

UMA ABORDAGEM PARA RECUPERAÇÃO SEMÂNTICA DE INFORMAÇÕES INSTITUCIONAIS COM REPRESENTAÇÕES VETORIAIS

ANTONIO PAULO MARTINS FERRI¹; GUILHERME DALLMANN LIMA²; MARILIA ROSA SILVEIRA²; LARISSA ASTROGILDO DE FREITAS³; ULISSES BRISOLARA CORRÊA³

¹Universidade Federal de Pelotas (UFPEL) – apmferri@inf.ufpel.edu.br

²Universidade Federal de Pelotas (UFPEL) – gdlima@inf.ufpel.edu.br, mrsilveira@inf.ufpel.edu.br

³Universidade Federal de Pelotas (UFPEL) – larissa@inf.ufpel.edu.br, ulisses@inf.ufpel.edu.br

1. DESCRIÇÃO DA INOVAÇÃO

Modelos de linguagem de grande escala (do inglês, *Large Language Models* — LLMs) têm ampliado as possibilidades de desenvolvimento de sistemas capazes de interagir com usuários por meio de linguagem natural. Entretanto, informações específicas, privadas ou atualizadas podem não estar presentes nos dados de treinamento, o que limita sua aplicação em contextos institucionais, nos quais consultas frequentemente dependem de informações particulares de uma organização.

Nesse contexto, a técnica de geração aumentada via recuperação (do inglês, *Retrieval-Augmented Generation* — RAG) permite complementar o conhecimento do modelo por meio da recuperação de informações provenientes de bases externas (LEWIS et al., 2020). Uma das formas de realizar essa recuperação é por meio de *embeddings*, representações vetoriais que possibilitam comparar semanticamente consultas e documentos e recuperar conteúdos relacionados mesmo sem correspondência lexical exata (KARPUKHIN et al., 2020).

A abordagem investigada neste trabalho consiste na recuperação semântica de informações institucionais em português por meio de representações vetoriais, considerando conjuntamente a dimensionalidade dos *embeddings* e a granularidade da fragmentação dos documentos. A configuração do processo de recuperação pode influenciar tanto a qualidade dos documentos recuperados quanto os recursos computacionais necessários. Parte-se da hipótese de que representações vetoriais de menor dimensionalidade, combinadas a uma granularidade adequada de *chunks*, podem manter uma recuperação semanticamente relevante com menor custo computacional. Assim, o diferencial a ser investigado está na análise sistemática dessas variáveis no contexto institucional da UFPEL, considerando a relevância dos resultados, o armazenamento e a eficiência das consultas. Trabalhos anteriores indicam que tanto a dimensionalidade das representações quanto a segmentação dos documentos podem afetar sistemas de recuperação semântica (WANG; ZHANG; YU, 2023; WANG et al., 2025).

2. ANÁLISE DE MERCADO

Neste trabalho, o contexto de aplicação é institucional, tendo como estudo de caso a Universidade Federal de Pelotas (UFPEL). O público diretamente relacionado à abordagem é formado por usuários que necessitam consultar informações acadêmicas e institucionais, enquanto o corpus considerado será restrito aos cursos de graduação e pós-graduação em Computação disponíveis no site institucional da universidade¹. Serão contemplados cinco cursos: Ciência da Computação, Engenharia de Compu-

¹<https://institucional.ufpel.edu.br/>

tação, Mestrado Acadêmico em Computação, Doutorado em Computação, ambos vinculados ao Programa de Pós-Graduação em Computação, e Especialização em Computação na Educação Básica.

Os dados a serem coletados abrangerão informações sobre os cursos, disciplinas, docentes, turmas ofertadas e projetos vigentes. Para os cursos, serão obtidos dados de caracterização, currículo, formas de ingresso, vagas e indicadores. Para as disciplinas, serão coletados dados sobre identificação, carga horária, ementa, objetivo, conteúdo programático, bibliografia, pré-requisitos e equivalências. Em relação aos docentes, serão coletadas informações sobre identificação, formação acadêmica, titulação, áreas de atuação, vínculo institucional e disciplinas ministradas. Também serão obtidos dados das turmas ofertadas, incluindo vagas, matrículas, docentes responsáveis e horários. Por fim, serão coletadas informações sobre os projetos vigentes de ensino, pesquisa e extensão, incluindo sua caracterização, vigência e participantes.

