PLN aplicada à revisão sistemática de biomarcadores sanguíneos para diagnóstico da Doença de Alzheimer

Aluno: Denis Kalleb Oliveira Costa

Orientador: Prof. Eduardo Palhares Júnior

DEFESAS PÚBLICAS DE PÓS-GRADUAÇÃO DO PROJETO ARANOUÁ

INSTITUTO FEDERAL Amazonas

SAMSUNG

FAEPI

ARANOUÁ

Introdução / Contextualização

Doença de Alzheimer e o desafio do diagnóstico precoce

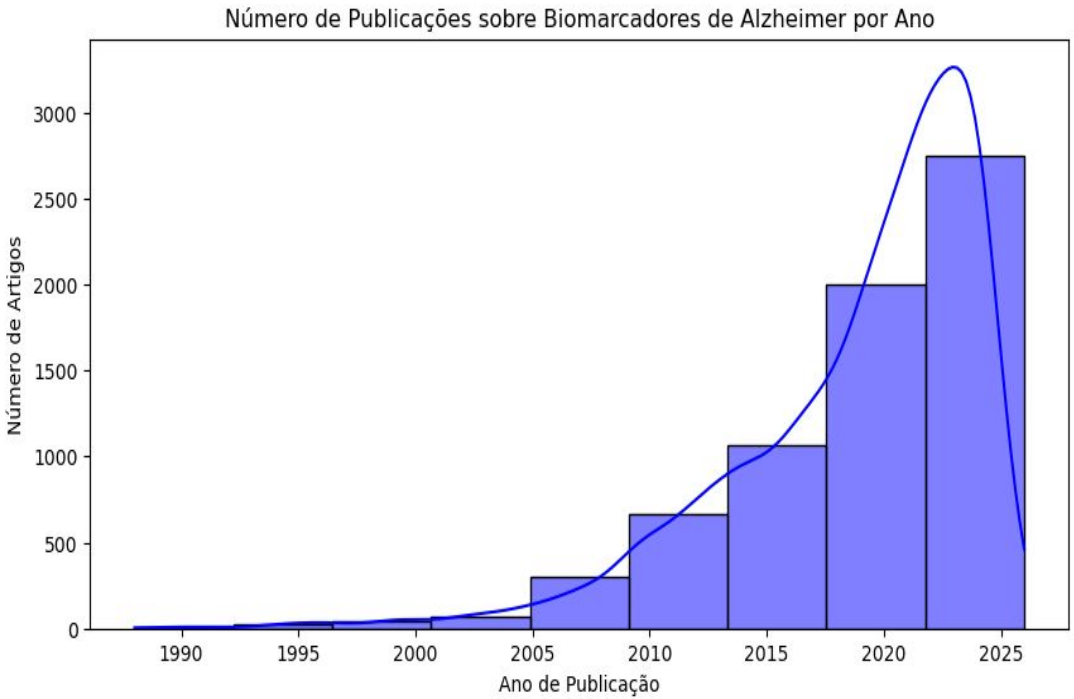
- A Doença de Alzheimer (DA) é a principal causa de demência no mundo, afetando milhões de pessoas.
- O diagnóstico precoce é essencial para aumentar a eficácia de tratamentos e intervenções.
- Os métodos convencionais, como PET scan e punção lombar, são invasivos e pouco acessíveis.
- Nos últimos anos, **biomarcadores sanguíneos** têm ganhado destaque como alternativa promissora e menos invasiva.

👉 O crescimento exponencial da literatura científica torna inviável uma análise manual.

Distribuição Temporal das Publicações

Distribuição Temporal das Publicações

- Maior volume de artigos a partir de 2019
- Reforça a crescente relevância dos biomarcadores sanguíneos no diagnóstico precoce da DA



Objetivos

Objetivo Geral

Desenvolver uma abordagem automatizada baseada em PLN para revisar e analisar artigos científicos sobre biomarcadores sanguíneos no diagnóstico da Doença de Alzheimer.

Objetivos Específicos

- Coletar e consolidar artigos científicos da base Web of Science;
- Realizar o pré-processamento textual utilizando o modelo spaCy SM;
- Aplicar TF–IDF e LDA para extração de termos e modelagem de tópicos;
- Comparar o desempenho entre os modelos SM e TRF usando métricas de similaridade;
- Identificar padrões e correlações entre biomarcadores e métodos diagnósticos.

