

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
Programa de Pós-Graduação em Biotecnologia

**“MINERAÇÃO DE TEXTO, INTELIGÊNCIA ARTIFICIAL E APLICAÇÕES EM
BIOTECNOLOGIA”**

Vítor Eulálio Reis

Dissertação de Mestrado apresentada ao
Programa de Pós-Graduação em
Biotecnologia da Universidade Federal de
São Carlos.

Orientadora:

Prof.^a Dr.^a Ignez Caracelli

São Carlos - SP
2024

VÍTOR EULÁLIO REIS

“MINERAÇÃO DE TEXTO, INTELIGÊNCIA ARTIFICIAL E APLICAÇÕES EM BIOTECNOLOGIA”

Dissertação de Mestrado em Biotecnologia do Programa de Pós-Graduação em Biotecnologia da Universidade Federal de São Carlos.

São Carlos - SP, abril de 2024.

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Programa de Pós-Graduação PPGBiotec

Folha de aprovação

Assinatura dos membros da comissão examinadora que avaliou e aprovou a Defesa de Mestrado do candidato Vítor Eulálio Reis, realizada em 26/04/2024:

BANCA EXAMINADORA

Prof.^a Dr.^a Ignez Caracelli

Departamento de Física - UFSCar

Prof. Dr. Alan Demétrius Baria Valejo

Departamento de Computação - UFSCar

Dr. Lucas Sousa Madureira,

IAE - Instituto da Aeronáutica e do Espaço

DEDICATÓRIA

À minha família,
Pelo amor, pela compreensão
e incentivo durante todo o percurso acadêmico.

AGRADECIMENTOS

Em nome do Criador, o Clemente e Misericordioso o Misericordador, expresso minha sincera gratidão à minha família, cujo apoio constante, incentivo e compreensão foram a minha fonte constante de inspiração.

Agradeço ao Criador, o Senhor dos Mundos, por guiar meus passos e por conceder sabedoria durante esta jornada acadêmica. Sua graça foi minha luz em momentos de desafio. Que a paz e as bênçãos estejam sobre minha família e a de todos que contribuíram para este esforço.

Sou profundamente grato à minha orientadora, Prof.^a Dr.^a Ignez Caracelli por sua mentoria. Sua sabedoria e direcionamento foram instrumentais na formação da minha abordagem e compreensão do assunto. Seu comprometimento com a excelência foi fundamental para alcançar os objetivos do projeto.

Um reconhecimento sincero também é estendido ao Julio Zukerman Schpector por sua orientação e a toda a equipe do projeto por seus esforços colaborativos e contribuições.

Estendo meus sinceros agradecimentos a Sérgio Vitório por sua orientação editorial e *feedback* perspicaz. Sua atenção aos detalhes e sugestões construtivas aprimoraram a qualidade deste trabalho.

Apreciação especial vai para a equipe da John Snow Labs, cuja experiência técnica e dedicação contribuíram significativamente para a implementação bem-sucedida deste projeto. Suas soluções inovadoras e espírito colaborativo foram inestimáveis.

Agrado aqui também à PDF2GO, que disponibilizaram uma conta estudantil para o processamento dos arquivos.

A cada pessoa mencionada, e a todos que tiveram um papel, obrigado por serem parte integrante desta jornada. Seu apoio e expertise tornaram não apenas possível, mas também gratificante, este empreendimento.

"O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Código de Financiamento

Eulálio Reis, Vítor **MINERAÇÃO DE TEXTO, INTELIGÊNCIA ARTIFICIAL E APLICAÇÕES EM BIOTECNOLOGIA**. 2024. Dissertação (Programa de Pós-graduação em Biotecnologia) – UFSCar, São Carlos, 2024

RESUMO

Foi feito um estudo utilizando a abordagem *Retrieval Augmented Generation* (RAG) para analisar e extrair *respostas* de textos científicos, e gira em torno da investigação da relação intrincada entre a tríade: severidade, recuperação e infecção da COVID-19, insuficiência e deficiência de vitamina D, e a relação do uso de protetor solar. Os métodos e técnicas são baseadas em IA de processamento de linguagem natural para recuperar, racionalizar e confrontar o conhecimento presente em uma literatura científica limitada. Nesse sentido, integra-se técnicas de extração da semântica em documentos relevantes, enquanto modelos de geração de texto são empregados para traduzir, resumir e destilar fatos a partir de uma base de conhecimento semântico. O estudo busca elucidar as potenciais correlações e interações negligenciadas ou inéditas presentes na literatura e contribuem para o contínuo debate em torno dos aspectos multifacetados do COVID-19 e sua correlação com os níveis de vitamina D. Para tanto, há o fornecimento de informações sobre possíveis direções para futuras pesquisas e estratégias de saúde pública, no controle e prevenção de doenças, e até mesmo no desenho de ensaios clínicos eficientes. Assim sendo, a abordagem RAG serve como uma forma de extração de conhecimento a partir de uma extensa literatura interdisciplinar, possibilitando uma exploração detalhada em escala sem precedentes de todas as conexões complexas documentadas entre os conceitos-chave envolvidos na conjuntura total da pandemia.

Palavras-chave: Vitamina D, Covid-19, Protetor solar, RAG, IA, Inteligência Artificial

Eulálio Reis, Vítor. TEXT MINING, ARTIFICIAL INTELLIGENCE AND APPLICATIONS IN BIOTECHNOLOGY.. 2024. Qualificação de Mestrado (Graduate Program in Biotechnology) – UFSCar, São Carlos, 2024.

ABSTRACT

This master's research employs the Retrieval Augmented Generation (RAG) approach to analyze and extract answers from scientific texts, focusing on investigating the intricate relationship among the triad: severity, recovery, and infection of COVID-19, vitamin D insufficiency and deficiency, and the correlation with sunscreen use. The methods and techniques are based on natural language processing AI to retrieve, rationalize, and confront knowledge present in a limited scientific literature. In this context, semantics extraction techniques are integrated into relevant documents, while text generation models are employed to translate, summarize, and extract facts from a semantic knowledge base. The study aims to elucidate potential correlations and overlooked or novel interactions present in the literature, providing valuable insights into the ongoing debate surrounding the multifaceted aspects of COVID-19 and its correlation with vitamin D levels. To achieve this, information is provided on possible directions for future research and public health strategies in disease control and prevention, and even in designing efficient clinical trials. Therefore, the RAG approach will serve as a knowledge extraction method from extensive interdisciplinary literature, enabling an unprecedented, detailed exploration of all complex documented connections between the key concepts involved in the overall context of the pandemic.

Keywords: Vitamin D, Covid-19, Sunscreen, RAG, AI, Artificial Intelligence

LISTA DE FIGURAS

	página
Capítulo 1 - Introdução	
Capítulo 2 - Materiais e Métodos	
Figura 1.1 Abordagem <i>RAG</i>	
Figura 1.2 Exemplo da estrutura interna de um stream de PDF	
Figura 2.3 Ilustração da <i>tokenização</i>	
Figura 2.4 Composição conceitual de Grafos de Conhecimento	33
Figura 2.5 Carregamento da $S(1)$ na base de dados.	34
Figura 2.6 Palavras e suas dependências	36
Figura 2.7 NER obtido através do <i>displaCy</i> .	37
Capítulo 3 - Resultados	
Figura 3.1 Publicações em números absolutos por disciplina para a Busca 1.	44
Figura 3.2 Publicações em números absolutos por disciplina para a Busca 2.	46
Figura 3.3 Publicações em números absolutos por disciplina para a Busca 3.	48
Figura 3.4 Publicações em números absolutos por disciplina para a Busca 4.	50
Figura 3.5 Perfil histórico de publicações e citações para a Busca 1.	53

Figura 3.6	Perfil histórico de publicações e citações para a Busca 2.	54
Figura 3.7	Perfil histórico de publicações e citações para a Busca 3.	55
Figura 3.8	Perfil histórico de publicações e citações para a Busca 4.	55
Figura 3.9	Ordem lógica do texto 3d0e16b69e5759e745a50af3b16b34b0.pdf.	67
Figura 3.10	Primeiro passo para a conversão de PDF para DOCX na nuvem .	
Figura 3.11	Segundo passo para conversão de PDF para DOCX na nuvem .	
Figura 3.12	Terceiro passo para conversão de PDF para DOCX na nuvem .	
Figura 3.13	Quarto passo para conversão de PDF para DOCX na nuvem .	
Figura 3.14	Legibilidade dos DOCX do 3d0e16b69e5759e745a50af3b16b34b0.pdf.	61
Figura 3.15	Extração de palavras e separação de linhas no DOCX do f813ee0787090b8ba7460f681d06d93a.pdf.	
Figura 3.16	Extração de palavras e separação de linhas no f813ee0787090b8ba7460f681d06d93a.pdf.	
Figura 3.17	<i>ChatGPT</i> no navegador; https://chat.openai.com	
Figura 3.18	<i>Gemini</i> no navegador; https://gemini.google.com	
Figura 3.19	Bloco de 4000 caracteres do 0a337a62de92b3e2717d0f7a454dbe4c.txt.	
Figura 3.20	Exemplo da interação pergunta e resposta.	
Figura 3.21	Performance em referenciamento de texto.	62
Figura 3.22	Grafo de Conhecimento Produzido pelos <i>Pipelines</i> da John Snow Labs.	64

LISTA DE TABELAS

	página
Capítulo 1 - Introdução	
Capítulo 2 - Materiais e Métodos	
Tabela 2.1 Desenho Experimental	23
Tabela 2.2 Pipelines de Extração utilizados	30
Tabela 2.3 Configuração da Máquina	37
Tabela 2.4 Softwares utilizados	39
Tabela 2.5 Bibliotecas utilizadas no processamento de PDF	
Tabela 2.6 <i>LLMs</i> disponíveis para geração	41
Capítulo 3 - Resultados	
Tabela 3.1 Tópicos utilizados no levantamento bibliográfico	43
Tabela 3.2 Origem das publicações em números absolutos e porcentagem para a Busca 1.	44
Tabela 3.3 Origem das publicações em números absolutos e porcentagem para a Busca 2.	46
Tabela 3.4 Origem das publicações em números absolutos e porcentagem para a Busca 3.	49
Tabela 3.5 Origem das publicações em números absolutos e porcentagem para a Busca 4.	50
Tabela 3.6 Contagem de Publicações por área	52
Tabela 3.7 Arquivos não alcançados da Busca 1.	57

Tabela 3.8	Arquivos não alcançados da Busca 2.	59
Tabela 3.9	Arquivos não alcançados da Busca 3.	59
Tabela 3.10	Benchmark das bibliotecas para processamento de PDFs.	
Tabela 3.11	Expressões Regulares empregadas no corpus.	

LISTA DE ENTRADAS

Abreviação	Significado	Página
Entrada 1 - E(1)	Bloco de 4000 caracteres do 0a337a62de92b3e2717d0f7a454dbe4c.txt	
Entrada 2 - E(2)	Base de dados (CSV)	

LISTA DE SAÍDAS

Abreviação	Significado	Página
Saída 1 - S(1)	Arquivo CSV oriundo do <i>pipeline</i> de Extração	26
Saída 2 - S(2)	Contagem de arquivos baixados	60
Saída 3 - S(3)	Experimento (512, 0, <i>clean</i>)	60
Saída 4 - S(4)	Experimento (512, 0, <i>raw</i>)	61
Saída 5 - S(5)	Resultado da PyMUPDF	73
Saída 6 - S(6)	Resultado Entrada 1	98

LISTA DE PROMPTS

		Página
Prompt 1 - P(1)	<i>Prompt</i> de extração de Entidades e Relações I	28
Prompt 2 - P(2)	<i>Prompt</i> de extração de Entidades e Relações II	32
Prompt 3 - P(3)	Aplicação do Prompt 1	65

LISTA DE FRAGMENTOS DE CÓDIGO		
Abreviação	Significado	Página
Fragmento 1 - F(1)	Composição de uma chain em <i>Langchain</i> .	46
Fragmento 2 - F(2)	Inicialização do motor de consulta básico	67
Fragmento 3 - F(3)	Inicialização do motor de consulta baseado em palavras chave	69

Fragmento 4 - F(4)	Inicialização do motor de consulta híbrido baseado em entidades	73
Fragmento 5 - F(5)	Inicialização do motor de consulta KnowledgeGraphQueryEngine	84
Fragmento 6 - F(6)	Inicialização do motor de consulta KnowledgeGraphRAGRetriever	87
Fragmento 7 - F(7)	Inicialização do motor de consulta GraphCypherQChain	93
Fragmento 8 - F(8)	Extração de texto de DOCx	95

LISTA DE ABREVIATURAS, SIGLAS E TERMOS

Abreviação	Significado	Tradução
<i>RAG</i>	Retrieval Augmented Generation	Geração de retorno aumentada
<i>LLM</i>	Large Language Model	Grande modelo de linguagem
<i>Prompt</i>	Prompt	Instrução
<i>GC</i>	Grafo de Conhecimento	Gráfico de Conhecimento
<i>QChain</i>	Question and Answer Chain	Cadeia de perguntas e respostas
<i>IA</i>	Inteligência Artificial	
<i>UI</i>	User Interface	Interface do Usuário

SUMÁRIO

	página
Resumo	5
Abstract	6
Lista de Figuras	7
Lista de Tabelas	8
Lista de Saídas	9
Lista de Prompts	9
Lista de Abreviaturas, Siglas e Termos	9
Introdução	12
Objetivos	13
Gerais	13
Específicos	14
Capítulo 1 - Dos Fundamentos	16
1.1 Da Covid 19	16
1.2 Da Vitamina D	17
1.3 Do Protetor Solar	18
1.4 Da Inteligência Artificial	19
Capítulo 2 - Materiais e Métodos	22
2.1 Metodologia	22
2.1.1 Aquisição dos dados	23
2.1.2 Trabalhando com os dados	27
2.1.3 A estrutura interna do PDF	

2.1.4	Extração do conteúdo de PDFs	
2.1.5	Pré-processamento dos dados	
2.1.6	Pós-processamento dos dados	
2.1.7	Desenho Experimental	
2.1.8	Extração de Entidades e Relações	
2.1.9	Criação do Grafo de Conhecimento	
2.1.10	Perguntas e Respostas	
2.1.11	Validação	
2.2	Materiais	
2.2.1	Recursos Computacionais	37
2.2.2	Recursos Digitais	39
2.2.3	Aquisição de Dados	40
2.2.4	Abordagem RAG	41

Capítulo 3 - Resultados e Discussão

3.1	Aquisição dos Dados	43
3.2	Análise de Dados	56
3.3	Preparo dos Dados	60
3.4	Extração do conteúdo de PDFs	64
3.5	Pós-processamento dos dados	65
3.6	Integração dos dados	67
3.7	Reduções	
3.8	Ingestão dos Dados	

3.9	Construção do GC	
3.10	Interação com a IA	61
3.11	Metodologia	62
3.12	Empregabilidade	62
3.13	Modelos e <i>Pipelines</i>	63
3.14	Generalização	344
Capítulo 4 - Conclusões		135
Referências		139

Introdução

A pandemia de COVID-19, iniciada em 2019, foi causada pelo novo coronavírus SARS-CoV-2 e representa uma das maiores crises sanitárias da história global recente, manifestando-se com uma ampla variedade de sintomas que vão de leves a graves - pneumonia, insuficiência respiratória aguda, óbito (Ministério da Saúde, 2021a).

Em adição, estudos recentes destacam a importância da vitamina D na ativação das células T, fundamentais para a resposta imunológica adaptativa contra o SARS-CoV-2 (Valentine et al., 2021). Além de suas funções clássicas na regulação da imunidade, a vitamina D também modula a resposta inflamatória, desempenhando um papel crucial na patogênese da COVID-19 (GUPTA et al., 2021). Diante disso, indivíduos deficientes em vitamina D mostram maior propensão a infecções, incluindo as respiratórias.

O uso de protetor solar, embora importante para a eventual proteção da pele contra os danos causados pela radiação ultravioleta (UV), pode impactar os níveis de vitamina D no organismo. Sendo assim, atua bloqueando a radiação UV, necessária para desencadear as reações químicas envolvidas na síntese da vitamina D (PONSONBY et al., 2006). Portanto, a redução da exposição à radiação UV pode resultar na diminuição da disponibilidade de vitamina D no corpo.

Levando em conta as considerações, o levantamento bibliográfico deste trabalho destaca a correlação entre deficiência e insuficiência de vitamina D, o uso de protetor solar e aspectos como a severidade, recuperação e infecção por COVID-19. No entanto, até o momento, não existem publicações que tenham investigado especificamente o papel do uso de protetor solar e a gravidade, recuperação e incidência da COVID-19, ou seja, uma respectiva relação.

A premissa subjacente à abordagem RAG repousa na ideia de que a análise aprofundada de estudos que exploram relações entre conceitos específicos pode revelar novas conexões indiretas, previamente não exploradas, utilizando a inteligência artificial para a interpretação semântica dos dados fornecidos. Assim sendo, o desbloqueio do entendimento destas relações pode ser feito através de

questionamentos, comparações entre estudos e testes de hipóteses preliminares com base no conhecimento representado nos estudos utilizados. Portanto, este enfoque não apenas oferece uma abordagem automatizada e escalável para ampliar o conhecimento, mas também tem o potencial de proporcionar *insights* valiosos sobre a complexidade e interconexão dos conceitos ainda não explorados.

Por fim, ao focar nas nuances das relações entre conceitos em estudos já existentes, é possível desvendar aspectos anteriormente negligenciados, enriquecendo, assim, a compreensão e a relevância do tema abordado. Para tanto, esta abordagem investigativa utilizando Inteligência Artificial promete contribuições significativas para o avanço do conhecimento e o desenvolvimento de perspectivas mais abrangentes em pesquisas acadêmicas.

Objetivos

Gerais

Este estudo visa investigar e compreender, por meio da aplicação da abordagem *Retrieval Augmented Generation (RAG)*, a possível relação entre COVID-19 com a vitamina D. A deficiência de vitamina D pode estar relacionada com a severidade do COVID-19. Um dos possíveis fatores para deficiência da vitamina D poderia ser o uso de protetor solar. Este esforço insere-se na disciplina da Inteligência Artificial e, como foco central, permitir que o usuário explore territórios inexplorados, buscando estabelecer conexões verídicas ou inéditas entre o emprego do protetor solar e os impactos na manifestação e propagação do vírus.

Além disso, o trabalho propõe a implementação de uma *UI* interativa através da qual o usuário interage com o assistente virtual. O Assistente virtual coordena todos os passos subjacentes à geração da resposta, visando disponibilizar as descobertas de maneira acessível e compreensível para os usuários.

Específicos

1. Analisar a Produção Científica Existente:
 - Identificar lacunas no conhecimento, destacando áreas pouco exploradas e potenciais correlações não documentadas.
2. Desenvolver um Modelo *RAG* Específico:
 - Implementar a abordagem *RAG* personalizando-a com diferentes parâmetros e componentes e descobrir a melhor configuração através de *benchmarks*.
3. Construir e Validar Grafos de Conhecimentos:
 - Criar um Grafos de Conhecimentos por experimento.
 - Validar a performance do *RAG* sob cada um dos grafos envolvidos no experimento.
4. Desenvolver uma *UI* Interativa com Assistente Virtual:
 - Projetar e implementar uma *UI* interativo que permita aos usuários explorar e compreender as relações descobertas.
 - Introduzir um assistente virtual, capaz de responder a consultas, fornecer informações detalhadas e orientar os usuários na navegação pelo conteúdo.
5. Compartilhar o Conhecimento:
 - Disponibilizar o código fonte em um repositório público no github, acessível em https://github.com/vriez/doctor_rag.

Ao atingir esses objetivos, este trabalho busca contribuir significativamente para o entendimento da relação entre o uso de protetor solar e a COVID-19, proporcionando uma plataforma interativa e informativa para o público em geral.

As publicações relativas aos três temas subjacentes serão assimiladas pela IA, formando uma base de dados disponível para uma interação em formato de perguntas e respostas que permitirá questões profundas, cujas respostas serão

retornadas com base nas argumentações contrastadas com os dados de seus correspondentes ensaios clínicos e estudos mapeados no grado de conhecimento alvo. Logo, a interação em formato de perguntas e respostas será realizada por meio de um assistente inteligente virtual (*ChatBot*).

Este estudo visa contribuir para o avanço do conhecimento, proporcionando perspectivas valiosas sobre a complexidade e a interconexão dos conceitos relacionados ao uso de protetor solar, à deficiência de vitamina D e à dinâmica do COVID-19, utilizando uma abordagem inovadora e automatizada.

Capítulo 1 - Dos Fundamentos

1.1 Da Covid-19

A COVID-19 é uma doença infecciosa originada pelo coronavírus SARS-CoV-2, membro da família *Coronaviridae*, que também inclui os vírus responsáveis pelas síndromes respiratórias agudas graves (SARS) e pelo síndrome respiratório do Oriente Médio (MERS) . Identificada inicialmente em dezembro de 2019 em Wuhan, China, essa patologia rapidamente se disseminou globalmente, desencadeando uma pandemia que afeta todos os continentes (PAHO, 2022).

O SARS-CoV-2, embora seja um vírus recentemente identificado, compartilha semelhanças com seus predecessores, como o SARS-CoV-1 e o MERS-CoV, que causaram epidemias nas décadas anteriores (Cambaza 2022).

O impacto social global da COVID-19 foi significativo, levando a *lockdowns*, quarentenas e medidas de distanciamento social, resultando em repercussões adversas na economia e no cotidiano das pessoas. Além disso, a pandemia instigou mudanças nos comportamentos, como a adoção frequente da lavagem de mãos, uso de álcool em gel e de máscaras faciais em locais públicos (UNICEF, 2022).

A manifestação da COVID-19 varia em intensidade, com a maioria dos infectados apresentando sintomas leves a moderados, enquanto alguns desenvolvem formas graves com potencial letal. A recuperação, em geral, estende-se por 2 a 3 semanas, embora alguns pacientes possam persistir com sintomas por um período prolongado (Santalucia, 2020).

O SARS-CoV-2, devido à sua alta transmissibilidade, propaga-se por meio do contato direto com gotículas respiratórias de indivíduos infectados ou pelo contato indireto com superfícies contaminadas (Ministério da Saúde, 2021b).

Os impactos nos sistemas imunológico, endocrinológico (Szczerbiński et al., 2023), digestivo (Cao et al., 2021), muscular, ósseo e cognitivo revelam a complexidade da resposta do organismo ao vírus (Greve et al., 2020). O sistema imunológico, em particular, é suscetível a uma resposta inflamatória exacerbada

desencadeada pelo SARS-CoV-2, podendo resultar em danos a tecidos e órgãos (Greve et al., 2020). Esta resposta, embora natural, pode se tornar prejudicial devido à ligação a receptores celulares, produção de citocinas e ativação de células imunológicas.

No sistema endocrinológico, o SARS-CoV-2 impacta a produção de hormônios do estresse, hormônios da tireoide e hormônios sexuais, influenciando aspectos como metabolismo (Shen, 2021), crescimento (Araújo et al., 2021), desenvolvimento (Araújo et al., 2021) e reprodução (Lee, 2021). Essa complexidade reforça a necessidade de estudos contínuos para compreender completamente os efeitos da COVID-19 em diversos sistemas do corpo humano.

1.2 Da Vitamina D

A vitamina D, um hormônio lipossolúvel crucial para a saúde humana, desempenha diversas funções vitais, regula o metabolismo de cálcio e fósforo, mantém ossos e músculos saudáveis, fortalece o sistema imunológico e protege contra doenças crônicas (Harvard T.H. Chan School of Public Health, 2022).

Existem duas formas principais de vitamina D: D2 (produzida por plantas e fungos) e D3 (sintetizada pela pele humana com exposição à luz solar). A D3 é produzida na pele a partir do colesterol, enquanto a D2 é obtida na dieta. A vitamina D é absorvida no intestino delgado, auxiliada pela bile e ácidos graxos, ela passa por um processo de metabolização no fígado e nos rins, onde se converte em formas ativas e inativas (American Academy of Dermatology, 2022a).

Essa vitamina desempenha um papel essencial em processos biológicos, regulando o metabolismo de cálcio e fósforo, mantendo ossos e músculos saudáveis, fortalecendo o sistema imunológico e associando-se à redução do risco de doenças crônicas (Bikle 2015).

Insuficiência e deficiência de vitamina D são comuns, podendo resultar de exposição solar inadequada, dieta deficiente ou condições médicas que afetam a absorção. As consequências incluem doenças ósseas, fraquezas musculares, maior

suscetibilidade a infecções e risco aumentado de doenças crônicas. O tratamento envolve a reposição da vitamina D, seja por exposição solar, dieta ou suplementação, variando de acordo com a idade e o estado de saúde (MedlinePlus, 2022).

Além das funções conhecidas, a vitamina D influencia os sistemas imunológico, endócrino, nervoso, muscular e cognitivo. Sua atuação abrangente destaca sua importância vital na manutenção da saúde (NIH, 2022).

1.3 Do Protetor Solar

O protetor solar é um indispensável produto cosmético, projetado para resguardar a pele contra os potenciais danos dos raios ultravioleta (UV) provenientes do sol. Essa forma de radiação eletromagnética pode ocasionar malefícios como queimaduras solares, envelhecimento prematuro e até mesmo o desenvolvimento de câncer de pele (American Academy of Dermatology, 2022b).

Existem dois tipos principais de protetor solar, cada um com características distintas (Latha et al. 2013):

- **Protetor Solar Físico:** Composto por substâncias como o óxido de zinco e o dióxido de titânio, esse tipo reflete os raios UV, tornando-os inofensivos. Sua formulação é geralmente segura para todos os tipos de pele, inclusive as mais sensíveis.
- **Protetor Solar Químico:** Formulado com substâncias que absorvem os raios UV, convertendo-os em calor. Esses protetores solares são, em geral, mais leves e menos espessos que seus equivalentes físicos, embora possam ser irritantes para algumas pessoas.

O funcionamento básico do protetor solar é estabelecer uma barreira na pele, impedindo a absorção dos raios UV. Sendo assim, os mecanismos de ação específicos variam entre os tipos (Latha et al. 2013):

- Protetor Solar Físico: O óxido de zinco e o dióxido de titânio refletem os raios UV, direcionando-os de volta ao espaço.
- Protetor Solar Químico: As substâncias químicas presentes absorvem os raios UV, convertendo-os em calor.