A análise proposta não tem como foco, nesta etapa, a comercialização da solução ou a estimativa de TAM, SAM e SOM. O recorte é a avaliação tecnológica da recuperação semântica em um ambiente institucional real, comparando configurações capazes de apoiar mecanismos de acesso às informações da universidade. Dessa forma, a comparação com alternativas existentes será realizada no nível técnico das configurações de recuperação, e não como análise comercial de concorrentes.

3. ESTRATÉGIA DE DESENVOLVIMENTO E IMPLEMENTAÇÃO

A estratégia de desenvolvimento corresponde ao processo de obtenção do corpus, seleção dos modelos de *embedding*, definição dos parâmetros de fragmentação (*chunk*) e sobreposição (*overlap*), pré-processamento e vetorização dos textos e, por fim, avaliação das diferentes configurações. Como o trabalho se encontra em etapa de investigação e validação experimental, não é proposto, neste momento, um modelo de negócios ou registro de propriedade intelectual; o desenvolvimento está concentrado na construção e avaliação do mecanismo de recuperação.

Seleção dos Modelos de *Embeddings*

A dimensão do *embedding* desempenha um papel importante na busca semântica, pois determina o número de componentes utilizados para representar cada documento no espaço vetorial (REIMERS; GUREVYCH, 2019). Nesse sentido, serão avaliados modelos pertencentes à mesma família arquitetural, mas com diferentes dimensões de *embedding*. Essa escolha permitirá comparar de forma mais justa o impacto da dimensionalidade no processo de representação vetorial dos documentos, mantendo características relacionadas à arquitetura e ao treinamento dos modelos, que influenciam a forma como os dados são organizados no espaço latente. Serão avaliadas as dimensões de 768, 1024 e 2048 (WANG; ZHANG; YU, 2023).

Definição dos Parâmetros de Fragmentação e Sobreposição

O corpus será fragmentado em trechos de 150, 300 e 500 caracteres. Esses valores foram definidos a partir de uma análise preliminar dos textos disponíveis no portal institucional, buscando contemplar diferentes níveis de fragmentação e permitir a avaliação de seus impactos na busca semântica.

Para a sobreposição, serão avaliados os valores de 20, 30 e 50 caracteres. A utilização de sobreposição tem como objetivo preservar parte do contexto entre fragmentos consecutivos, reduzindo a perda de informações quando uma unidade de sig-

nificado for dividida em diferentes trechos. Dessa forma, serão avaliadas nove combinações entre tamanho de fragmento e sobreposição. A análise buscará verificar como essas configurações influenciam a representação vetorial e a recuperação semântica dos documentos, considerando também a quantidade de fragmentos gerados em cada configuração. A relevância da segmentação para sistemas de recuperação aumentada também é discutida por Wang et al. (2025).

Pré-Processamento e Vetorização dos Textos

Os dados coletados pelo *crawler* serão submetidos a uma etapa de pré-processamento antes da vetorização. Como as informações do portal apresentam diferentes formatos e níveis de detalhamento, os conteúdos serão organizados em textos autocontidos, separados por entidade e por escopo semântico. Dessa forma, uma disciplina, por exemplo, poderá possuir representações distintas para sua ementa, objetivo, conteúdo programático e bibliografia.

Para preservar o contexto dos textos, cada fragmento será acompanhado de informações que identifiquem a entidade e o respectivo escopo. Também serão tratados campos ausentes e removidos conteúdos considerados insuficientes ou irrelevantes para a representação semântica. Textos longos serão fragmentados conforme os parâmetros definidos anteriormente, buscando preservar fronteiras naturais, como parágrafos e frases, e utilizando sobreposição entre fragmentos consecutivos.

Após o pré-processamento, os textos serão transformados em vetores utilizando modelos de *embedding*. Os vetores serão armazenados em um banco de dados PostgreSQL com a extensão *pgvector*, juntamente com os identificadores, textos e metadados correspondentes. Essa organização permitirá relacionar os resultados da busca vetorial às entidades originais e aplicar filtros por atributos, permitindo a avaliar as configurações de fragmentação, sobreposição e dimensionalidade dos *embeddings*.

Protocolo de Avaliação das Configurações

A avaliação considerará as nove combinações de tamanho de fragmento e sobreposição definidas anteriormente. Cada configuração será mantida de forma independente, permitindo a comparação dos resultados sem interferência da fragmentação.