Fundamentação: Alzheimer e Biomarcadores

Doença de Alzheimer e Biomarcadores Sanguíneos

- A DA é uma doença neurodegenerativa progressiva, sem cura, que compromete memória e cognição.
- O diagnóstico precoce é crucial para intervenção e controle da progressão.
- **Biomarcadores** são **substâncias biológicas** detectáveis, como **proteínas** ou **peptídeos**, que indicam **processos** fisiológicos, patológicos ou respostas a intervenções terapêuticas. Esses marcadores têm se mostrado alternativas promissoras aos métodos diagnósticos invasivos.

Biomarcadores mais relevantes:

- **p-Tau181 / p-Tau217** → associados a depósitos de tau no cérebro.
- **NfL (Neurofilamento de Cadeia Leve)** → marcador inespecífico de neurodegeneração.
- **GFAP** → indica resposta inflamatória glial.
- **Beta-amiloide** → acúmulo de placas senis, um dos principais sinais patológicos da DA.

Fundamentação: PLN e spaCy

Processamento de Linguagem Natural (PLN)

- Conjunto de técnicas que permitem interpretar e analisar linguagem humana por meio de algoritmos.
- Utilizado em tarefas como lematização, análise gramatical, extração de entidades e vetorização.

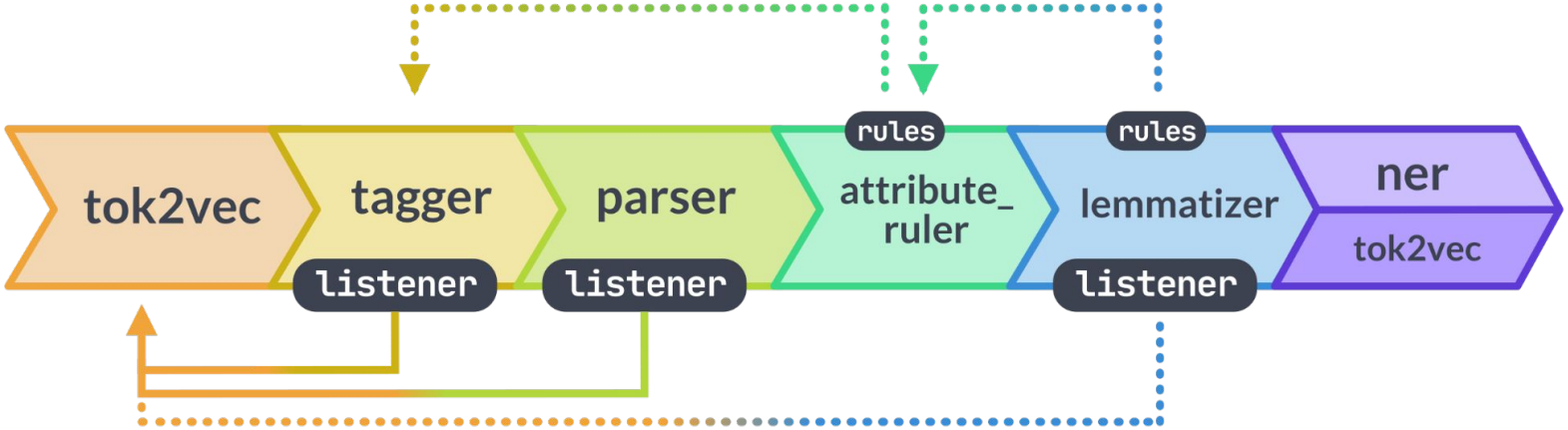
Por que spaCy?

- Biblioteca moderna, eficiente e modular.
- Modelos pré-treinados otimizados para inglês.

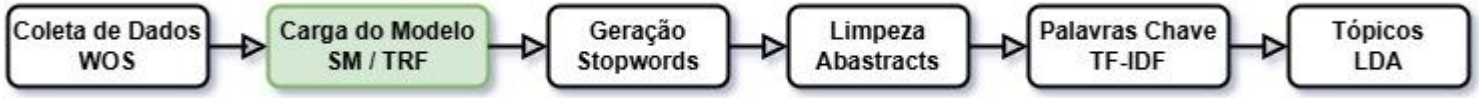
Modelos usados:

- **en_core_web_sm** (CNNs): usado como base da análise.
- **en_core_web_trf** (Transformers): usado para comparação semântica.