O uso diário do protetor solar é crucial, mesmo em dias nublados, pois os raios UV conseguem penetrar nas nuvens. Sua reaplicação a cada duas horas, ou com maior frequência em casos de natação ou transpiração intensa, é recomendada.

O fator de proteção solar (FPS) é um indicador essencial, representando a capacidade do protetor em proteger contra os raios UVA/B. Um FPS de 30 significa que a pele demora até 30 vezes mais para sofrer radiação solar e danos com o protetor do que sem ele (Typology, 2022).

Conforme às recomendações da Sociedade Brasileira de Dermatologia (SBD), o uso diário do protetor solar com FPS 30 ou superior é vital, mesmo em dias nublados. O protetor solar, assim, consolida-se como um elemento essencial na preservação da saúde da pele, salvaguardando contra danos como queimaduras, envelhecimento precoce e potenciais riscos de câncer cutâneo (Sander, 2020).

1.4 Da Inteligência Artificial

No contexto de sistemas computacionais, a Inteligência Artificial (IA), é a disciplina que tem como objetivo a criação de máquinas inteligentes capazes de imitar funções cognitivas humanas, como aprendizado e resolução de problemas.

Como campo na ciência da computação, é dedicada à criação e cultivo de entidades inteligentes capazes de interpretar seus ambientes e iniciar ações para cumprir objetivos especificados. Traçando suas origens às fases iniciais da tecnologia computacional, a IA catalisou marcos significativos desde meados do século 20, sinalizando uma era de progresso sem precedentes no domínio da inteligência digital (Our World in Data, 2022).

Na fase inicial de sua evolução, a disciplina testemunhou o nascimento de algoritmos de aprendizado de máquina e o desenvolvimento de redes neurais artificiais, anunciando uma era transformadora na computação. Este período crucial permitiu que sistemas computacionais aprendessem a cumprir objetivos específicos a partir de vastos conjuntos de dados, facilitando processos de tomada de decisão desprovidos de instruções de programação explícitas (Roberts, s.d.). Os anos 1960 ainda destacaram a utilidade da IA em aplicações práticas, notavelmente em diagnósticos médicos e processamento de linguagem natural (PLN), este último preocupando-se com as intrincadas tarefas de entender e gerar linguagem humana.

Apesar de enfrentar um período de estagnação nos anos 1970 devido a desafios técnicos e conceituais, os anos 1980 marcaram um renascimento na pesquisa de IA, impulsionado pelo advento de novas metodologias e avanços tecnológicos (DataVersity, 2022).

Com o advento do novo milênio, a IA adentra empreendimentos comerciais, particularmente nos domínios de reconhecimento de fala e processamento de imagens. O início do século 21 foi caracterizado por um surto de inovação, primariamente impulsionado por avanços em técnicas de aprendizado profundo que revolucionaram as aplicações de PLN (DataVersity, 2022). Através da aplicação de aprendizado profundo e redes neurais artificiais, computadores obtiveram uma capacidade aprimorada de processar e interpretar representações complexas de dados, possibilitando conquistas como tradução de linguagens e a geração de conteúdo textual criativo, uma vez consideradas além do reino da possibilidade (DataVersity, 2022).

O desenvolvimento acelerado da IA nos anos recentes tem sido notavelmente marcado pelo surgimento de Modelos de Linguagem de Grande Escala (*LLMs*). Tais modelos, treinados em extensos corpora de dados textuais, são adeptos na execução de uma miríade de tarefas de PLN, variando de tradução e geração de texto a sumarização e resposta a perguntas. Estas inovações mantêm a promessa de transformar indústrias inteiras, fomentando maior eficiência, produtividade e inovação. No setor de PLN, avanços estão posicionados para realçar a qualidade da tradução automática, possibilitar a geração de textos criativos, facilitar a

sumarização de textos volumosos e fornecer respostas detalhadas e informativas a consultas (DataVersity, 2023).

Além de seu impacto em tecnologias específicas, a IA é instrumental na liderança de descobertas novas em vários domínios. Ao analisar grandes conjuntos de dados para desenterrar padrões, simular cenários experimentais e gerar hipóteses, a IA contribui significativamente para avanços em campos tão diversos quanto medicina, ciência, engenharia e negócios.

Entre as inovações recentes neste campo está a abordagem de Geração Aprimorada por Recuperação, em inglês Retrieval-Augmented Generation (RAG), um método que sustenta os *LLMs* para assegurar que o processo generativo permaneça factual, preciso e atualmente relevante. RAG aborda especificamente o desafio de "alucinações" nos resultados dos modelos, um termo referindo-se a instâncias quando o texto gerado é factualmente incorreto ou sem sentido (Lewis et al., 2020; Huang et al., 2023). Para mitigar este efeito, bases de dados vetoriais e baseadas em grafos são empregadas, não como uma panaceia, mas como ferramentas poderosas para compelir maior precisão, consistência e explicabilidade das saídas dos *LLMs* (Zapier, 2023).

Mergulhando mais fundo nos anais de PLN e agentes conversacionais, o trabalho seminal de Josef Weizenbaum ainda nos anos 1960 se destaca. A criação de Weizenbaum, ELIZA, um programa que simulava um psicoterapeuta rogeriano (Psicologias do Brasil, 2021), marcou um momento crucial no desenvolvimento de agentes conversacionais, estabelecendo as bases para avanços subsequentes no campo. A capacidade de ELIZA de imitar conversas semelhantes às humanas teve uma influência profunda na evolução dos *ChatBots* de IA, sublinhando décadas de pesquisa visando diminuir a lacuna entre as capacidades linguísticas humanas e os processos computacionais (Weizenbaum, 1966).

Este rico mosaico da evolução da IA, marcado tanto por seus triunfos quanto por seus desafios, sublinha a natureza dinâmica do campo e seu impacto profundo na esfera tecnológica. À medida que a IA continua a evoluir, promete redefinir os limites do que é possível, anunciando um futuro onde máquinas inteligentes se tornam uma parte integrante do empreendimento e da criatividade humanos.

Aprendizagem, no contexto de IA, é uma fase importante do ciclo de vida de modelos, e a aprendizagem de máquina é essencial para isso. Dada uma tarefa que resolve um problema, o processo de aprendizagem consiste em alimentar o modelo com dados de forma iterativa, visando ajustar seus parâmetros internos para que ele possa aprender padrões e fazer previsões ou tomar decisões com base nesses dados. Essa iteração é fundamental para que o modelo melhore sua capacidade de generalização e desempenho em novos dados não vistos anteriormente. Assim, a aprendizagem de máquina permite que os modelos se adaptem e evoluam com base na experiência, tornando-se mais eficazes ao longo do tempo.

Os principais paradigmas de a aprendizagem na IA, são:

- **Aprendizagem supervisionada:** O sistema de IA é treinado com um conjunto de dados rotulados, que contém exemplos de entrada e saída.
- **Aprendizagem não supervisionada:** O sistema de IA aprende com dados não rotulados, sem a necessidade de exemplos de saída.
- **Aprendizagem por reforço:** O sistema de IA aprende a realizar uma tarefa através de tentativa e erro, recebendo recompensas por ações positivas e punições por ações negativas.

As estratégias de aprendizagem de interesse são:

- **Aprendizagem por Transferência:** É uma estratégia poderosa onde um modelo desenvolvido para uma tarefa é reutilizado como ponto de partida para outra tarefa relacionada. Esta metodologia capitaliza o conhecimento adquirido durante o treinamento inicial, permitindo uma adaptação eficiente a novos contextos com menor necessidade de dados de treinamento específicos. A aprendizagem por transferência, em inglês *transfer learning*, é particularmente valioso por sua habilidade de atravessar diferentes domínios ou tarefas, oferecendo um caminho acelerado para a generalização e a autonomia em sistemas de IA. É uma ponte entre o aprendizado existente e novas situações, facilitando a aplicação de experiências passadas a desafios inéditos.
- **Ajuste Fino:** É a técnica utilizada na fase de ajuste de um modelo pré-treinado. O ajuste fino, em inglês *fine-tuning*, ajusta e otimiza um modelo

pré-treinado em um conjunto de dados, geralmente amplo e genérico, para que ele possa desempenhar melhor em uma tarefa específica ou adaptar-se a um novo conjunto de dados que pode ser mais específico ou diferente dos dados originais de treinamento. Embora frequentemente associado à aprendizagem supervisionada, esta estratégia também pode ser aplicada em contextos de aprendizagem semi-supervisionada e até mesmo não supervisionada, dependendo da natureza da tarefa e dos dados. O *fine tuning* é particularmente útil à *transfer learning*, onde o conhecimento aprendido em uma tarefa é transferido e adaptado para melhorar o desempenho em outra tarefa relacionada.

O objetivo da aprendizagem na IA é desenvolver sistemas que sejam capazes de:

- **Aprender com a experiência:** Os sistemas de IA devem ser capazes de aprender com seus erros e acertos, e melhorar seu desempenho ao longo do tempo.
- **Generalizar para novas situações:** Os sistemas de IA devem ser capazes de aplicar o que aprenderam em uma situação para outras situações semelhantes.
- **Ser autônomo:** Os sistemas de IA devem ser capazes de tomar decisões e realizar tarefas de forma autônoma, sem a necessidade de intervenção humana constante.

A aprendizagem é fundamental para a IA por vários motivos:

- **Permite que os sistemas de IA se adaptem a novos dados e situações:** O mundo real é complexo e dinâmico, e os sistemas de IA precisam ser capazes de se adaptar a novas informações e situações para serem eficazes.
- **Melhora o desempenho dos sistemas de IA:** A aprendizagem permite que os sistemas de IA identifiquem padrões e façam previsões mais precisas, o que pode levar a um melhor desempenho em diversas tarefas.
- **Torna os sistemas de IA mais autônomos:** A aprendizagem permite que os sistemas de IA tomem decisões mais autonomamente, sem a necessidade de intervenção humana constante.

Capítulo 2 - Materiais e Métodos

2.1 Metodologia

A abordagem *RAG* aumenta a veracidade de *LLMs*, restringindo sua argumentação e racionalização ao conhecimento armazenado na Base de dados curada, específica e controlável.

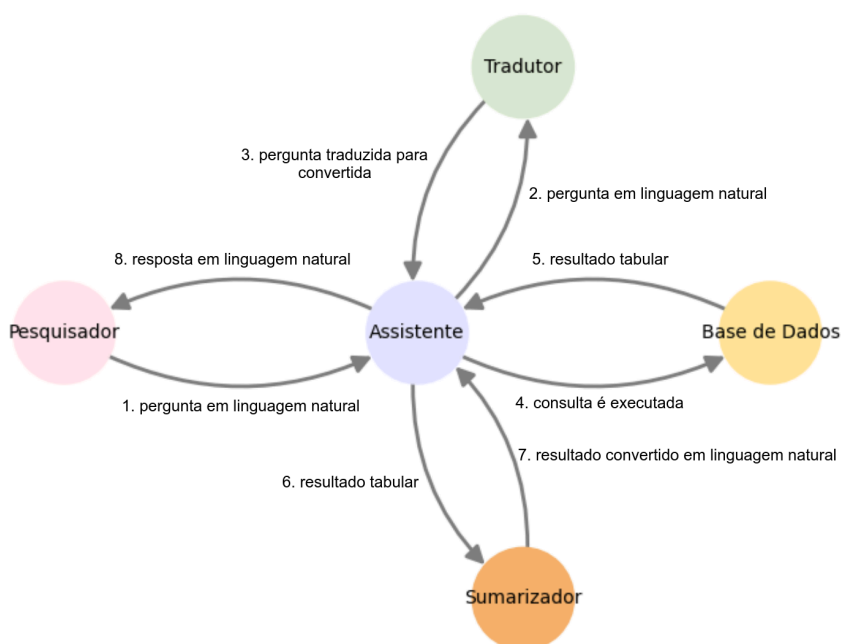


Figura 1.1: Abordagem *RAG*.

Na Figura 1.1, apresenta-se a interação homem-máquina permitida pela abordagem *RAG*. O pesquisador faz uma pergunta em inglês para o assistente, (1). Em seguida, o assistente coordena a tradução da pergunta em inglês (2) para uma consulta em *Cypher* (3). O Assistente submete a consulta em *Cypher* à base de dados (4) e obtém os respectivos resultados em formato tabular (5). O Assistente submete os resultados obtidos ao sumarizador(6.), que retorna a resposta formulada em linguagem natural, inglês (7); a qual é subsequentemente retornada ao pesquisador.

RAG e Aprendizagem

RAG é uma técnica específica para tarefas de geração de texto que combina modelos de linguagem com sistemas de recuperação de informação. Durante a geração, o modelo consulta um banco de dados externo para buscar informações relevantes, que são então usadas para enriquecer a resposta gerada. Em outras palavras, não se usa a memória do modelo, usa-se apenas seu raciocínio para discernir os dados contidos na base de dados.

Nesta abordagem, o modelo previamente treinado num grande e geral conjunto de dados é utilizado sob modo de inferência, não havendo nenhum processo de aprendizagem relacionado à abordagem.

A abordagem RAG utiliza o conhecimento armazenado em uma base de dados curada, enquanto o ajuste fino utiliza conhecimento adquirido em uma tarefa específica.

Enquanto a aprendizagem está limitada, tipicamente, limitada à uma tarefa para a qual existe um conjunto de dados para treinamento, a RAG mais flexível, podendo ser utilizada para qualquer tarefa que possa ser representada em linguagem natural; uma outra condição é que o modelo utilizado entenda também o modelo/linguagem da base de dados.

2.1.3 A estrutura interna do PDF

O formato PDF (Portable Document Format) foi criado pela Adobe em 1991 e tornou-se um padrão ISO (3200) em 2008. O PDF foi criado para mostrar documentos de precisamente por toda e qualquer dispositivos, plataformas, bibliotecas e impressoras.

Um arquivo PDF é uma descrição autossuficiente de um documento plano com layout fixo, capaz de incorporar objetos dos seguintes tipos:

1. **Texto:** O conteúdo real do documento, que pode incluir desde parágrafos simples até textos complexos formatados.

2. **Fontes:** Fontes embutidas ou referenciadas usadas para exibir o conteúdo de texto dentro do documento. As fontes garantem que o texto apareça igual em diferentes plataformas de visualização.
3. **Gráficos Vetoriais:** Gráficos que usam caminhos, pontos, linhas e formas baseadas em equações matemáticas para representar imagens nos documentos PDF. Gráficos vetoriais são independentes de resolução, o que significa que podem ser escalados para qualquer tamanho sem perder qualidade.
4. **Imagens Rasterizadas:** Imagens que são representadas por uma grade de pixels ou pontos. Estas são dependentes de resolução, o que significa que sua qualidade pode diminuir se forem ampliadas demais.
5. **Anotações e Hiperlinks:** Elementos interativos, como notas, destaques ou links para outras seções do documento ou para sites externos.
6. **Campos de Formulário:** Campos interativos onde os usuários podem inserir informações, como campos de texto, caixas de seleção e botões de rádio.
7. **Metadados:** Informações sobre o próprio documento PDF, incluindo autor, data de criação, data de modificação e palavras-chave para indexação.
8. **Camadas (Grupos de Conteúdo Opcionais):** Conteúdo que pode ser tornado visível ou oculto pelo usuário, útil para documentos com múltiplos propósitos ou que precisam apresentar informações de forma seletiva.
9. **Favoritos e Estrutura do Documento:** Auxílios de navegação que fornecem acesso rápido a seções específicas do documento, atuando como uma tabela de conteúdos.
10. **Assinaturas Digitais:** Recurso de segurança para autenticar a origem e integridade do conteúdo do documento.
11. **Criptografia e Configurações de Segurança:** Protegem o documento contra acesso ou modificação não autorizados, com restrições à impressão, edição ou cópia.
12. **Etiquetas e Numeração de Páginas:** Etiquetas personalizadas para páginas, que podem incluir diferentes estilos de numeração, como numerais romanos para páginas introdutórias.
13. **Arquivos Embutidos:** Outros documentos ou arquivos que estão incluídos dentro do PDF, acessíveis para visualização ou download.

14. **Espaços de Cor e Perfis:** Informações para garantir a reprodução consistente de cores em diferentes dispositivos, especificando como as cores devem ser renderizadas.
15. **JavaScript:** Código embutido que pode adicionar interatividade, automatizar tarefas ou validar entradas de formulários dentro do documento.
16. **Elementos Multimídia:** Clipes de áudio e vídeo embutidos que podem ser reproduzidos dentro do documento, enriquecendo a experiência interativa para os espectadores.

Neste trabalho, foca-se na extração do texto de PDFs, que, a baixo nível, é um processo complexo que envolve várias etapas técnicas. Entender esse processo requer alguma familiaridade com a estrutura dos documentos PDF e as informações são armazenadas nestes.

Analisando a Estrutura do PDF

O primeiro passo é analisar esses objetos para construir uma representação da estrutura do documento. Isso envolve decodificar o formato binário do PDF para identificar os diferentes objetos, seus tipos e seus relacionamentos.

Dos objetos de maior interesse aqui, destacam-se:

- a. **Objetos:** Blocos de construção básicos dos arquivos PDF, tais como dicionários, arrays, fluxos (que podem conter dados binários) e tipos simples (números, strings, nomes, etc.).
- b. **Objetos Indiretos:** Objetos que podem ser referenciados em todo o PDF, permitindo reutilização e modularidade.

Estes representam o texto que se deseja extrair dos arquivos ao longo deste trabalho.

Localizando e Interpretando Streams de Conteúdo

Texto e outro conteúdo da página são armazenados em streams de conteúdo dentro de objetos de página. Esses streams contêm uma série de instruções (ou operadores) para desenhar a página, incluindo:

- Objetos de Texto: Delimitados por operadores *BT* (*begin of text*; início do texto) e *ET* (*end of text*; finalizar texto), contendo os comandos de desenho do texto.
- Operadores de Exibição de Texto: Como *Tj*, *TJ*, *""*, e *'-'*, que especificam as strings de texto a serem desenhadas e como elas são desenhadas.
- Operadores de Posicionamento de Texto: Incluindo *Td* e *TD*, que movem a posição atual para onde o texto subsequente será desenhado. O operador *Td* ajusta a posição do texto adicionando as coordenadas especificadas à posição atual de texto, facilitando a colocação do texto em pontos específicos na página.
- Operadores de Estado do Texto: Controlam a aparência do texto, incluindo fonte (*Tf*), tamanho da fonte, espaçamento entre caracteres e modo de renderização de texto.

Streams de PDFs são sequências de bytes compostas por dicionário denominado pelas tags “<< >>”, uma sequência de bytes envolta pelas palavras chave *stream* e *endstream*, que, respectivamente, denotam o início e o fim do conteúdo do stream.

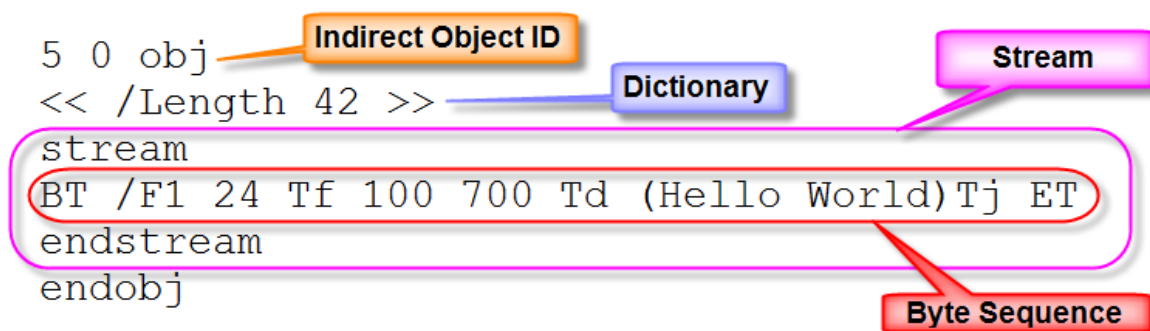


Figura 2.2 (1.2): Exemplo da estrutura interna de um stream de PDF. Obtido de <https://blog.didierstevens.com/2008/05/19/pdf-stream-objects/> (Acesso em: 1/4/2024).

A Figura 2.2 ilustra como o texto de um PDF é armazenado na estrutura interna do documento. Este trecho, quando renderizado corresponde a um documento no qual, na coordenada $(x, y) = (100, 700)$ (**Tf 100 700 Td**), “**Hello World**” com fonte **F1** de tamanho 24 (**/F1 24 Tf**); onde (0,0) corresponde ao canto inferior esquerdo. **5 0 obj** informa que o começo da definição do bloco de stream se dá no objeto número **5** com número de geração **0**. **/Length 42** indica o comprimento

do stream em bytes.

Extraindo e Decodificando Texto

Fragmentos de texto extraídos geralmente são codificados usando vários esquemas de codificação, dependendo das fontes usadas (INFOSEC INSTITUTE, n.d.). O processo de extração envolve:

- Decodificação de Fonte: Cada fonte em um PDF pode ter sua própria codificação, mapeando códigos de caracteres para caracteres ou glifos reais. Esse mapeamento deve ser interpretado para decodificar o texto corretamente.
- Arquivos CMap (Mapa de Caracteres): Para fontes que usam codificações CID (Character ID), como muitas fontes não latinas, arquivos CMap são usados para mapear códigos de caracteres para valores Unicode.

Trabalhando com Layout e Ordem do Texto

A ordem lógica do texto em um PDF pode não corresponder à aparência visual devido à flexibilidade de layout do formato (SAVE EMAILS AS PDF, n.d.). Etapas adicionais geralmente são necessárias para inferir a ordem correta das strings de texto, incluindo:

- Operadores de Posicionamento de Texto: Comandos como Td e Tm movem o cursor de texto ou transformam a matriz de texto, afetando o posicionamento do texto. Eles devem ser interpretados para entender onde o texto aparece na página.
- Análise de Layout: Para layouts complexos (por exemplo, texto em várias colunas, tabelas), algoritmos podem analisar as posições e sequências de blocos de texto para reconstruir a ordem de leitura pretendida.
- Abordagens de Aprendizado de Máquina: Ferramentas avançadas de extração de texto podem empregar modelos de aprendizado de máquina treinados em uma variedade de layouts de documentos para prever a ordem de leitura mais provável dos blocos de texto.
- Caixas Delimitadoras: Calcular as caixas delimitadoras ao redor dos blocos de texto ajuda a entender sua relação espacial e a ordená-los com base nas convenções de layout do documento (por exemplo, da esquerda para a direita, de cima para baixo para o inglês).

Desafios e Considerações

- Elementos Não Textuais: O texto pode estar intercalado ou sobreposto por elementos não textuais (imagens, gráficos vetoriais), complicando a extração.
- Gráficos Codificados como Texto: Às vezes, o que aparece como texto pode ser na verdade gráficos vetoriais ou imagens, exigindo OCR (Reconhecimento Óptico de Caracteres) para extração.
- Subconjunto de Fontes: PDFs geralmente incluem apenas um subconjunto dos caracteres usados em uma fonte, complicando o mapeamento de códigos de caracteres para o texto real.

A extração de texto em um nível mais baixo envolve decodificar a estrutura do PDF, interpretar streams de conteúdo, decodificar texto de acordo com codificações de fonte e reconstruir a ordem lógica do texto. Dadas as complexidades de formatação e codificação de PDF, a extração precisa de texto pode ser desafiadora e frequentemente requer algoritmos sofisticados de análise e análise sintática. Ferramentas e bibliotecas que realizam extração de texto de PDF encapsulam esses processos, fornecendo interfaces para acesso mais fácil ao texto contido em documentos PDF.

2.1.4 Extração do conteúdo de PDF

Aqui discorre-se sobre as três estratégias de extração do conteúdo de PDFs acima mencionadas na seção 2.1.2.

Extração no ecossistema Unix

Esta abordagem consistiu em, programaticamente, extrair o conteúdo de todos os arquivos em PDF utilizando todas as ferramentas acima mencionadas.

Conversão para docx

Este procedimento foi executado na nuvem, através do conversor de PDF para Word do *PDF2GO*, disponível em <https://www.pdf2go.com/pt/pdf-para-word> (Acesso em: 5/2/2024).

Extração em nuvem:

Esta abordagem consistiu em extrair o texto dos arquivos PDF utilizando o serviço do PDF2GO, disponível no endereço <https://www.pdf2go.com/pdf-to-text#> (Acesso em: 5/2/2024).

2.1.6 Pós-processamento dos dados

A limpeza do texto consiste em duas etapas: correções manual e automática.

Correção

Como nenhuma das abordagens de conversão de PDF para txt manteve a ordem original de leitura humana, foi necessário corrigir o conteúdo extraído de forma a garantir a ordem lógica de leitura.

Assim, faz-se necessário limpar o texto espúrio e conectar o texto de uma forma que faça sentido.

Nesta etapa, a fim de aumentar a reprodutibilidade, extrai-se o conteúdo textual dos arquivos pdf em texto pleno, o qual é salvo em um arquivo de texto (.TXT).

Uma vez salvos, os arquivos de texto passam por duas etapas:

1. Remoção de dados não significativos ou não legíveis
 - a. tabelas
 - b. imagens
 - c. notas de rodapé
 - d. tudo que antecede a introdução
 - e. declarações burocráticas, formais e legais
 - f. seção de referências
 - g. marcas d'água de instituições
 - h. informações às margens do documento
2. Correção da ordem do texto

Além de corrigir a ordem lógica do texto, fez-se necessário corrigir erros tipográficos e de pontuação, remoção de dados tabulares e qualquer resquício de layout.

Para tanto, utilizou-se três abordagens; correção programática, correção assistida por IA e correção manual.

Correção Programática

Para atacar esta tarefa, utilizou-se as bibliotecas: *spaCy*, *wordsegment*, *spellchecker*. As quais apesar de funcionarem pontualmente, não desempenharam a tarefa de forma satisfatória no geral. Devido à presença de objetos numéricos e simbólicos, expressões, quantidades e referências bibliográficas, as bibliotecas utilizadas não foram capazes de detectar as sentenças.

Correção Assistida por IA

A correção assistida por IA foi atacada com um único prompt, que foi utilizado de duas formas, via API e via modo playground; para três modelos distintos, *mistral*, *gemini-1.0-pro* e *gpt-4.0*.

Para o uso de *LLMs*, é importante respeitar o *context window* do modelo correspondente; que é a quantidade (medida em tokens) de texto que o modelo consegue tratar de uma vez, no processamento de uma pergunta. Isto inclui tamanho do *Prompt*, da entrada, da pergunta e da resposta gerada.

