação para resultados de Re-ranking

Tradicionalmente, os sistemas de RI retornam resultados de maneira ordenada de acordo com sua relevância em relação à consulta do usuário. Esse paradigma de interação com o usuário pode ser visto nos motores de busca mais utilizados atualmente. A fim de avaliar o desempenho num cenário como esse, a métrica, além de contabilizar a relevância dos resultados retornados, deve levar em consideração em que ordem eles foram apresentados para o usuário. Uma premissa aceita é que resultados mais relevantes deveriam ficar mais adiante no ordenamento apresentado para o usuário.

Quando se atribui relevância a um resultado, pode-se fazer de maneira binária, indicando 1 para relevante e 0 para não relevante. Muitas vezes essa categorização pode não ser a melhor escolha, pois não captura algumas nuances, como o nível de relevância de um documento.

Nos casos em que o nível de relevância de um documento importa, podem ser usadas duas estratégias. A primeira é atribuir um valor numérico de relevância, por exemplo, atribuir um número real de $[0, 1]$ de maneira que quanto mais relevante o documento, maior seu escore. Essa estratégia apresenta limitações e pode gerar resultados inconsistentes. A segunda estratégia é usar um conjunto de classes ordinais para medir a relevância de um documento, por exemplo: nada Relevante (0), Marginalmente Relevante (1), Muito Relevante (2). O número de classes de relevância pode variar conforme o caso de uso, porém essas classes possuem um ordenamento total.

2.11.1 Ganho Acumulado Descontado

Ganho Acumulado Descontado do inglês *Discounted Cumulative Gain* (DCG) é uma métrica que pode ser utilizada para avaliar a qualidade de um ordenamento de relevância. Sendo consistente com a definição utilizada em (MOREIRA, 2023) cada documento possui um escore de relevância dentro de quatro classes na função rel_i:

(3) Muito Relevante

(2) Moderadamente Relevante

(1) Pouco Relevante

(0) Nada Relevante

$$DCG = \sum_{i=1}^{n} \frac{2^{rel_i} - 1}{log_2(i+1)}$$

A expressão é o ***somatório do ganho descontado*** $GD(i)$ de cada resultado i retornado: $DCG = \sum_{i=1}^{n} GD(i)$. Em relação ao ganho descontado $GD(i)$ ele é definido pela expressão $GD(i) = \frac{2^{rel_i}-1}{log_2(i+1)}$ onde: $2^{rel_i} - 1$ é o ganho $G(i) \in \{0, 1, 3, 7\}$ de um documento, e $log_2(i+1)$ é o fator de desconto de acordo com a posição i no ranking.

O DCG é uma métrica sensível ao conjunto de documentos que está sendo ordenado, dessa forma não é possível comparar a métrica de resultados provenientes de conjuntos de documentos diferentes, pois o conjunto que contém documentos mais relevantes tenderá a sempre ter um DCG maior.

2.11.2 Ganho Acumulado Descontado Normalizado

O Ganho Acumulado Descontado Normalizado (do inglês *Normalized Discounted Cumulative Gain* (NDCG)) se apresenta como uma métrica mais adequada para comparar resultados de conjuntos de documentos diferentes, levando em consideração a relevância dos documentos de cada conjunto. No **NDCG** o **DCG** é normalizado pelo **IDCG**. O IDCG é o ganho cumulativo descontado ideal para um conjunto de documentos, ou seja, é DCG para o ordenamento ótimo. Ele é o valor máximo que o resultado de um reordenamento poderia assumir para um dado conjunto de documentos. como representado na Figura 2.5.

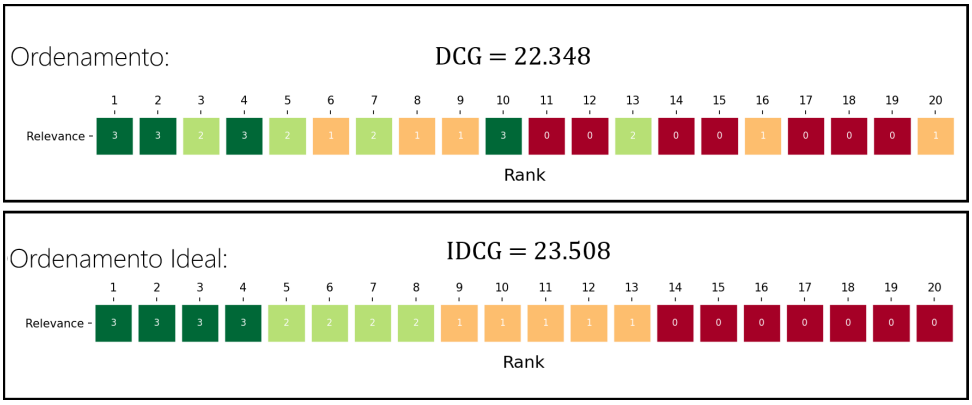


Figura 2.5 – DDG do Ordenamento e do Ordenamento Ideal(IDCG)

$$NDCG = \frac{DCG}{IDCG}$$

2.11.3 Variações das métricas

Existe uma variação de cada uma das métricas DCG,IDCG,NDCG nas quais só são considerados os top K resultados do ranking, sendo que K é um parâmetro. Essas métricas são:

NDCG@K: é o NDCG para os top K resultados

DCG@K: é o DCG para os top K resultados

IDCG@K: é o IDCG para os top K resultados

2.12 Soluções tecnológicas Out-of-the-box

Soluções tecnológicas *out-of-the-box* são produtos, serviços ou sistemas que podem ser usados imediatamente após a compra ou o download, sem a necessidade de configurações ou personalizações extensas. O termo *out-of-the-box* literalmente significa **"fora da caixa"**, o que implica que o usuário pode começar a usar o produto assim que o tira da caixa, ou seja, quase sem preparação prévia.

Essas soluções são projetadas para funcionar com configurações que atendem às necessidades comuns de diversos usuários ou empresas. Elas são populares porque simplificam o processo de implementação de novas tecnologias, economizando tempo e recursos que, de outra forma, seriam necessários para a configuração e personalização.

3 TRABALHOS RELACIONADOS

A área de RI vem apresentando relevância de longa data no meio acadêmico, sendo uma tarefa central quando se trata de linguística computacional. Mesmo com os avanços em linguística computacional proporcionados pelos LLM, a área de RI ainda apresenta margem para avanços significativos. No artigo *Large Language Models for Information Retrieval: A Survey* (ZHU et al., 2024) são abordadas várias técnicas utilizadas em sistemas de RI para obter melhores resultados, dando assim uma visão mais ampla da área, a fim de consolidar metodologias existentes e fornecer *insights* sutis por meio de uma visão geral abrangente. Sendo essencial para direcionar futuras pesquisas, principalmente quando se trata de abordagens baseadas em aprendizado de máquina e redes neurais.

Mesmo com o avanço dos modelos de redes neurais na área da linguística computacional, as abordagens tradicionais de RI ainda são relevantes e não se tornaram obsoletas mesmo com décadas de pesquisa (CHEN et al., 2022). E são especialmente relevantes em abordagens não-supervisionadas.Nesse contexto, existem trabalhos recentes como (KAMALLOO et al., 2023), que comparam a efetividade de modelos neurais para recuperação de informação com modelos como BM25, ambos inseridos numa abordagem de *retrieve and rerank*. Esses trabalhos continuam apontando o BM25 como um *baseline* competitivo, em muitos casos de uso, mesmo que quando comparado com abordagens supervisionadas. Além disso, quando modelos neurais são aplicados em um domínio para o qual não foram ajustados apresentam desempenho inferior ao BM25 na tarefa de recuperação (IZACARD et al., 2021) (MAILLARD et al., 2021), dessa forma, reafirmando a flexibilidade e generalidade do BM25, uma vez que é uma abordagem não supervisionada, e robusta a variação de domínio.

Em contraste, é amplamente aceito (MA et al., 2023), (SASAZAWA et al., 2023), (ZHU et al., 2024), (REDDY et al., 2023) que abordagens de recuperação de informação de dois estágios (*Retrieve and Rerank*) são superiores às suas contrapartes. Independentemente do modelo usado, como o *first stage retriever*, adicionar modelo para re-ranquear melhora a qualidade dos escores de relevância gerados. Embora produzam resultados com maior qualidade, modelos de ranqueamento usualmente são mais custosos computacionalmente. Dessa forma, é necessário combiná-los com um modelo mais "leve"utilizado como uma "peneira de granularidade mais grossa", e usar o modelo de ranqueamento para refinar os resultados da primeira etapa. Motivado por esses fatos, a escolha da abordagem utilizada na recuperação de informação foi a abordagem de duas etapas: *retrieve and*

rerank.

A segmentação de documentos é uma estratégia muito utilizada atualmente em vários sistemas de PLN incluindo em RI (BAI et al., 2023; MICULICICH; HAN, 2023; AHMAD, 2024; BALAGUER et al., 2024). Uma das motivações da segmentação é o limite de tamanho da janela de contexto imposto por modelos baseados na arquitetura *transformer*. A segmentação é especialmente importante em sistemas de RI que utilizam modelos neurais, já que a qualidade dos segmentos produzidos interfere diretamente na qualidade dos resultados retornados pelo sistema de RI. É relativamente trivial que estratégias de segmentação semântica tendem a produzir resultados melhores que estratégias que não dão importância ao significado dos segmentos.

