PPGCC - UFSCar

Relatório Final

Tema - Correlação de Palavras de Artigos Sobre Saúde Mental

Processamento de Linguagem Natural

Profa. Dra. Helena Caseli

Grupo:

Michel Ribeiro Koba - RA: 792211

Sorocaba
08/07/2025

Introdução

A saúde mental tem se tornado uma das áreas mais discutidas e pesquisadas na atualidade, especialmente diante do aumento significativo nos índices de transtornos como ansiedade, depressão, bipolaridade e outras condições psicológicas que impactam milhões de pessoas no Brasil e no mundo. Esse cenário tem impulsionado a produção científica voltada ao entendimento, diagnóstico e tratamento dessas condições.

Com o crescimento do volume de publicações científicas sobre saúde mental, torna-se cada vez mais desafiador explorar, organizar e identificar conexões relevantes entre os diversos conceitos abordados nos artigos, como sintomas, tratamentos, fatores genéticos, intervenções psicossociais, entre outros. Nesse contexto, ferramentas baseadas em Processamento de Linguagem Natural (PLN) oferecem soluções eficazes para extrair padrões, identificar relações semânticas e apoiar a análise de grandes corpora textuais.

Este projeto propõe o desenvolvimento de um sistema de correlação de palavras-chave com base em artigos científicos escritos em português sobre saúde mental, utilizando técnicas de PLN e aprendizado de máquina. O sistema busca identificar agrupamentos e relações entre termos relevantes, contribuindo para a compreensão do vocabulário e dos temas mais recorrentes na literatura especializada.

Tarefa Escolhida

A tarefa consiste na identificação e visualização de correlações semânticas entre palavras-chave presentes em artigos científicos sobre saúde mental, escritos em português. Para isso, será testada técnicas para representação dos corpus, comparação entre similaridades semânticas entre os termos e correlação entre os clusters formados pelos grupos de artigos.

DataSet

Base de Dados

Para a realização deste projeto, foi construído um *corpus* de artigos científicos sobre saúde mental a partir da Biblioteca Virtual em Saúde (BVS), uma plataforma mantida pelo Ministério da Saúde.

A coleta dos artigos foi realizada utilizando os seguintes critérios de filtragem:

- **Assunto principal:** Saúde Mental
- **Idioma:** Português
- **Ano de publicação:** 2022, 2023 e 2024
- **Tipo de documento:** Artigo Científico

Como resultado da aplicação desses filtros, foi obtido um total de 438 artigos exportados no formato CSV, contendo metadados como título, resumo, autor, periódico e data de publicação.

Durante o processo de ingestão e validação inicial do dataset, foram identificados 5 registros com problemas de formatação, como campos ausentes ou dados corrompidos. Além disso, dentre os relatórios selecionados, 110 estavam em línguas diferentes do portugês. Esses registros foram excluídos, resultando em um conjunto final de 323 artigos com títulos e resumos legíveis e aptos para uso nas etapas de pré-processamento e modelagem.

Esse corpus será utilizado como base para o treinamento dos modelos de representação semântica de palavras e para a análise de correlação de termos relevantes no contexto da saúde mental.

Filtragem de Artigos	
Total Artigos	438
Corrompidos	5
Não Português	110
Restantes	323

Análise dos Dados

O corpus coletado será utilizado como base para as análises de similaridade semântica entre palavras-chave. As principais colunas selecionadas para essa tarefa são: Title (título do artigo), Keyword(s) (palavras-chave atribuídas pelos autores) e Abstract (resumo do artigo).

O objetivo inicial é realizar experimentos comparando diferentes composições do texto de entrada, comparando, por exemplo, os resultados obtidos utilizando os títulos dos artigos e os resumos.

Essa comparação permitirá avaliar o impacto da densidade e diversidade textual na formação dos embeddings e na identificação de correlações semânticas entre os termos.

Pré-Processamento

Como os campos utilizados são compostos por texto livre, é necessário aplicar um processo de pré-processamento linguístico para preparar os dados para as etapas de modelagem. Essa etapa visa reduzir ruídos, padronizar o vocabulário e otimizar a qualidade das representações vetoriais geradas.

Etapas implementadas de pré-processamento são:

- Remoção de elementos irrelevantes: exclusão de números, citações, links e outros caracteres que não contribuem semanticamente.

- Remoção de stopwords: eliminação de palavras sem valor semântico relativo.

- Normalização ortográfica + Lematização: conversão de todas as palavras para minúsculas e redução de palavras flexionadas à sua forma base.

Essas etapas são fundamentais para evitar a dispersão semântica causada por ruídos textuais e para garantir a consistência lexical do corpus, contribuindo para uma melhor performance dos modelos de embeddings (Word2Vec, FastText) utilizados nas análises.