2.1.7 Desenho Experimental

Ao longo deste trabalho, iniciou-se utilizando um GC como a estrutura de dados responsável pela operação dos dados. No entanto, devido às dificuldades na construção do GC, discutidas na seção resultados e discussões, utilizou-se uma abordagem alternativa.

Considerações gerais

Por limitações computacionais, é necessário dividir a entrada em blocos menores. Como não se sabe o tamanho do bloco para o qual a extração é otimizada, faz-se necessário aplicar o método científico para descobrir como dividir os documentos em blocos.

Tabela 2.1: Desenho Experimental

Etapa	Variáveis	Propósito
Ingestão	PDF	disponibilidade do dado
Parametrização	chunk_size chunk_overlap	Especificar os tamanhos dos blocos de texto a serem processados, a sobreposição entre um bloco e o seu posterior bem como sua representação numérica
Formatação	clean replace	Remoção de caracteres específicos, tokens e pontuações específicas e alguns padrões a fim de trazer mais consistência ao processamento
Extração de Entidades e Relações	<i>pipelines</i> disponibilizados pela John Snow Labs.	Obtenção das entidades e relações em determinados através de <i>pipelines</i> domínio-específicos
Armazenamento	csv	Armazenamento das entidades e relações correspondentes ao formato da saída esperada do pipeline
Construção do Grafo de Conhecimento	Neo4j	Os dados do csv são arranjados no formato esperado do Neo4j e carregado.

A Tabela 2.1 apresenta uma síntese do método experimental adotado nesta pesquisa, visando descobrir uma abordagem reprodutível para a obtenção dos melhores Grafos de Conhecimento (GCs). Esses GCs são fundamentais para

proporcionar as respostas mais eficazes e relevantes, pautadas por critérios de factualidade, veracidade, verificabilidade, utilidade e clareza, que são os objetivos centrais desta investigação.

No âmbito desta abordagem, as variáveis independentes que regem o experimento são as seguintes: *chunk_size*, *chunk_overlap*, *clean* e *pipeline*. Estas variáveis são cuidadosamente manipuladas com o propósito de otimizar o processo e obter GCs que atendam aos critérios estabelecidos.

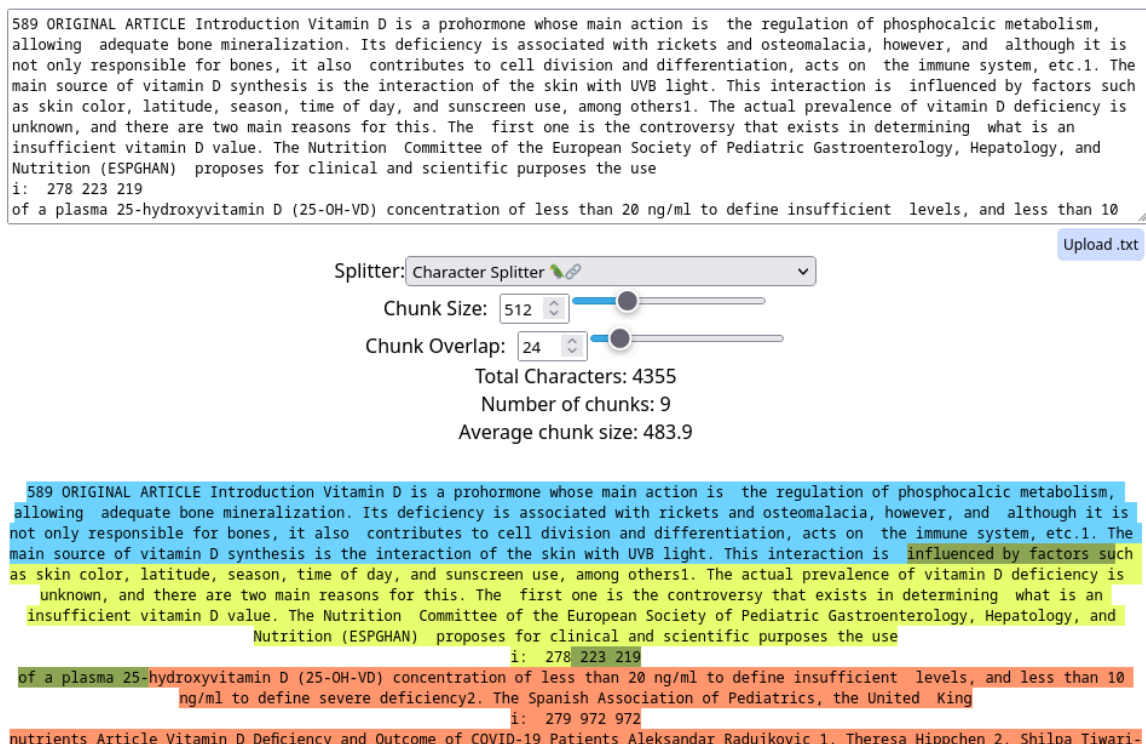


Figura 2.3: Ilustração da *tokenização* (fonte: <https://chunkviz.up.railway.app/>)

A Figura 2.3 proporciona uma visualização da *tokenização* de um texto genérico. Para esse exemplo, optou-se por um tamanho de bloco (*chunk_size*) de 512 caracteres, com uma sobreposição entre blocos (*chunk_overlap*) de 24 caracteres. Esses parâmetros são selecionados para a divisão do texto em blocos de texto menores.

O *chunk_size* é um valor contido no conjunto [512, 1024, 2048, 4096] e o *chunk_overlap* é um valor contido no conjunto [0, 24, 56]. O parâmetro *clean* indica se o arquivo de entrada é limpo ou não, isto é, tem alguns caracteres removidos. O

parâmetro *pipelines* descreve as tarefas que foram executadas no experimento em questão. A combinação destes parâmetros resulta em 96 experimentos, cada qual com seu GC.

Por outro lado, a variável dependente, que reflete os resultados do experimento, é a quantidade de relações não redundantes. A quantidade de relações é um indicativo que uma base de conhecimento profunda, já a qualidade das relações fica para a validação das respostas, que envolve considerações objetivas e subjetivas. Essas métricas são cruciais para avaliar a eficácia e a qualidade das respostas geradas pelos GCs.

Cada iteração do experimento utiliza como entrada um único documento no formato PDF, acompanhado de um conjunto específico de parâmetros para particionamento e pré-processamento, além da definição do *pipeline*. Como resultado, o experimento gera consistentemente um único arquivo CSV como saída, e os dados resultantes são inseridos em uma única base de dados no Neo4j.

O Reconhecimento de Entidades Nomeadas (NER, do inglês "Named Entity Recognition") é uma técnica de processamento de linguagem natural (PLN) que visa identificar e classificar entidades nomeadas em texto em categorias predefinidas, como nomes de pessoas, organizações, locais, datas, quantidades, valores monetários, percentuais, entre outras. O objetivo é extrair informações estruturadas de dados não estruturados, facilitando a compreensão, organização e utilização desses dados em diversas aplicações, como sistemas de busca, análise de sentimentos, assistentes virtuais, entre outros.

Os arquivos passam pelo passo de extração de entidades e relações. Este passo será feito por múltiplos processos de extração de entidades. Destes processos, um é um *prompt multiple shot* e os outros são os modelos que executam a *NER* providenciados pela John Snow Labs.

Um exemplo de experimento é:

```
(2048, 24, true, true, 'clinical_temp_events_re_pipeline')
```

Este conjunto de parâmetros indica que o texto de entrada será dividido em blocos de aproximadamente 2048 caracteres, com uma sobreposição de 24 caracteres entre cada bloco adjacente. O texto será submetido a um processo de limpeza, e alguns caracteres serão substituídos conforme necessário.

Posteriormente, será aplicado o *pipeline* denominado '*clinical_temp_events_re_pipeline*', que foi treinado para a tarefa específica de extrair relações e entidades relacionadas a eventos temporais em textos relacionados a ensaios clínicos. O resultado final será um arquivo CSV formatado de acordo com as expectativas para essa tarefa específica.

```
relation,entity1,entity1_begin,entity1_end,chunk1,entity2,entity2_begin,entity
2_end,chunk2,confidence
TrAP,TREATMENT,130,137,delivery,PROBLEM,142,153,the
organism,0.9999888
TeRP,TEST,322,342,vitamin D3 absorption,PROBLEM,423,438,mainly
dependent,1.0
TrAP,TREATMENT,652,660,vitamin
D,PROBLEM,709,715,altered,0.9999774
TeRP,TEST,667,681,its metabolites,PROBLEM,709,715,altered,1.0
TeRP,TREATMENT,777,832,"The digestion and absorption processes,on
the other hand",PROBLEM,842,856,greatly altered,1.0
TeRP,TREATMENT,952,968,complex
enzymatic,PROBLEM,977,1001,physicochemicalmechanisms,0.8664989
```

S(1)

Resumidamente, o processo consiste em carregar todos os arquivos na memória, seguido pela fragmentação destes em blocos menores. Cada bloco passa por um processo de extração das entidades e relações contidas, resultando na exportação dessas informações para um formato CSV, como no S(1). Esses dados são então armazenados e catalogados com base no tamanho do bloco e nas etapas de pré-processamento realizadas.

2.1.8 Extração de Entidades e Relações

A etapa de extração de entidades e relações envolve processamento e análise mais denso de todas as outras. O êxito desta tarefa pauta a experimentação e determina o sucesso da abordagem. Esta tarefa é requisito para a criação do GC. Esta tarefa é 100% específica ao domínio e também a tarefa mais computacionalmente intensa. A qualidade desta etapa determina a qualidade de todas as etapas subsequentes.

Neste trabalho utiliza-se múltiplas alternativas para a tarefa em questão. Uma chain própria, que é a composição da entrada com o *prompt* com o *LLM* e um parser,

que analisa o texto gerado. Assim, o LLM é alimentado com a entrada (dado a ser processado) e o *prompt* (instruções que o LLM executará na dado de entrada) e o parser que garante que o resultado é válido.

Domain:

- biology
- biochemistry
- endocrinology
- dermatology
- medical sciences
- clinical trials
- virology
- nephrology
- gastroenterology
- hepatology
- immunology
- osteology
- posology
- statistics

Task:

- Consider Domain: Consider the context given to accurately determine triples.
- Extract key Concepts and Relationships: Identify key concepts and relationships in the given input text.
- Prioritize Saliency: Focus on extracting the most important and meaningful concepts and relationships.
- Generate triples: A triple, `subject-predicate-object`, that represents the relationship between two concepts.
- Triple expansion: If any triple component contains a conjunction, expand it.

P(1)

Format:

Structured Output: The output must be a valid csv 5-elements object, where each triple contains the following list of attributes: subject_name, "subject_type", "relationship", "object_name", "object_type"; in this order.

Example:

Input [plain text]: "Albert Einstein is best known for developing the theory of relativity, which revolutionized physics."

Output [list of triples]:

- 1. - `Albert Einstein`_-.`person`_-.`developed`_-.`Theory of Relativity`_-.`theory` <@>
- 2. - `Theory of

```
Relativity`_-.`theory`_-.`revolutionized`_-.`Physics`_-.`field of study` <@>
```

Additional Considerations:

- each triple must contain 5 elements
- each sub-sentence can have up to 7 triples
- revisit the triples with more than 5 elements, if not comply remove it
- each triple object must be a valid csv object
- the output must be an enumerated list of triples
- concepts can not be a conjunction, null, None or empty string
- if a concept or relationship contain a conjunction, expand the whole triple into two conjunction-free triples
- relationships can not be a conjunction, null, None or empty string

Format Instructions: {{ format_instructions }}

Input: {{ unstructured_text }}

O $P(1)$ define o que são as entidades e relações, como extraí-las e, por último, como formatá-las. Garantir o desempenho eficiente de um *Prompt* é frequentemente desafiador devido à natureza não estruturada do texto de entrada e à abordagem estatística do *LLM*. Isso torna a extração de 100% das entidades e relações uma tarefa complexa.

Além desta técnica, utiliza-se *pipelines* já prontos que são responsáveis por realizar a mesma tarefa, porém focados em um domínio mais específico.

Tabela 2.2: *Pipelines* de Extração utilizados

Pipeline	Tarefa
Clinical Temporal Events RE	Extração de relações temporais em ensaios clínicos
Clinical RE	Extração de relações relativas a ensaios clínicos
Posology Relation Extraction	Extração de relações numéricas e temporais de posologias

A exploração da Tabela 2.2, que detalha os *pipelines* de extração utilizados, oferece uma perspectiva elucidativa sobre as distintas tarefas abordadas por cada modelo.

O *pipeline Clinical Temporal Events RE* é dedicado à tarefa de extração de eventos temporais clínicos e destaca-se a especialização na identificação e análise de relações temporais em ensaios clínicos. O intuito é que as entidades e relações dele extraídas contribuam para a compreensão dos eventos clínicos ao longo do tempo.

O *pipeline Clinical RE* desempenha a tarefa de extração de relações clínicas, isto é, concentra-se na identificação e interpretação de relações específicas relacionadas a ensaios clínicos. O intuito que guia o uso deste *pipeline* é a compreensão das complexidades de interações variadas em ensaios clínicos, contribuindo para uma compreensão mais profunda dos contextos e relacionamentos nesse domínio.

O *pipeline Posology Relation Extraction*, direcionado à extração de relações de posologia, destaca-se por sua habilidade em extrair relações numéricas e temporais em posologias. Essa tarefa específica sugere uma abordagem focada na compreensão das dosagens e temporalidades associadas a tratamentos, fornecendo informações cruciais para a análise e otimização de protocolos de administração.

A extração de entidades e relações poderia também ser obtida através de uma *chain* customizada.

2.1.9 Criação do Grafo de Conhecimento

Antes de construir o GC, convém explicar o contexto teórico no qual ele surge, na ontologia. Uma ontologia é um modelo conceitual que define os conceitos, relacionamentos e regras dentro de um domínio específico. Ela especifica o que existe neste domínio, como as coisas se relacionam e as propriedades que elas possuem.

As ontologias são agnósticas de domínio, o que significa que podem ser aplicadas a diferentes domínios. Elas se preocupam principalmente em capturar o próprio conhecimento do domínio, em vez de serem adaptadas a uma tarefa específica.

Um caso específico de ontologias é o GC. Um GC é uma instância específica de uma ontologia, preenchida com dados reais sobre entidades e seus relacionamentos. A ontologia aqui utilizada foi implementada através da tríade sujeito - relação- objeto.

Os GCs geralmente são construídos para tarefas ou aplicativos específicos. Eles aproveitam a estrutura e as regras definidas em uma ontologia para representar e analisar dados relevantes a esse contexto.

Os GCs são dinâmicos e em evolução. Eles são constantemente atualizados com novos dados e podem até mesmo adaptar sua estrutura com base em análises e raciocínios contínuos.

Para ilustrar a diferença entre ontologias e GCs, podemos pensar em uma ontologia como um dicionário para um campo específico. Ela define todos os termos, conceitos e seus relacionamentos relevantes. Uma ontologia define o que pode existir em um domínio e como as coisas se relacionam. Um GC representa fatos e relacionamentos específicos dentro desse domínio.

Compreender a diferença entre esses dois conceitos é crucial em vários campos, incluindo inteligência artificial e recuperação de informações. As ontologias fornecem a base para estruturar o conhecimento, enquanto os GCs aplicam essa estrutura para representar e raciocinar sobre os dados reais.

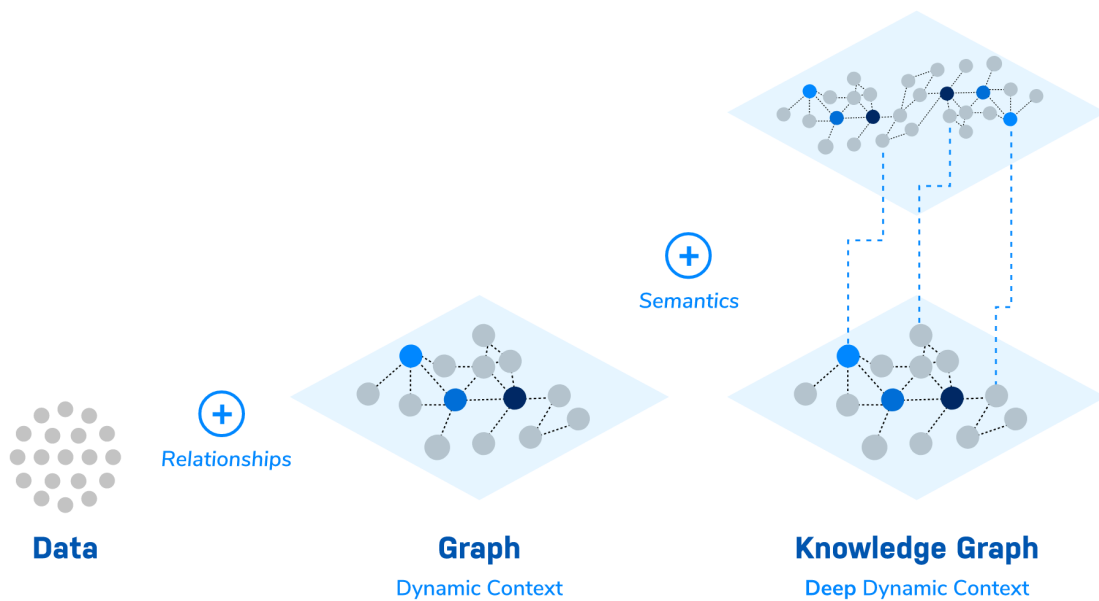
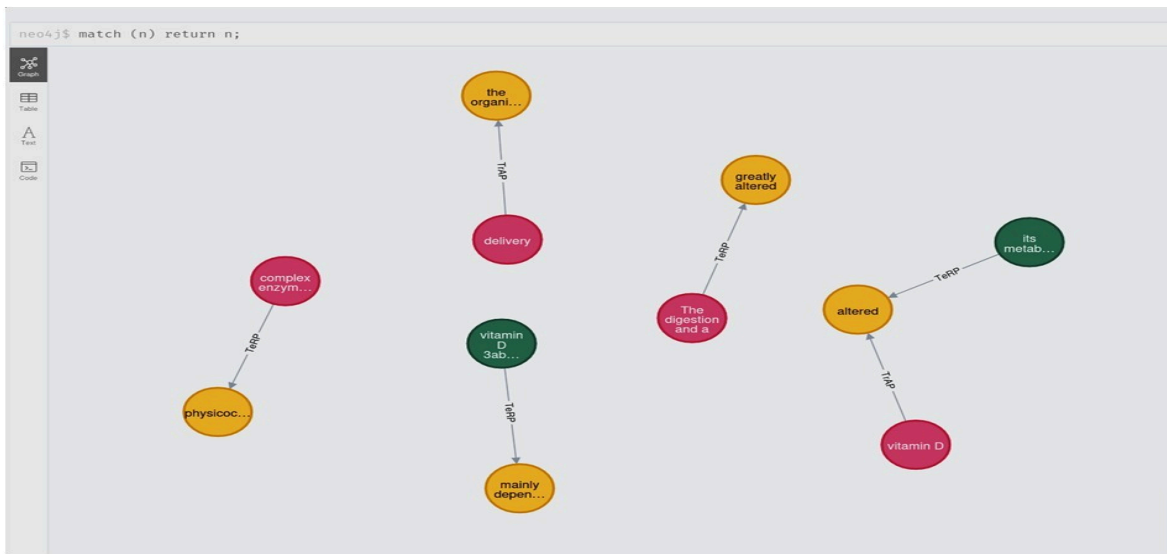
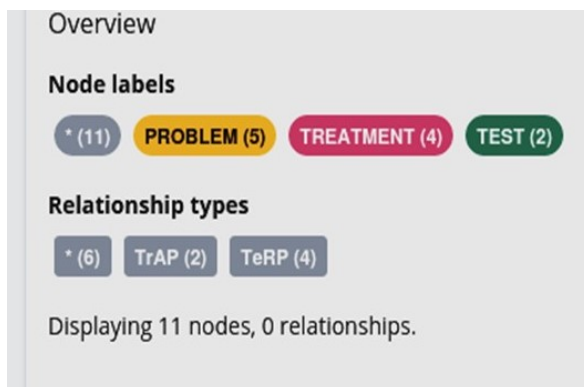


Figura 2.4: Composição conceitual de Gcs (fonte: <https://neo4j.com/use-cases/knowledge-graph/>).

A Figura 2.4 ilustra as camadas de abstrações que compõem GCs sob a perspectiva da arquitetura da informação. Os pontos representam os dados, enquanto que as linhas pontilhadas representam relações entre dados, dando origem a grafos. A atribuição de rótulos e atributos a entidades e relações de um grafo equivale a adicionar semântica a este, formando GCs.



a



b

Figura 2.5: Carregamento da S(1) na base de dados.

a) representa os conceitos (*Nodes*)

b) nomes dos tipos dos conceitos e das relações (*Relationships*).

A base de dados é gradualmente enriquecida à medida que cada arquivo CSV é carregado, contribuindo para a formação do GC. A Figura 2.5 representa o processo de carregamento do S(1), ilustrando a incorporação progressiva dessas informações à estrutura do grafo.

A construção do GC foi feita de algumas maneiras, dentre as quais:

Prompt puro

A ideia aqui foi utilizar uma cadeia (*chain*) capaz de, por si só, extrair as entidades e relações de um bloco de texto e criar o GC. Isso funciona esporadicamente.

O prompt foi utilizado da seguinte maneira:

```
prompt = PromptTemplate(
    template=PROMPT_MESSAGE,
    input_variables=["unstructured_text"],
    partial_variables={
        "format_instructions":
            output_parser.get_format_instructions()},
    format="json",
    template_format="jinja2",
    validate_template=True,
)

chain = (
    {"unstructured_text": RunnablePassthrough()}
    | camel_case_prompt
    | llm
    | output_parser
)
```

F(1)

O F(1) mostra como a *chain* executa as instruções contidas no PROMPT_MESSAGE, contido no P(1), o qual recebe como contexto o texto a ser processado e formatado de acordo com a *format_instructions*. O resultado é então analisado, retornado e o efeito colateral é efetivado; o efeito é o armazenamento dos dados no Neo4j.

NaLLM

NaLLM (<https://github.com/neo4j/NaLLM>) é um projeto, criado pela Neo4j Inc., que visa facilitar a criação de grafos de conhecimento utilizando o Neo4j como base de dados e os modelos da OpenAI.

spaCy

displaCy é um método/função oferecida pela biblioteca *spaCy* que classifica palavras e traça a relação entre elas. De forma que, a partir de uma frase, obtém-se um objeto representando as relações de dependências entre palavras. Uma outra funcionalidade da *displaCy* é performar a *NER*.

“Coronaviruses are enveloped and single-stranded RNA viruses.”

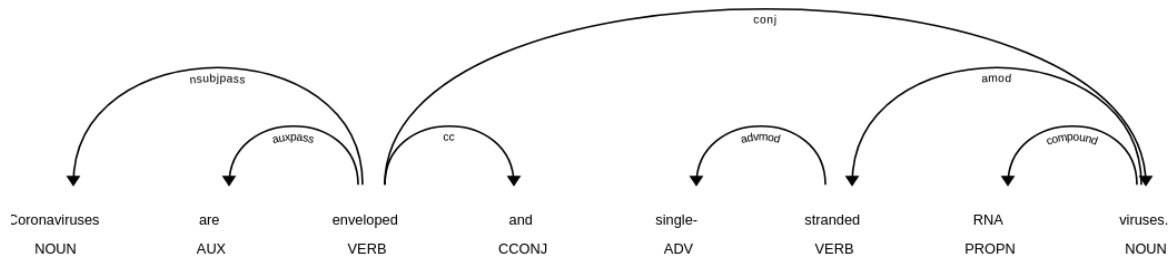


Figura 2.6: Palavras e suas dependências.

Coronaviruses are enveloped and single-stranded **RNA ORG** viruses.

Figura 2.7: NER obtido através do *displaCy*.

Onde as classes mostradas na Figura 2.6 são mostradas abaixo, com outros exemplos:

- **NOUN**, substantivo: Essa tag identifica palavras que representam pessoas, lugares, coisas ou ideias. Exemplos: "gato", "livro", "felicidade".
- **AUX**, verbo auxiliar: Essa tag identifica verbos que ajudam outros verbos a formar tempos ou modos. Eles não funcionam de forma independente. Exemplos: "tem", "é", "estavam".
- **VERB**, verbo: Essa tag identifica palavras de ação ou verbos que descrevem o que está acontecendo ou o que alguém ou algo faz. Exemplos: "correr", "pensar", "dormir".
- **CCONJ**, conjunção coordenativa: Essa tag identifica conjunções que unem palavras, frases ou orações de igual valor gramatical. Exemplos: "e", "mas", "ou".

- **ADV**, advérbio: Essa tag identifica palavras que modificam verbos, adjetivos ou outros advérbios. Elas fornecem mais informações sobre como, quando, onde ou por que algo acontece. Exemplos: "rapidamente", "muito", "amanhã".
- **PROPN**, nome próprio: Essa tag identifica nomes específicos de pessoas, lugares ou organizações. Exemplos: "Alice", "Londres", "Microsoft".

KnowledgeGraphIndex

A construção de um GC utilizável foi alcançada através do componente *KnowledgeGraphIndex*, da *LlamaIndex*, que funciona de forma reprodutível, programática e com baixo esforço e consumo de recursos computacionais.

O termo "índice", no contexto da recuperação e representação dos dados, refere-se a uma estrutura de dados usada para melhorar a velocidade das operações de recuperação de dados. Um índice pode ser pensado como uma maneira de organizar dados que permite uma busca rápida, recuperação e acesso de informações dentro de um banco de dados, um sistema de arquivos ou outra coleção de dados.

A natureza exata de um índice pode variar amplamente dependendo de sua implementação e das necessidades específicas que visa atender, variando de tabelas de pesquisa simples a estruturas de dados mais complexas como B-trees, tabelas de hash ou índices invertidos.

Armazenamento Vetorial vs. Grafo de Conhecimento

- **Armazenamento Vetorial:** Isso refere-se ao método de armazenar dados como vetores, que são essencialmente arranjos de números. No contexto de machine learning e motores de busca, o armazenamento vetorial é frequentemente usado para representar dados complexos como texto, imagens e mais em um espaço matemático. O armazenamento vetorial permite operações eficientes de busca por similaridade, onde o objetivo é encontrar pontos de dados (vetores) que estão mais próximos a um vetor de

consulta dado. Isso é particularmente útil em aplicações como busca semântica, sistemas de recomendação e detecção de anomalias.