Existem trabalhos que abordam a segmentação semântica, como (LUKASIK et al., 2020) e (LI et al., 2022). Estes trabalhos propõem arquiteturas híbridas baseadas em transformers e redes neurais recorrentes para endereçar esse problema. Essas abordagens possuem um bom desempenho, porém requerem que os modelos sejam treinados para essa tarefa específica.

Quando se trata de recuperação de informação aplicada em dados privados, geralmente se pensa em abordagens que utilizam *fine-tunning*, uma vez que a mudança no domínio em que os modelos estão inseridos implica em uma degradação do seu desempenho. Todavia, em contextos de pequena escala, o ajuste dos modelos muitas vezes é inviável, seja pela escassez de dados anotados, falta de *knowhow*, e baixo retorno sobre o investimento de capital monetário e humano requerido pela abordagem.

Outra maneira de abordar a segmentação semântica é utilizando *in-context-learning* com LLM pré-treinados. Essa abordagem não requer treinamento e pode gerar bons resultados dependendo do modelo de linguagem usado. No artigo (LI et al., 2022), é utilizada uma abordagem de *in-context-learning* para segmentar códigos. Todavia, essa abordagem não possui muitos casos de uso na literatura, sendo no que âmbito de atas de reunião essa abordagem nunca foi utilizada.

3.1 Análise

Dado o bom desempenho das soluções *retrive and rerank* quando comparadas com recuperação de um estágio, foi optado por utilizar essa estratégia de dois estágios para buscar os trechos das atas. Estes trechos são gerados em uma abordagem zero-shot de segmentação semântica com o GPT-4, em vista que os LLM vêm mostrando altas

capacidades com *in-context learning*, além de que essa é a abordagem mais flexível entre todas as abordagens disponíveis.

4 METODOLOGIA

Este capítulo visa discutir o pipeline empregado na recuperação de informações para atas de reuniões. O pipeline abrange a *Preparação dos Dados*, como mostrado na Figura 4.1, *Busca Semântica* representada na Figura 4.2, e a apresentação dos resultados para o usuário.

4.1 Preparação dos dados

A preparação dos dados é a etapa do pipeline que visa transformar os Documentos PDF das Atas em dados recuperáveis. Para isso, é necessário extrair o texto dos documentos em PDF, segmentar o texto em pedaços menores e inserir em um banco de dados vetorial. Dessa forma, os dados ficam prontos para a etapa da Busca Semântica.

Na extração do texto é utilizada a biblioteca de manipulação de PDF em Python PyMuPDF (ARTIFEX, 2015-2024). O texto extraído das atas é salvo em um DataFrame Pandas no formato Comma Separated Values (CSV), junto com o nome do arquivo do qual foram retirados, e outros metadados relevantes para cada caso de uso. Além disso, o texto extraído de cada documento foi salvo no formato TXT em um diretório local.

A unidade básica utilizada na busca não são os documentos completos, mas sim os segmentos gerados a partir da segmentação semântica.

Um *script* iterou pelos arquivos TXT extraídos das atas. Para cada arquivo TXT, o *script* realizou a segmentação usando a API OpenAI, como será descrito na próxima seção. As respostas da API foram salvas em formato TXT em outro diretório local. Posteriormente, um *script* extraiu os conteúdos salvos no formato TXT, e outros metadados pertencentes aos documentos das atas. As informações foram salvas em um arquivo CSV, sendo elas:

1. Quantidade de *tokens* passados no *prompt*
2. Quantidade de *tokens* gerados pelo modelo
3. Motivo de parada da API
4. Texto gerado pelo modelo
5. Tempo de processamento da resposta
6. Órgão ao qual pertence a ata
7. Nome do arquivo do qual foi extraído o texto

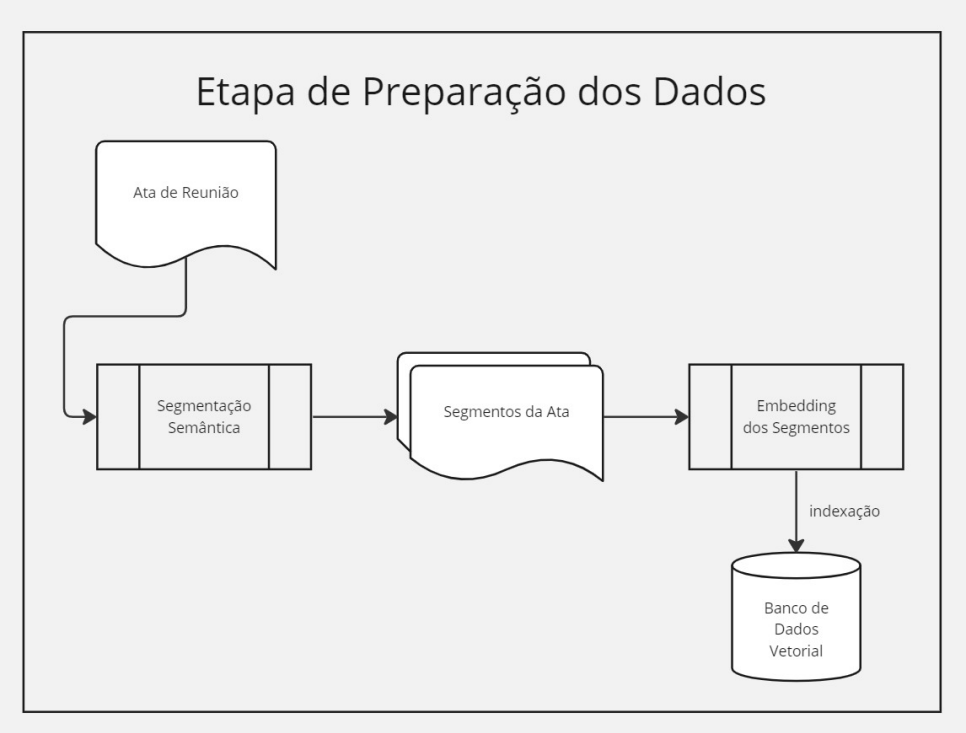


Figura 4.1 – Fluxograma ilustrando a preparação dos dados

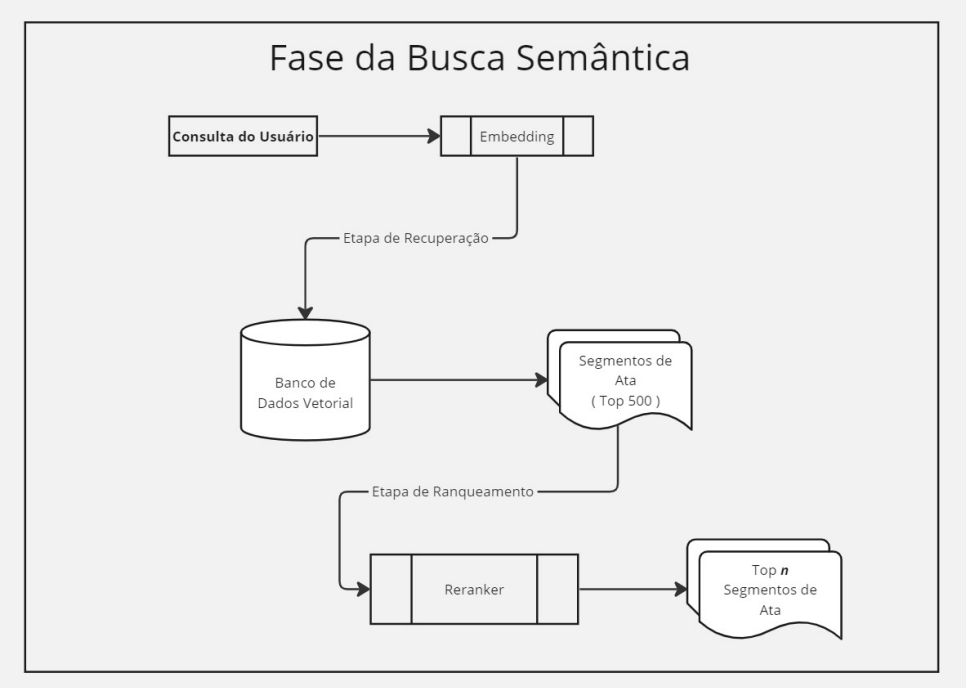


Figura 4.2 – Fluxograma ilustrando a fase da busca semântica

O arquivo CSV resultante passou por um processo no qual o texto com os marcadores <*division*> foi dividido em vários trechos de texto separados e salvos em outro CSV, contendo referência ao texto pelo qual foi gerado.

4.1.1 Segmentação do Texto

Como limitação imposta pela arquitetura de Transformers, os modelos utilizados nessa abordagem conseguem trabalhar apenas com uma janela de contexto limitada. Dessa forma, é necessário fragmentar o texto de uma ata em pedaços menores que estejam dentro dos limites estabelecidos pela arquitetura. Além disso, fragmentar o texto em partes semanticamente relevantes e independentes, tende a melhorar os resultados da busca semântica, uma vez que a unidade básica da busca não é um documento inteiro, mas sim um trecho do documento que contém apenas um tópico importante.

A abordagem mais comum de segmentação é dividir o texto em segmentos de tamanho fixo, também conhecida como *fixed size chunking*. Essa abordagem resolve o problema do tamanho de contexto limitado, porém resulta em uma divisão que não é semanticamente coesa. Dessa forma, um tópico importante pode ser dividido em dois ou mais segmentos, sendo que cada segmento possui informação incompleta sobre aquele tópico.