Construção da Pipeline

Modelos de Representação

A tarefa central deste projeto consiste em identificar correlações semânticas entre palavras-chave extraídas de artigos científicos sobre saúde mental, com o objetivo de revelar agrupamentos e relações relevantes entre termos como transtornos, sintomas e tratamentos. A literatura aponta que a melhor forma de realizar essa correlação é por meio do uso das embeddings, realizando a correlação e semelhança entre os vetores formados.

Assim, para representar essas palavras em um espaço vetorial e analisar sua similaridade, serão investigados os métodos baseados em **word embeddings**: **Word2Vec** e **FastText**.

Comparação entre Modelos:

Os modelos serão comparados quanto à sua capacidade de identificar palavras semanticamente próximas, a qualidade dos agrupamentos gerados e sua robustez com termos menos frequentes. Os vetores gerados serão visualizados com técnicas de redução de dimensionalidade como PCA ou UMAP, permitindo a inspeção visual dos agrupamentos semânticos.

Processamento do DataSet

Na fase inicial do projeto, foi realizada a aplicação direta do modelo Word2Vec sobre os dados pré-processados a partir dos títulos e resumos dos artigos presentes no dataset selecionado. Essa abordagem permitiu obter uma visão preliminar dos dados, possibilitando identificar possíveis desafios que poderiam comprometer a qualidade dos resultados ao longo do projeto.

Com o primeiro processamento finalizado, foi realizada uma projeção inicial com PCA (Análise de Componentes Principais) para redução de dimensionalidade, seguida de um plot das 50 palavras mais frequentes (Figura 1). No entanto, ao analisar o resultado, foi possível notar a presença de termos em inglês, evidenciando que o filtro de idioma aplicado inicialmente não havia sido totalmente eficaz.

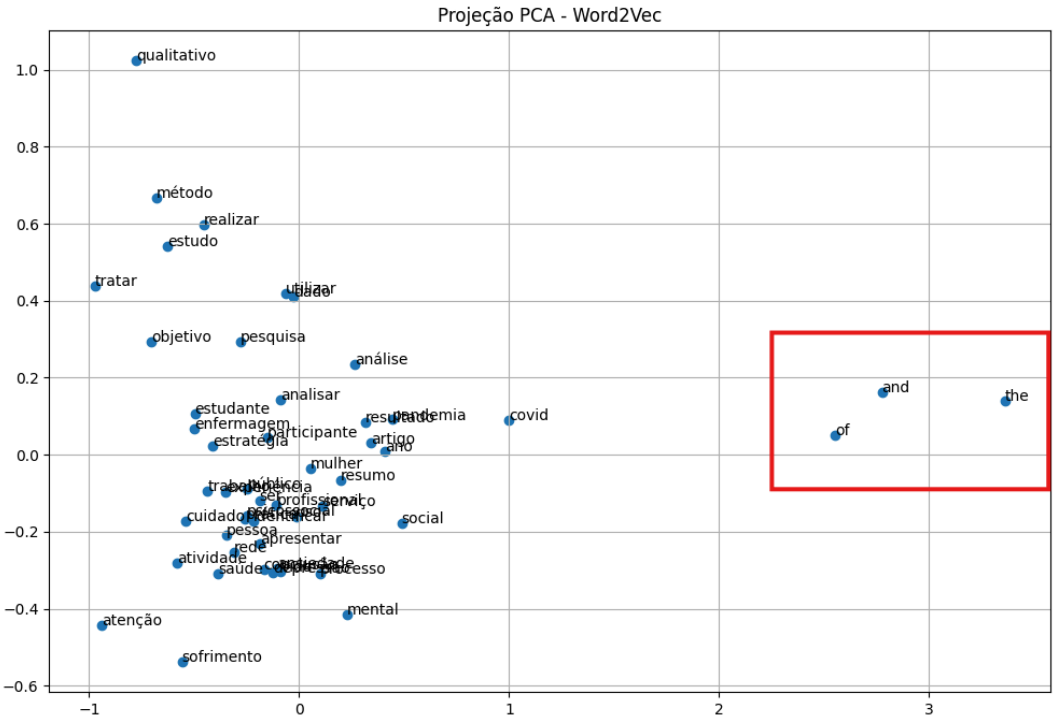

Figura 1: Plot inicial das 50 palavras

A investigação posterior do dataset revelou que alguns artigos estavam mesclados com versões em espanhol ou inglês, juntamente com sua tradução em português. Isso resultava na presença de palavras em múltiplos idiomas nos campos de texto analisados.