O conjunto de avaliação será composto por dez consultas distribuídas entre diferentes entidades, como cursos, disciplinas, docentes, turmas ofertadas e projetos vigentes. As consultas serão formuladas de maneira parafraseada em relação aos documentos, buscando avaliar a capacidade dos modelos de realizar recuperação semântica mesmo sem correspondência direta entre os termos da consulta e do corpus.

Para cada configuração e consulta, o texto será vetorizado com o mesmo modelo utilizado na indexação, sendo recuperados os dez documentos mais próximos com base na similaridade de cosseno. Serão analisados o *score* dos resultados, a margem entre os documentos recuperados, a quantidade de resultados plausíveis e a sobreposição dos documentos recuperados entre diferentes configurações.

4. RESULTADOS ESPERADOS E IMPACTO

Espera-se que os resultados permitam identificar a influência do tamanho dos fragmentos, da sobreposição e da dimensionalidade dos *embeddings* na recuperação semântica dos documentos. A comparação entre as diferentes configurações possibilitará verificar quais combinações apresentam melhor equilíbrio entre qualidade da recuperação e quantidade de fragmentos gerados.

A análise também permitirá discutir a relação entre a granularidade dos fragmentos e a preservação do contexto semântico, bem como avaliar se o aumento da dimensionalidade dos vetores resulta em melhorias significativas na recuperação. Além da qualidade dos documentos recuperados, pretende-se observar aspectos relacionados ao armazenamento e à eficiência das consultas no contexto da UFPEL. Dessa forma, o impacto esperado é a identificação de configurações mais adequadas para a aplicação de busca semântica no contexto institucional, contribuindo para mecanismos de acesso mais eficientes e integração às aplicações baseadas em RAG.

5. CONCLUSÕES

Este trabalho propõe investigar como diferentes configurações de dimensionalidade dos *embeddings*, tamanho de *chunks* e sobreposição entre fragmentos podem influenciar a recuperação semântica de informações. A hipótese considerada é de que configurações mais compactas podem obter uma recuperação semanticamente adequada, ao mesmo tempo em que reduzem a quantidade de dados armazenados e os recursos necessários para a busca.

Como continuidade do trabalho, pretende-se ampliar o conjunto de consultas utilizadas na avaliação e realizar uma análise manual da relevância dos documentos recuperados, possibilitando o emprego de métricas mais robustas de recuperação de informação. Espera-se que os resultados possam contribuir para o desenvolvimento de mecanismos de acesso mais eficientes e, posteriormente, para sua integração a aplicações baseadas em RAG.

6. REFERÊNCIAS BIBLIOGRÁFICAS

- KARPUKHIN, V.; OGUZ, B.; MIN, S.; LEWIS, P.; WU, L.; EDUNOV, S.; CHEN, D.; YIH, W. Dense Passage Retrieval for Open-Domain Question Answering. In: **CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP)**, 2020, Online. **Proceedings...** Association for Computational Linguistics, 2020. p. 6769–6781.
- LEWIS, P.; PEREZ, E.; PIKTUS, A.; PETRONI, F.; KARPUKHIN, V.; GOYAL, N.; KÜTLER, H.; LEWIS, M.; YIH, W.; ROCKTÄSCHEL, T.; RIEDEL, S.; KIELA, D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: **CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NEURIPS)**, 2020. **Advances in Neural Information Processing Systems**, v. 33, p. 9459–9474, 2020.
- WANG, H.; ZHANG, H.; YU, D. On the Dimensionality of Sentence Embeddings. In: **CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP)**, Singapore, 2023. **Findings of the Association for Computational Linguistics: EMNLP 2023**. Association for Computational Linguistics, 2023. p. 10344–10354.
- WANG, Z.; GAO, C.; XIAO, C.; et al. Document Segmentation Matters for Retrieval-Augmented Generation. In: **ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (ACL)**, Vienna, 2025. **Findings of the Association for Computational Linguistics: ACL 2025**. Association for Computational Linguistics, 2025. p. 8063–8075.
- REIMERS, N.; GUREVYCH, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: **CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP)**, 2019, Hong Kong, China. **Proceedings...** Association for Computational Linguistics, 2019. p. 3982–3992.