Uma abordagem simples e eficaz é a utilização de expressões regulares para segmentar o texto conforme os tópicos discutidos. Essa estratégia de segmentação traz resultados excelentes com um baixo custo computacional e monetário. Todavia, essa abordagem requer que a estrutura da ata seja conhecida a priori, dessa forma, é uma abordagem com alto grau de interferência humana. Além disso, essa abordagem não é generalizável para atas que possuam uma estrutura textual diferente, então precisa ser adaptada para cada caso de uso.

Para realizar uma segmentação que leve em consideração o conteúdo semântico dos segmentos e que não necessite interferência humana, foi utilizada uma abordagem de *in-context learning*. Nessa abordagem, é utilizado um modelo de linguagem generativo, instruído em como realizar a segmentação, e recebe como entrada o texto da ata. O modelo deve segmentar o texto em várias partes semanticamente coesas.

Fragmentação com GPT4-Turbo É utilizada uma abordagem *zero-shot in-context-learning*, para fragmentar o texto com o modelo generativo GPT-4-Turbo por meio da API da openAI. O modelo é utilizado para fragmentar o texto das atas em partes semantica-

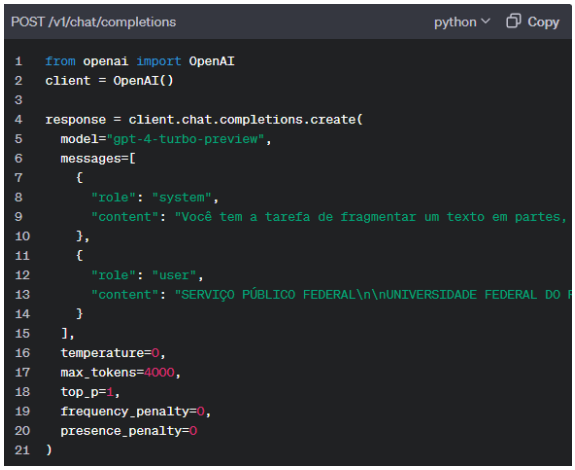
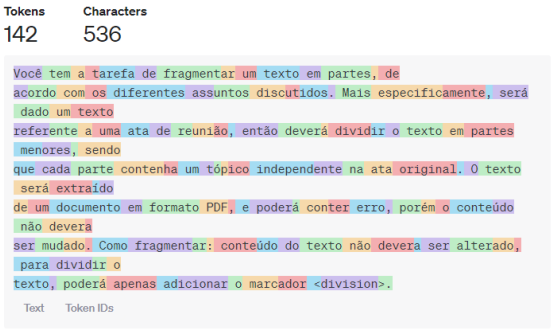


Figura 4.3 – Captura de tela que mostra a configuração das chamadas do modelo de *chat completion* GPT4-Turbo na API OpenAI

Figura 4.4 – Captura de tela que ilustra quantidade de tokens na Instrução de como realizar a fragmentação



Fonte: https://platform.openai.com/tokenizer

mente relevantes. Cada fragmento resultante engloba um tópico descrito na ata.

Detalhes da chamada da API: Na chamada da API para o serviço de *chat completion* são passados dois textos como parâmetro: **1.** Instrução de como realizar a fragmentação e **2.** Texto da Ata.

1. Você tem a tarefa de fragmentar um texto em partes, conforme os diferentes assuntos discutidos. Mais especificamente, será dado um texto referente a uma ata de reunião, então deverá dividir o texto em partes menores, sendo que cada parte contenha um tópico independente na ata original. O texto será extraído de um documento em formato PDF e poderá conter erro, porém o conteúdo não devera ser mudado. Para fragmentar, o conteúdo do texto não deve ser alterado. Para dividir o texto, poderá apenas adicionar o marcador <division>.
2. Texto extraído do PDF da Ata. Que varia conforme a ata segmentada. É passado como *"user"* na API.

Um exemplo de entrada e saída esperada da tokenização num exemplo hipotético.

Entrada: Joãozinho atravessou a rua para buscar duas sacolas para sua mãe. O cachorro é o animal considerado como o melhor amigo do homem. Saída: Joãozinho atravessou a rua para buscar duas sacolas para sua mãe.**<division>** O cachorro é o animal considerado como o melhor amigo do homem.

O custo da abordagem: O modelo GPT4-Turbo possuí um custo para Entrada e Saída de 0,01 e 0,03 dólares a cada 1000 tokens respectivamente. Além disso, conforme a Figura 4.4 a frase de instrução possui 142 tokens, e cada marcador do tipo <division> é representado por 3 tokens. Levando em consideração essas informações, o custo em dólar para fragmentar um documento que possui *n* tokens, que possui *m* fragmentos(tópicos) pode ser calculado pela seguinte expressão:

$$custo = (142 + n) * 0,01/1000 + (n + (m - 1) * 3) * 0,03/1000$$

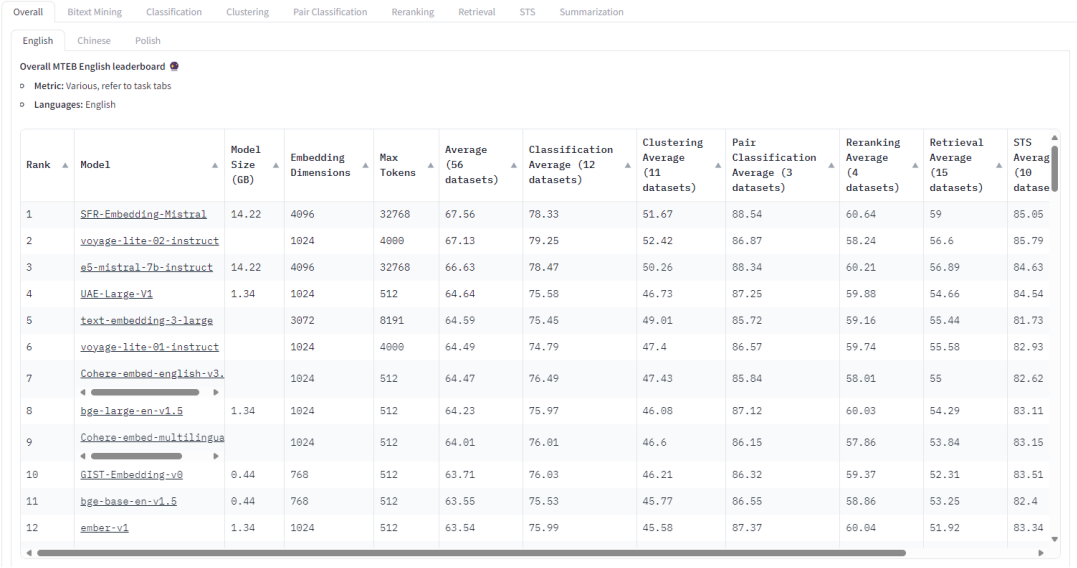
Com essa abordagem de fragmentação, a relevância de um trecho de texto retornado ao usuário é muito maior quando comparado com retornar o texto inteiro, ou com alguma abordagem mais simples que não considera o conteúdo semântico do fragmento - como, por exemplo, dividir o texto em *chunks* de tamanho igual.

4.1.2 Indexação dos Segmentos

Para indexar cada segmento é usado um banco de dados vetorial, nesse caso o ChromaDB, no qual o índice de cada segmento é o *embedding* a nível de sentença daquele segmento. O modelo utilizado para a geração de *embeddings* a nível de sentença é o modelo *text-embedding-3-large* (OPENAI, 2024) - lançado em janeiro de 2024. Esse modelo produz *embeddings* de dimensionalidade igual a 3072 por padrão, mas pode ser configurado para produzir *embeddings* de dimensionalidade = 1024 ou 256. A dimensionalidade padrão foi escolhida, pois nessa abordagem, o desempenho computacional da recuperação não é um fator crítico.

O modelo *text-embedding-3-large* apresenta um bom desempenho no *benchmark* MTEB para a língua inglesa, ficando na quinta posição como mostrado na Figura 4.5. É importante ressaltar que esse modelo possui capacidades multilíngues.

Figura 4.5 – Captura de tela do *leaderboard* no *benchmark* MTEB no dia 13/2/2024



Overall | Bitext Mining | Classification | Clustering | Pair Classification | Reranking | Retrieval | STS | Summarization

English | Chinese | Polish

Overall MTEB English leaderboard

- **Metric:** Various, refer to task tabs
- **Languages:** English

Rank	Model	Model Size (GB)	Embedding Dimensions	Max Tokens	Average (56 datasets)	Classification Average (12 datasets)	Clustering Average (11 datasets)	Pair Classification Average (3 datasets)	Reranking Average (4 datasets)	Retrieval Average (15 datasets)	STS Averag (10 datase
1	SFR-Embedding-Mistral	14.22	4096	32768	67.56	78.33	51.67	88.54	60.64	59	85.05
2	voyage-lite-02-instruct		1024	4000	67.13	79.25	52.42	86.87	58.24	56.6	85.79
3	e5-mistral-7b-instruct	14.22	4096	32768	66.63	78.47	50.26	88.34	60.21	56.89	84.63
4	UAE-Large-V1	1.34	1024	512	64.64	75.58	46.73	87.25	59.88	54.66	84.54
5	text-embedding-3-large		3072	8191	64.59	75.45	49.01	85.72	59.16	55.44	81.73
6	voyage-lite-01-instruct		1024	4000	64.49	74.79	47.4	86.57	59.74	55.58	82.93
7	Cohere-embed-english-v3.		1024	512	64.47	76.49	47.43	85.84	58.01	55	82.62
8	bge-large-en-v1.5	1.34	1024	512	64.23	75.97	46.08	87.12	60.03	54.29	83.11
9	Cohere-embed-multilingua		1024	512	64.01	76.01	46.6	86.15	57.86	53.84	83.15
10	GIST-Embedding-v0	0.44	768	512	63.71	76.03	46.21	86.32	59.37	52.31	83.51
11	bge-base-en-v1.5	0.44	768	512	63.55	75.53	45.77	86.55	58.86	53.25	82.4
12	ember-v1	1.34	1024	512	63.54	75.99	45.58	87.37	60.04	51.92	83.34

Fonte: https://huggingface.co/spaces/mteb/leaderboard

4.2 Busca Semântica

A busca semântica utiliza os segmentos gerados pela etapa anterior como unidade básica de informação. A busca ocorre em dois estágios: (i) recuperação (retrieval) e (ii) re-ranqueamento (reranking), como mostrado na Figura 4.2. Utilizando a abordagem de dois estágios, é possível combinar o menor custo computacional do estágio de recuperação com os scores de relevância com qualidade superior produzidos pelo modelo de ranqueamento.