Para resolver esse problema, foi incluída uma nova etapa de filtragem utilizando a biblioteca *langdetect*, com o objetivo de identificar e remover tokens que não estivessem em português. No entanto, aplicar o filtro diretamente sobre os resumos ou títulos completos mostrou-se inadequado, pois o funcionamento da biblioteca considera a proporção de palavras em cada idioma, o que poderia levar à exclusão de textos mistos, mas majoritariamente relevantes. Uma aplicação rigorosa desse filtro, por exemplo, reduziria o número de textos utilizáveis de 323 para apenas 81 artigos, perdendo uma porção significativa da base de dados.

Dessa forma, a abordagem adotada foi aplicar uma filtragem das palavras após o pré-processamento e geração das *embeddings* dos textos, o *langdetect* foi utilizado individualmente sobre cada palavra dentro do vocábulo, permitindo a remoção apenas dos termos detectados como pertencentes a outros idiomas, preservando assim a maior parte do conteúdo original em português.

O Resultado após a após a aplicação dessa técnica pode ser observado na figura 2.

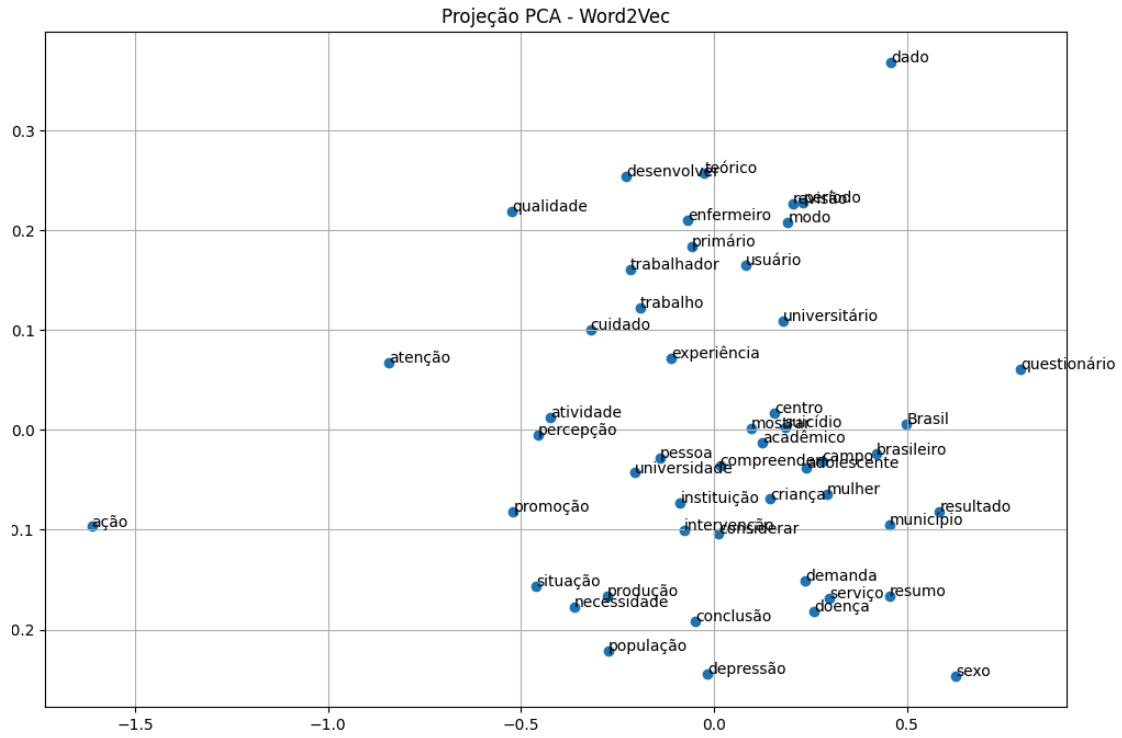

Figura 2: Plot das 50 palavras após filtro de linguagem

Parametrização dos Modelos

Para o treinamento do modelo Word2Vec e FastText, foram definidos os seguintes parâmetros, com o objetivo de balancear desempenho computacional e qualidade semântica dos vetores gerados:

- **vector_size** = 10: Define a dimensionalidade dos vetores de palavra (embeddings). Por serem textos pequenos, é um valor razoável para evitar o overfitting.

- **window** = 5: determina o tamanho da janela de contexto, considerando até 5 palavras antes e depois da palavra central para aprender suas associações semânticas.

- **min_count** = 3: Especifica o número mínimo de ocorrências para que uma palavra seja considerada no treinamento. Esse parâmetro foi utilizado para eliminar palavras com baixa frequência, que geralmente apresentam pouco valor semântico e podem introduzir ruídos no modelo.