Pipeline do modelo SM



Metodologia Geral (Pipeline Visual)



Etapas da Metodologia

1. **Coleta de Artigos** — Extração de mais de 6.900 publicações da Web of Science com filtro por termos-chave.
2. **Pré-processamento** — Limpeza textual com spaCy SM: lematização, tokenização e remoção de stopwords.
3. **Extração de Palavras-chave** — TF–IDF padrão e customizado.
4. **Modelagem de Tópicos** — Aplicação de LDA para identificar temas emergentes.
5. **Indicadores e Visualizações** — Frequência de biomarcadores, coocorrência e evolução temporal.
6. **Métricas de Similaridade** — Análise entre modelos SM e TRF (Cosine, Jaccard, Levenshtein, KL).

Coleta de Artigos Científicos

Coleta dos Artigos

- Foram extraídas mais de 6.900 publicações da base de dados *Web of Science (WOS)*, acessada por meio do Portal de Periódicos da CAPES.
- Os arquivos foram baixados em lotes, uma vez que a plataforma permite a exportação de, no máximo, mil registros por vez.

String de Busca:

- Utilizou-se a ferramenta de busca avançada da plataforma, aplicando-se a seguinte string de consulta:

```
("Alzheimer") AND ("p-Tau217"OR "p-Tau181"OR
"GFAP"OR "Beta-amyloid"OR "Neurofilament light"OR
"NfL"OR "biomarker") AND ("blood"OR "plasma") AND
("diagnosis"OR "early detection"OR "screening")
```

Pré-processamento com spaCy SM

Função clean_text( )

Etapas aplicadas:

- Conversão para letras minúsculas
- Remoção de números, pontuação e caracteres especiais
- Tokenização com spaCy
- Lematização (forma base da palavra)
- Filtro de stopwords padrão e customizadas

Modelos Utilizados:

- Modelo SM (base principal)
- Modelo TRF (para comparação)

Resultado: Novas colunas no DataFrame:

- clean_default
- clean_custom

Tabela 1. Exemplos de palavras removidas como *stopwords*

Padrão spaCy	Contrações	Customizadas
about	's	alzheimer
above	're	biomarker
among	'm	study
after	've	methods
an	'll	patients

Tabela 2. Colunas geradas no pré-processamento textual

Coluna	Descrição
texto_completo	Junção do campo Abstract com Author Keywords, usada como base para análise textual.
clean_default	Versão limpa do texto utilizando somente as *stopwords* padrão do spaCy.
clean_custom	Versão limpa utilizando *stopwords* padrão e termos personalizados do domínio biomédico.

Pré-processamento com spaCy TRF

Modelo Transformer (spaCy TRF)

Após a análise com o modelo SM, o mesmo pré-processamento foi repetido com o modelo **spaCy en_core_web_trf**, baseado em arquitetura Transformer.

Características:

- Maior sensibilidade a contexto semântico.
- Baseado no modelo RoBERTa.
- Requer mais memória e tempo de execução.

Etapas aplicadas:

- Tokenização e lematização com maior preservação de estrutura gramatical.
- Remoção de stopwords padrão + customizadas.
- Geração da coluna **clean_custom_trf** para comparações com a versão SM.

📌 Utilizado apenas para comparação entre resultados, métricas de similaridade e análise de desempenho linguístico.

Comparação entre os modelos

O modelo **SM** tende a gerar termos mais literais e frequentes, como: `blood`, `cognitive impairment`, `beta`, `fluid`, `participant`.

TRF captura melhor os **bigramas** e variações gramaticais como: `early diagnosis`, `cognitive decline`, `cerebrospinal fluid`.

📌 Isso reforça que o modelo **TRF preserva melhor o contexto**, enquanto o **SM é mais eficiente e direto**.

Técnica TF–IDF: Extração de Termos

TF–IDF (Term Frequency–Inverse Document Frequency)

- Mede a relevância de um termo em um conjunto de documentos.
- Combina frequência local (no texto) com frequência global (na base).