O estágio de recuperação (*retrieval*) é o primeiro estágio da busca. Ele é usado para retornar os 500 resultados mais relevantes conforme a consulta do usuário. Para isso, é utilizado um banco de dados vetorial no qual cada trecho extraído dos documentos está indexado. A consulta do usuário é transformada em um *embedding* com o modelo *text-embedding-3-large*, o mesmo utilizado para gerar os índices dos segmentos. O banco de dados se encarrega de retornar os 500 vizinhos mais próximos, conforme a métrica de similaridade de cosseno. Os 500 segmentos recuperados são passados de forma textual para o próximo estágio da busca - a etapa de ranqueamento.

Nessa segunda etapa da busca, os resultados da recuperação são reordenados por um modelo que calcula a relevância de um trecho de texto conforme a consulta feita pelo usuário. Para calcular os escores de relevância, é utilizado um modelo de ranqueamento com pesos fechados em uma API disponibilizada gratuitamente (quando para fins de pesquisa) pela (COHERE, 2024). O modelo apresenta capacidades multilíngue e pode ser

utilizado com o português. Os 500 segmentos em formato textual da etapa anterior, e também a consulta do usuário, são passados como parâmetro da chamada da API. Também é passado um parâmetro n, que indica quantos resultados serão retornados por essa etapa. A API retorna o resultado do ranqueamento com os top n documentos. As informações retornadas pela API para cada documento são: (i)texto do documento, (ii) escore de relevância, (iii) índice que o documento foi passado. O índice (iii) pode ser cruzado com os dados salvos anteriormente para retornar outras informações de cada segmento.

4.3 Interação do usuário

A interação do usuário é construída com a ferramenta Streamlit, que permite o desenvolvimento de interfaces gráficas em Python de maneira simplificada. A interface segue o estilo das interfaces dos motores de busca do Google e Bing. O usuário digita a consulta que quer fazer em uma barra de pesquisa, localizada no topo da tela. Os segmentos retornados pela etapa de busca são apresentados de maneira ordenada para o usuário, conforme a relevância. Para cada segmento, além do seu conteúdo textual, podem ser apresentados o título do documento, escore de relevância do segmento, data da reunião e qualquer informação que se julgar relevante para aquela ata de reunião. Na Figura 4.6 está representado um exemplo de como essa interface poderia ser.

4.4 Sensibilidade dos dados e implicações da abordagem

O uso de APIs de terceiros em produtos de software é uma estratégia frequentemente utilizada, já que, em muitos casos, oferece boa eficiência operacional e agilidade no desenvolvimento. O uso de APIs permite que uma empresa utilize uma solução sem precisar desenvolvê-la. Todavia, essa abordagem também carrega consigo uma série de riscos, especialmente no que diz respeito à segurança e à privacidade dos dados. Ao integrar APIs de terceiros, as empresas expõem suas aplicações a códigos e infraestruturas sobre as quais têm controle limitado. Um dos riscos mais significativos é a possibilidade de revelação não intencional de dados sensíveis para terceiros. Dependendo do tipo de API e dos dados que ela processa, informações confidenciais podem ser expostas se as políticas de privacidade e segurança da API não forem rigorosamente avaliadas e monitoradas.

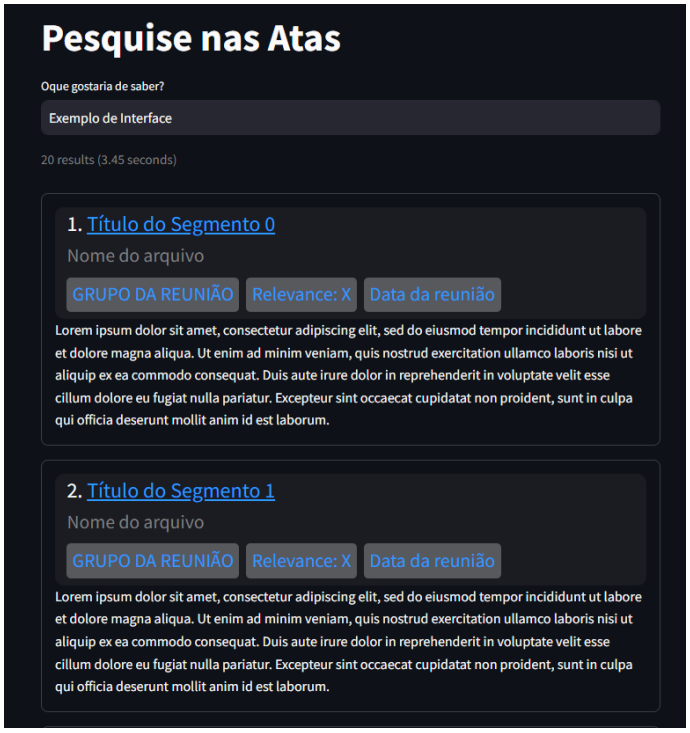


Figura 4.6 – Uma interface genérica - Os top 2 resultados retornados (título,nome do arquivo, grupo, escore de relevância, data da reunião, texto

Um sistema de recuperação de informações para atas de reuniões, quando integrado com APIs de terceiros, deve tratar com extrema cautela a manipulação de informações que são muitas vezes sensíveis ou confidenciais. Essas informações podem incluir detalhes estratégicos da empresa, dados pessoais dos participantes, decisões financeiras, entre outros aspectos que, se divulgados indevidamente, podem prejudicar a integridade e a competitividade da organização. A abordagem descrita nesse capítulo faz uso de APIs de terceiros. Dessa forma, as informações contidas nas atas são fornecidas para essas organizações. Essa abordagem é especialmente inadequada para atas que contêm informações confidenciais, como, por exemplo, em cenários corporativos.

As alternativas ao uso de APIs envolvem abordagens que permitem às organizações implementar ou utilizar capacidades de IA sem depender de serviços de terceiros disponibilizados por meio de APIs. Empresas e equipes de pesquisa podem optar por desenvolver seus próprios modelos de IA, isso requer investimento significativo em especialização, dados e infraestrutura computacional, mas oferece controle total sobre os algoritmos, os dados e a aplicação da tecnologia. Outra opção é o uso de soluções *open source*, essas soluções oferecem a vantagem de utilizar modelos de IA desenvolvidos por terceiros, que podem ser hospedados na infraestrutura da própria organização, o que garante maior controle e segurança sobre os dados. Essas alternativas são muito atrativas, porém demandam uma quantidade muito maior de recursos (financeiros e humanos)

quando comparados com o uso de APIs. Além disso, quando se trata de IA, são poucas as organizações que possuem real acesso aos recursos mais poderosos de IA. Atualmente, os modelos com maior capacidade não são *open source*, como, por exemplo, o GPT4, que só é acessível por meio de serviços da OpenAI. Dessa forma, é preciso abrir mão de usar modelos mais poderosos quando se utiliza soluções *open source*.

Por fim, cabe ao desenvolvedor da abordagem de RI ponderar os prós e contras da utilização de APIs de terceiros. Geralmente, os modelos *open source* apresentam capacidades suficientes para serem adequados em uma abordagem de RI.

5 EMPREGANDO A ABORDAGEM NO CONTEXTO DO INSTITUTO DE INFORMÁTICA

A abordagem de busca semântica foi aplicada no contexto do Instituto de Informática (INF) da UFRGS, em vista da importância que as reuniões apresentam para a boa gestão em diversos âmbitos no INF, que aliada à grande quantidade de atas geradas como documentação dessas reuniões, implicam em dificuldade, principalmente para procurar por informações dentro do conjunto gerado. O instituto mantém um acervo contendo as atas documentando reuniões de longa data conduzidas por diversos órgãos, incluindo o (i) CONSELHO DO INSTITUTO DE INFORMÁTICA (CONINF), o (ii) PROGRAMA DE PÓS GRADUAÇÃO (PPGC), e o (iii) COLEGIADO DO DEPARTAMENTO DE INFORMÁTICA. Esses documentos se encontram todos em formato PDF, podendo ser nativamente digitais ou digitalizados.