- **workers** = 4: Define o número de threads paralelas usadas durante o treinamento, acelerando o processamento em máquinas com múltiplos núcleos.

- **sg** = 1: Modelo Skip-gram, que é mais eficaz para identificar relações semânticas envolvendo palavras menos frequentes — especialmente adequado para este projeto, dado que os textos utilizados são resumos e, portanto, relativamente curtos.

Avaliação Quantitativa

Tabela 1: Similaridade de cosseno W2V vs FastText treinados com resumos

Similaridade de Cosseno (Resumos)		
Pares de Palavras	Word2Vec	FastText
sintoma + ansiedade	9.644	9.976
sintoma + estresse	9.925	9.958
sintoma + depressão	9.908	9.893
sintoma + tristeza	7.915	8.411
sintoma + sono	9.336	9.783
feliz + depressão	3.918	6.555
política + prevenção	9.397	9.828
ansiedade + biblioteca	4.666	6.712

Tabela 2: Similaridade de cosseno W2V vs FastText treinados com títulos

Similaridade de Cosseno (Títulos)		
Pares de Palavras	Word2Vec	FastText
sintoma + ansiedade	8.795	9.998
sintoma + estresse	8.711	9.998
sintoma + depressão	8.889	9.998
sintoma + tristeza	N/A	9.986
sintoma + sono	8.722	9.979
feliz + depressão	N/A	9.804
política + prevenção	1.595	9.997
ansiedade + biblioteca	N/A	9.963

Para comparar o desempenho dos dois modelos treinados, foi realizada uma análise quantitativa baseada na similaridade de cosseno entre pares de palavras. A Tabela 1 apresenta os resultados obtidos treinando os modelos com os resumos, com destaque para conjuntos de palavras que, conceitualmente, deveriam apresentar alta ou baixa correlação.

No caso das comparações realizadas utilizando apenas os títulos dos artigos como base para o treinamento, observou-se uma baixa capacidade de associação semântica entre as palavras. Isso se deve, provavelmente, ao fato de que os títulos, por serem curtos e condensados, não oferecem contexto suficiente para que o modelo aprenda relações significativas entre os termos. Além disso, foi possível notar uma diferença importante entre os modelos analisados: palavras ausentes no vocabulário do Word2Vec estavam presentes no FastText. Essa divergência se deve à diferença entre as implementações dos dois modelos. Enquanto o Word2Vec só é capaz de gerar vetores para palavras vistas exatamente durante o treinamento, o FastText utiliza subpalavras (n-gramas de caracteres), o que permite a geração de vetores mesmo para termos inéditos, conferindo maior flexibilidade ao modelo.

Assim, para o caso do treinamento com resumos, ambos os modelos mostraram bom desempenho ao identificar relações semânticas esperadas, como no

caso de termos relacionados a sintomas de transtornos de saúde mental, que apresentaram alta similaridade (como por exemplo sintoma + estresse).

No entanto, observou-se que o modelo Word2Vec demonstrou maior capacidade de desassociação, ou seja, conseguiu atribuir valores de similaridade mais baixos entre termos que não deveriam estar semanticamente relacionados. Um exemplo é a associação entre as palavras “feliz” e “depressão”, que apresentou menor similaridade no Word2Vec, indicando uma melhor separação conceitual entre os termos.

Esses resultados indicam que, além de capturar bem as correlações esperadas, o modelo Word2Vec foi eficaz em evitar associações indevidas, contribuindo para uma representação semântica mais precisa dos dados. Por conta disso, este modelo foi escolhido como principal para a composição das visualizações.