- **Grafo de Conhecimento:** Um GC representa dados como uma rede de entidades interconectadas e suas relações. Cada nó no grafo representa uma entidade (como uma pessoa, lugar ou conceito), e as arestas representam as relações entre elas. Grafos de conhecimento são usados para armazenar dados complexos e interrelacionados de uma maneira que preserva o contexto e as relações entre diferentes peças de informação. Eles são amplamente usados em aplicações que requerem consultas semânticas ricas, como busca semântica, sistemas de resposta a perguntas e plataformas de integração de dados.

A construção do GC satisfatória foi alcançada através do componente *KnowledgeGraphIndex*, que funciona de forma reprodutível, programática e com baixo esforço e consumo de recursos computacionais.

***KnowledgeGraphIndex* como uma Estrutura Híbrida**

KnowledgeGraphIndex é uma estrutura de dados híbrida que combina características de ambos grafos de conhecimento e mecanismos de indexação. Neste contexto, o objetivo seria aproveitar os dados ricos e semanticamente conectados de um GC enquanto também fornece capacidades de busca e recuperação rápidas através da indexação. Isso envolve a criação de índices que permitem um acesso rápido a nós e relações dentro de um GC baseado em certos atributos chave ou padrões de consulta.

Assim, esta se beneficia da ontologia empregada pelo GC, enquanto também garante que consultas contra esses dados possam ser realizadas rapidamente, de forma semelhante à eficiência do armazenamento vetorial no manuseio de buscas por similaridade. Tal estrutura híbrida é particularmente poderosa em cenários que requerem um entendimento semântico profundo dos dados combinado com capacidades de consulta de alta performance, como em aplicações avançadas de IA, motores de busca semântica e tarefas complexas de análise de dados.

2.1.10 Perguntas e Respostas

Aqui são utilizadas 6 estratégias de perguntas e respostas, sendo estas:

1. Recuperação de entidades baseada em vetores
2. Recuperação de entidades baseada em palavras chave
3. Recuperação de entidades híbrida
4. KnowledgeGraphQueryEngine
5. KnowledgeGraphRAGRetriever
6. Cadeia de Perguntas e Respostas

Recuperação de entidades baseada em vetores

Utiliza-se o próprio índice (*LlamaIndex*) como motor de consulta. Este é o método padrão como os motores de consulta de GCs construídos com base no índice. Já está pronto para uso e funciona sem qualquer parâmetro adicional.

```
query_engine = index.as_query_engine() F(2)
```

Este motor de consulta emprega similaridade vetorial para buscar entidades do GC, trazendo trechos de texto vinculados, podendo até mesmo explorar relacionamentos.

Recuperação de entidades baseada em palavras chave

Este motor de consulta emprega palavras-chave da consulta para recuperar entidades relevantes do GC, extraindo fragmentos de texto relacionados e, opcionalmente, explora outros relacionamentos para coletar mais contexto.

```
query_engine = index.as_query_engine(  
    include_text=False,  
    retriever_mode="keyword",  
    response_mode="tree_summarize",  
) F(3)
```

F(3) apresenta uma variação do F(2), porém como um foco ligeiramente distinto como especificado nos parâmetros abaixo.

- `retriever_mode="keyword"`: busca por palavras-chave
- `include_text=False`: não inclui o texto do Node na resposta.
- `response_mode="tree_summarize"`: resumir a resposta sob uma estrutura em árvore do GC. A árvore será construída recursivamente, tendo a consulta inicial como o nó principal e as respostas mais pertinentes como nós secundários.

Esse modo de resposta funciona bem na tarefa de sumarização, oferecendo um panorama geral sobre um assunto ou esclarecendo questões que demandam explicações detalhadas. Este modo, permite a elaboração de respostas mais complexas, como listar motivos que justificam uma afirmação ou descrever procedimentos.

Recuperação de entidades híbrida

Similar ao motor de consulta apresentado no F(3), porém utilizando outro modo de embedding. Aqui também indica-se outros 3 parâmetros distintos para tornar a busca mais precisa.

```
query_engine = index.as_query_engine(
    include_text=True,
    response_mode="tree_summarize",
    embedding_mode="hybrid",
    similarity_top_k=3,
    explore_global_knowledge=True,
)
```

F(4)

O parâmetro `embedding_mode="hybrid"`, indica para o motor de consulta utilizar **recuperação vetorial de entidades** e **recuperação baseada em palavras-chave** para buscar informações do GC.

Este método de recuperação usa palavras-chave para encontrar tríades relevantes e o de recuperação vetorial de entidades para encontrar tríades similares através de similaridade semântica. Esta estratégia combina busca por palavras-chave e busca semântica, unindo os pontos fortes de ambos os métodos para melhorar a precisão e a relevância dos resultados da busca.

- `include_text=True`: utiliza o texto dos *Nodes* da resposta.

- *similarity_top_k=3*: com base nos *embeddings*, recuperar os três resultados mais similares.
- *explore_global_knowledge=True*: diz ao motor de consulta este deve considerar o contexto global do GC ao recuperar informações. Fazendo com que o motor de consulta não limite a busca ao contexto local (na vizinhança imediata de um dado nó), senão inclui o contexto global mais amplo do GC. É utilizado quando se deseja recuperar informações que não estão diretamente conectadas à consulta, mas são relevantes no contexto geral do GC.

O diferencial principal entre o método de busca por palavras-chave e o método híbrido para localizar informações no GC reside na técnica de pesquisa utilizada: enquanto o primeiro se apoia exclusivamente em termos-chave, o segundo adota uma estratégia que integra tanto *embeddings* quanto palavras-chave.

KnowledgeGraphQueryEngine

O *KnowledgeGraphQueryEngine* é um motor de consulta avançado que permite consultar um GC usando linguagem natural. Aqui, um *LLM* gera consultas Cypher, que são então executadas no GC.

```
query_engine = KnowledgeGraphQueryEngine(
    storage_context=storage_context,
    llm=llm,
    verbose=True,
)
```

F(5)

O F(5) apresenta a inicialização de um objeto *KnowledgeGraphQueryEngine*, onde o parâmetro *storage_context* provisiona o armazenamento tanto das entidades extraídas pelo *LLM*, quanto dos índices.

KnowledgeGraphRAGRetriever

O *KnowledgeGraphRAGRetriever* é um *recuperador* que executa consultas RAG em um GC. Portanto, dada uma pergunta ou tarefa como entrada e realiza as seguintes etapas:

1. Busca entidades relacionadas no GC usando extração de palavras-chave ou *embeddings*.

2. Obtenção de um subgrafo dessas entidades do GC.
3. Construção de um contexto baseado no subgrafo da etapa 2.

```
graph_rag_retriever = KnowledgeGraphRAGRetriever(  
    storage_context=storage_context,  
    llm=llm,  
    verbose=True,  
)  
  
kg_rag_query_engine = RetrieverQueryEngine.from_args(  
    graph_rag_retriever, service_context=service_context  
)
```

F(6)

Esta estratégia busca entidades do GC com similaridade vetorial, buscando pedaços de texto relacionados e podendo também estudar os relacionamentos. Esta é a forma padrão, sem parâmetros adicionais, de construir os motores de consulta para um *KnowledgeGraphIndex*.

Question and Answer Chain

Com o GC montado, resta apenas inicializar a Cadeia de Perguntas e Respostas, em inglês, *Question and Answer Chain (QAChain)*, o qual é o componente da *Langchain* e está representado pelo Assistente, na Figura 1.1.

```
cypher_chain = GraphCypherQAChain.from_llm(  
    graph=graph,  
    cypher_llm=cypher_model,  
    qa_llm=qa_model,  
    validate_cypher=True,  
    return_intermediate_steps=True,  
    verbose=True,  
)
```

F(7)

O Fragmento 7 apresenta a inicialização da *chain* provida pela *LangChain*, onde:

- `graph`: Base de dados de onde as entidades e relações serão consultadas; GC.
- `cypher_llm`: modelo responsável por traduzir os a pergunta em linguagem natural para uma consulta na linguagem Cypher.
- `qa_llm`: modelo que interpreta as respostas da consulta e gera uma resposta clara, concisa e informativa.
- `validate_cypher=True`; validar as consultas traduzidas antes de executá-las.
- `verbose=True`; mostra a verbosidade do processo de obtenção da resposta.

Todos os componentes apresentados nesta seção são equivalentes ao Assistente da Figura 1.1. No que toca às estratégias da LlamaIndex, uma análise compreensiva foi inspirada em ().

2.1.11 Validação

Grafo de Conhecimento

Para a avaliação da qualidade da estrutura do GG, cogitou-se utilizar as métricas apresentadas na seção 4.2 de (Sumin et al.). Contudo, devido a estrutura topologia simplificada, i.e., nós e relações sem propriedades e apenas uma classe de nós, a empregabilidade destas métricas não agregaria nas análises aqui feitas.

Assim, o que se fez foi verificar se, algumas das tríades geradas pelo LLMs, tomadas aleatoriamente, estavam corretamente representadas no Grafo de conhecimento.

RAG

No âmbito da *RAG*, a validação assume um papel crucial, representando a fase na qual a intervenção humana é essencial. Ela transcende a mera confirmação de resultados, evoluindo para uma interação mais profunda, na qual a máquina é questionada, justificada e desafiada com contra-argumentos.

Nesse processo, busca-se não apenas validar, mas também compreender as nuances subjacentes.

Questionário

1 Pergunta sobre Relações Diretas

Qual é a relação entre a severidade, recuperação e infecção à COVID 19, insuficiência e deficiência de Vitamina D e o uso de protetor solar?

2 Pergunta sobre Interação com Protetor Solar

Trace a relação entre a severidade, recuperação e infecção à COVID 19, insuficiência e deficiência de Vitamina D e o uso de protetor solar trazendo referências bibliográficas conhecidas que suportem a resposta.

3 Fundamentação do *LLM*

Quem é Silvio Santos?

O questionário acima foi pensado de forma a explorar o desempenho da abordagem em articular a argumentação com base nas relações e conceitos do GC. Além de testar a argumentação. Em seguida, verifica-se se o modelo consegue referenciar ou pelo menos apontar quais fatos foram utilizados na resposta. Por fim, verifica-se se o modelo consegue pautar a racionalização nos exclusivamente aos fatos apresentados na base de dados.

O questionário será submetido a todos os 96 GCs gerados. A validação não se restringe à simples aceitação ou rejeição de uma resposta gerada pela máquina; ao contrário, é uma exploração ativa. Envolve a exploração de evidências, a confirmação por meio de fontes alternativas e a busca por lacunas ou contradições

nos resultados obtidos. Este diálogo entre máquina e usuário não apenas verifica a precisão, mas também enriquece o entendimento, proporcionando um contexto mais amplo.

Além disso, a validação é um estágio dinâmico que se entrelaça com o processo de formulação de contra perguntas. O questionamento crítico durante a validação não apenas influencia a aceitação ou rejeição de uma resposta, mas também serve como um catalisador para novas descobertas e aprimoramentos no modelo de geração.

Dentro desse contexto, a validação transcende a simples confirmação; ela é uma jornada de exploração, questionamento e refinamento. É através desse processo iterativo que se abre espaço para discussões enriquecedoras, promovendo uma colaboração mais eficaz entre a capacidade de geração da máquina e a perspicácia humana.

2.1.12 Avaliação

A avaliação das resposta foi feita de duas formas, uma utilizando a métrica chamada *FaithfulnessEvaluator*, da módulo `llama_index.core.evaluation` da própria *LlamaIndex* e outra analisando todos os metadados e respostas geradas.

2.2.1 Recursos Computacionais

A pesquisa usou integralmente uma máquina com as configurações indicadas na Tabela 2.3.

Tabela 2.3: Configuração da Máquina utilizada

Propriedade	Nome
GPU	NVIDIA GeForce RTX 2070 / 8 GiB
Memória RAM	64 GB

Armazenamento em Disco	224GB
Processadores	12

A Tabela 2.3 apresenta atributos essenciais que descrevem as características da máquina utilizada durante a realização do projeto.

O sistema operacional instalado na máquina é o Ubuntu, na versão 20.4. O sistema operacional desempenha um papel fundamental ao gerenciar os recursos do computador e facilitar a execução de programas.

No que se refere à GPU, a máquina está equipada com uma NVIDIA GeForce RTX 2070, oferecendo 8 gigabytes (GiB) de memória dedicada para processamento gráfico.

A máquina possui uma capacidade de 64 *gigabytes* de memória RAM. Essa memória é utilizada para armazenar temporariamente dados em uso ativo, contribuindo para o desempenho eficiente do sistema.

Quanto ao armazenamento em disco, a máquina conta com 224 *gigabytes* de espaço. Essa capacidade de armazenamento é destinada a guardar permanentemente dados, incluindo o sistema operacional, aplicativos e arquivos do usuário.

A máquina é equipada com 12 processadores. Esses componentes são responsáveis por executar operações computacionais, desempenhando um papel crucial no poder de processamento geral do sistema. Essas configurações fornecem uma visão abrangente das características técnicas da máquina utilizada no contexto do trabalho.

2.2.2 Recursos Digitais

O levantamento fez uso de recursos digitais que estão detalhadamente elencados na Tabela 2.4, ela oferece uma visão abrangente dos softwares

empregados durante a realização do projeto, destacando seus tipos e objetivos específicos.

Tabela 2.4: Softwares utilizados

Recurso	Versão	Tipo	Objetivo
Ubuntu	Ubuntu 23.0	Sistema Operacional	
<u>LangChain</u>	0.1.13	código fonte	Prova de Conceito
<u>LlamaIndex</u>	0.10.23	biblioteca	Framework de desenvolvimento de assistentes inteligentes.
<u>OpenAI</u>		serviço	LLMs para processar texto
<u>Google</u>		serviço	LLMs para processar texto
<u>Spark-NLP</u>	5.3.3	código fonte	Extração de entidades e suas relações
<u>User Interface</u>		código fonte	Interface gráfica para a Chat
Python	3.12.2	ferramenta	linguagem de programação
Conda	23.11.0	ferramenta	Virtualizador
Pip	24.0	ferramenta	Gerenciador de pacotes
Neo4j	5.18.0	ferramenta	Base de dados em Grafos
Neo4j Cloud		serviço	Base de dados em Grafos
Docker	24.0.5	ferramenta	Containerização
Jupyter Notebook	7.1.2	ferramenta	desenvolvimento

"LangChain" refere-se ao *Jupyter notebook* que foi utilizado como parte integrante da Prova de Conceito da abordagem RAG, evidenciando seu potencial e utilidade prática. Serviu como o componente fundamental no entendimento do funcionamento, efeitos indesejados dos processos subjacentes e resultados esperados, bem como um guia prático da experiência do usuário.

"LlamaIndex" é uma ferramenta avançada de indexação e busca, projetada para otimizar a recuperação de informações em grandes conjuntos de dados. Utilizando algoritmos modernos de inteligência artificial, o LlamaIndex facilita o acesso rápido e eficiente a dados específicos, tornando-se um componente essencial em sistemas de busca semântica e análises complexas de dados.

"OpenAI", é a empresa que tem liderado o avanço em inteligência artificial com suas versões do modelo de linguagem GPT, incluindo o "gpt-4.0" e o "gpt-3.5-turbo". O gpt-4.0, a última geração, apresenta melhorias significativas em compreensão e geração de texto, oferecendo uma capacidade notável de produzir conteúdo coerente e contextualmente relevante em diversas aplicações, desde chatbots até sistemas avançados de geração de texto. O "gpt-3.5-turbo", por sua vez, é uma versão otimizada que oferece um balanço entre desempenho e custo operacional, permitindo a implementação eficiente em uma ampla gama de serviços que necessitam de processamento de linguagem natural em tempo real.

"Google" lançou o "gemini-1.0-pro", uma ferramenta de inteligência artificial focada em aprimorar as capacidades de análise e processamento de dados em escala. Integrando-se perfeitamente com a plataforma de computação em nuvem da Google, o gemini-1.0-pro utiliza algoritmos avançados de aprendizado de máquina para analisar, interpretar e prever tendências a partir de grandes volumes de dados. Esse modelo é particularmente eficaz em ambientes corporativos onde a precisão e a rapidez na tomada de decisões baseadas em dados são críticas.

O "*Spark-NLP*" é um *Jupyter notebook* que foi utilizado como complemento à extração de entidades e relações. Este recurso faz uso de pipelines proprietários de

extração de entidades e relações, providenciados pela John Snow Labs através de uma licença acadêmica.

A "User Interface" refere-se ao código fonte de uma interface gráfica desenvolvida para a interação do usuário com o assistente, facilitando a experiência do usuário.

"Python" é a linguagem de programação utilizada para o desenvolvimento e execução de código no projeto.

"Conda" é uma ferramenta de virtualização utilizada para criar ambientes isolados e gerenciar dependências, garantindo a consistência do ambiente necessário para a aplicação funcionar.

"Pip" é um gerenciador de pacotes, simplificando o processo de instalação e atualização de bibliotecas e módulos necessários para a aplicação.

"Neo4j" é uma base de dados em Grafos, desempenhando um papel fundamental na representação e manipulação de dados de forma estruturada.

"Neo4j Cloud" é a mesma base de dados em Grafos Neo4j, porém utilizado via API, não necessitando nenhuma instalação local.

"Docker" *containeriza* a solução, aumentando a portabilidade da solução.

"Jupyter Notebook" é uma ferramenta que permite criar documentos com código, visualizações e texto. É muito usada por cientistas de dados e pesquisadores para analisar dados e compartilhar resultados.

"Web of Science" é uma plataforma que oferece acesso a artigos científicos, livros e conferências em diversas áreas. Ajuda pesquisadores a encontrar e acessar pesquisas relevantes facilmente.

Esses recursos digitais, devidamente categorizados e explicados, constituem a infraestrutura tecnológica adotada para a execução eficiente do trabalho, evidenciando a diversidade de ferramentas e linguagens empregadas no contexto do projeto.

Tabela 2.5: Bibliotecas utilizadas no processamento de PDF.

Nome	Características	Documentação	Versão
PyPDF2	Permite unir e dividir documentos PDF. Extrai texto, imagens e metadados de PDFs. Manipulação da ordem das páginas em documentos.	https://pypdf2.readthedocs.io/en/3.0.0/	3.0.0
pdfminer.six	Extrai texto, imagens e metadados de PDFs. Converte PDFs para HTML e XML. Ideal para análises detalhadas de conteúdo PDF.	https://pdfminersix.readthedocs.io/en/latest	20231228
tika	Extrai metadados e texto de diversos formatos de arquivo. Suporta uma ampla variedade de tipos de documentos. Facilita a análise de conteúdo digital em grande escala.	https://tika.apache.org/	3.0.0-BETA
PyMuPDF	Visualização, extração e edição de conteúdo de documentos PDF. Manipulação de imagens, textos e metadados dentro de PDFs. Funcionalidades avançadas para trabalhar com documentos PDF.	https://pymupdf.readthedocs.io/en/latest/	1.24.0

pdfplumber	<p>Extraí texto e tabelas de documentos PDF.</p> <p>Análise da organização espacial do texto em páginas.</p> <p>Especialmente útil para entender a disposição do texto.</p>	https://pypi.org/project/pdfplumber/	0.11.0
tabula	<p>Converte tabelas de PDFs para CSV ou Excel.</p> <p>Ideal para extrair dados tabulares de documentos.</p> <p>Facilita a análise de dados tabulares contidos em PDFs.</p>	https://readthedocs.org/projects/tabula-py/downloads/pdf/latest/	2.9.0
stavrakis	<p>Focado em extrair texto de arquivos PDF.</p> <p>Oferece ferramentas úteis para manipulação de texto.</p> <p>Facilita o acesso ao conteúdo textual de documentos PDF.</p>	https://github.com/g-stavakis/PDF_Text_Extraction	f61e155
pdf2text	<p>Converte documentos PDF em texto puro.</p> <p>Simplifica a extração de texto para análise.</p> <p>Útil para processamento de linguagem natural ou análise de dados.</p>	https://pypi.org/project/pdf2text/	1.0.0

A Tabela 2.5 apresenta os recursos utilizados especificamente na extração do conteúdo de PDFs, suas características principais e a versão que é empregada no processamento dos dados.

2.2.3 Aquisição de Dados

As publicações fornecidas pelo *Web of Science* foram baixadas manualmente de seus respectivos canais, quando disponíveis. Esse procedimento resultou em um total de 185 arquivos em formato PDF, dos quais 2 são compartilhados entre as buscas 2 e 4.

Cabe destacar que alguns artigos não estão acessíveis de forma gratuita ou através do acesso via rede CAFe (Acesso CAFe). Estes específicos registros são detalhadamente apresentados nas tabelas subsequentes.

Tabela 2.6: LLMs disponíveis para geração

Nome	Max Tokens	Tarefa
gpt-3.5 turbo	16385	extração de relações entre entidades, geração e sumarização textual
gemini-1.0-pro	32000	extração de relações entre entidades, geração e sumarização textual
mistral	32000	extração de relações entre entidades, geração e sumarização textual

2.2.4 Abordagem RAG

A Tabela 2.4 mostra os três trabalhos sobre os quais este trabalho foi construído. A prova de conceito, foi rodada e testada, verificando-se seu potencial e constatando suas limitações. O principal inconveniente foi financeiro; o uso do *LLM* utilizado, i.e., *gpt-3.5-turbo* que tem um custo de US \$0,0015/1000 *tokens* de entrada e US \$0,002/1,000 *tokens* de saída. Apesar de seu potencial (WILLIAMS et al., 2023) para este trabalho, o custo é inviabilizante. Um *LLM* alternativo é o *gemini-1.0-pro*, da família Gemini, oferecido pela Google a um custo de US\$ 0,00025 a cada 1000 caracteres de entrada e US\$ 0,0005 a cada 1000 caracteres de saída, (GOOGLE. Gemini API for developers in Cloud. *Google Blog*).

Diante dessa limitação, considerou-se a alternativa do modelo *mistral*, um *LLM* de geração de texto gratuito, como mostrado na Tabela 2.6. Embora tenha apresentado algumas inconsistências na formatação dos resultados, sua validade conceitual foi comprovada. No entanto, observou-se que a velocidade de execução e as relações geradas, ainda que semanticamente coerentes e valiosas, poderiam ser mais específicas.

Afortunadamente, há alternativas para o *gpt-3.5-turbo*. O *mistral* será utilizado neste trabalho, posto que este é gratuito e performa as tarefas que necessitamos.

O *gemini-1.0-pro*, está disponível gratuitamente para desenvolvedores através de uma API limitada a 60 requisições por minuto.

mistral é um modelo de linguagem grande (*LLM*) que foi treinado em um conjunto de dados de texto e código. O modelo *mistral* possui pequeno porte (7,3 bilhões de parâmetros) e performa bem nas seguintes tarefas:

1. geração de texto
2. tradução de idiomas
3. geração de diferentes tipos de conteúdo criativo
4. geração de código fonte em múltiplas linguagens
5. perguntas e respostas

Neste trabalho, o *LLM* é utilizado tanto como extrator de relações, como também o tradutor e sumarizador da Figura 1.1.

Nesta pesquisa utiliza-se o *GC* como a estrutura de dados que representa relações entre entidades. O *GC* per se é somente um modelo abstrato, os dados de fato ficam armazenados no *Neo4j*, que é otimizado para as consultas em grafos,

Para uma extração mais específica ao domínio biomédico, utilizou-se os modelos de extração de entidades e relações de textos da empresa *John Snow Labs*, que providencia uma enorme gama de tarefas. As tarefas que utilizamos neste trabalho são apresentadas na Tabela 2.2, que mostra os *pipelines* e modelos disponíveis para a tarefa específica de extração de entidades e relações, uma

escolha alinhada com a proposta deste trabalho, que visa extrair conhecimento semântico de artigos científicos, incluindo dados clínicos e de posologias.

2.1.2 Trabalhando com os Dados

A construção do GC depende intrinsecamente de sentenças/frases coesas, íntegras, corretas e completas, de forma que:

1. sejam legíveis.
2. possuam, dentro de si, o máximo de significado e semântica o possível.
3. formem, uma a uma, um encadeamento lógico e ordenado de ideias, conceitos e argumentação.

A presença das características acima descritas é condicional para a criação do GC. Outra condição para que a máquina processe o texto mais profundamente é necessário que o texto por ela ingerido não possua hierarquia, isto é; que o texto seja fluído.

Convém ilustrar a diferenciação entre extração e conversão no contexto de arquivos PDF.

Extração: A Essência do Conteúdo

A extração de texto de um PDF se concentra na captura da essência textual do documento, ignorando a formatação original. O conteúdo é extraído e retornado em sua forma bruta, pronto para ser utilizado em diversas aplicações.

Conversão: Reimaginando o PDF como Texto

A conversão de um PDF em texto vai além da extração, recriando o documento original com a máxima fidelidade possível. De forma a preservar não apenas as palavras, mas também a formatação, incluindo fontes, estilos, margens e colunas. O resultado final é um texto que se assemelha visualmente ao PDF original, apesar de ser teoricamente possível converter um PDF em texto, este trabalho investiga somente a conversão se dá de PDF para DOCX.

Para a obtenção de txts de PDFs, foram tentadas 3 abordagens:

- Extração utilizando ferramentas disponíveis no ecossistema UNIX.
- Conversão para docx em nuvem, seguida de extração.
- Extração feita em nuvem.

Como LLMs processam texto pleno, o procedimento utilizado neste trabalho é a extração.

Capítulo 3 - Resultados e Discussão

3.1 Aquisição dos Dados

Para a realização deste trabalho utilizou-se o portal de busca de publicações científicas Web of Science, da *Clarivate* para realizar as consultas. Após realizar as consultas desejadas, os dados gratuitamente disponíveis foram baixados em formato PDF, os metadados relacionados à busca também foram obtidos em seus formatos correspondentes; imagens em .png e tabelas em .xlsx.