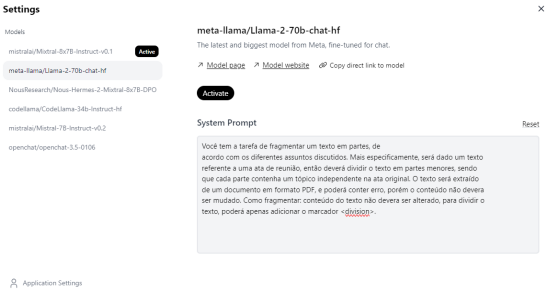
Neste trabalho, foi optado por utilizar apenas os documentos nativamente digitais, uma vez que não haveria problema de eventual ruído dos dados durante a fase de pré-processamento. Para trabalhar com documentos digitalizados, seria necessário utilizar técnicas relacionadas à área de *Optical Character Recognition* (OCR). Esta abordagem tornaria o problema mais complexo, além de introduzir eventuais erros que dificultariam o seu processamento com as técnicas de RI utilizadas.

Originalmente, o corpus é composto de um total de 414 documentos de atas de reunião em formato PDF, conduzidas por esses três grupos: CONINF (131 documentos), PPGC (159 documentos), Colegiado (124 documentos) porém com a restrição de utilizar apenas documentos nativamente digitais o Corpus foi reduzido para 245 documentos sendo: CONINF (57 documentos), PPGC (109 documentos), e Colegiado (79 documentos).

Seguindo a abordagem de segmentação descrita na metodologia, ao total foram gerados 1939 segmentos a partir dos 245 documentos contidos no corpus reduzido. Além disso, o uso da API resultou em um custo total de 19,32 dólares, como reportado abaixo com outras estatísticas:

- **Média de tokens por documento:** 1781,39 tokens
- **Média de tempo de processamento por documento:** 178,64s
- **Tempo total de processamento:** 43766,8s (12,15h)
- **Custo médio:** 7,88 centavos de dólar

Figura 5.1 – Exemplo da configuração utilizada para testar a capacidade dos modelos



- **Custo total:** 19,32 dólares

5.1 Experimentos e Validação

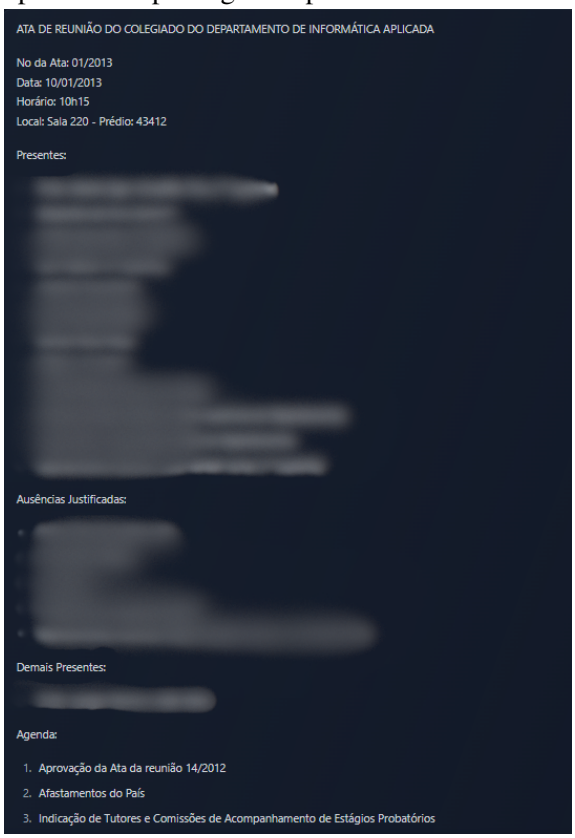
Nesta seção, são descritos três experimentos realizados a fim de avaliar alguns aspectos da abordagem utilizada no contexto das atas de reunião do INF. O Experimento 1 serve para verificar se outros LLM possuem a capacidade de executar a segmentação semântica com *zero-shot in-context-learning*. O segundo experimento serve para avaliar o resultado da segmentação, por meio de uma inspeção manual. O terceiro experimento faz uma breve avaliação quantitativa dos resultados da busca.

5.1.1 Experimento 1: Substituindo o GPT4 por outros LLM na etapa de Segmentação Semântica do Texto

A abordagem de segmentação Semântica do texto foi testada com outros modelos que não o GPT4 a fim de averiguar se tais modelos possuem a capacidade para executar essa tarefa em um contexto zero-shot. Para avaliar a capacidade de tais modelos, foi utilizada a interface de chat do Hugging Chat, que permite interagir com vários modelos de linguagem generativos. Foi usada a mesma configuração utilizada na Segmentação com o GPT4-Turbo discutida na metodologia, como o exemplo da Figura 5.1 ilustra. Os modelos utilizados foram: meta-llama/Llama-2-70b-chat-hf, mistralai/Mixtral-8x7B-Instruct-v0.1, NousResearch/Nous-Hermes-2-Mixtral-8x7B-DPO, openchat/openchat-3.5-0106.

Resultados e discussão Os modelos *open source* testados não apresentaram a capacidade de realizar a tarefa designada. Assim, os resultados obtidos indicam que os modelos não possuem a capacidade de seguir instruções de maneira muito precisa. Os modelos geraram resumos do texto fornecido, como visto na Figura 5.2, embora essa não

Figura 5.2 – Captura de parte da resposta gerada pelo modelo meta-llama/Llama-2-70b-chat-hf



tenha sido a instrução dada aos modelos.

A hipótese é que, por se tratar de modelos *open source* ordens de grandeza menores que o GPT4, não possuem as mesmas capacidades do modelo de maior escala. Isso vai de encontro às pesquisas feitas em PLN que estudam *Scalling Laws* dos modelos de linguagem. A tendência que essas leis vêm mostrando é que o tamanho dos modelos é proporcional às habilidades que eles são capazes de adquirir. Além do tamanho desses modelos, a quantidade de dados em que os modelos foram pré-treinados também impacta nas habilidades adquiridas.

5.1.2 Experimento 2: Inspeção manual do resultado da Segmentação Semântica com GPT4

Para analisar de forma qualitativa os resultados da abordagem de segmentação semântica, foi realizada uma inspeção manual dos resultados da segmentação em três documentos do corpus. Mais especificamente, foi escolhido aleatoriamente um documento de cada um dos três subconjuntos de documentos (CONINF, PPGC, e COLEGIADO), tendo em vista que cada grupo possui um modelo de escrita próprio.

A Figura 5.3 apresenta a segmentação da ata 12/2012 do Colegiado. Pode-se ver que nessa reunião foram discutidos seis tópicos: (i) Aprovação da ata anterior, (ii) Afastamentos do País, (iii) progressão funcional da classe de professor adjunto, (iv) Relatório final de estágio probatório,(v) Abertura de concursos para professor adjunto, (vi) Pontuação nas progressões de classe de professor ajunto. Após a segmentação, o texto da ata 12/2012 ficou dividido em oito segmentos, sendo um para o cabeçalho da reunião, um para a finalização da reunião e um para cada tópico citado anteriormente.

Na Figura 5.4 é apresentada a segmentação da ata 03/2020 do CONINF. Nessa reunião, foram enumerados quatro tópicos: (i) Aprovação da ata anterior, (ii)Texto a ser encaminhado com indicação para professor emérito do professor, (iii) texto a ser encaminhado com indicação para professor emérito do professor, (v) Assuntos gerais. A segmentação dessa ata resultou em 5 segmentos: uma para a introdução da ata, uma para cada um dos tópicos (i), (ii) e (iii), um segmento para o tópico (iv) com a finalização da reunião.

Na Figura 5.5 está representada a segmentação da ata 001/2021 do PPGC. Nessa ata foram discutidos 4 tópicos principais: (i) Ata anterior, (ii) decisões da Subcomissão Executiva, (iii) recredenciamento de orientador, (iv) desligamento do curso. O resultado da segmentação possui 4 segmentos, sendo: uma para a introdução da ata com o tópico (i), um para cada um dos tópicos (ii) e (iii), um segmento para o tópico (iv) com a finalização da reunião.

Discussão dos resultados Embora os segmentos de texto possuam tópicos distintos e estejam conforme a tarefa dada ao modelo, os segmentos podem não ser produzir os melhores resultados para um sistema de busca semântica, em vista que o modelo apresentou uma tendência de ser demasiadamente conservador ao dividir os trechos. Dessa forma, cada segmento apresentou vários subtópicos que não foram subdivididos, mas que provavelmente poderiam gerar melhores resultados se fossem divididos. Todavia, é importante ressaltar que o modelo GPT4-Turbo executou a tarefa totalmente conforme a instrução passada ao modelo, já que no Contexto de Instrução está explicito que é para dividir o texto nos diferentes tópicos discutidos em uma reunião. Para uma divisão diferente, seria necessário deixar explicito no contexto da instrução como segmentar de outra maneira.