Avaliação Qualitativa e Visualizações

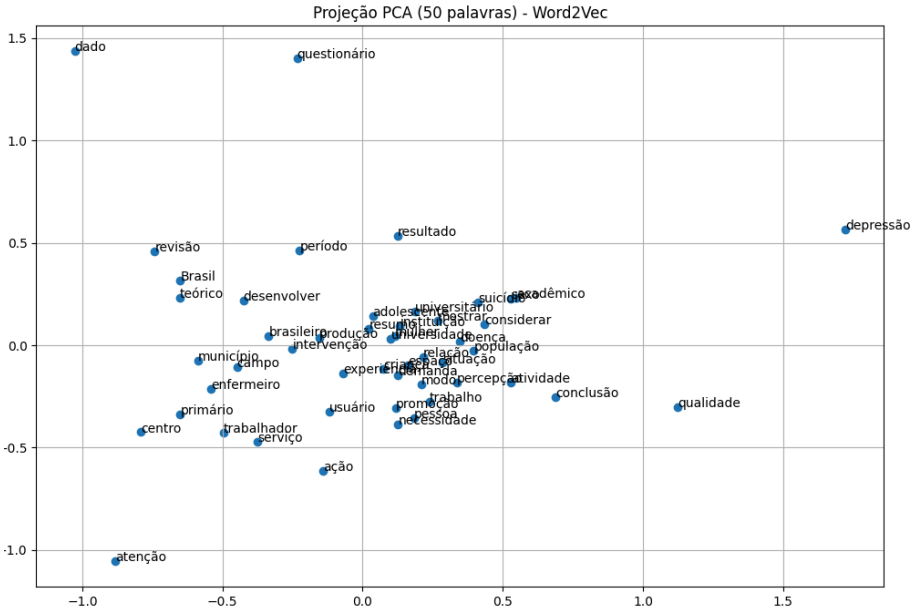

Figura 4: Visualização PCA 50 palavras

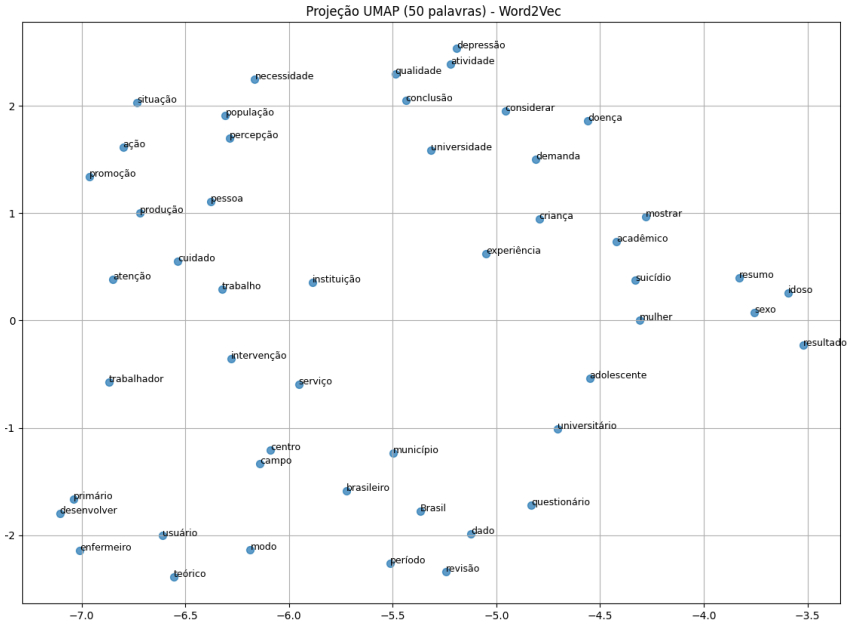

Figura 5: Visualização UMAP 50 palavras

Com o desenvolvimento da pipeline e a definição dos modelos e parâmetros de treinamento, iniciou-se a etapa de avaliação qualitativa, cujo objetivo é fornecer uma análise visual e interpretativa dos resultados obtidos. A partir das representações vetoriais geradas pelos modelos, foram aplicadas técnicas de redução de dimensionalidade — PCA (Principal Component Analysis) e UMAP (Uniform Manifold Approximation and Projection) — para projetar os dados em duas dimensões. Nas Figuras 3 e 4, são apresentados exemplos dos gráficos resultantes, que ilustram a organização espacial das palavras no espaço vetorial, permitindo observar agrupamentos, proximidades e possíveis correlações semânticas de forma mais intuitiva.

Apesar de ser uma visualização inicial e ser possível visualizar alguns agrupamentos, essas representações ainda não tem um valor analítico muito rico.

A análise gráfica mais expressiva foi obtida com o uso do UMAP, especialmente ao definir palavras centrais e observar as palavras mais próximas semanticamente em torno delas, como ilustrado na Figura 6. Esse tipo de visualização permitiu uma exploração mais rica das relações aprendidas pelo modelo. Na mesma figura, é possível extrair *insights* semânticos relevantes, como a identificação de agrupamentos relacionados a sintomas de saúde mental, incluindo termos como depressão, estresse, ansiedade e sono, que foram extraídos com base nos resumos dos artigos. Esses agrupamentos reforçam a capacidade do modelo em capturar contextos específicos e sugerem uma representação consistente do vocabulário em torno de temas clínicos recorrentes nos textos analisados.

Abaixo estão algumas visualizações com centros com associações interessantes.