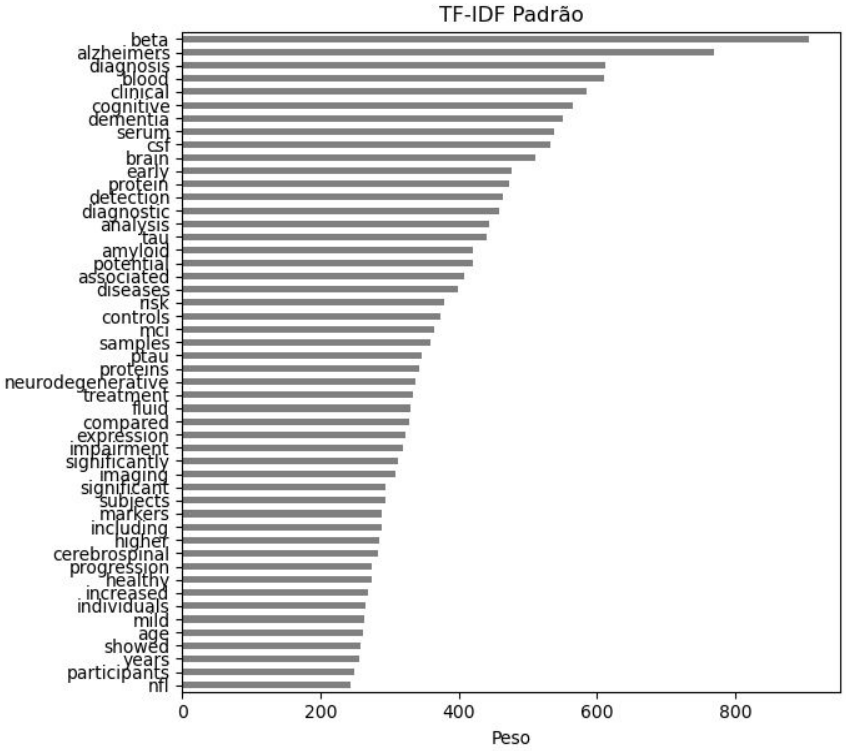
Variações Utilizadas:

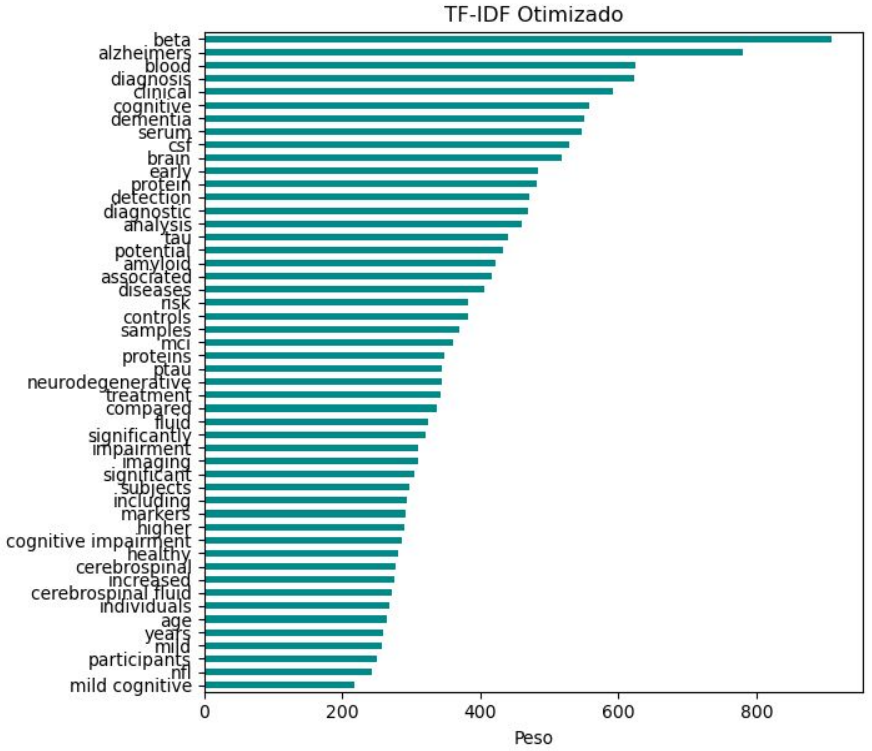
- TF–IDF padrão: max_features até 50 palavras-chave.
- TF–IDF customizado: bigramas, min/max frequência, stopwords reforçadas.

Resultado:

- Identificação de termos mais importantes como: progression, diagnosis, cognitive, brain, biomarkers, outros.

Gráfico de barras comparando TF-IDF padrão e customizado





Modelagem de Tópicos com LDA

LDA (Latent Dirichlet Allocation)

- Algoritmo probabilístico para identificar temas latentes em documentos.
- Agrupa palavras que aparecem frequentemente juntas em diferentes artigos.

Configuração utilizada:

- n_components = 5
- Baseado em TF–IDF customizado

Tópicos encontrados:

- Tópico 1: risk, beta, tau, diagnosis.
- Tópico 2: amyloid, detection, cerebrospinal.
- Tópico 3: cognitive, blood, brain.