Figura 5.3 – Segmentos Semânticos em Ata do Colegiado

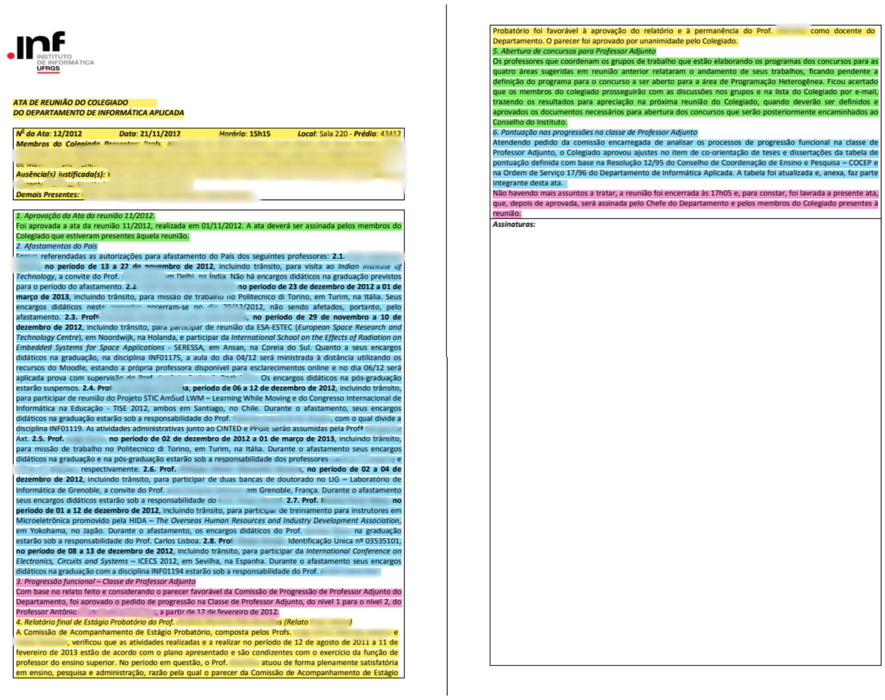
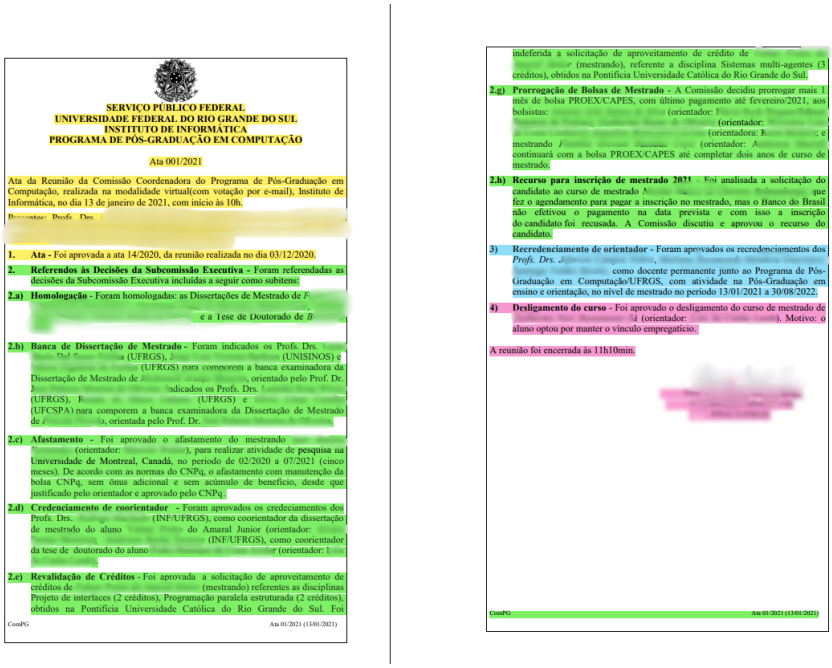


Figura 5.4 – Segmentos Semânticos em Ata do CONINF

Figura 5.5 – Segmentos Semânticos em Ata do PPGC



SERVIÇO PÚBLICO FEDERAL
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

Ata 001/2021

Ata da Reunião da Comissão Coordenadora do Programa de Pós-Graduação em Computação, realizada na modalidade virtual(com votação por e-mail), Instituto de Informática, no dia 13 de janeiro de 2021, com início às 10h.

1. **Ata** - Foi aprovada a ata 14/2020, da reunião realizada no dia 03/12/2020.
2. **Referendos às Decisões da Subcomissão Executiva** - Foram referendadas as decisões da Subcomissão Executiva incluídas a seguir como subitens:

2.a) Homologação - Foram homologadas: as Dissertações de Mestrado de *F* e a Tese de Doutorado de *B*

2.b) Banca de Dissertação de Mestrado - Foram indicados os Profs. Drs. (UFRGS), (UNISINOS) e (UFRGS) para comporem a banca examinadora da Dissertação de Mestrado de , orientado pelo Prof. Dr. Indicados os Profs. Drs. (UFRGS), (UFRGS) e (UFCSPA) para comporem a banca examinadora da Dissertação de Mestrado de , orientada pelo Prof. Dr.

2.c) Afastamento - Foi aprovado o afastamento do mestrando (orientador:), para realizar atividade de **pesquisa na Universidade de Montreal, Canadá**, no período de 02/2020 a 07/2021 (cinco meses). De acordo com as normas do CNPq, o afastamento com manutenção da bolsa CNPq, sem ônus adicional e sem acúmulo de benefício, desde que justificado pelo orientador e aprovado pelo CNPq.

2.d) Credenciamento de coorientador - Foram aprovados os credenciamentos dos Profs. Drs. (INF/UFRGS), como coorientador da dissertação de mestrado do aluno do Amaral Junior (orientador: ; (INF/UFRGS), como coorientador da tese de doutorado do aluno (orientador:

2.e) Revalidação de Créditos - Foi aprovada a solicitação de aproveitamento de créditos de (mestrando) referentes as disciplinas Projeto de interfaces (2 créditos), Programação paralela estruturada (2 créditos), obtidos na Pontifícia Universidade Católica do Rio Grande do Sul. Foi

ComPG Ata 01/2021 (13/01/2021)

indeferida a solicitação de aproveitamento de crédito de (mestrando), referente a disciplina Sistemas multi-agentes (3 créditos), obtidos na Pontifícia Universidade Católica do Rio Grande do Sul.

2.g) Prorrogação de Bolsas de Mestrado - A Comissão decidiu prorrogar mais 1 mês de bolsa PROEX/CAPES, com último pagamento até fevereiro/2021, aos bolsistas: (orientador: (orientador: (orientadora: ; e mestrando (orientador: continuará com a bolsa PROEX/CAPES até completar dois anos de curso de mestrado.

2.h) Recurso para inscrição de mestrado 2021 - Foi analisada a solicitação do candidato ao curso de mestrado que fez o agendamento para pagar a inscrição no mestrado, mas o Banco do Brasil não efetivou o pagamento na data prevista e com isso a inscrição do candidato foi recusada. A Comissão discutiu e aprovou o recurso do candidato.

3) Recredenciamento de orientador - Foram aprovados os recredenciamentos dos *Profs. Drs.* como docente permanente junto ao Programa de Pós-Graduação em Computação/UFRGS, com atividade na Pós-Graduação em ensino e orientação, no nível de mestrado no período 13/01/2021 a 30/08/2022.

4) Desligamento do curso - Foi aprovado o desligamento do curso de mestrado de (orientador:). Motivo: o aluno optou por manter o vínculo empregatício.

A reunião foi encerrada às 11h10min.

ComPG Ata 01/2021 (13/01/2021)

5.1.3 Análise dos resultados das Consultas

Para analisar de forma quantitativa e qualitativa os resultados conforme as consultas, foram utilizadas as métricas DCG@K, IDCG@K, NDCG@K, onde K representa o ponto de corte no ranking dos resultados analisados. Além disso, foi utilizada uma representação visual dos top K resultados, a fim de gerar mais *insights* sobre os resultados.

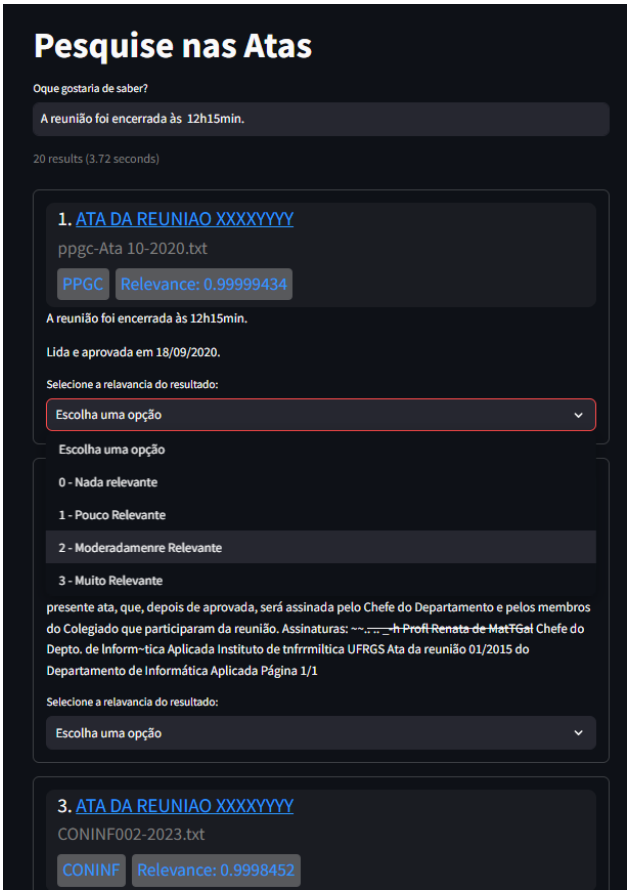
Para anotar os dados, foi necessário adaptar a interface como mostrado na Figura 5.6. A fim de quantificar a relevância dos resultados em relação à consulta, utilizou-se um conjunto ordinal de quatro classes, consistente com (MOREIRA, 2023, chap 16):

- 0 - Nada Relevante
- 1 - Pouco Relevante
- 2 - Moderadamente Relevante
- 3 - Muito Relevante

5.1.3.1 ***Exemplo sobre Créditos de Extensão***

Nesse exemplo, o usuário gostaria de recuperar informações importantes sobre créditos de extensão. Dessa maneira, foram atribuídos os seguintes escores de relevância conforme o resultado: Muito Relevante (3), para documentos em que o principal assunto

Figura 5.6 – Interface adaptada para anotação de relevância



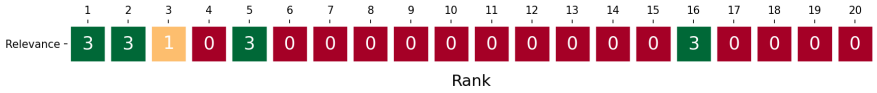
são créditos de extensão. Moderadamente Relevante (2), resultados que de alguma forma possuem informações sobre Atividades de Extensão, mas não têm esse assunto como um dos tópicos principais. Pouco relevante (1), resultados em que créditos complementares são citados. Nada relevante (0): demais documentos.