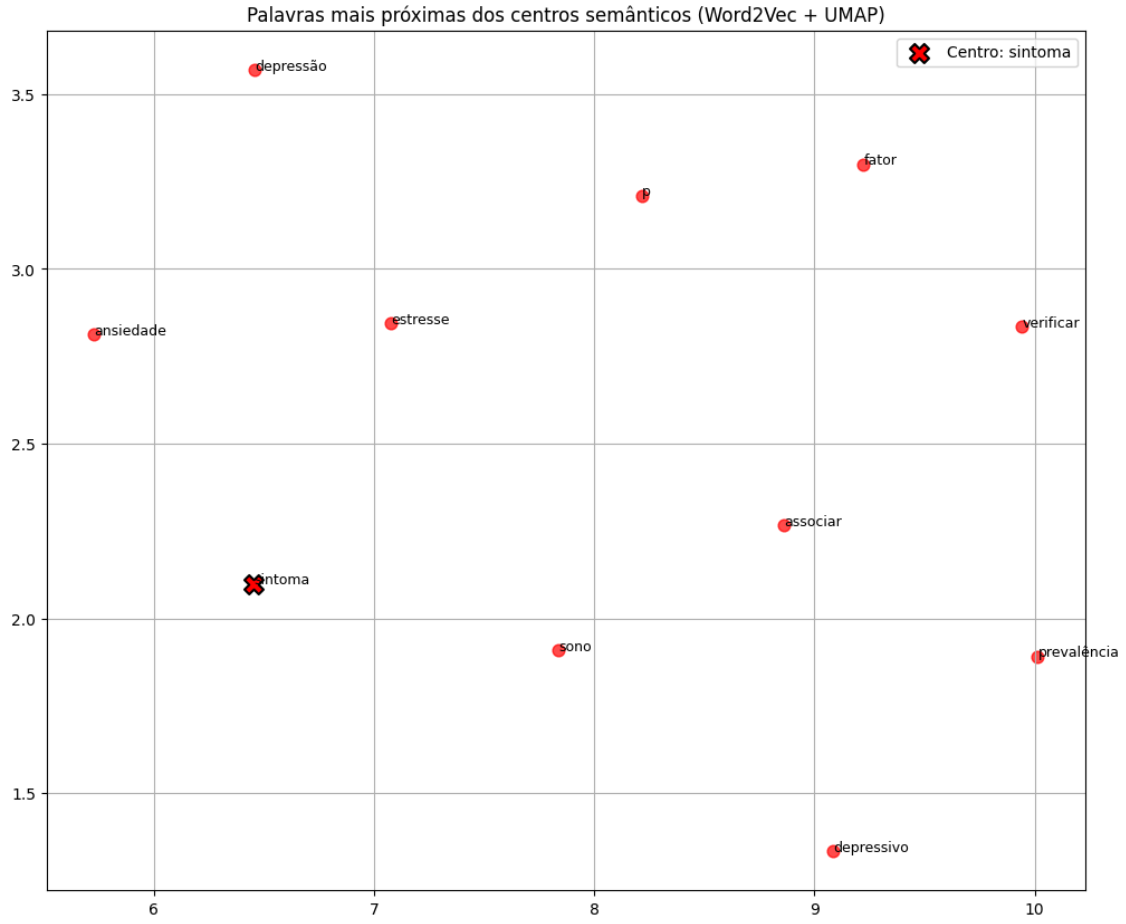

Figura 6: Visualização espacial centro: Sintoma

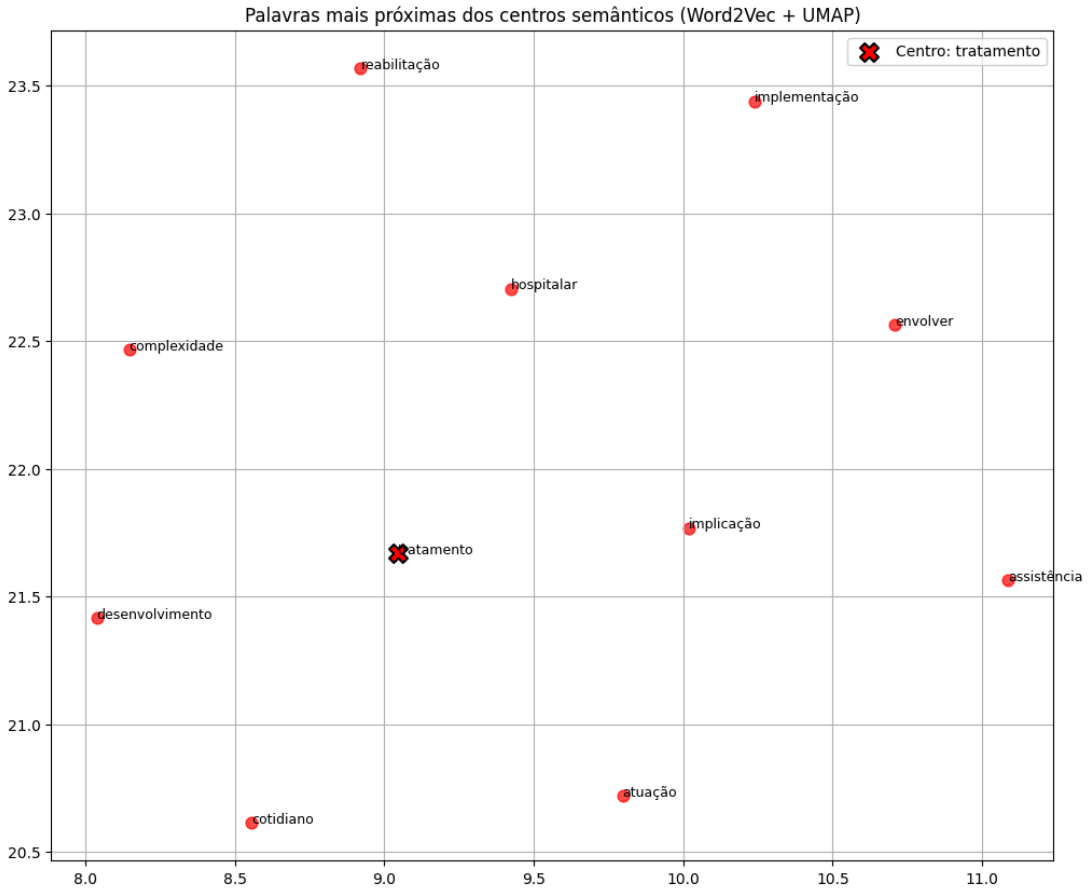

Figura 7: Visualização espacial centro: Tratamento

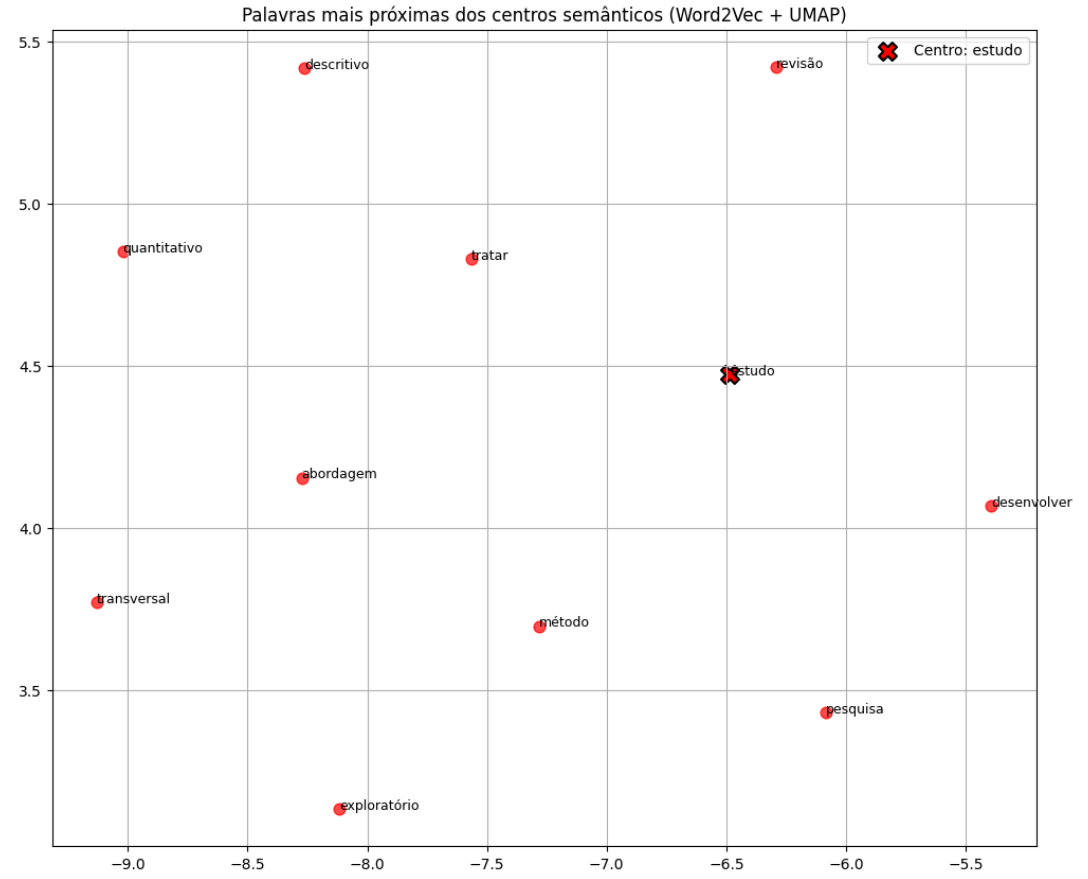

Figura 7: Visualização espacial centro: Estudo

Conclusão

A elaboração deste projeto, desde o planejamento inicial até a implementação prática, proporcionou uma compreensão abrangente sobre a construção de uma pipeline de produção para tarefas de Processamento de Linguagem Natural (PLN). A combinação entre revisão teórica e exploração prática foi essencial para selecionar ferramentas adequadas e lidar com os desafios encontrados durante o desenvolvimento da pipeline.

Durante o processo, ficou evidente que o tratamento de dados provenientes de fontes não estruturadas, como resumos e títulos de artigos científicos, envolve inevitavelmente imprevistos e ruídos. Por isso, compreender e aplicar estratégias de limpeza, filtragem e normalização se mostrou crucial para garantir uma base sólida e confiável para o treinamento dos modelos (mesmo que não tenha ficado perfeita no final do projeto).

O projeto também contribuiu significativamente para o aprofundamento do entendimento sobre o funcionamento dos modelos de embeddings, em especial na distinção entre Word2Vec e FastText. Embora parecerem idênticos, os dois modelos adotam abordagens diferentes: enquanto o Word2Vec gera vetores apenas para palavras vistas durante o treinamento, o FastText consegue generalizar e criar vetores para palavras não vistas por meio do uso de subpalavras (n-gramas de caracteres), o que oferece vantagens importantes em contextos com vocabulário variado ou incompleto.