Tópicos extraídos com LDA

Tabela 3: Tópicos extraídos com LDA e suas palavras mais representativas

Tópico	Palavras-chave
1	years, associated, samples, compared, tau, diagnosis, progression, beta, potential, risk
2	diseases, proteins, amyloid, alzheimers, clinical, cerebrospinal, tau, healthy, detection, beta
3	beta, blood, fluid, treatment, diseases, cognitive, amyloid, alzheimers, impairment, brain
4	dementia, proteins, protein, serum, diseases, expression, diagnostic, blood, analysis, showed
5	dementia, subjects, alzheimers, diagnosis, nfl, neurodegenerative, controls, including, mild, compared

Tabela 4: Descrição interpretativa dos tópicos gerados via LDA

Tópico	Descrição Interpretativa
Tópico 1	Relacionado a risco, progressão da doença e avaliação de amostras. Indica estudos focados em identificação precoce.
Tópico 2	Agrupamento centrado em biomarcadores clássicos como beta-amiloide e tau, além de exames como líquor e PET.
Tópico 3	Foco em tratamentos e sintomas cognitivos, como comprometimento e perda de memória.
Tópico 4	Estudos laboratoriais envolvendo proteínas no soro, expressão gênica e análises moleculares.
Tópico 5	Investigações clínicas com foco em amostras de pacientes, grupos controle e degeneração neuronal.

Síntese dos tópicos extraídos com LDA e seus principais termos

Tabela 5: Síntese dos tópicos extraídos com LDA e seus principais termos

Tópico	Tema interpretado	Principais termos
1	Biomarcadores e risco clínico	samples, progression, diagnosis, risk, potential, associated, beta, years
2	Fluídos e biomarcadores clássicos	proteins, amyloid, tau, cerebrospinal, clinical, diseases, healthy, detection
3	Declínio cognitivo e terapias	cognitive, treatment, impairment, brain, alzheimers, fluid, blood, diseases
4	Expressão e análises laboratoriais	serum, protein, diagnostic, expression, analysis, showed, subjects
5	Neurodegeneração e diagnóstico precoce	nfl, neurodegenerative, mild, controls, diagnosis, dementia, including

Ambiente Computacional

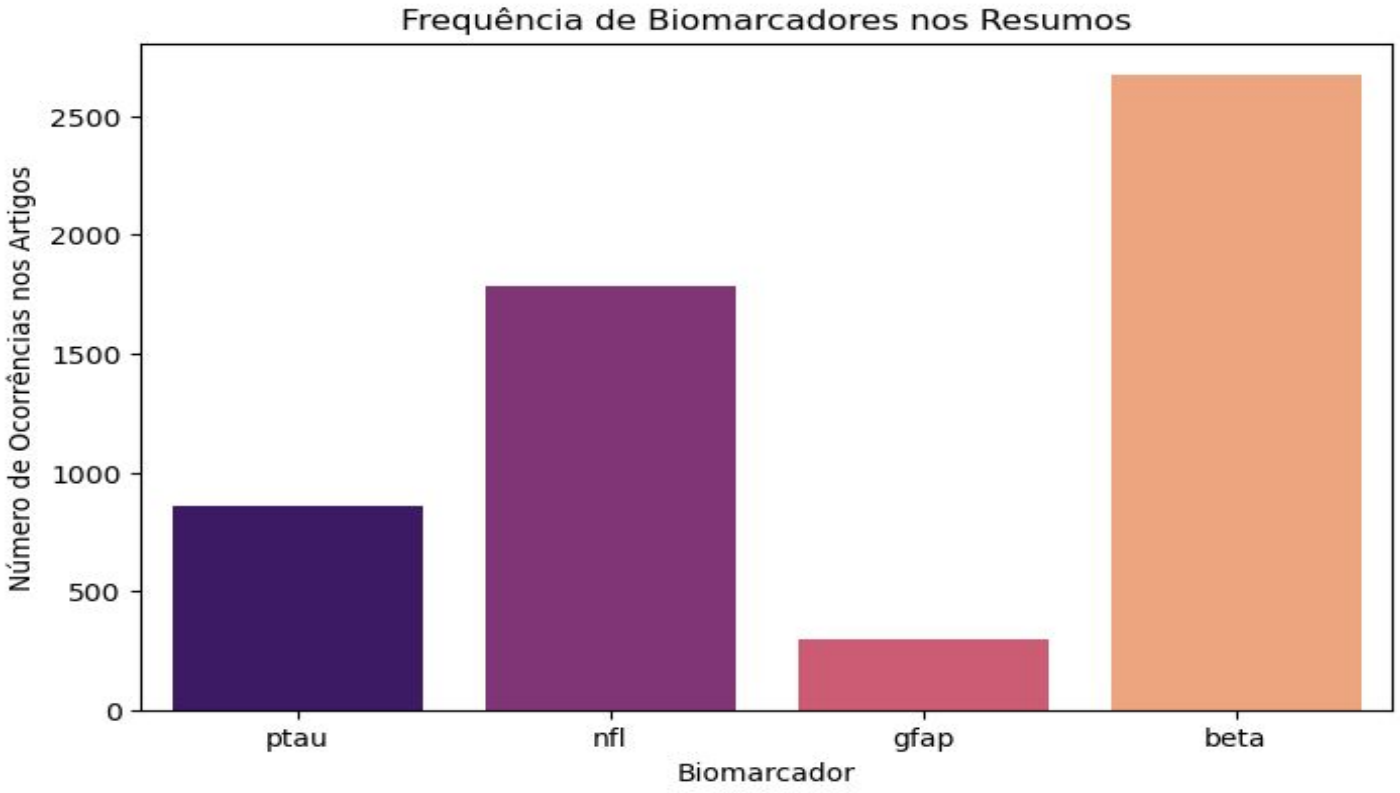
Tabela 6: Tempo de execução das células do notebook (via `%%time`)