Consulta: "Discussões, decisões ou menções relacionadas a créditos de extensão em atividades acadêmicas"

- NDCG@20 = 0.9759884319660093
- DCG@20: 28.965616052847604
- IDCG@20: 29.67823706117081
- Resultado Visual:

Consulta: "Créditos de Extensão"

- NDCG@20 = 0.8918520155507559
- DCG@20: 16.337031720469575
- IDCG@20: 18.318096988748493
- Resultado Visual:



5.1.3.2 ***Exemplo CEI***

Nesse exemplo, o usuário gostaria de procurar por documentos que contenham informações relevantes a respeito do Centro de Empreendimentos em Informática (CEI). Dessa forma, resultados que continham algum trecho de texto com informações sobre o CEI foram considerados como muito Relevante (3). Resultados que continham algum trecho de texto que cita o CEI, porém não como principal tópico, foram considerados moderadamente relevantes (2). Todos os outros resultados foram considerados nada relevantes (0), incluindo citações de membros do CEI em contextos totalmente irrelevantes para essa consulta, como, por exemplo, no trecho dos membros presentes na reunião, em que muitas vezes é citada a professora Luciana (Diretora do CEI).

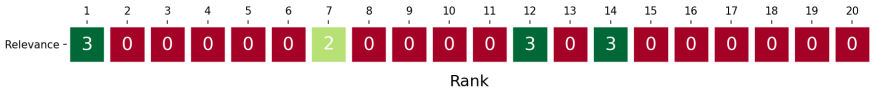
Consulta: "Centro de Empreendimento em Informática"

- NDCG@20 = 0.9957812160106706

- DCG@20: 36.28341128456305
- IDCG@20: 36.43713167227915
- Resultado Visual:

Consulta: "CEI"

- NDCG@20 = 0.7208159854554872
- DCG@20: 11.683373254659948
- IDCG@20: 16.20853794922038
- Resultado Visual:

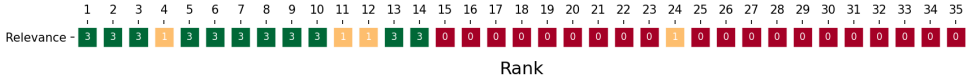


5.1.3.3 ***Exemplo sobre o afastamento do professor***

Nesse exemplo, foi escolhido um professor do INF, que é citado diversas vezes no conjunto de dados. A fim de não expor informações sobre o professor, o nome do professor é representado por <nome> na consulta. Além disso, afastamento é um tópico presente no conjunto de dados. Dessa forma, com a consulta: "Afastamento do professor <nome>", queremos verificar se os modelos conseguem ser suficientemente precisos na busca, para retornar resultados que sejam a intersecção entre dois tópicos comuns do conjunto de dados. Na atribuição do escore de relevância, foram considerados muito relevantes (3) apenas trechos que continham informação sobre afastamento do professor. Resultados pouco relevantes (1) foram trechos de texto que citavam algum afastamento no qual o professor teve alguma relevância, mas não foi o indivíduo afastado. Como, por exemplo: casos em que o professor substituiu algum professor afastado. Os demais resultados foram considerados como nada relevantes (0).

Consulta: "Afastamento do Professor <nome>"

- NDCG@35 = 0.96606733214257
- DCG@35: 33.61562831622226
- IDCG@35: 34.79636170045065
- Resultado Visual:



5.1.4 Discussão dos resultados

A comparação dos resultados mostradas em 5.1.3.1 apresenta um *insight* sobre como a forma de escrita da consulta pode influenciar nos resultados. Na consulta (i) "Discussões, decisões ou menções relacionadas a créditos de extensão em atividades acadêmicas"o usuário escreve de forma mais detalhada e extensa sobre o tipo de documento que deseja procurar, dando um contexto mais rico na consulta. Por outro lado, na consulta (ii) "Créditos de Extensão"o usuário apenas provém o termo isolado sobre o que gostaria de procurar. A consulta (i), na qual é fornecido um contexto mais rico, gerou resultados melhores em todas as métricas. Isso sugere que a busca semântica pode se beneficiar de consultas mais detalhadas, e que forneçam um contexto mais rico.

Já na segunda comparação (5.1.3.2), foram comparadas as consultas (iii) "Centro de Empreendimento em Informática"e (iv) "CEI", uma vez que CEI é a sigla para Centro de Empreendimentos em Informática. Esta sigla, no contexto do instituto em questão, é bem conhecida. Porém, como é utilizado um modelo que não é ajustado para o contexto do INF, o modelo não consegue associar o nome com a sigla. A qualidade dos resultados apresentou uma diferença ainda maior que no primeiro exemplo. Isso está relacionado com o fato dos modelos não possuírem o conhecimento de que CEI e Centro de Empreendimentos em Informática se referem à mesma organização. Para os modelos, "CEI"é apenas uma sigla inserida no texto. Muitos dos resultados retornados na consulta (iv) eram aberturas de atas nas quais a sigla CEI aparecia apenas na especificação do cargo de um membro presente. Resultados como esse claramente não são relevantes para a consulta. A hipótese gerada a partir dessas observações é que: os modelos não entendem o significado da sigla CEI, a relevância atribuída a um documento é exclusivamente com a presença ou não da sigla.

No terceiro exemplo 5.1.3.3 não foi feita uma comparação entre duas consultas, mas foi analisado o resultado da consulta "Afastamento do Professor <nome>". Para a análise, foram considerados os 35 resultados mais relevantes, e as métricas indicam um bom resultado, assim como a representação visual também mostra um *feedback* positivo. Isso mostra que os modelos conseguiram ser precisos na busca e retornar a intersecção

entre dois tópicos frequentes no conjunto de dados.

6 CONCLUSÃO

Este trabalho apresenta uma abordagem para Recuperação de Informações de atas de Reunião focando no uso de modelos *out-of-the-box*, ou seja, modelos que não são ajustados para o domínio de dados em que são utilizados, utilizando alternativas mais genéricas e acessíveis do que modelos com ajuste fino.

O *pipeline* utilizado abrange a extração e segmentação semântica do texto, seguida da busca semântica de dois estágios (*Retrieve and Rerank*), e posterior apresentação dos resultados para o usuário. A aplicação da abordagem apresentou um desempenho razoável na tarefa de recuperar as informações presentes nas atas. Mesmo que utilizando um pipeline simples, a avaliação qualitativa mostrou certo grau de precisão e revocação com o uso da abordagem, ainda que limitados. A abordagem, sem dúvidas, poderia se beneficiar da utilização de mais técnicas para extrair mais desempenho do sistema de RI, já que existem diversas técnicas para RI consolidadas, as quais não foram exploradas.

6.1 Limitações

Nessa sessão são apresentadas e discutidas algumas das principais limitações desse trabalho.

Uma das limitações é em relação à abordagem de Segmentação Semântica utilizando in-context-learning com o GPT4. Embora essa abordagem tenha desempenho adequado para o caso de uso, não possui viabilidade em sistemas de larga escala devido ao custo. A utilização do modelo GPT4-Turbo foi capaz de realizar a tarefa de segmentação *out-of-the-box*, embora a avaliação qualitativa dos resultados sugere que a segmentação poderia ser melhor. Além disso, esse método de segmentação não pode ser utilizado com documentos longos, uma vez que utiliza a arquitetura Transformer, implicando em uma janela de contexto limitada. Com o uso do GPT4-Turbo, o tamanho máximo de documentos é 4096 tokens, já que é a quantidade máxima de tokens permitidos na geração via API. No contexto aplicado ao INF, esse método foi viável devido ao tamanho das atas. Porém, em casos como segmentação de livros, artigos científicos, esse método pode não ser aplicável devido à natureza mais extensa dos documentos.

Outra limitação desse trabalho é a metodologia empregada na análise e validação dos resultados. Para um nível maior de significância, a abordagem necessita de uma avaliação mais extensa e robusta, já que muitas das análises foram empíricas e os experimentos

conduzidos não são tão significativos, já que carecem de uma metodologia mais rigorosa. Os experimentos também foram conduzidos apenas no corpus de documentos do INF, e devido à falta de dados anotados e à dificuldade de gerá-los, não foi possível conduzir experimentos mais extensos. Os experimentos foram pensados para endereçar aspectos específicos do *pipeline*, mas infelizmente utilizaram um tamanho amostral pequeno, que por sua vez possivelmente contém viés cognitivo do autor.

Uma das limitações do trabalho é a qualidade dos resultados. Embora a avaliação qualitativa mostrou certo um grau de precisão e revocação com o uso da abordagem, esse desempenho é adequado apenas nos casos de uso em que o sistema serve apenas como uma heurística para procurar pelas informações. Não sendo adequado em cenários que necessitam uma maior precisão e revocação. Além disso, existem várias técnicas descritas na área de Recuperação de Informação que poderiam ter sido utilizadas para extrair um maior desempenho do sistema de RI, porém, não foram utilizadas no pipeline, como, por exemplo: **Query Augmentation**, **Document Augmentation**.