Embora este projeto não tenha sido direcionado para uma aplicação prática específica, os resultados obtidos indicam possíveis aplicações futuras, como o uso da correlação entre palavras para auxiliar na categorização automática de artigos científicos da área de saúde mental. Além disso, as representações semânticas podem servir de base para identificar temáticas já exploradas e auxiliar na descoberta de novos sintomas ou conceitos ainda não amplamente discutidos, contribuindo para o avanço de pesquisas nessa área.

Por fim, vale destacar que a execução do projeto seguiu uma trajetória levemente distinta do planejamento inicial, com mudanças que foram detalhadas na subseção “Desafios / Mudanças no Planejamento”.

Desafios / Mudanças no Planejamento

A ideia inicial de utilizar as palavras-chave fornecidas nos artigos como base para o treinamento do modelo foi descontinuada ao longo do desenvolvimento. Durante os testes, observou-se que essas palavras, por estarem isoladas e fora de um contexto textual contínuo, apresentaram baixa correlação semântica entre si. O uso direto dessas palavras no treinamento do Word2Vec resultava em uma representação dissociada, dificultando a aprendizagem de relações significativas entre os termos. Diante disso, optou-se por utilizar exclusivamente os títulos e resumos dos artigos, que oferecem um contexto linguístico mais rico, favorecendo a geração de *embeddings* mais coerentes com os objetivos do projeto.