Célula	Tempo de Execução
Carga dos Artigos	26.9 s
Limpeza dos Abstracts – Modelo SM	29 min 17 s
Limpeza dos Abstracts – Modelo TRF	3 h 28 min 9 s

Tabela 7: Configuração do computador utilizado para os experimentos

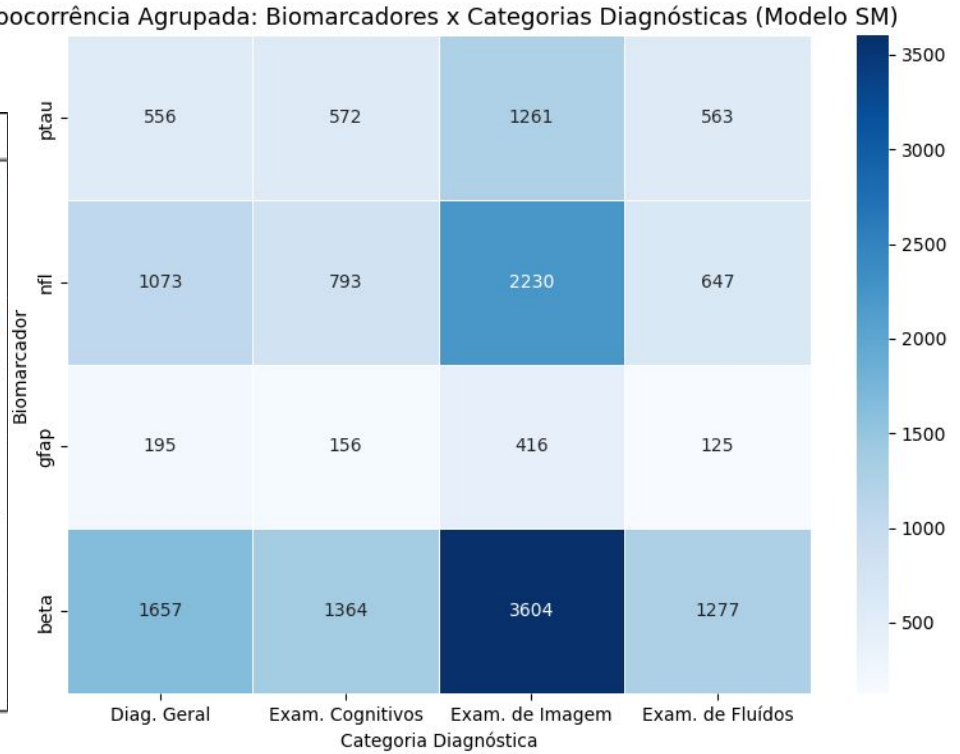
Componente	Especificação
Placa-mãe	ASUS Prime H510M-A, Intel Socket LGA1200
Armazenamento	SSD Hikvision C100, 480GB, Sata III
Memória RAM	Kingston Fury Beast, 16GB, 3200MHz, DDR4
Processador	Intel Core i5-10400F, 2.9GHz, 6 Núcleos, 12 Threads
Fonte de alimentação	Cougar VTC500, 500W, 80 Plus White