Com respeito às questões no que diz respeito à segurança e à privacidade dos dados com a utilização da API de terceiros. Em muitos cenários corporativos, atas de reuniões possuem informações confidenciais, e ao utilizar a API de terceiros, pode-se colocar em risco a segurança e confidencialidade desses dados, levantando questionamentos sobre a aplicabilidade dessa abordagem em alguns cenários.

6.2 Trabalhos Futuros

Trabalhos futuros podem servir primeiramente para ajustar as limitações do trabalho atual. A validação poderia ser conduzida em um corpus mais extenso e anotado. Existem vários *benchmarks* para avaliar o desempenho em tarefas de Recuperação de Informação como, por exemplo, o MTEB (MUENNIGHOFF et al., 2023), BEIR(THAKUR et al., 2021), e a utilização desses *benchmarks* possibilitaria uma comparação mais direta com outros *pipelines* de RI.

Além disso, para melhorar o desempenho da abordagem, pode ser explorada a técnica de *Query Augmentation* como em (CHEN; WISEMAN, 2023; MAO et al., 2021), essa técnica é amplamente descrita e validada na literatura de RI(HAMBARDE; PROENCA, 2023). Também seria valiosa a tentativa de utilizar outros paradigmas de interação com o usuário na Recuperação de Informação, como, por exemplo, um *chatbot*, permitindo ao usuário especificar e refinar melhor a sua intenção de busca. Além disso, seria

valioso explorar a aplicação de OCR para a extração do texto das atas, já que o trabalho atual utiliza apenas documentos nativamente digitais.

REFERÊNCIAS

AHMAD, S. R. **Enhancing Multilingual Information Retrieval in Mixed Human Resources Environments: A RAG Model Implementation for Multicultural Enterprise**. 2024.

AMATI, G. Bm25. In: ____. **Encyclopedia of Database Systems**. Boston, MA: Springer US, 2009. p. 257–260. ISBN 978-0-387-39940-9. Available from Internet: <https://doi.org/10.1007/978-0-387-39940-9_921>.

ARTIFEX. **PyMuPDF**. [S.l.], 2015–2024. Available from Internet: <pymupdf.readthedocs.io>.

BAI, H. et al. Segformer: A topic segmentation model with controllable range of attention. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 37, n. 11, p. 12545–12552, Jun. 2023. Available from Internet: <https://ojs.aaai.org/index.php/AAAI/article/view/26477>.

BALAGUER, A. et al. **RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture**. 2024.

BAUDRU, J.; ROSELLO, L. B.; BERSINI, H. Ace: Adaptative chatgpt for entreprise.

BRIGGS, J. **Natural Language Processing for Semantic Search**. 2024. Available from Internet: <https://www.pinecone.io/learn/series/nlp/dense-vector-embeddings-nlp/>.

BROWN, T. B. et al. **Language Models are Few-Shot Learners**. 2020.

CHEN, X. et al. **Salient Phrase Aware Dense Retrieval: Can a Dense Retriever Imitate a Sparse One?** 2022.

CHEN, X.; WISEMAN, S. **BM25 Query Augmentation Learned End-to-End**. 2023.

CHOR, B. et al. Private information retrieval. **Journal of the ACM (JACM)**, ACM New York, NY, USA, v. 45, n. 6, p. 965–981, 1998.

COHERE. **Rerank: Improve search performance with a single line of code**. 2024. Available from Internet: <https://cohere.com/rerank>.

DEVLIN, J. et al. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. 2019.

DURRHEIM, K. et al. Using word embeddings to investigate cultural biases. **British Journal of Social Psychology**, v. 62, n. 1, p. 617–629, 2023. Available from Internet: <https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1111/bjso.12560>.

EL-KHAIR, I. A. Tf*idf. In: ____. **Encyclopedia of Database Systems**. Boston, MA: Springer US, 2009. p. 3085–3086. ISBN 978-0-387-39940-9. Available from Internet: <https://doi.org/10.1007/978-0-387-39940-9_956>.

FINARDI, P. et al. **The Chronicles of RAG: The Retriever, the Chunk and the Generator**. 2024.

FREITAS, C. Dataset e corpus. In: CASELI, H. M.; NUNES, M. G. V. (Ed.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. BPLN, 2023. book chapter 14. ISBN 978-65-00-80693-9. Available from Internet: <https://brasileiraspln.com/livro-pln/1a-edicao/parte7/cap14/cap14.html>.

HAMBARDE, K. A.; PROENCA, H. **Information Retrieval: Recent Advances and Beyond**. 2023.

INDYK, P.; MOTWANI, R. Approximate nearest neighbors: towards removing the curse of dimensionality. In: **Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing**. New York, NY, USA: Association for Computing Machinery, 1998. (STOC '98), p. 604–613. ISBN 0897919629. Available from Internet: <https://doi.org/10.1145/276698.276876>.

IZACARD, G. et al. Unsupervised dense information retrieval with contrastive learning. **Trans. Mach. Learn. Res.**, v. 2022, 2021. Available from Internet: <https://api.semanticscholar.org/CorpusID:249097975>.

KAMALLOO, E. et al. Evaluating embedding apis for information retrieval. In: **Annual Meeting of the Association for Computational Linguistics**. [s.n.], 2023. Available from Internet: <https://api.semanticscholar.org/CorpusID:258587920>.

KAPLAN, J. et al. **Scaling Laws for Neural Language Models**. 2020.

KOBAYASHI, M.; TAKEDA, K. Information retrieval on the web. **ACM computing surveys (CSUR)**, ACM New York, NY, USA, v. 32, n. 2, p. 144–173, 2000.

KOJIMA, T. et al. Large language models are zero-shot reasoners. In: KOYEJO, S. et al. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2022. v. 35, p. 22199–22213. Available from Internet: <https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf>.

LAY, M. Information retrieval. In: **Papers from the Australian Road Research Board 23rd Regional Symposium and local Government Engineers Association of Western Australia. Third State Conference, Perth, 1985: Technical Papers.** [S.l.: s.n.], 1985.

LI, J. et al. Neural text segmentation and its application to sentiment analysis. **IEEE Transactions on Knowledge and Data Engineering**, v. 34, n. 2, p. 828–842, 2022.

LUKASIK, M. et al. **Text Segmentation by Cross Segment Attention**. 2020.

MA, X. et al. Fine-tuning llama for multi-stage text retrieval. **ArXiv**, abs/2310.08319, 2023. Available from Internet: <https://api.semanticscholar.org/CorpusID:263908865>.

MAILLARD, J. et al. **Multi-task Retrieval for Knowledge-Intensive Tasks**. 2021.

MAO, Y. et al. **Generation-Augmented Retrieval for Open-domain Question Answering**. 2021.

MICULICICH, L.; HAN, B. **Document Summarization with Text Segmentation**. 2023.

MIKOLOV, T. et al. **Efficient Estimation of Word Representations in Vector Space**. 2013.

MOREIRA, V. Recuperação de informação. In: CASELI, H. M.; NUNES, M. G. V. (Ed.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. BPLN, 2023. book chapter 16. ISBN 978-65-00-80693-9. Available from Internet: <https://brasileiraspln.com/livro-pln/1a-edicao/parte8/cap16/cap16.html>.

MUENNIGHOFF, N. et al. **MTEB: Massive Text Embedding Benchmark**. 2023.

OPENAI. **Embeddings**. 2024. Accessed on 12/3/2024. Available from Internet: <https://platform.openai.com/docs/guides/embeddings>.

PENNINGTON, J.; SOCHER, R.; MANNING, C. GloVe: Global vectors for word representation. In: MOSCHITTI, A.; PANG, B.; DAELEMANS, W. (Ed.). **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Doha, Qatar: Association for Computational Linguistics, 2014. p. 1532–1543. Available from Internet: <https://aclanthology.org/D14-1162>.

PIVETTA, M. V. L. **An Information Retrieval and Extraction Tool for Covid-19 Related Papers**. 2024.

REDDY, R. G. et al. **Inference-time Re-ranker Relevance Feedback for Neural Information Retrieval**. 2023.

REIMERS, N.; GUREVYCH, I. **Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks**. 2019.

SASAZAWA, Y. et al. **Text Retrieval with Multi-Stage Re-Ranking Models**. 2023.

SETH, R.; SHARAFF, A. Sentiment data analysis for detecting social sense after covid-19 using hybrid optimization method. **SN Computer Science**, v. 4, 07 2023.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: **9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)**. [S.l.: s.n.], 2020.

THAKUR, N. et al. **BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models**. 2021.

VASWANI, A. et al. Attention is all you need. In: . [s.n.], 2017. Available from Internet: <https://arxiv.org/pdf/1706.03762.pdf>.

ZHAO, W. X. et al. Dense text retrieval based on pretrained language models: A survey. **ACM Trans. Inf. Syst.**, Association for Computing Machinery, New York, NY, USA, v. 42, n. 4, feb 2024. ISSN 1046-8188. Available from Internet: <https://doi.org/10.1145/3637870>.

ZHOU, Y. et al. **Large Language Models Are Human-Level Prompt Engineers**. 2023.

ZHU, Y. et al. **Large Language Models for Information Retrieval: A Survey**. 2024.