Um dos principais desafios enfrentados durante o desenvolvimento do projeto foi a tentativa de utilizar o modelo Grove, inicialmente previsto como parte da abordagem metodológica. No entanto, surgiram problemas de compatibilidade entre o modelo e a versão do Python utilizada (3.12.4), a qual era necessária para a geração

das visualizações dos resultados. Como a integração do Grove exigiria alterações significativas no ambiente e nos arquivos do projeto, optou-se por descontinuar essa abordagem. Essa decisão visou preservar a consistência e a clareza da pipeline desenvolvida, mantendo o projeto coeso e de fácil compreensão.

Ferramentas

O desenvolvimento do projeto foi realizado em Python. A manipulação do dataset usou da biblioteca pandas, enquanto o pré-processamento linguístico recursos oferecidos por bibliotecas como NLTK e spaCy. Para a criação dos embeddings, escolhemos os modelos Word2Vec e FastText, por meio da biblioteca Gensim. As análises de similaridade e agrupamento usadas foram proporcionadas por ferramentas como Scikit-learn e UMAP, além da aplicação de técnicas de redução de dimensionalidade. A visualização dos resultados usou a biblioteca Matplotlib.

Referências

https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/s/saude-mental

https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response

https://jair.org/index.php/jair/article/view/11259

https://brasileiraspln.com/livro-pln/3a-edicao/parte-dominios/cap-saude-mental/cap-saude-mental.html

https://arxiv.org/abs/1902.08691?.com

https://www.datacamp.com/tutorial/explainable-ai-understanding-and-trusting-machine-learning-models