Citações de Biomarcadores em Contexto Diagnóstico: Frequência nos Resumos



Coocorrência: Biomarcadores x Exames

Categoria	Termos Utilizados	Descrição
Diag. Geral	diagnosis, screening	Termos amplos relacionados ao processo de diagnóstico ou triagem de pacientes.
Exam. Cognitivos	cognitive	Avaliações do funcionamento cognitivo, como testes de memória e raciocínio.
Exam. de Imagem	pet, mri, neuroimaging, fmri, ct	Técnicas de imagem cerebral como PET, Ressonância Magnética e Tomografia.
Exam. de Fluídos	csf, blood test, plasma, elisa	Exames laboratoriais baseados em fluídos corporais (sangue, líquor, plasma).



Comparação entre SM e TRF - Métricas

Cosine Similarity

Mede o “ângulo” entre dois vetores TF-IDF. Quanto mais próximos os textos no sentido semântico (mesmo com palavras diferentes), maior o valor.

Interpretação: proximidade semântica ponderada.

Jaccard Similarity

Mede a interseção sobre a união dos tokens únicos. Foca na **presença ou ausência de palavras**, sem considerar pesos.

Interpretação: o quanto os textos compartilham os mesmos termos.

Levenshtein Distance

Conta quantas edições (inserções, remoções ou trocas de caracteres) são necessárias para transformar um texto no outro.

Interpretação: distância literal entre os textos.

KL Divergence (Kullback–Leibler)

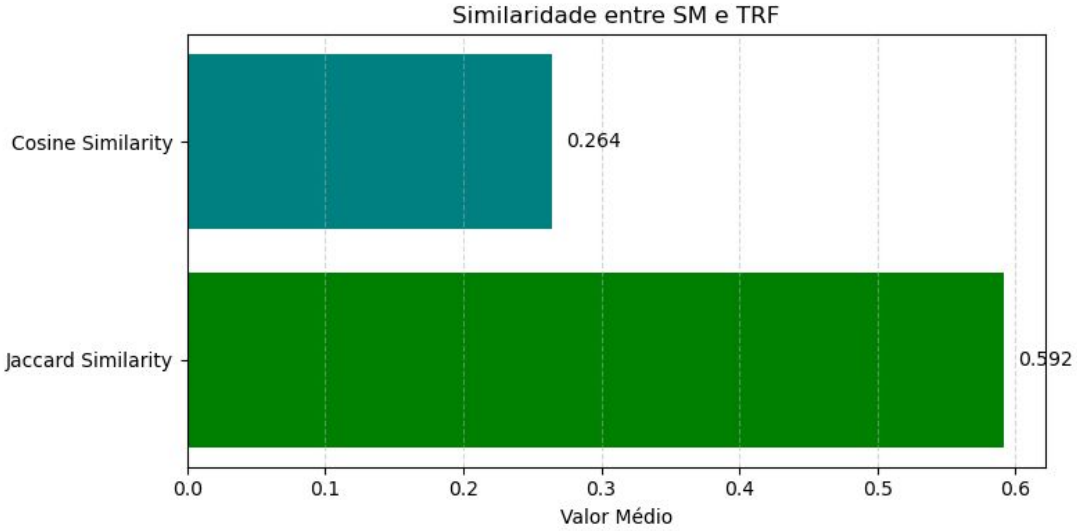
Compara a diferença entre duas distribuições de probabilidade (ex: frequência de palavras).

Interpretação: quanto um texto “destoa” do outro na distribuição de termos.

Comparação entre SM e TRF - Métricas

Pontos-chave:

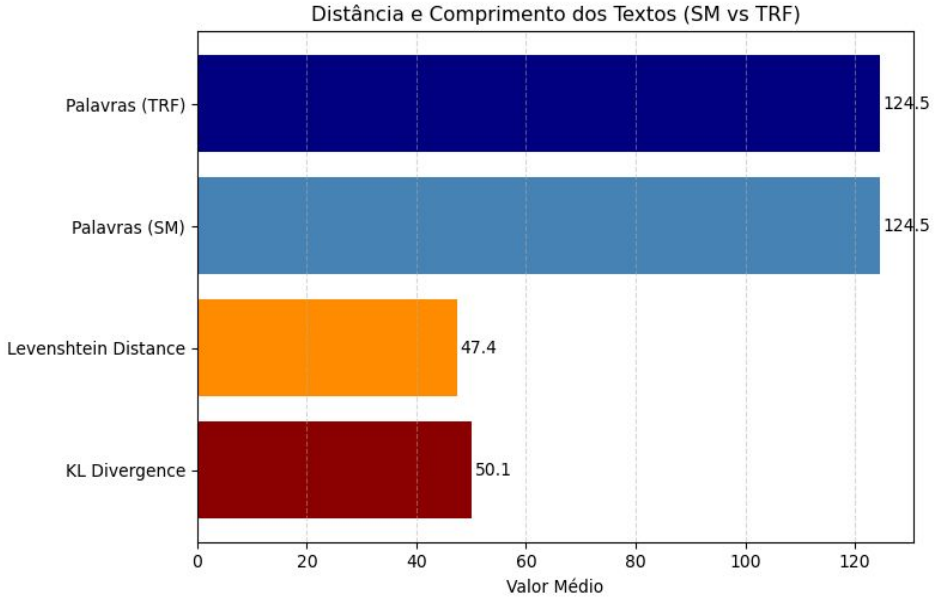
- **Jaccard**: 0.592 - indica boa sobreposição de tokens únicos entre os modelos.
- **Cosine**: 0.264 - baixa semelhança quando considerados os pesos dos termos (TF-IDF).



Comparação entre SM e TRF - Métricas

Pontos-chave no slide:

- Ambos os modelos produziram textos com mesmo comprimento médio (**124.5 tokens**).
- A **KL Divergence** foi maior (50.1), indicando maior diferença na distribuição de palavras entre os textos.
- A **Levenshtein Distance** (47.4) revela um número considerável de edições necessárias para transformar uma versão em outra, mesmo com tamanhos idênticos.



Conclusão

Encerramento e Contribuições

Este trabalho propôs uma abordagem automatizada de **revisão sistemática** com **PLN** aplicada à Doença de Alzheimer.

Utilizando modelos **spaCy SM** e **TRF**, foi possível:

- Pré-processar e analisar mais de 6 mil artigos científicos.
- Extrair palavras-chave relevantes com **TF–IDF**.
- Identificar temas recorrentes com **LDA**.
- Detectar forte correlação entre biomarcadores e exames clínicos.

O modelo TRF mostrou-se mais preciso em tarefas contextuais, porém com maior custo computacional.

A análise demonstrou a viabilidade de aplicar técnicas de PLN para apoiar pesquisas biomédicas em larga escala.

Trabalhos Futuros

Aprimoramento da análise semântica:

- Explorar o uso de embeddings contextuais como BERT ou BioBERT.
- Aplicar o NER biomédico com dicionários especializados.

Expansão da base de dados:

- Incluir outras fontes como PubMed, Scopus e bases em português.

Refinamento do pipeline:

- Adicionar classificadores supervisionados por tipo de estudo.
- Melhorar o visual analytics com dashboards interativos.

Aplicação prática:

- Integrar a ferramenta com revisões sistemáticas de pesquisadores da área da saúde.

Referências

- Honnibal, M. et al. (2023). *spaCy: Industrial-strength NLP in Python*.
  🔗 https://spacy.io

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). *Latent Dirichlet Allocation*.
  🔗 https://jmlr.org/papers/v3/blei03a.html

- Ramos, J. (2003). *Using TF-IDF to Determine Word Relevance*.
  🔗 https://www.cs.rutgers.edu/~mlittman/courses/ml03/iCML03/papers/ramos.pdf

- Chen, D. et al. (2023). *Blood-based biomarkers for Alzheimer's disease*.
  🔗 https://www.nature.com/articles/s41582-023-00799-1

- Bandeira, L., Ferreira, H., de Almeida, J. M., de Paula, A. J., & Dalpian, G. M. (2024).**CO_2 Reduction Beyond Copper-Based Catalysts: A Natural Language Processing Review From the Scientific Literature**.https://pubs.acs.org/doi/10.1021/acssuschemeng.3c06920

AGRADECIMENTO

Agradeço a todos que contribuíram para a realização deste trabalho:

- Ao meu orientador **Prof. Eduardo Palhares Júnior** pela orientação técnica e acadêmica.
- Ao **Instituto Federal do Amazonas – IFAM**, pelo suporte institucional e científico.
- À **SAMSUNG Eletrônica da Amazônia** e ao projeto **Aranouá**, pelo financiamento da pesquisa aplicada.
- À **CAPES** pelo apoio por meio do Programa de Excelência Acadêmica (PROEX).
- Aos colegas, amigos e minha família pelo incentivo e apoio ao longo dessa jornada.