Tabela 3.1: Tópicos utilizados no levantamento bibliográfico

Busca	Tópicos						Publicações	Baixados	
1	vitamin d	insufficiency	deficiency		sunscreen	link	64	45	
2	vitamin d	insufficiency	deficiency		covid 19	link	98	93	
3	sunscreen				covid 19	link	18	15	
4	vitamin d		covid 19	severity	recovery	infection	link	34	34
5	sunscreen		covid 19	severity	recovery	infection	0	0	

A Tabela 3.1 mostra em perspectiva, todas as buscas realizadas. Essas buscas abordam tópicos relacionados à interação entre a vitamina D, a insuficiência/deficiência de vitamina D e a COVID-19. A Tabela 3.1 apresenta quatro colunas distintas. A coluna "Busca" identifica cada consulta realizada, enquanto a coluna "Tópicos" expõe os elementos constituintes de cada busca. Na representação visual, a cor azul está associada à causa, e a cor vermelha está relacionada ao efeito.

A coluna "Publicações" indica a quantidade de registros resultantes de cada busca, enquanto a coluna "Baixados" revela quantos desses registros foram

acessados para análise mais aprofundada. Essa abordagem sistemática tem como objetivo explorar e analisar criticamente a literatura científica disponível sobre a interseção entre vitamina D, deficiência dessa vitamina e a COVID-19.

Os resultados do levantamento mostraram que havia um forte interesse na potencial conexão entre a vitamina D insuficiente ou deficiente e a COVID-19. A busca pela combinação de palavras-chave "vitamina D" + "insuficiência" + "deficiência" foi a mais popular, com 64 publicações encontradas e 45 downloads.

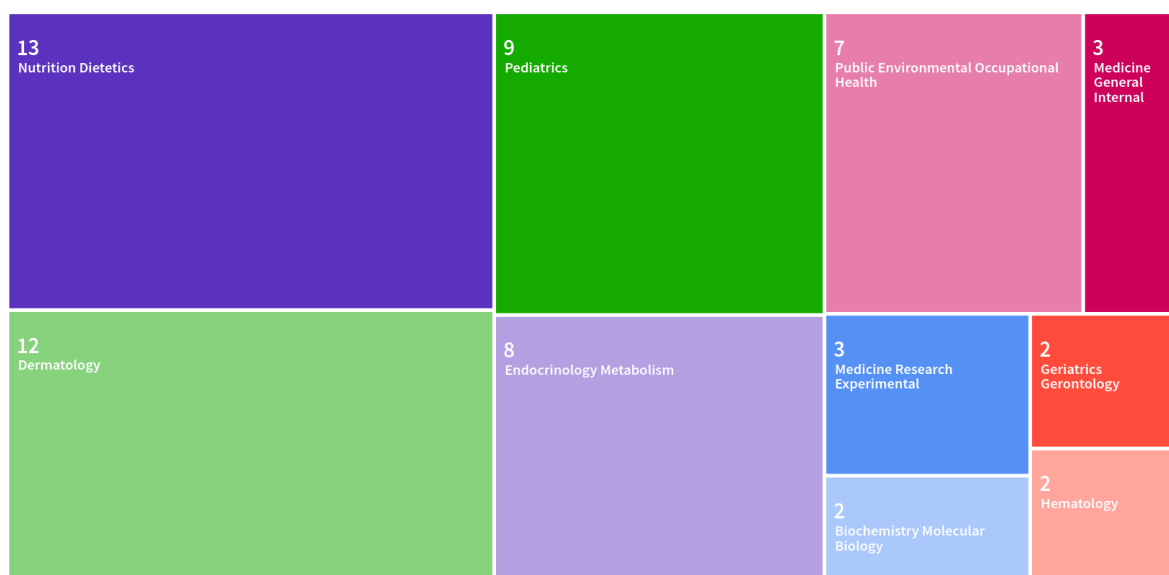


Figura 3.1: Publicações em números absolutos e porcentagem por disciplina para a Busca 1. As áreas não são estritamente proporcionais aos valores de cada entrada . Os resultados foram obtidos pela consulta no Web of Science.

A Figura 3.1 mostra as 10 disciplinas que mais contribuíram em publicações selecionadas. Cada disciplina é representada por uma, bem como o número de publicações e o nome da disciplina. Sendo nutrição, dermatologia, pediatria e endocrinologia as áreas com maior contribuição.

Tabela 3.2: Publicações em números absolutos e porcentagem por disciplina para

a Busca 1. Os resultados foram obtidos pela consulta no Web of Science.

Web of Science Categories	Record Count	% of 64
Nutrition Dietetics	13	20.31
Dermatology	12	18.75
Pediatrics	9	14.06
Endocrinology Metabolism	8	12.50
Public Environmental Occupational Health	7	10.94
Medicine General Internal	3	4.69
Medicine Research Experimental	3	4.69
Biochemistry Molecular Biology	2	3.13
Geriatrics Gerontology	2	3.13
Hematology	2	3.13
Multidisciplinary Sciences	2	3.13
Oncology	2	3.13
Pharmacology Pharmacy	2	3.13
Rheumatology	2	3.13
Biology	1	1.56
Biophysics	1	1.56
Cardiac Cardiovascular Systems	1	1.56
Education Educational Research	1	1.56
Education Scientific Disciplines	1	1.56

Environmental Sciences	1	1.56
Genetics Heredity	1	1.56
Immunology	1	1.56
Medical Laboratory Technology	1	1.56
Nuclear Science Technology	1	1.56
Nursing	1	1.56

A Tabela 3.2 complementa a Figura 3.1, mostrando o resto das contribuições, em contagem e porcentagem, por cada disciplina.

A Busca 2 pela combinação de palavras-chave "VITAMINA D" + "INSUFICIÊNCIA" + "DEFICIÊNCIA" + "COVID-19" também foi popular, com 97 publicações encontradas e 87 downloads.

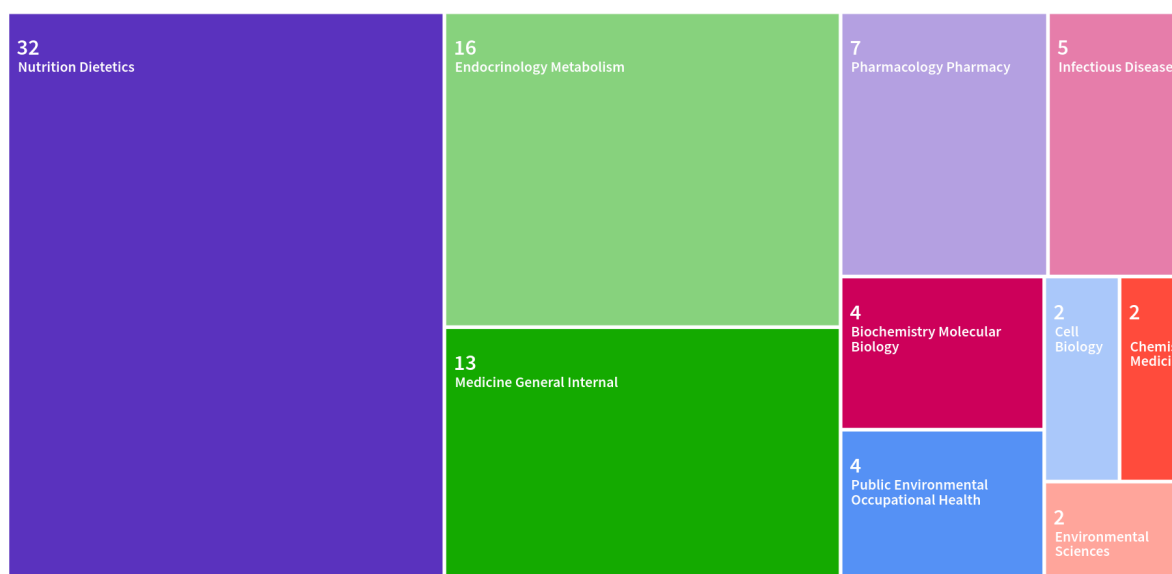


Figura 3.2: Publicações em números absolutos e porcentagem por disciplina para a Busca 2. As áreas não são estritamente proporcionais aos valores de cada entrada. Os resultados foram obtidos pela consulta no Web of Science.

A Figura 3.2 mostra as 10 disciplinas que mais contribuíram em publicações selecionadas. Cada disciplina é representada por uma, bem como o número de publicações e o nome da disciplina. Sendo nutrição, endocrinologia, medicina geral e farmacologia as áreas com maior contribuição.

Tabela 3.3: Publicações em números absolutos e porcentagem por disciplina para a Busca 2. Os resultados foram obtidos pela consulta no Web of Science.

Web of Science Categories	Record Count	% of 97
Nutrition Dietetics	31	31.96
Endocrinology Metabolism	16	16.49
Medicine General Internal	13	13.40
Pharmacology Pharmacy	7	7.22
Infectious Diseases	5	5.16
Biochemistry Molecular Biology	4	4.12
Public Environmental Occupational Health	4	4.12
Cell Biology	2	02.06
Chemistry Medicinal	2	02.06
Environmental Sciences	2	02.06
Food Science Technology	2	02.06
Immunology	2	02.06
Integrative Complementary Medicine	2	02.06
Medicine Research Experimental	2	02.06

Multidisciplinary Sciences	2	02.06
Respiratory System	2	02.06
Rheumatology	2	02.06
Cardiac Cardiovascular Systems	1	01.03
Chemistry Multidisciplinary	1	01.03
Clinical Neurology	1	01.03
Genetics Heredity	1	01.03
Geriatrics Gerontology	1	01.03
Hematology	1	01.03
Mathematical Computational Biology	1	01.03
Medical Laboratory Technology	1	01.03

A Tabela 3.3 complementa a Figura 3.2, mostrando o resto das contribuições, em contagem e porcentagem, por cada disciplina.

A pesquisa relacionada ao protetor solar e à COVID-19 também mostrou algum interesse, mas com foco em subtópicos específicos. Por exemplo, a busca pela combinação de palavras-chave "protetor solar" + "COVID-19" foi a segunda mais popular, com 18 publicações encontradas e 13 downloads.

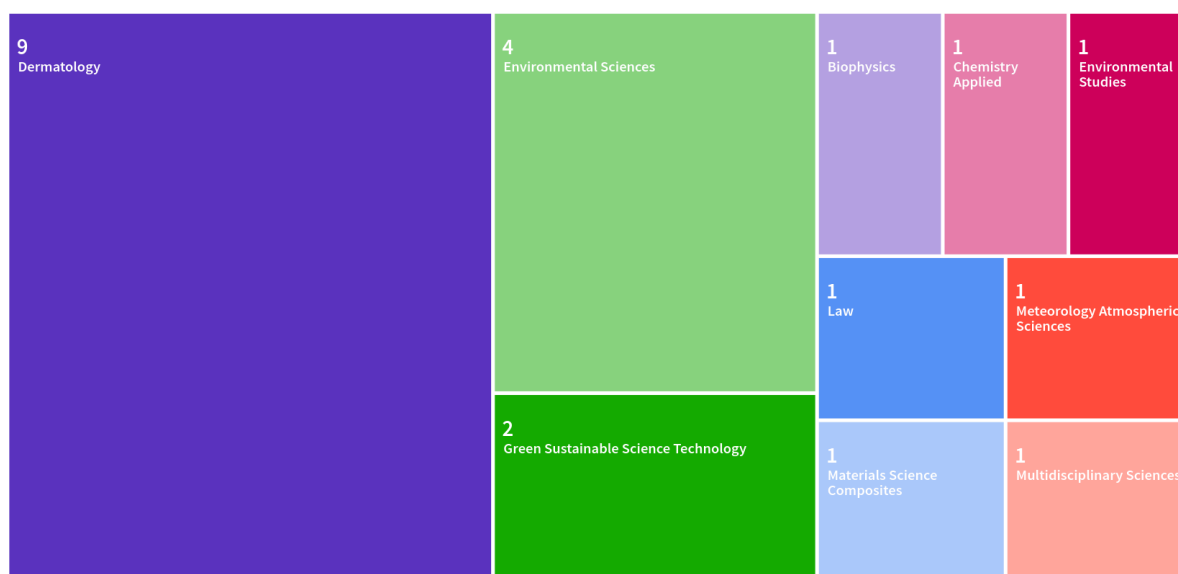


Figura 3.3: Publicações em números absolutos e porcentagem por disciplina para a Busca 3. As áreas não são estritamente proporcionais aos valores de cada entrada . Os resultados foram obtidos pela consulta no Web of Science.

A Figura 3.3 mostra as 10 disciplinas que mais contribuíram em publicações selecionadas. Cada disciplina é representada por uma, bem como o número de publicações e o nome da disciplina. Sendo dermatologia, ciências ambientais e sustentáveis as áreas com maior contribuição.

Tabela 3.4: Publicações em números absolutos e porcentagem por disciplina para a Busca 3. Os resultados foram obtidos pela consulta no Web of Science.

Web of Science Categories	Record Count	% of 18
Dermatology	9	50.00
Environmental Sciences	4	22.22
Green Sustainable Science Technology	2	11.11
Biophysics	1	5.56

Chemistry Applied	1	5.56
Environmental Studies	1	5.56
Law	1	5.56
Materials Science Composites	1	5.56
Meteorology Atmospheric Sciences	1	5.56
Multidisciplinary Sciences	1	5.56
Nutrition Dietetics	1	5.56
Pharmacology Pharmacy	1	5.56
Physiology	1	5.56
Polymer Science	1	5.56

A Tabela 3.4 complementa a Figura 3.3, mostrando o resto das contribuições, em contagem e porcentagem, por cada disciplina.

A busca pela combinação de palavras-chave "VITAMINA D" + "COVID-19" + "GRAVIDADE" + "RECUPERAÇÃO" + "INFECÇÃO" foi a terceira mais popular, com 34 publicações encontradas e 34 downloads.



Figura 3.4: Publicações em números absolutos e porcentagem por disciplina para a Busca 4. As áreas não são estritamente proporcionais aos valores de cada entrada . Os resultados foram obtidos pela consulta no Web of Science.

A Figura 3.4 mostra as 10 disciplinas que mais contribuíram em publicações selecionadas. Cada disciplina é representada por uma, bem como o número de publicações e o nome da disciplina. Sendo nutrição, endocrinologia, medicina geral e farmacologia as áreas com maior contribuição.

Tabela 3.5: Publicações em números absolutos e porcentagem por disciplina para a Busca 4. Os resultados foram obtidos pela consulta no Web of Science.

Web of Science Categories	Record Count	% of 34
Nutrition Dietetics	12	35.29
Endocrinology Metabolism	5	14.71
Medicine General Internal	3	8.82
Biochemistry Molecular Biology	2	5.88

Food Science Technology	2	5.88
Medicine Research Experimental	2	5.88
Cell Biology	1	2.94
Chemistry Medicinal	1	2.94
Chemistry Multidisciplinary	1	2.94
Clinical Neurology	1	2.94
Mathematics	1	2.94
Microbiology	1	2.94
Multidisciplinary Sciences	1	2.94
Nanoscience Nanotechnology	1	2.94
Pediatrics	1	2.94
Pharmacology Pharmacy	1	2.94
Primary Health Care	1	2.94
Rheumatology	1	2.94
Virology	1	2.94

A Tabela 3.5 complementa a Figura 3.4, mostrando o resto das contribuições, em contagem e porcentagem, por cada disciplina.

A Busca 1 empregou os termos "VITAMINA D", "INSUFICIÊNCIA", "DEFICIÊNCIA" e "PROTETOR SOLAR", encontrando 64 publicações, das quais 45 estavam gratuitamente disponíveis.

A Busca 2, substituindo "PROTETOR SOLAR" por "COVID-19", identificou 97+1 publicações, tendo sido 93 acessadas.

A Busca 3, utilizando "PROTETOR SOLAR" e "COVID-19", resultou em 18 publicações, com 15 baixadas.

A Busca 4, com os termos "vitamina D", "COVID-19", "SEVERIDADE", "RECUPERAÇÃO" e "INFECÇÃO", descobriu 34 publicações, todas baixadas.

Por fim, a Busca 5, incorporando "PROTETOR SOLAR", não trouxe nenhum resultado.

Tabela 3.6: Contagem de Publicações por disciplina

Domínio	Publicações	% 213	Baixados	% 185
Nutrição	56	26,29	51	27,56
Endocrinologia e metabolismo	29	13,62	18	9,72
Dermatologia	21	9,86	14	7,56

Na Tabela 3.6, a coluna "Domínio" indica a área do conhecimento.

A coluna "Publicações" indica o número de artigos científicos publicados no domínio.

A coluna "% 213" indica a porcentagem de publicações do domínio em relação ao total de 213 publicações.

A coluna "Baixados" indica o número de downloads de artigos científicos do domínio.

A coluna "% 185" indica a porcentagem de downloads do domínio em relação ao total de 185 downloads.

Com base nos dados apresentados na Tabela 3.6, podemos observar que a área da nutrição é a que apresenta o maior número de publicações (56), representando 26,29% do total de publicações. Em seguida, vem a área da

endocrinologia e metabolismo (29 publicações, representando 13,62%), seguida da área da dermatologia (21 publicações, representando 9,86%).

Em relação aos *downloads*, a área da nutrição também é a que apresenta o maior número (51 downloads, representando 27,56%). Em seguida, vem a área da dermatologia (14 downloads, representando 7,56%), seguida da área da endocrinologia e metabolismo (18 downloads, representando 9,72%).

Os resultados da busca sugerem que existe uma relação entre vitamina D e COVID-19. As Buscas 1 e 2, que utilizaram os termos "vitamina D" e "insuficiência/deficiência", encontraram um número significativo de publicações. A Busca 4, com os termos "vitamina D", "COVID-19", "severidade", "recuperação" e "infecção", também resultou em uma quantidade expressiva de publicações, indicando possível associação entre vitamina D, gravidade, recuperação e infecção por COVID-19.

No entanto, a Busca 5, que utilizou os termos "COVID-19", "severidade", "recuperação", "infecção" e "protetor solar", não resultou em nenhuma publicação. Isso sugere que a proteção solar não está associada à relação entre vitamina D e COVID-19.

No que toca ao intervalo de tempo das publicações utilizadas. Obteve-se o perfil histórico das publicações e citações de cada busca ao longo do tempo.

Abaixo, as figuras, as barras cor-de-rosa refletem a quantidade de publicações para cada ano, enquanto as barras roxas indicam o número correspondente de citações. Essa diferenciação de cores oferece uma visão clara e distinta da produção e impacto das publicações ao longo do período analisado.

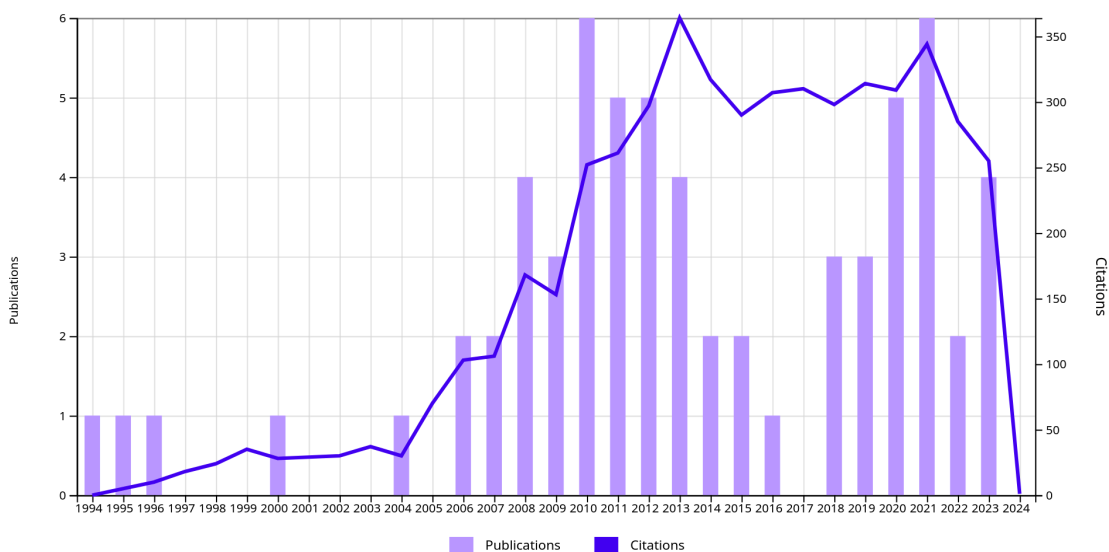


Figura 3.5: Perfil histórico de publicações e citações para a Busca 1. Os resultados foram obtidos pela consulta no Web of Science.

A Figura 3.5, apresenta a contagem de publicações e citações relativas à Busca 1, ano a ano, desde 1994. Sugerindo que a relação em questão já possui uma consistência ao longo do tempo, indicando um entendimento aprofundado e um extenso estudo sobre o tema.

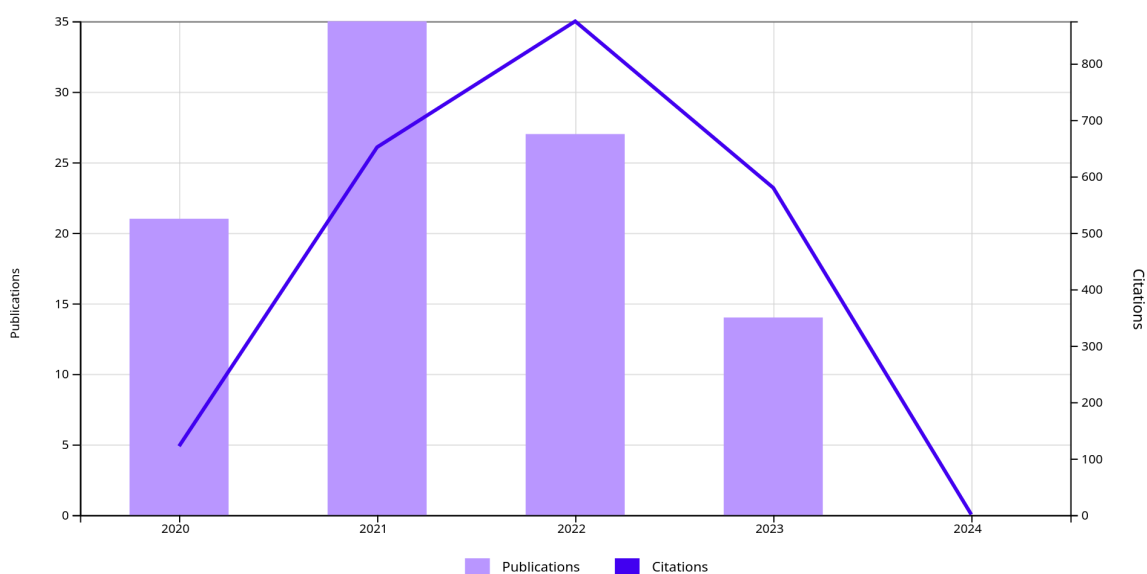


Figura 3.6: Perfil histórico de publicações e citações para a Busca 2. Os resultados foram obtidos pela consulta no Web of Science.

A Figura 3.6 apresenta a produção de publicações e citações relativas à Busca 2, ano a ano, desde 2020. Observa-se um aumento notável de publicações de 2020 para 2021, possivelmente associada à pandemia, seguida de uma diminuição gradual até atingir uma produção inferior a de 2020 em 2023. Este comportamento é similar ao das citações, embora estejam defasadas em 1 ano.

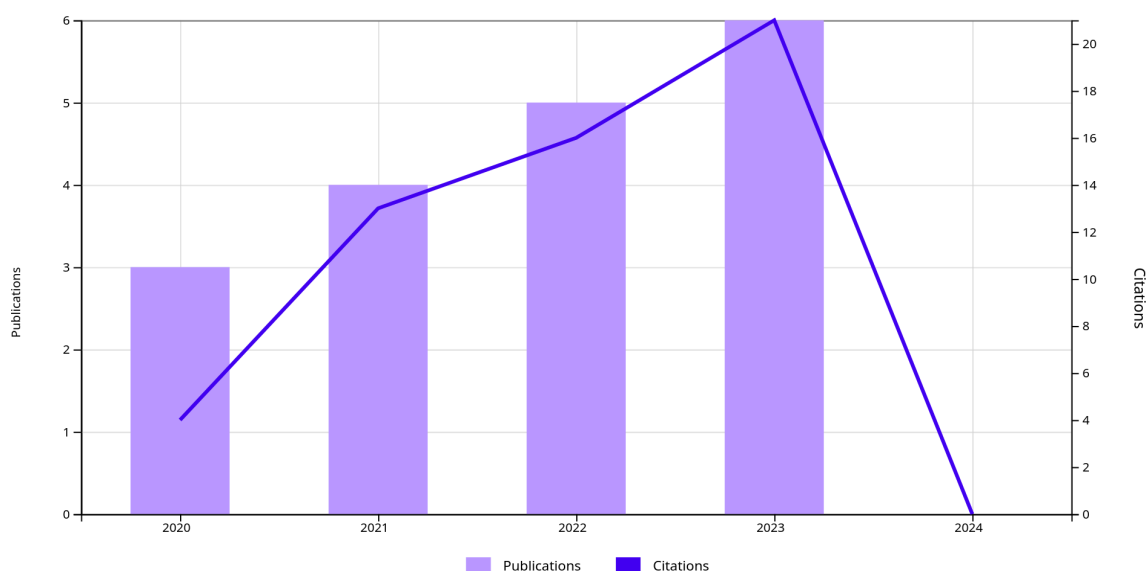


Figura 3.7: Perfil histórico de publicações e citações para a Busca 3. Os resultados foram obtidos pela consulta no Web of Science.

A Figura 3.7 apresenta a produção de publicações e citações relativas à Busca 3 ano a ano, desde 2020. Evidencia-se um aumento constante tanto na produção quanto nas citações ao longo deste período.

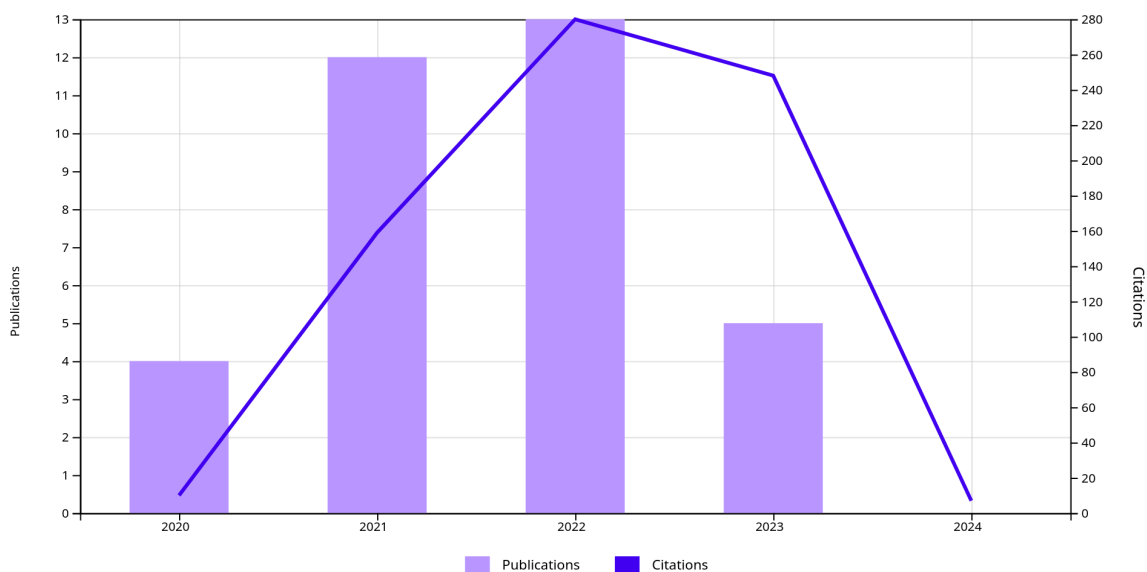


Figura 3.8: Perfil histórico de publicações e citações para a Busca 4. Os resultados foram obtidos pela consulta no Web of Science.

A Figura 3.8 representa a produção de publicações e citações relacionadas à Busca 4, ano a ano, a partir de 2020. Destaca-se um aumento significativo de 2020 para 2021, seguido por uma estagnação até 2022 e uma subsequente diminuição, retornando a valores anteriores à pandemia.

3.2 Análise de Dados

A Tabela 3.6 sugere que a área da nutrição é uma área de pesquisa ativa, com um número significativo de publicações e *downloads*. A área da endocrinologia e metabolismo também é uma área ativa, mas com um número menor de publicações e *downloads*. A área da dermatologia é uma área menos ativa, com um número ainda menor de publicações e *downloads*.

Esses resultados, além de oferecerem uma compreensão aprofundada da produção científica em diferentes áreas, também sugerem possíveis interconexões entre disciplinas aparentemente diversas. A riqueza desses dados ressalta a

dinâmica e a complexidade do cenário científico, incentivando a exploração contínua e o aprimoramento do conhecimento em diversas frentes do saber.

A análise das tabelas Tabela 3.2, Tabela 3.3, Tabela 3.4 e Tabela 3.5 proporciona uma visão detalhada da produção científica correspondente, agrupada pelas disciplinas que as originaram.

Explorando os resultados da primeira busca, destacam-se as áreas de nutrição, dermatologia, pediatria e endocrinologia como as mais prolíficas, contabilizando 13, 12, 9 e 8 publicações, respectivamente. Esse panorama inicial revela o vigor e o engajamento dessas disciplinas na pesquisa científica.

A segunda busca aprofunda nossa compreensão, revelando um equilíbrio notável entre diversas áreas. Nutrição, dermatologia, endocrinologia, medicina geral e farmacologia emergem como protagonistas, apresentando 31, 16, 13, 9 e 7 publicações, respectivamente. Essa diversidade de focos indica uma abordagem abrangente e multifacetada na produção científica.

A terceira busca, por sua vez, nos conduz a um cenário mais específico. Dermatologia, ciências do ambiente e energias sustentáveis se destacam, embora com um número mais reduzido de publicações, registrando 9, 4 e 2, respectivamente. Essa concentração em áreas específicas pode refletir a natureza especializada e aprofundada das pesquisas nesses campos.

A quarta busca revela um panorama intermediário, com áreas como nutrição, endocrinologia, medicina geral e bioquímica molecular contribuindo de maneira equilibrada para a produção científica, apresentando 12, 5, 3 e 2 publicações, respectivamente. Esse equilíbrio sugere uma interseção de interesses e colaborações entre essas disciplinas, promovendo um diálogo interdisciplinar na pesquisa.

Abaixo, apresenta-se as publicações que não fizeram parte do trabalho.

Tabela 3.7: Arquivos não alcançados da Busca 1

Título da Publicação

1	Optimal Vitamin D Status for the Prevention and Treatment of link Osteoporosis
2	Environmental factors that influence the cutaneous production of link vitamin D
3	Vitamin D Levels, Dietary Intake, and Photoprotective Behaviors link Among Patients With Skin Cancer
4	Vitamin D and the Skin: An Update for Dermatologists link
5	Increased Body Mass Index (BMI) and Sunscreen Use Are Associated with Inadequate Vitamin D Status in Greek Adults in Winter link
7	Serum 25-hydroxyvitamin D in erythropoietic protoporphyria link
8	The Influence of Sunlight Exposure on Serum Vitamin D Concentration link and Bone Turnover; a controlled clinical trial
9	The resurgence of the importance of vitamin D in bone health link
10	25-Hydroxyvitamin D serum levels in rheumatic female patients in link southeast Spain: The paradigm of daily optimal sunshine levels and inadequate vitamin D status
11	Does chronic sunscreen use reduce vitamin D production to link insufficient levels?
13	Prevalence and predictors of vitamin D deficiency in non-supplemented women with systemic lupus erythematosus in the link Mediterranean region: a cohort study
14	Prevalence and predictors of vitamin D insufficiency in women of reproductive age living in northern latitude Get access Arrow link
15	Sunlight, UV-radiation, vitamin D and skin cancer: How much sunlight link do we need?

16	Vitamin D status in children and adolescents: A prospective cohort study link
18	McCullum Award Lecture, 1994: Vitamin D—new horizons for the 21st century link
19	Serum Vitamin D and Facial Aging: Is There a Link? link

Tabela 3.8: Arquivos não alcançados da Busca 2.

Título da Publicação	
1	Vitamin D deficiency aggravates COVID-19: systematic review and meta-analysis link
2	Prevalence of vitamin D deficiency among psychiatric inpatients: a systematic review link
3	Covid-19 Susceptibility and Severity Might be Modified by Vitamin D Status: Theoretical and Practical Considerations link

Tabela 3.9: Arquivos não alcançados da Busca 3.

Título da Publicação	
1	A case report of hydroxychloroquine-induced auditory and visual link

	hallucination
3	Should Consumers Transfer Inorganic Sunscreens Into Travel-Size link Containers? Evaluation of Inorganic Sunscreen Emulsion Performance, Quality, and Stability in a 12-Week Study

As tabelas Tabela 3.7, Tabela 3.8 e Tabela 3.9 mostram as publicações não alcançadas nas buscas 1, 2 e 3, respectivamente. As tabelas apresentam o Título das publicações que não estavam disponíveis gratuitamente um link para a sua localização.

3.3 Preparo dos Dados

Os arquivos obtidos são armazenados e organizados de maneira sistemática, onde os dados brutos são agrupados em um diretório denominado "busca":

```
$ du -h -d 1 busca; tree busca | tail -1
164M busca
0 directories, 185 files
```

S(2)

A S(2) mostra a presença de 185 arquivos, após passarem pelo processo de limpeza, são processados e, posteriormente, estruturados em um diretório com uma nomenclatura autoexplicativa:

```
$ du -h -d 1 cleaned_data__512_0; tree cleaned_data__512_0 | tail
-1
85M cleaned_data__512_0/clinical_re_pipeline
85M cleaned_data__512_0/posology_relation_extraction_pipeline
```

S(3)

```
85M    cleaned_data__512_0/clinical_temp_events_re_pipeline
255M   cleaned_data__512_0
3 directories, 63918 files
```

O diretório "cleaned_data__512_0" armazena blocos limpos de 512 caracteres, com 0 sobreposição, distribuídos em três subdiretórios correspondentes aos diferentes modos de extração: *clinical_re_pipeline*, *posology_relation_extraction_pipeline* e *clinical_temp_events_re_pipeline*, no total 638918 arquivos de relações são gerados, como mostrado na S(3).

Similarmente, os dados que não passaram pelo processo de limpeza são processados e organizados em um diretório com uma nomenclatura autoexplicativa:

```
$ du -h -d 1 data__512_0; tree data__512_0 | tail -1
85M    data__512_0/clinical_re_pipeline
85M    data__512_0/posology_relation_extraction_pipeline
85M    data__512_0/clinical_temp_events_re_pipeline
255M   data__512_0
3 directories, 63918 files
```

S(4)

O diretório "data__512_0" armazena blocos não limpos de 512 caracteres, com 0 sobreposição, divididos em três subdiretórios correspondentes aos diferentes modos de extração. Como mostrado na S(4), esses processos culminam na geração de 638918 arquivos entidades e relações para o GC. Com o grafo pronto, torna-se possível empregar a abordagem RAG para formular perguntas, cujas respostas são obtidas mantendo as interações.

3.4 Extração do conteúdo de PDFs

Extração no ecossistema Unix:

Esta abordagem consistiu em, programaticamente, extrair o conteúdo de todos os arquivos em PDF utilizando todas as bibliotecas apresentadas na seção 2.2.2.

Os arquivos foram processados por todas as bibliotecas acima, cujos resultados variam de ferramenta para ferramenta; aspectos gerais dos dados obtidos são apresentados na tabela abaixo.

Tabela 3.10: Benchmark das bibliotecas para processamento de PDFs.

Aspecto	PyPDF2	pdfminer	tika	PyMuPDF	pdfplumber	tabula	stavrakis	pdf2text
Extração de texto	Sim*	Sim	Sim	Sim	Sim	Sim	Sim	Sim
Extração de tabelas	Não	Sim	Sim	Sim	Sim	Não	Não	Não
Análise de layouts	Não	Sim	Limitado	Sim	Limitado	Limitado	Limitado	Limitado
Formatação	Não	Parcial	Limitado	Sim	Limitado	Limitado	Limitado	Limitado
ML	Não	Não	Não	Não	Não	Não	Não	Não
Usabilidade	Fácil	Moderada	Moderada	Moderada	Fácil**	Fácil**	Moderada	Fácil
Velocidade	+ Rápido	+ Lenta	Variável	Bom	Bom**	Bom**	Moderado	Rápido
Acurácia	Baixo	Alto	Variável	Alto	Alto**	Alto**	Moderado	Baixo
Legibilidade	Parcial	Parcial	Parcial	Parcial	Parcial	Parcial	Parcial	Parcial
Licença	Aberto	MIT	Apache 2.0	AGPLv3	MIT	MIT	MIT	MIT
Comunidade/suporte	Grande	Grande	Grande	Grande	Ativa	Grande	Moderada	Pequena

* **Básico** Quando empregado para o processamento de documentos simples.

** **Para tabelas** Quando empregado para o processamento de dados tabulares.

Algumas das métricas apresentadas na Tabela 3.10, são definidas como:

Acurácia: Detecção de palavras, linhas de texto e blocos de texto.

Legibilidade: Ordem lógica do texto extraído

Usabilidade: Facilidade em utilizar e integrar a ferramenta em um sistema

Velocidade: Tempo necessário para extrair arquivos de um PDF

ML: Emprego de funcionalidades melhoradas por Machine Learning

A Tabela 3.10 apresenta um panorama do comportamento geral em cada caso de empregabilidade das bibliotecas disponíveis gratuitamente. Enquanto algumas focam na extração do texto, outras focam na extração de tabelas. De forma geral, pode-se afirmar:

- **PyPDF2**: Funciona para extração básica de texto de PDFs com formatação simples.
- **pdfminer.six**: Biblioteca poderosa para extração de texto com recursos de análise de layout. Também pode lidar com extração básica de tabelas.
- **tika**: Kit de ferramentas do lado do servidor para vários formatos de arquivo, incluindo extração de PDF de texto, metadados e algumas tabelas.
- **PyMuPDF**: Biblioteca versátil para extração avançada de texto, extração de imagem, manipulação de página e potencialmente preenchimento de formulário/extração de anotação; requer configuração.
- **plumber**: Biblioteca fácil de usar, projetada especificamente para extração de tabelas de PDFs com alta precisão.
- **tabula**: Outra biblioteca popular conhecida por sua boa precisão e facilidade de uso na extração de tabelas de PDFs.

Fatores como velocidade, precisão e complexidade de uso são baseadas em outros benchmarks (Pradeep)(STATISTICS CANADA) e são apenas aqui mostrados. Precisão e acurácia, dizem respeito à extração do que está disponível na estrutura do PDF, enquanto a legibilidade diz respeito à capacidade de um humano ler o texto.

A vantagem desta abordagem está na flexibilidade de poder aplicar procedimentos intermediários para melhoria dos dados. No entanto, o gerenciamento de 185 arquivos pode adicionar uma complexidade técnica alta.

A grande desvantagem desta abordagem é a baixa e/ou parcial legibilidade do texto obtido. Frequentemente são encontradas descontinuidades e outras interrupções que acabam por comprometer o significado semântico do texto como um todo.

A exemplo, toma-se o arquivo **3d0e16b69e5759e745a50af3b16b34b0.txt**, entre a primeira e segunda páginas, apresentado na Figura 13.9 Em verde, tem-se o texto grifado em verde mostrando a continuidade do texto, na ordem que um humano lê; aqui considera-se apenas texto. Ao final da primeira página, o trecho “...*Vitamin D negatively modulates the RAS*” é completado por “*system [21], inhibiting renin expression/ generation...*”.

No entanto, as bibliotecas disponíveis, discernem a legibilidade do texto, retornando algo similar a:

How vitamin D insufficiency may negatively impact disease such as COVID-19 is unclear; however, vitamin D negatively modulates the RAS * Corresponding author. E-mail addresses: abutler@rcsi.com (A.E. Butler), amoin@rcsi.com (A.S.M. Moin), Sathyapalan@hymn.ac.uk (T. Sathyapalan), satkin@rcsi.com (S.L. Atkin). <https://doi.org/10.1016/j.jsbmb.2021.105965> Received 16 May 2021; Received in revised form 8 August 2021; Accepted 10 August 2021 0960-0760/©2021ElsevierLtd.Allrightsreserved. A.E. Butler et al. <https://doi.org/10.1016/j.jsbmb.2021.105965> Table 1 Demographic and biochemical characteristics of the PCOS and non-PCOS control women. Data are presented as mean (SD). Age (years) BMI (kg/m [2]) Weight (kg) Waist circumference (cm) Hip circumference (cm) Systolic blood pressure (mmHg) Diastolic blood pressure (mmHg) AMH (pmol/l) Sex Hormone Binding Globulin (SHBG) (nmol/l) Cholesterol (mmol/l) Triglycerides (mmol/l) HDL (mmol/l) LDL (mmol/l) CRP (mmol/l) Testosterone (nmol/l) Androstenedione (nmol/l) FAI Glucose (mmol/l) Insulin (µU/ml) Total vitamin D (nmol/l) CONTROL (n = 68)

28 (1) 27 (1) 74 (2) 81 (2) 101 (1) 115 (1) 74 (1) 22 (2) 73 (10) 4.7 (0.1) 1.0 (0.1) 1.4 (0.1) 2.7 (0.1) 2.1 (0.5) 1.1 (0.1) 8.1 (0.7) 2.1 (0.2) 4.7 (0.1) 7.2 (0.7) 62 (3) PCOS (n = 99) 30 (1) 35 (1) 98 (2) 102 (2) 119 (2) 122 (2) 77 (1) 46 (3) 43 (5) 4.8 (0.1) 1.6 (0.2) 1.2 (0.03) 2.9 (0.1) 4.8 (0.6) 1.5 (0.1) 11.3 (0.7) 5.7 (0.6) 4.9 (0.1) 17.0 (1.9) 43 (3) P value 0.03 <0.0001 <0.0001 <0.0001 <0.0001 0.0008 0.03 <0.0001 0.002 0.2 0.003 <0.0001 0.3 <0.0001 0.001 0.003 <0.0001 0.3 <0.0001 <0.0001 BMI, Body mass index; AMH, anti-mullerian hormone; HDL, high density lipo-protein; LDL, low density lipoprotein, CRP, C-reactive protein; FAI, free androgen index. system [21], inhibiting renin expression/ generation, induces ACE2 and suppresses the ang/Ang II/AT1R axis [22], and it has been shown that plasma ACE2 levels are predictors of COVID-19 disease outcome [23]. We hypothesized that vitamin D status would be related to plasma levels of the RAS-related proteins Angiotensin-converting enzyme 2 (ACE2), renin and angiotensinogen.

S(5)

1. Introduction

Angiotensin-converting enzyme 2 (ACE2) is part of the Renin-Angiotensin System (RAS) system, which is involved in the regulation of blood pressure and body fluid control mechanisms. RAS over-activation is a notable risk factor in renal [1] and cardiovascular disease [2] development. RAS overactivity is present in metabolic syndrome [3], type 2 diabetes (T2D) and obesity [4]. *in vitro* and *in vivo* studies show vasoconstriction and sodium retention in cardiac myocytes and mesangial cells as well as in skeletal muscle myoblasts, evidence that supports the clinical and epidemiological data, and tissue RAS activity likely accompanies and aggravates metabolic disorders [5].

Polycystic ovary syndrome (PCOS) encompasses many characteristics of metabolic syndrome, women with PCOS frequently exhibiting insulin resistance [6] and obesity [7], underscored by the significant

proportion of these women who develop type 2 diabetes [8]. Therefore, PCOS women are similarly prone to develop infections as those with diabetes [9–11].

Vitamin D deficiency is known to be common in women with PCOS, 67–85 % of PCOS women having circulating concentrations of less than 20 ng/ml (50 nmol/l), and low 25-hydroxyvitamin D [25(OH)D] levels correlate with increased insulin resistance, obesity and androgen levels [12–14]. Moreover, a recent systematic review of randomized controlled trials determined that vitamin D supplementation improved metabolic dysfunction in PCOS women [15]. Vitamin D status has also been identified as a potential modulating factor in infection risk, and notably SARS-CoV-2 infection, with vitamin D insufficiency being linked to increased COVID-19 incidence [16,17] and severity risk [16,18–20]. How vitamin D insufficiency may negatively impact disease such as COVID-19 is unclear; however, vitamin D negatively modulates the RAS

* Corresponding author.

E-mail addresses: abutler@rcsi.com (A.E. Butler), amoin@rcsi.com (A.S.M. Moin), Sathyapalan@hums.ac.uk (T. Sathyapalan), satkin@rcsi.com (S.L. Atkin).

<https://doi.org/10.1016/j.jsmb.2021.105965>

Received 16 May 2021; Received in revised form 8 August 2021; Accepted 10 August 2021

Available online 5 October 2021

0960-0760/© 2021 Elsevier Ltd. All rights reserved.

A.E. Butler et al.

Journal of Steroid Biochemistry and Molecular Biology 214 (2021) 105965

Table 1
Demographic and biochemical characteristics of the PCOS and non-PCOS control women. Data are presented as mean (SD).

	CONTROL (n = 68)	PCOS (n = 99)	P value
Age (years)	28 (1)	30 (1)	0.03
BMI (kg/m ²)	27 (1)	35 (1)	<0.0001
Weight (kg)	74 (2)	98 (2)	<0.0001
Waist circumference (cm)	81 (2)	102 (2)	<0.0001
Hip circumference (cm)	101 (1)	119 (2)	<0.0001
Systolic blood pressure (mmHg)	115 (1)	122 (2)	0.0008
Diastolic blood pressure (mmHg)	74 (1)	77 (1)	0.03
AMH (pmol/l)	22 (2)	46 (3)	<0.0001
Sex Hormone Binding Globulin (SHBG) (nmol/l)	73 (10)	43 (5)	0.002
Cholesterol (mmol/l)	4.7 (0.1)	4.8 (0.1)	0.2
Triglycerides (mmol/l)	1.0 (0.1)	1.6 (0.2)	0.003
HDL (mmol/l)	1.4 (0.1)	1.2 (0.03)	<0.0001
LDL (mmol/l)	2.7 (0.1)	2.9 (0.1)	0.3
CRP (mmol/l)	2.1 (0.5)	4.8 (0.6)	<0.0001
Testosterone (nmol/l)	1.1 (0.1)	1.5 (0.1)	0.001
Androstenedione (nmol/l)	8.1 (0.7)	11.3 (0.7)	0.003
FAI	2.1 (0.2)	5.7 (0.6)	<0.0001
Glucose (mmol/l)	4.7 (0.1)	4.9 (0.1)	0.3
Insulin (μU/ml)	7.2 (0.7)	17.0 (1.9)	<0.0001
Total vitamin D (nmol/l)	62 (3)	43 (3)	<0.0001

BMI, Body mass index; AMH, anti-mullerian hormone; HDL, high density lipoprotein; LDL, low density lipoprotein, CRP, C-reactive protein; FAI, free androgen index.

system [21], inhibiting renin expression/ generation, induces ACE2 and suppresses the ang/Ang II/AT1R axis [22], and it has been shown that plasma ACE2 levels are predictors of COVID-19 disease outcome [23].

We hypothesized that vitamin D status would be related to plasma levels of the RAS-related proteins Angiotensin-converting enzyme 2

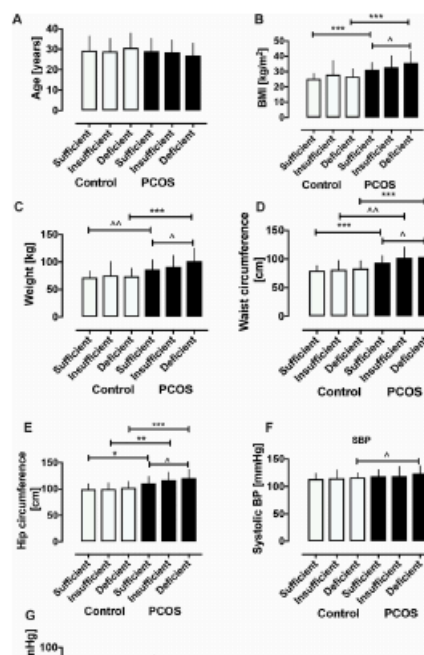


Figura 3.9: Ordem lógica do texto 3d0e16b69e5759e745a50af3b16b34b0.pdf.

A saída S(5), foi obtida através da *PyMUPDF*, não possuindo diferença significativa entre as outras ferramentas analisadas; de forma que este exemplo ilustra significativamente o comportamento das bibliotecas.

Observa-se na S(5), que o conteúdo da tabela está todo contido no texto, porém organizada em colunas, ao invés de em linhas.

Conversão para docx:

Este procedimento foi executado na nuvem, através do conversor de PDF para Word do *PDF2GO*, disponível em <https://www.pdf2go.com/pt/pdf-para-word>, utilizando-se o seguinte passo a passo:

1. Sobe-se os arquivos PDFs.



Figura 3.10: Primeiro passo para a conversão de PDF para DOCX na nuvem .

2. Solicita-se a conversão com a opção de melhoramento do *layout* por Reconhecimento Óptico de Caracteres (OCR em inglês).

Figura 3.11: Segundo passo para conversão de PDF para DOCX na nuvem

Páginas digitalizadas serão imagens.

Converter com OCR Premium
Páginas digitalizadas serão convertidas em texto editável. ⓘ

Método OCR ⓘ
Layout

Idioma fonte do seu arquivo
Para obter os melhores resultados, selecione todos os idiomas que seu arquivo contém.

Inglês x Espanhol x

Melhorar OCR ⓘ

Configurações opcionais

Converter arquivo(s) PDF para:
Microsoft Word (*.docx)

INICIAR →

3. Baixa-se os arquivos processados em (.docx).

PDF2GO Todas ferramentas EDUCATION Menu

Baixar

Exportar como Compartilhar Excluir

✓ Concluído
Obtenha uma assinatura para evitar a fila, aproveitar os recursos PRO e processar seus arquivos com mais rapidez.
[Seja Premium](#)

Current Task

Baixar Arquivos

Figura 3.12: Terceiro passo para conversão de PDF para DOCX na nuvem .

4. Carrega-se os dados processados na nuvem

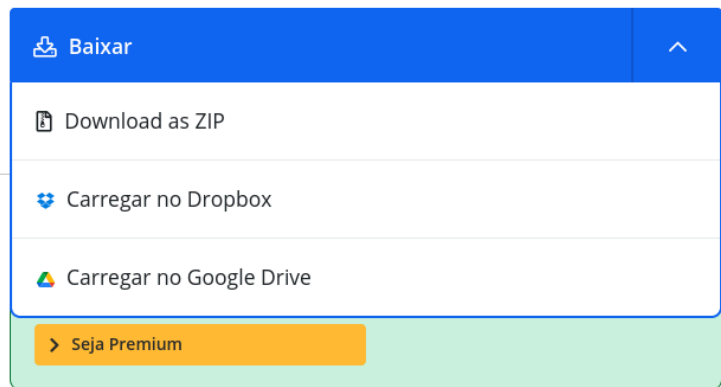


Figura 3.13: Quarto passo para conversão de PDF para DOCX na nuvem .



A sequência de Figuras 3.10-3.13 indica o passo a passo para a execução da conversão de PDF em DOCX em massa.

Todos os 185 documentos em PDF foram convertidos e seus DOCX estão disponíveis no endereço;
<https://drive.google.com/drive/folders/1d6ApF3ecpphf2rDeNCzP4Ko5N8yvNIWG?usp=sharing>.

O melhoramento do layout por OCR resulta em arquivos (.docx) com o layout muito similar, à primeira vista, ao do arquivo original com todas as imagens do arquivo original presentes e as tabelas formatadas de forma similar a original; texto resultante é legível para um humano.

1. Introduction

Angiotensin-converting enzyme 2 (ACE2) is part of the Renin-Angiotensin System (RAS) system, which is involved in the regulation of blood pressure and body fluid control mechanisms. RAS overactivation is a notable risk factor in renal [1] and cardiovascular disease [2] development. RAS overactivity is present in metabolic syndrome [3], type 2 diabetes (T2D) and obesity [4]. *In vitro* and *in vivo* studies show vasoconstriction and sodium retention in cardiac myocytes and mesangial cells as well as in skeletal muscle myoblasts, evidence that supports the clinical and epidemiological data, and tissue RAS activity likely accompanies and aggravates metabolic disorders [5].

Polycystic ovary syndrome (PCOS) encompasses many characteristics of metabolic syndrome, women with PCOS frequently exhibiting insulin resistance [6] and obesity [7], underscored by the significant

proportion of these women who develop type 2 diabetes [8]. Therefore, PCOS women are similarly prone to develop infections as those with diabetes [9–11].

Vitamin D deficiency is known to be common in women with PCOS, 67–85 % of PCOS women having circulating concentrations of less than 20 ng/ml (50 nmol/l), and low 25-hydroxyvitamin D [25(OH)D] levels correlate with increased insulin resistance, obesity and androgen levels [12–14]. Moreover, a recent systematic review of randomized controlled trials determined that vitamin D supplementation improved metabolic dysfunction in PCOS women [15]. Vitamin D status has also been identified as a potential modulating factor in infection risk, and notably SARS-CoV-2 infection, with vitamin D insufficiency being linked to increased COVID-19 incidence [16,17] and severity risk [16,18–20]. How vitamin D insufficiency may negatively impact disease such as COVID-19 is unclear; however, vitamin D negatively modulates the

* Corresponding author.

E-mail address: a.butler@hull.ac.uk (A.E. Butler), a.moia@hull.ac.uk (A.S.M. Moia), S.sathyapalan@hull.ac.uk (T. Sathyapalan), s.kim@hull.ac.uk (S.L. Adiga).

<https://doi.org/10.1016/j.ajme.2021.105965>

Received 16 May 2021; Received in revised form 8 August 2021; Accepted 10 August 2021

Available online 5 October 2021

0960-0760/© 2021 Elsevier Ltd. All rights reserved.

A.E. Butler et al.

Journal of Science of Metabolism and Endocrinology 2 (4) (2021) 105965

Table 1

Demographic and biochemical characteristics of the PCOS and non-PCOS control women. Data are presented as mean (SD).

	CONTROL (n = 68)	PCOS (n = 99)	P value
Age (years)	28 (1)	30 (1)	0.03
BMI (kg/m ²)	27 (1)	33 (1)	<0.0001
Weight (kg)	74 (2)	88 (2)	<0.0001
Waist circumference (cm)	81 (2)	102 (2)	<0.0001
Hip circumference (cm)	101 (1)	119 (2)	<0.0001
Systolic blood pressure (mmHg)	115 (1)	122 (2)	0.0008
Diastolic blood pressure (mmHg)	74 (1)	77 (1)	0.03
AMH (ng/ml)	22 (2)	46 (3)	<0.0001
Sex Hormone Binding Globulin (SHBG) (nmol/l)	73 (10)	43 (5)	0.002
Cholesterol (mmol/l)	4.7 (0.1)	4.8 (0.1)	0.2
Triglycerides (mmol/l)	1.0 (0.1)	1.0 (0.2)	<0.0001
HDL (mmol/l)	1.4 (0.1)	1.2 (0.03)	0.3
LDL (mmol/l)	2.7 (0.1)	2.9 (0.1)	<0.0001
CRP (mg/dl)	2.1 (0.2)	4.8 (0.6)	0.001
Testosterone (nmol/l)	1.1 (0.1)	1.5 (0.1)	0.003
Androstenedione (nmol/l)	8.1 (0.7)	11.3 (0.7)	<0.0001
FAI	2.1 (0.2)	5.7 (0.6)	0.3
Glucose (mmol/l)	4.7 (0.1)	4.9 (0.1)	<0.0001
Insulin (μU/ml)	7.2 (0.7)	17.0 (1.9)	<0.0001
Total vitamin D (nmol/l)	62 (3)	43 (3)	

BMI, Body mass index; AMH, anti-müllerian hormone; HDL, high density lipoprotein; LDL, low density lipoprotein; CRP, C-reactive protein; FAI, free androgen index.

system [21], inhibiting renin expression/ generation, induces ACE2 and suppresses the ang/ang II/AT1R axis [22], and it has been shown that plasma ACE2 levels are predictors of COVID-19 disease outcome [23].

We hypothesized that vitamin D status would be related to plasma levels of the RAS-related proteins Angiotensin-converting enzyme 2 (ACE2), renin and angiotensinogen. To test this hypothesis, we determined plasma levels of RAS proteins and concentrations of 25(OH)D in a population of women with and without PCOS.

2. Methods

2.1. Study population

99 PCOS and 68 non-PCOS control Caucasian women who presented sequentially to the Department of Endocrinology, Hull and East Yorkshire Hospitals NHS Trust were recruited to the local PCOS

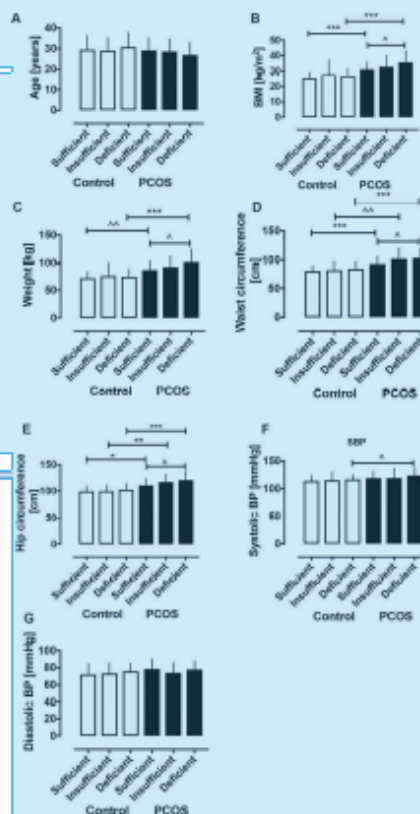


Fig. 1. Demographic data for the PCOS and non-PCOS control women stratified according to vitamin D status. Demographic data of age (A), BMI (B), weight (C), waist circumference (D), hip circumference (E), systolic blood pressure (F) and diastolic blood pressure (G) for the PCOS and non-PCOS control women stratified according to vitamin D status. Of the 99 PCOS women, 16 (16

A Figura 2x contrasta a ordem lógica do DOCX obtido da conversão do arquivo **3d0e16b69e5759e745a50af3b16b34b0.pdf**, com a esperada (Figura 3.14). Na primeira página, tem-se os blocos de cada coluna do texto corretamente delineados, com as notas de rodapé em um bloco separado.

Na segunda página, observa-se um recorte que junta partes das duas colunas em um único bloco, apesar de não serem objetos concatenados; este resultado foi obtido da seleção do texto desde “*Angiotensin-converting enzyme...*” até o “*COVID-19 disease outcome[23].*”, como nos exemplos das Figuras 1, 2 e da Saída 1.

O esperado de um texto legível é que o primeiro objeto da segunda coluna (à direita) somente seja acessado após o último objeto da primeira coluna (à esquerda) ter sido acessado; o que não é observado na Figura 3.14.

As vantagens desta abordagem são o processamento em nuvem e a seleção do conteúdo de interesse, através do atributo `paragraph` do documento. Também, acessando o texto via o atributo `paragraph`, não se leva em conta a estrutura do texto, permitindo a seleção específica do texto do documento.

Entretanto, o texto obtido através deste método não é legível para a máquina, isso porque o layout é construído por partes, inviabilizando a coesão do texto. Uma outra inconveniência é que o processo de OCR, por vezes, acaba conectando palavras erroneamente e comprometendo a pontuação como ilustrado na Figura 3.15.

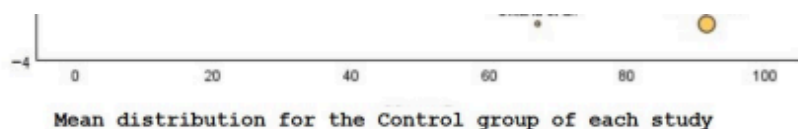


Figure 5. Bubble plot showing the differences in SD between control groups [47-65].

Figures 4 and 5 contain bubble plots, which is a type of data visualization that displays three variables simultaneously on a two-dimensional graph [76]. The three variables in-cluded in these plots are: Cohen's d, patient's mean distribution in the intervention/control groups, and the additional variable is represented by the size of the bubbles plotted on the graph. Each's bubble position in the chart represents where that study is in terms of mean distribution when compared with the others, and the bubble size represents the sample size of each of the 19 intervention or control groups, respectively. It could be argued here that the studies with bigger bubbles carry more weight as the number patients is wider and it covers more of the population sample.

Cohen's d indicates the magnitude of the difference between the groups in standard deviation units [77]. In Figure 5, Cohen's d shows that for the intervention or control groups, several studies greatly differed from the ones inside the bubble [59-61,63].

Figura 3.15: Extração de palavras e separação de linhas no DOCX do f813ee0787090b8ba7460f681d06d93a.pdf.



Figure 5. Bubble plot showing the differences in SD between control groups [47-65].

Figures 4 and 5 contain bubble plots, which is a type of data visualization that displays three variables simultaneously on a two-dimensional graph [76]. The three variables in-cluded in these plots are: Cohen's d, patient's mean distribution in the intervention/control groups, and the additional variable is represented by the size of the bubbles plotted on the graph. Each's bubble position in the chart represents where that study is in terms of mean distribution when compared with the others, and the bubble size represents the sample size of each of the 19 intervention or control groups, respectively. It could be argued here that the studies with bigger bubbles carry more weight as the number patients is wider and it covers more of the population sample.

Cohen's d indicates the magnitude of the difference between the groups in standard deviation units [77]. In Figure 5, Cohen's d shows that for the intervention or control groups, several studies greatly differed from the ones inside the bubble [59-61,63].

Figura 3.16: Extração de palavras e separação de linhas no f813ee0787090b8ba7460f681d06d93a.pdf.

As Figuras 3.15 e 3.16 mostram um mesmo trecho de um mesmo arquivo em DOCX e PDF respectivamente. Verifica-se que o DOCX não consegue resolver a formatação da palavra *included*, isto se deve à OCR para o layout e não conteúdo. Como este tipo de situação surge da formatação do PDF, e não do texto contido no

PDF em si, bibliotecas são capazes de resolver a maioria destas ocorrências enquanto que a conversão com otimização de layout não.

O conteúdo dos DOCX foi extraído utilizando uma variação do Snippet abaixo:

```
from docx import Document
```

```
document = Document(FILE_PATH)
```

```
lines = []
```

```
for paragraph in document.paragraphs:
```

```
    for run in paragraph.runs:
```

```
        lines.append(run.text)
```

```
content = " ".join(lines)
```

F(8)

Como se verifica no F(8), a o conteúdo final do texto extraído do DOCX é o concatenamento de todas as linhas detectadas, o que não garante a ordem lógica.

A plataforma PDF2GO, oferece também a opção de OCR para otimização de detecção de conteúdo, no entanto, frequentemente verificou-se a ocorrência de múltiplas palavras concatenadas, como por exemplo a *`JournalofSteroidBiochemistryandMolecularBiology214`* observada na S(5); o que desencorajou o seu uso.

Extração em nuvem:

Similarmente à conversão para docx, no endereço <https://www.pdf2go.com/pdf-to-text#>, sobe-se todos arquivos PDFs, solicita-se a conversão e baixa-se os arquivos processados, em (.txt).

Este procedimento resulta em arquivos txt, com layout simplificado, com tabelas organizadas de forma as colunas serem concatenadas uma seguida da outra, como uma linha vazia entre duas colunas consecutivas. O texto resultante apresenta ordem lógica de leitura humana parcial, comparável aos das bibliotecas apresentadas na Tabela 2.5.

A vantagem desta abordagem é o processamento em nuvem, que não necessita nenhuma configuração prévia e permite a manipulação de arquivos em massa de forma coordenada, instantânea e integrada, posto que é possível mover estes dados diretamente para o *Google Drive*, armazenamento em nuvem.

As inconveniências aqui encontradas são as mesmas que as encontradas através das ferramentas de código aberto. Esta é a abordagem escolhida e utilizada ao longo deste trabalho, devido a conveniência do processamento e gerenciamento na nuvem e uma melhor indentação do conteúdo do arquivo txt gerado.

3.5 Pós processamento dos dados

O pós-processamento dos dados tem como objetivo o aumento da qualidade do texto extraído dos PDF a fim de fomentar a criação do melhor GC o possível. Este procedimento consiste em duas etapas: correções manual e automática.

Como nenhuma das abordagens de conversão de PDF para txt manteve a ordem original de leitura humana, foi necessário corrigir o conteúdo extraído de forma a garantir a ordem lógica de leitura.

Assim, faz-se necessário limpar o texto espúrio e consertar o texto de uma forma que este siga sua ordem lógica.

Nesta etapa, a fim de aumentar a reprodutibilidade, extrai-se o conteúdo textual dos arquivos pdf em texto pleno (*plaintext*), o qual é salvo em um arquivo de texto (.txt).

Uma vez salvos, os arquivos de texto passam por duas etapas:

1. Remoção de dados não significativos ou não legíveis
 - a. tabelas
 - b. imagens
 - c. notas de rodapé
 - d. tudo que antecede a introdução
 - e. declarações burocráticas, formais e legais
 - f. seção de referências
 - g. marcas d'água de instituições
 - h. informações às margens do documento
2. Correção da ordem do texto

Além de corrigir a ordem lógica do texto, fez-se necessário corrigir erros tipográficos e de pontuação, remoção de dados tabulares e qualquer resquício de layout.

Para tanto, utilizou-se três abordagens; correção programática, correção assistida por IA e correção manual.

Correção Programática

Para atacar esta tarefa, utilizou-se as bibliotecas: *spaCy*(3.7.4), *wordsegment*(1.3.1), *spellChecker*(0.4). As quais apesar de funcionarem pontualmente, não desempenharam a tarefa de forma satisfatória no geral. Devido à presença de objetos numéricos e simbólicos, expressões, quantidades e referências bibliográficas, as bibliotecas utilizadas não foram capazes de detectar as sentenças.

Também, utilizou-se o pipeline descrito em `GIT_LINK`, para as correções necessárias nos textos obtidos através dos DOCX, porém os resultados não foram satisfatórios. Isso porque, programaticamente, as ocorrências padrões são dependentes do contexto, nos quais estas aparecem.

Correção Assistida por IA

A correção assistida por IA foi atacada com um único prompt, que foi utilizado de duas formas, via API e via modo playground; para três modelos distintos, *mistral*, *gemini-1.0-pro* e *gpt-4.0*.

Para o uso de LLMs, é importante respeitar o *context window* do modelo correspondente; que é a quantidade (medida em tokens) de texto que o modelo consegue tratar de uma vez, no processamento de uma pergunta. Isto inclui tamanho do Prompt, da entrada, da pergunta e da resposta gerada.

API

Este procedimento foi realizado através de requisições feitas programaticamente as APIs disponíveis. O prompt utilizado foi:

```
given the text below:

{{ unstructured_text }}

correct the punctuation, spelling, orthography and typos.
then, place the blocks of text that prevent the sentences to be
semantically coherent, to the end of the text after a line break followed
by the `gobbledygook:` pause.
do not paraphrase the input text.
the corrected text should contain no line breaks, no symbols that are not
present in the original text.
the corrected text is a standard text file, no markdown language should
be present.
get rid of any table or tabular data, e.g.: 2 (11.11) 16 (88.89) 5
(27.78) 2 (11.11) 11 (61.11) 2 (11.11) 16 (88.89)
make sure that no information is not or added, but re-arranged.
```

P(2)

No P(2), a variável ``unstructured_text`` assume o valor do texto a ser processado. As instruções nele descritas são diretas, assertivas e concisas. O *prompt* solicita a correção de pontuação, ortografia e de digitação. Em seguida, o *prompt* solicita a formação de frases coerentes e completas de forma a colocar elementos que previnem tal operação, em uma seção chamada de gobbledygook. As instruções também descrevem como o resultado deve ser formatado, com exemplos do que é gobbledygook e as condições para o resultado ser entendido como correto.

Devido ao preço do consumo da API do *gpt-3.5-turbo* ou *gpt-4.0*, optou-se por seguir com os dois outros modelos, *mistral* e *gemini-1.0-pro*. Para ambos, utilizou-se a temperatura 0, que força o modelo a comportar-se de forma determinística; garantindo a reprodutibilidade da tarefa.

O *mistral* foi capaz de entender e executar as tarefas com erros pequenos. Entretanto, seu uso foi desencorajado devido ao tempo de processamento de aproximadamente 60 segundos para blocos de texto de 512 caracteres.

O *gemini-1.0-pro* performou de forma insatisfatória, com apenas uma pequena fração de sucesso. Apesar de um tempo de processamento interessante, entre 3 e 6 segundos para o processamento de blocos de 512 caracteres, e da possibilidade de utilizá-lo gratuitamente, seu baixo desempenho impediu sua empregabilidade.

Os modelos entendem a tarefa e são capazes de executar, ainda que parcialmente, as instruções. Na totalidade dos dados, nenhum dos modelos foi capaz de processar todos os blocos de texto de forma plausível, i.e., com instruções integralmente executadas e que o texto resultante pudesse ser utilizado sem a necessidade supervisão antes de sua utilização.

Navegador

Realiza-se aqui o mesmo procedimento feito na API, porém através do navegador.

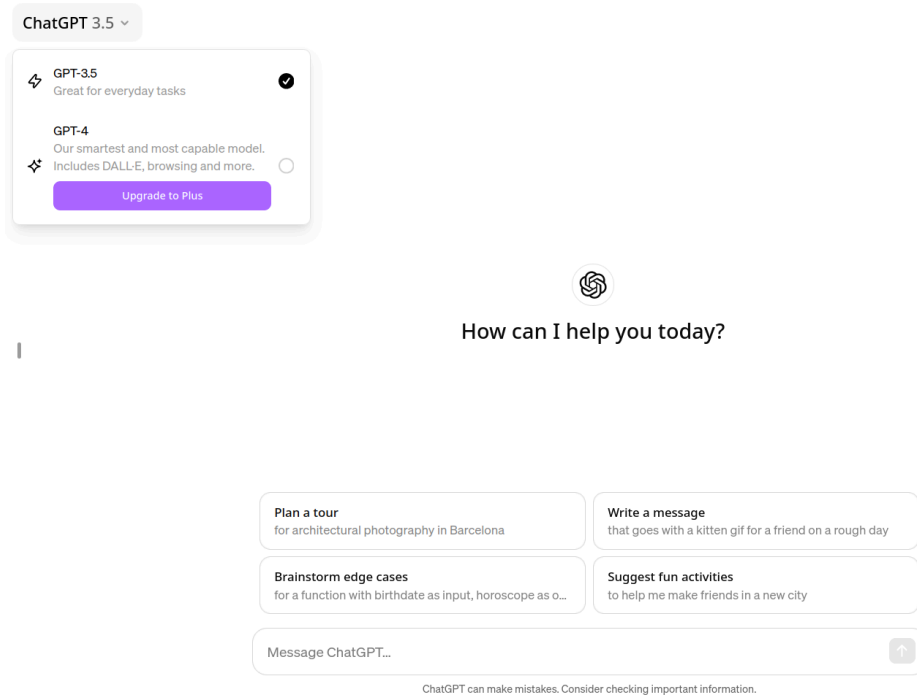


Figura 3.17: *ChatGPT* no navegador; <https://chat.openai.com/>.

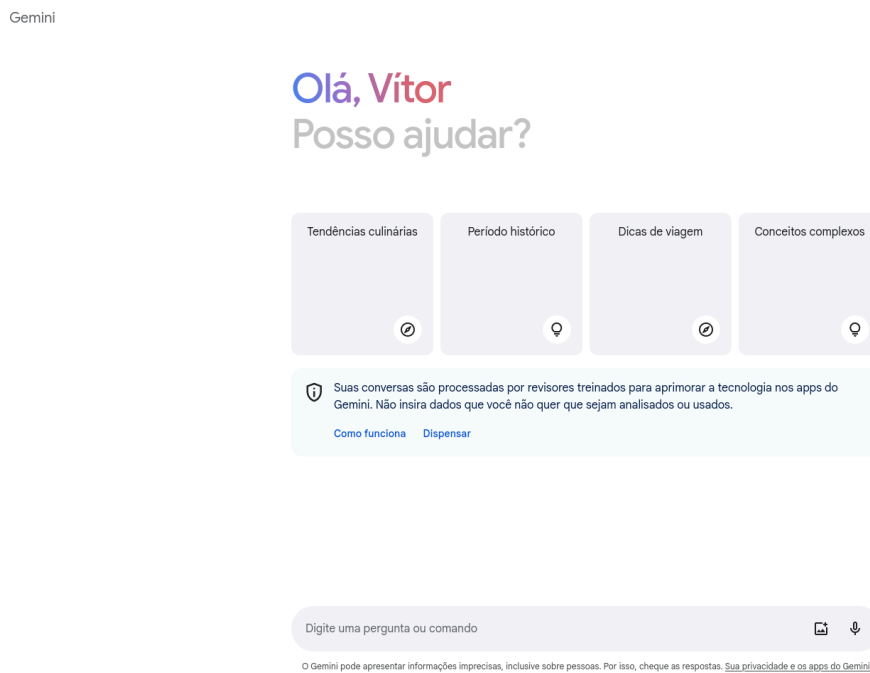


Figura 3.18: *Gemini* no navegador; <https://gemini.google.com>.

É importante esclarecer que ChatGPT refere-se à interface através do qual o usuário interage com os modelos. Assim como Gemini refere-se tanto à interface, quanto à família de modelos (gemini-1.0-pro, gemini-nano, gemini-ultra). Assim, o modelo utilizado no ChatGPT foi o gpt-4.0 e no Gemini, o gemini-1.0-pro.

As Figuras 3.17 e 3.18 mostram as interfaces das duas plataformas utilizadas. Ambas são *chats*, que disponibilizam um campo de entrada para que perguntas sejam feitas e, ao lado esquerdo da tela, o histórico de perguntas e o marketplace, permitem o armazenamento das interações como conversas, de forma que o usuário possa organizar as perguntas e respostas como queira.

```
given the text below:

How vitamin D insufficiency may negatively impact disease such as COVID-19 is unclear; however, vitamin D negatively modulates the RAS * Corresponding author. E-mail addresses: abutler@rcsi.com (A.E. Butler), amoin@rcsi.com (A.S.M. Moin), Sathyapalan@hyms.ac.uk (T. Sathyapalan), satkin@rcsi.com (S.L. Atkin). https://doi.org/10.1016/j.jsbmb.2021.105965 Received 16 May 2021; Received in revised form 8 August 2021; Accepted 10 August 2021 0960-0760/©2021ElsevierLtd.Allrightsreserved. A.E. Butler et al. JournalofSteroidBiochemistryandMolecularBiology214(2021)105965 Table 1 Demographic and biochemical characteristics of the PCOS and non-PCOS control women. Data are presented as mean (SD). Age (years) BMI (kg/m [2]) Weight (kg) Waist circumference (cm) Hip circumference (cm) Systolic blood pressure (mmHg) Diastolic blood pressure (mmHg) AMH (pmol/l) Sex Hormone Binding Globulin (SHBG) (mmol/l) Cholesterol (mmol/l) Triglycerides (mmol/l) HDL (mmol/l) LDL (mmol/l) CRP (mmol/l) Testosterone (nmol/l) Androstenedione (nmol/l) FAI Glucose (mmol/l) Insulin (µU/ml) Total vitamin D (nmol/l) CONTROL (n = 68) 28 (1) 27 (1) 74 (2) 81 (2) 101 (1) 115 (1) 74 (1) 22 (2) 73 (10) 4.7 (0.1) 1.0 (0.1) 1.4 (0.1) 2.7 (0.1) 2.1 (0.5) 1.1 (0.1) 8.1 (0.7) 2.1 (0.2) 4.7 (0.1) 7.2 (0.7) 62 (3) PCOS (n = 99) 30 (1) 35 (1) 98 (2) 102 (2) 119 (2) 122 (2) 77 (1) 46 (3) 43 (5) 4.8 (0.1) 1.6 (0.2) 1.2 (0.03) 2.9 (0.1) 4.8 (0.6) 1.5 (0.1) 11.3 (0.7) 5.7 (0.6) 4.9 (0.1) 17.0 (1.9) 43 (3) P value 0.03 <0.0001 <0.0001 <0.0001 <0.0001 0.0008 0.03 <0.0001 0.002 0.2 0.003 <0.0001 0.3 <0.0001 0.001 0.003 <0.0001 0.3 <0.0001 BMI, Body mass index; AMH, anti-mullerian hormone; HDL, high density lipo-protein; LDL, low density lipoprotein, CRP, C-reactive protein; FAI, free androgen index. system [21], inhibiting renin expression/ generation, induces ACE2 and suppresses the ang/Ang II/AT1R axis [22], and it has been shown that plasma ACE2 levels are predictors of COVID-19 disease outcome [23]. We hypothesized that vitamin D status would be related to plasma levels of the RAS-related proteins Angiotensin-converting enzyme 2 (ACE2), renin and angiotensinogen.

correct the punctuation, spelling, orthography and typos. then, place the blocks of text that prevent the sentences to be semantically coherent, to the end of the text after a line break followed by the `gobbledygook:` pause. do not paraphrase the input text. the corrected text should contain no line breaks, no symbols that are not present
```

P(3)

```
in the original text.  
the corrected text is a standard text file, no markdown language should be  
present.  
get rid of any table or tabular data, e.g.: 2 (11.11) 16 (88.89) 5 (27.78) 2  
(11.11) 11 (61.11) 2 (11.11) 16 (88.89)  
make sure that no information is not or added, but re-arranged.
```

O Prompt 3 mostra um exemplo do que foi submetido nos campos de entrada das plataformas, o que basicamente.

Visou-se corrigir as limitações e inconveniências encontradas durante o uso da *API*. Este procedimento consistiu em reproduzir manualmente, e não necessariamente sequencialmente, o procedimento realizado pela API. Ou seja, os prompts foram copiados e colados no campo de entrada e, em seguida, processados.

Os resultados obtidos foram classificados como correto e incorreto, de forma que:

Resultado correto: Obtido através da execução ordenada das instruções separando o texto em duas seções uma onde o texto fluido e outra contendo todo o texto que, quando removido, permite aproveitamento semântico total do texto.

Resultado incorreto: Obtido com resquícios de elementos que atrapalham o aproveitamento léxico, morfológico e semântico total do texto.

Estes foram processados da seguinte maneira:

Resultado correto: Salvo em um arquivo correspondente ao bloco executado.

Resultado incorreto: Dado o feedback para a o sistema, fez-se em seguida acompanhamentos (follow up-questions) a fim de guiar o modelo na execução das instruções de forma correta.

Um exemplo da complexidade desta operação é descrito abaixo:

O primeiro bloco de 4000 caracteres do arquivo **0a337a62de92b3e2717d0f7a454dbe4c.txt**, corresponde a o texto grifado na Figura 3.19:

Guoqiang Liu 
 Tianpei Hong 
 Jin Yang 

Department of Endocrinology and Metabolism, Peking University Third Hospital, Beijing 100191, People's Republic of China

Abstract: There is no specific and effective medication for coronavirus disease 2019 (COVID-19), and avaccine is not available in recent months. Here, we hypothesize that a single large dose of vitamin D (Vit D) could be an option for trial in COVID-19. Vit D deficiency or insufficiency is very common in the general population as well as in patients with COVID-19. It has been shown that low Vit D level is associated with viral infection, and Vit D supplementation is beneficial for people infected with viruses, such as HIV and hepatitis C virus. Although COVID-19 is a respiratory disease, the morbidity and mortality of this disease are driven by coagulopathy. Clinical studies have shown that Vit D can exert anticoagulant effects. Vit D, a lipid-soluble vitamin, can be administered as a draught. Vit D supplementation is safe and has rare toxic events. In addition, the cost of Vit D is fairly low. Based on these observations, we speculate that a single dose of 300,000 IU Vit D may have a role in the prevention and treatment of COVID-19.

Keywords: vitamin D, large dose, coronavirus disease 2019

Introduction

Coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is an unprecedented challenge for the whole world and modern medicine. According to reports of the World Health Organization (WHO), incidence and prevalence of this virus infection is still increasing worldwide, and as of August 5, 2020, more than 18 million people have been affected, including more than 696 thousand deaths. Since COVID-19 is a completely new disease, no specific drug has been proven to be efficacious and safe for treating it at present.¹ Moreover, vaccine development and clinical application may take years. Therefore, any low-risk treatment is worth trying to help control the global pandemic.

Vitamin D (Vit D) is an essential lipid-soluble vitamin and steroid prohormone. The benefits of Vit D are not limited to the regulation of calcium and phosphorus metabolism and the maintenance of bone health. It has been demonstrated that Vit D deficiency or insufficiency is associated with cardiovascular diseases,² tumors,³ respiratory infections,^{4,5} asthma,^{6,7} and other diseases^{8–11} (Table 1). Vit D deficiency is defined as a 25-hydroxyvitamin D [25(OH)D] below 20 ng/mL, insufficiency as a 25(OH)D of 21–29 ng/mL, and sufficiency as a 25(OH)D of 30–100 ng/mL.¹² Epidemiologically, Vit D deficiency or insufficiency has been reported to be very common in all age

Correspondence: Jin Yang
 Tel +86-10-82266722
 Email yangjin@bjmu.edu.cn

submit your manuscript | www.dovepress.com
 DovePress    
<https://doi.org/10.2147/DDDT.S217154>


Drug Design, Development and Therapy 2020:14 3429–3434 **3429**
 © 2020 Liu et al. This work is published and licensed by Dove Medical Press Limited. The full terms of this license are available at <http://www.dovepress.com/terms.php> and incorporate the Creative Commons Attribution – Non Commercial (cc-by-nc) 4.0 International license (<http://creativecommons.org/licenses/by-nc/4.0/>). By accepting this license agreement, the user agrees that all contents of this work are protected without any further permission from Dove Medical Press Limited, provided the work is properly attributed. For permission for commercial use of this work, please see paragraphs 4.2 and 5.1 of our terms (<http://www.dovepress.com/terms.php>).

Table 1 Vitamin D Deficiency or Insufficiency-related Diseases and Disorders

Category	Related Diseases and Disorders
Cardiovascular disease ²	Heart failure, myocardial infarction, and ischemic heart disease
Tumor ³	Colorectal cancer, breast cancer, prostate cancer, melanoma, lung cancer, and cutaneous cancers
Infectious diseases ^{4,5}	Acute respiratory tract infection
Respiratory disease ^{6,7}	Asthma
Neuropsychiatric	Alzheimer's disease, Parkinson's disease

Moreover, Vit D plays a role in the treatment of RNA viruses such as human immunodeficiency virus (HIV) and hepatitis C virus (HCV). Vit D deficiency is associated with lower absolute CD4⁺ T-cell count recovery in HIV-positive patients and Vit D supplementation may improve the CD4⁺ T cell count recovery.²⁰ A pilot, open-label, prospective phase I clinical trial in HIV-positive patients with Vit D deficiency showed that a single dose of 200,000 IU oral Vit D increased the frequencies of antigen-specific T cells expressing macrophage inflammatory protein 1β (MIP-1β), an important anti-HIV blocking chemokine. In addition, plasma cathelicidin, a Vit D response gene encoding product with broad antimicrobial activity, was also increased after the

Figura 3.19: Bloco de 4000 caracteres do 0a337a62de92b3e2717d0f7a454dbe4c.txt.

A Figura 3.19 contém um exemplo interessante que evidencia a complexidade da tarefa em questão. Primeiramente, a ordem natural do texto é mantida até “very common in all age”. Em seguida, as notas de rodapé são incluídas e então até a terceira linha da tabela “Respiratory disease 6,7 Asthma”, quando a seleção segue para a segunda coluna, onde o cursor está, sem passar por toda a primeira coluna anteriormente.

Abstract: There is no specific and effective medication for coronavirus disease 2019 (COVID-19), and **avaccine** is not available in recent months. Here, we hypothesize that a single large dose of vitamin D (Vit D) could be an option for trial in COVID-19. Vit D deficiency or insufficiency is very common in the general population as well as in patients with COVID-19. It has been shown that low Vit D level is associated with viral infection, and Vit D supplementation is beneficial for people infected with viruses, such as HIV and hepatitis C virus. Although COVID-19 is a respiratory disease, the morbidity and mortality of this disease are driven by coagulopathy. Clinical studies have shown that Vit D can exert anticoagulant effects. Vit D, a lipid-soluble vitamin, can be administered as a draught. Vit D supplementation is safe and has rare toxic events. In addition, the cost of Vit D is fairly low. Based on these observations, we speculate that a single dose of 300,000 IU Vit D may have a role in the prevention and treatment of COVID-19. Keywords: vitamin D, large dose, coronavirus disease 2019 Introduction Coronavirus disease 2019 (COVID-19) caused by severe acute respiratory **syn-drome** coronavirus 2 (SARS-CoV-2) is an unprecedented challenge for the whole world and modern medicine. According to reports of the World Health Organization (WHO), incidence and prevalence of this virus infection is still increasing world-wide, and as of August 5, 2020, more than 18 million people have been affected, including more than 696 thousand deaths. Since COVID-19 is a completely new disease, no specific drug has been proven to be efficacious and safe for treating it at present.¹ Moreover, vaccine development and clinical application may take years. Therefore, any low-risk treatment is worth trying to help control the global pandemic. Vitamin D (Vit D) is an essential lipid-soluble vitamin and steroid prohormone. The benefits of Vit D are not limited to the regulation of calcium and phosphorus metabolism and the maintenance of bone health. It has been demonstrated that Vit D deficiency or insufficiency is associated with cardiovascular **diseases,2 tumors,3 respiratory infections,4,5 asthma,6,7** and other **diseases8–11 (Table 1)**. Vit D deficiency is defined as a 25-hydroxyvitamin D [25(OH)D] below 20 ng/mL, insufficiency as a 25(OH)D of 21–29 ng/mL, and sufficiency as a 25(OH)D of 30–100 ng/mL.¹² Epidemiologically, Vit D deficiency or insufficiency has been reported to be very common in all age Drug Design, Development and Therapy 2020:14 3429–3434 3429 © 2020 Liu et al. This work is published and licensed by Dove Medical Press Limited. The full terms of this license are available at <https://www.dovepress.com/terms.php> and incorporate the Creative Commons Attribution – Non Commercial (unported, v3.0) License (<http://creativecommons.org/licenses/by-nc/3.0/>). By accessing the work you hereby accept the Terms. Non-commercial uses of the work are permitted without any further permission from Dove Medical Press Limited, provided the work is properly attributed. For permission for commercial use of this work, please see paragraphs 4.2 and 5 of our Terms (<https://www.dovepress.com/terms.php>). Liu et al Table 1 Vitamin D Deficiency or Insufficiency-related Diseases and Disorders Dovepress Moreover, Vit D plays a role in the treatment of RNA viruses such as human immunodeficiency virus (HIV) Category Cardiovascular **disease2 Tumor3 Infectious diseases4,5 Respiratory disease6,7 Related Diseases and Disorders Heart failure, myocardial infarction, and ischemic heart disease Colorectal cancer, breast cancer, prostate cancer, melanoma, lung cancer, and cutaneous cancers Acute respiratory tract infection Asthma and hepatitis C virus (HCV)**. Vit D deficiency is associated with lower absolute CD4+ T-cell count recovery in HIV-positive patients and Vit D supplementation may improve the CD4+ T cell count **recovery.20**

E(1)

]

Abstract: There is no specific and effective medication for coronavirus disease 2019 (COVID-19), and a vaccine is not available in recent months. Here, we hypothesize that a single large dose of vitamin D (Vit D) could be an option for trial in COVID-19. Vit D deficiency or insufficiency is very common in the general population as well as in patients with COVID-19. It has been shown that low Vit D level is associated with viral infection, and Vit D supplementation is beneficial for people infected with viruses, such as HIV and hepatitis C virus. Although COVID-19 is a respiratory disease, the morbidity and mortality of this disease are driven by coagulopathy. Clinical studies have shown that Vit D can exert anticoagulant effects. Vit D, a lipid-soluble vitamin, can be administered as a draught. Vit D supplementation is safe and has rare toxic events. In addition, the cost of Vit D is fairly low. Based on these observations, we speculate that a single dose of 300,000 IU Vit D may have a role in the prevention and treatment of COVID-19. Keywords: vitamin D, large dose, coronavirus disease 2019 Introduction Coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is an unprecedented challenge for the whole world and modern medicine. According to reports of the World Health Organization (WHO), incidence and prevalence of this virus infection is still increasing worldwide, and as of August 5, 2020, more than 18 million people have been affected, including more than 696 thousand deaths. Since COVID-19 is a completely new disease, no specific drug has been proven to be efficacious and safe for treating it at present. Moreover, vaccine development and clinical application may take years. Therefore, any low-risk treatment is worth trying to help control the global pandemic. Vitamin D (Vit D) is an essential lipid-soluble vitamin and steroid prohormone. The benefits of Vit D are not limited to the regulation of calcium and phosphorus metabolism and the maintenance of bone health. It has been demonstrated that Vit D deficiency or insufficiency is associated with cardiovascular diseases, tumors, respiratory infections, asthma, and other diseases. Vit D deficiency is defined as a 25-hydroxyvitamin D [25(OH)D] below 20 ng/mL, insufficiency as a 25(OH)D of 21–29 ng/mL, and sufficiency as a 25(OH)D of 30–100 ng/mL. Epidemiologically, Vit D deficiency or insufficiency has been reported to be very common in all age groups. Moreover, Vit D plays a role in the treatment of RNA viruses such as human immunodeficiency virus (HIV) and hepatitis C virus (HCV). Vit D deficiency is associated with lower absolute CD4+ T-cell count recovery in HIV-positive patients, and Vit D supplementation may improve the CD4+ T cell count recovery.

S(6)

gobbledygook: Drug Design, Development and Therapy 2020:14 3429–3434 3429 © 2020 Liu et al. This work is published and licensed by Dove Medical Press Limited. The full terms of this license are available at <https://www.dovepress.com/terms.php> and incorporate the Creative Commons Attribution – Non Commercial (unported, v3.0) License (<http://creativecommons.org/licenses/by-nc/3.0/>). By accessing the work you hereby accept the Terms. Non-commercial uses of the work are permitted without any further permission from Dove Medical Press Limited, provided the work is properly attributed. For permission for commercial use of this work, please see paragraphs 4.2 and 5 of our Terms (<https://www.dovepress.com/terms.php>). Liu et al Table 1 Vitamin D Deficiency or Insufficiency-related Diseases and Disorders Dovepress





A Entrada 1, retratada pelo texto destacado em verde na Figura 3.19, é processada pelo Prompt 1, resultando em S(6). Contudo, esta representação constitui um exemplo de resposta inadequada, uma vez que o texto corrigido apenas se aproxima do conteúdo grifado.

Uma discrepância é observada entre o segmento "very common in all age" e "Moreover, Vit D plays", devido a uma extração incorreta na etapa anterior do processo. Tal lacuna não pôde ser corrigida pelo modelo, a menos que este detivesse a totalidade do texto original em sua memória.

Em S(6), o modelo tentou abordar essa descontinuidade, complementando a frase para "very common in all age groups." Embora a completude automática não estivesse instruída, o modelo julgou-a pertinente. Esse comportamento, embora intrigante, contradiz os critérios objetivos de classificação dos resultados, pois diverge das instruções. Nesse contexto, a falta de conformidade com as diretrizes deveria incorrer em uma classificação de "gobbledygook".

Conforme estipulado nas instruções, notas de rodapé, por sua incapacidade de contribuir para a formação de uma frase coesa, são classificadas como "gobbledygook", assim como tabelas, cuja exclusão é explicitamente recomendada nas diretrizes.

Para ilustrar a complexidade da tarefa de processamento de texto, é importante destacar situações desafiadoras enfrentadas nesse contexto, conforme demonstrado na Entrada 1 e explicado detalhadamente a seguir.

	Erros ortográficos, decorrente de detecção errônea de palavras
	Erros ortográficos, decorrentes do pré-processamento
	Referências bibliográficas
	Valores agrupados

Primeiramente, há os erros ortográficos resultantes de detecções equivocadas de palavras, bem como aqueles advindos de etapas de pré-processamento. A correção desses erros, como observado na Entrada 1 e na Saída 6, é eficazmente realizada com o auxílio de IA. Entretanto, do ponto de vista programático, corrigir tais erros representa um desafio considerável. Exige-se um conhecimento abrangente do vocabulário e um entendimento profundo do contexto para resolver imprecisões entre palavras adjacentes.

Em PDF, há algumas maneiras como referências bibliográficas são escritas: fonte[2, 3], fonte[2–4], fonte(2, 4–7), fonte², fonte², fonte^{2,4–7}, dentre outras. Quando o texto é extraído do PDF, a formatação é perdida, resultando respectivamente em: fonte[2, 3], fonte[2–4], fonte(2, 4–7), fonte2, fonte2, fonte2,4–7. A remoção destas

referências vai muito além de ser capaz de identificar tais padrões numéricos, senão o contexto, para que estes não sejam confundidos com intervalos, índices, quantidades ou parte de substantivos.

Além disso, a extração de texto de arquivos em PDF apresenta suas próprias complicações, especialmente no que se refere às referências bibliográficas. Nas publicações acadêmicas (em PDF), as referências podem ser formatadas de diversas maneiras, como fonte[2, 3], fonte[2–4], fonte(2, 4–7), fonte², fonte², fonte^{2,4–7}. Quando o texto é extraído de um PDF, a formatação original é frequentemente perdida, resultando em uma simplificação que uniformiza os formatos para, fonte[2, 3], fonte[2–4], fonte(2, 4–7), fonte2, fonte2, fonte2,4–7, respectivamente. A remoção adequada dessas referências transcende a capacidade de identificar padrões numéricos, exigindo uma análise contextual detalhada para evitar a confusão com intervalos, índices, quantidades, ou partes integrantes de substantivos.

Outra questão significativa é a ocorrência de valores agrupados no texto extraído de PDFs. Esses podem indicar uma variedade de elementos, incluindo dados tabulares, números, ênfases, entre outros, o que complica ainda mais a tarefa de resolução automatizada, mesmo para a IA. Conforme indicado na Entrada 1, o texto destacado em vermelho não é retido na saída final; ele não é sequer transferido para a seção 'gobbledygook'. Isso exemplifica as dificuldades enfrentadas na tarefa de correção textual.

Através deste exame das complexidades inerentes ao processamento de texto, fica evidente a necessidade de avanços contínuos em tecnologias de IA para lidar efetivamente com estas questões multifacetadas.

3.6 Integração dos dados

A concatenação dos dados foi realizada de forma a respeitar a ordem lógica de cada documento, sendo posteriormente incorporados em um DataFrame do pandas. Esta escolha de estrutura de dados, adequada para a manipulação de conjuntos de dados tabulares programática, reflete a necessidade de uma análise completa e abrangente do corpus. Em particular, o DataFrame facilita a manipulação

simultânea de todo o conjunto de dados, graças à sua capacidade de tratamento de dados em formato tabular, algo parcialmente análogo à estrutura de outros sistemas de armazenamento de dados.

A estrutura resultante do DataFrame é organizada da seguinte maneira: cada linha é identificada por um índice único ('id'), associado ao nome do arquivo de origem ('fname') e ao conteúdo do texto extraído ('text'). Essa organização permite um tratamento detalhado de cada registro, habilitando a seleção, a identificação e a análise do conteúdo textual de maneira eficiente. A seguir, apresenta-se uma amostra desse DataFrame, exportada no formato Comma Separated Values (CSV), para ilustrar a sua estrutura e funcionamento.

id	fname	text
0	18bb960439a4c04ef42d37662b1af07d	"Evidence recommends that vitamin D might be a crucial supportive agent for the immune system, mainly in cytokine response regulation against COVID-19."
1	18bb960439a4c04ef42d37662b1af07d	"Hence, we carried out a systematic review and meta-analysis in order to maximise the use of everything that exists about the role of vitamin D in the COVID-19."
2	18bb960439a4c04ef42d37662b1af07d	"A systematic search was performed in PubMed, Scopus, Embase and Web of Science up to December 18, 2020."
3	18bb960439a4c04ef42d37662b1af07d	Studies focused on the role of vitamin D in confirmed COVID-19 patients were entered into the systematic review.

E(2)

Neste modelo, os 'ids' são incrementados sequencialmente dentro de cada arquivo ('fname'), correspondendo à posição do texto dentro da coluna 'text', conforme estabelecido pela lógica de leitura posta pelo pré-processamento. O arquivo CSV gerado, denominado 'corpus.csv', atua como uma base de dados central para análises subsequentes.

A consolidação dos dados extraídos em um formato tabular unificado traz múltiplas vantagens para a análise de dados. Primeiramente, essa estrutura

favorece a aplicação de técnicas de processamento de linguagem natural (PLN) de forma integrada, potencializando a extração de percepções detalhadas e aprofundadas acerca do conteúdo do corpus. Além disso, simplifica significativamente operações de limpeza, filtragem e agregação de dados, etapas cruciais para a preparação e análise de dados.

Além disso, a consolidação dos dados em um formato tabular único simplifica a realização de operações de limpeza, filtragem e agregação dos dados, essenciais para a preparação da análise. Isso também possibilita uma visão global do conjunto de dados, facilitando a identificação de padrões, tendências e discrepâncias que seriam difíceis de perceber em documentos isolados.

Outro benefício relevante é a eficiência no manuseio dos dados: ao invés de acessar e processar cada documento PDF individualmente, o pesquisador ou analista pode trabalhar com um único arquivo consolidado, otimizando tempo e recursos computacionais. Ademais, essa abordagem promove uma maior reprodutibilidade e compartilhamento da pesquisa, uma vez que outros pesquisadores podem facilmente acessar e manipular o conjunto de dados estruturado.

A estratégia de unificar os textos extraídos de documentos PDF em um DataFrame não somente eleva a eficácia e eficiência da análise de dados, mas também amplia as possibilidades de investigação e descoberta nos dados coletados, promovendo uma maior reprodutibilidade e facilitando o compartilhamento da pesquisa dentro da comunidade acadêmica.

3.7 Transformações

Correções

Expressão regular (*Regex*), é uma sequência especial de caracteres usada para localizar e/ou identificar padrões específicos dentro de textos. Funcionam como um filtro para localizar sequências de caracteres desejadas em texto

Nesta etapa do trabalho, expressões regulares foram utilizadas para automatizar/facilitar a limpeza e correção do texto, a maioria das quais é mostrada na tabela abaixo:

Tabela 3.11: Expressões Regulares empregadas no corpus.

	Expressão Regular	Substituto	Objetivos e Casos
1	<code>(\d+(\. \d+)?)\s*\(\d+(\. \d+)?)\s+</code>	vazio	Reconhece pares repetidos de números dentro de parênteses, Provavelmente representando dados tabulares.
2	<code>(?:\d+\. \d+\s){2},\d+\. \d+</code>	vazio	Localiza sequências de três números de pontos flutuantes consecutivos.
3	<code>(?:\s^)(\d+(\. \d+)?\s+){3},</code>	vazio	Identifica pares de números consecutivos separados por espaço.
4	<code>(?:\d+\. \d+\d+)(?:\s+(\d+\. \d+\d+)+){2},</code>	vazio	Utilizado novamente dois ou mais números consecutivos (de qualquer tipo) separados por espaço.
5	<code>(?:\s^)(-\d+(\. \d+)?\s+){3},</code>	vazio	Identifica sequências numéricas com mais de 3 números consecutivos separados por espaço.
6	<code>(?:\b\d+(?:\.\d+)?(?:\.\d+(\d+)?\d+(?:\.\d+)?%\d+(?:\d+)?)(?:\s+)\$){2},</code>	vazio	Novamente tabelas com percentuais.
7	<code>(?:\s+-?\d+(?:\.\d+)?(?:\s+\d+\s+(\d+(?:\.\d+)?\s*))){3},</code>	vazio	Localiza dados tabulares com números e parênteses opcionais, seguido de "de" e dois números. Pode exigir refinamento com base no contexto.

8	<code>(?:\s+~?\d+(?:\.\d+)?)~?(?:\s+\d+\s+\(\d+(?:\.\d+)?\))?\s*){2,}</code>	manual	Localiza dados sequenciais com mais de dois números. Usada para remover referências.
9	<code>\b\d+\s+of\s+\d+\b</code>	manual	Localiza "of" seguido por dois números. Usada na remoção de contagem de páginas.
10	<code>\d+, \d+(?:, \d+)*\.</code>	!	Localiza números com ponto flutuante que imediatamente sucedem vírgulas. Usada para remover referências.
11	<code>\. \d+, \d+(?:, \d+)*</code>	!	Localiza pares de números separados por vírgulas. Usada para remover referências.
12	<code>~\n</code>	vazio	Localiza palavras interrompidas por quebras de linha, removendo o hífen e a quebra de linha. Resultando na palavra original completa.
13	<code>\[s*\d+(?:, s*\d+)*s*\]</code>	vazio	Localiza sequências de números inteiros separados por vírgulas. Usada para remover referências.
14	<code>\[s*(\d+(?:, \d+)*)\]</code>	vazio	Localiza grupos de sequências de números inteiros separados por vírgulas. Usada para remover referências.
15	<code>(\d+)-(\d+)</code>	vazio	Localiza números separados por hífen. Usada para remover referências.
16	<code>\(\d+-\d+)\.</code>	!	Localiza números separados por hífen à direita de um `.`. Usada para remover referências.
17	<code>\[\d+-\d+\]</code>	vazio	Localiza números separados por hífen envoltos em []. Usada para remover referências.
18	<code>\(\d{2}\)</code>	vazio	Localiza números de dois dígitos dentro envoltos em (). Usada para remover referências.
19	<code>\((\d{2})-(\d{2})\)</code>	vazio	Localiza números de dois dígitos dentro envoltos em (). Usada para remover referências.

20	<code>\.(\d{2})-(\d{2})</code>	.	Localiza números de dois dígitos dentro envoltos em () separados por -, logo após . Usada para remover referências.
21	<code>(\d{2})-(\d{2})\.</code>	.	Localiza números de dois dígitos separados por hífen à direita de um ponto. Usada para remover referências.
22	<code>[A-Z](\d{2}),(\d{2})</code>	vazio	Localiza uma letra maiúscula seguida de dois dígitos, vírgula e outros dois dígitos dentro de colchetes. Usada para remover referências.
23	<code>[a-z](\d{2})</code>	vazio	Localiza uma letra minúscula seguida de dois dígitos dentro de colchetes. Usada para remover referências.
24	<code>[a-zA-Z]+\.[(\d+,)]+(\d+)</code>	vazio	Localiza uma palavra, ponto, sequência de dígitos e vírgulas, hífen e outro dígito. Usada para remover referências.
25	<code>[a-zA-Z]+\.[(\d+,)]+(\d+)</code>	vazio	Similar à regex 5, mas com hífen em vez de vírgula. Usada para remover referências.
26	<code>\s{2,}</code>	espaço	Localiza dois ou mais espaços em branco. Usada para remover espaçamento excessivo.
27	<code>[\d+([, -]\d+)+]</code>	vazio	Localiza um número, seguido de um ou mais conjuntos de vírgula ou hífen e dígitos, dentro de colchetes. Usada para identificar e extrair sequências numéricas.
28	<code>\b\d{2},\d{2}(?:,\d{2})*\b</code>	vazio	Localiza um par de dígitos, vírgula, outro par de dígitos e vírgulas opcionais com pares de dígitos, delimitado por "fronteiras de palavra". Usada para identificar datas no formato DD,MM(YY)
29	<code>[\d+([, -]\d+)+]</code>	vazio	Similar à regex 26, mas com hífen e vírgula como opções de separador. Usada para identificar e remover referências.
30	<code>et al(\d+)</code>	et al.	Localiza "et al." seguido de um número. Usada para identificar e extrair o número de autores.

31	<code>\b[a-zA-Z]{5}\d+\b</code>	espaço	Localiza uma palavra de 5 letras e um número, delimitado por "fronteiras de palavra". Usada para identificar códigos específicos.
32	<code>([,]\d{2})\n</code>	,	Localiza uma vírgula, dois dígitos e uma nova linha. Usada para remover referências.
33	<code>([,]\d{2})</code>	,	Localiza uma vírgula e dois dígitos. Usada para remover referências.
34	.	.	Correção de Erros morfológicos.
35	,	,	Correção de Erros morfológicos
36	α	-alpha-	Correção de Erros tipográficos.
37	SarsCov	Sars-Cov	Correção de Erros tipográficos.
38	metaa	meta-a	Correção de Erros tipográficos.
39	<code>^\.{1,3}\$</code>	.	Localiza números inteiros de até 3 dígitos que sucedem pontos finais. Usada na remoção de referências.
40	<code>\b[A-Za-z]+,\d+\b</code>	manual	Localiza palavras separadas por uma vírgula sem espaços em branco. Usada na correção de resolução.
41	.	.	Correção de pontuação e resolução.
42	<code>-\n</code>	—	Correção de formatação e resolução.
43	<code>, \n</code>	,	Correção de pontuação e resolução.
44	<code>;\n</code>	;	Correção de pontuação, formatação e resolução.
45	,,	,	Correção de pontuação, formatação e resolução.
46	<code>(\r?\n\r){3,}</code>	2 quebras de linha	Correção de pontuação, formatação e resolução.
47	.	.	Correção de pontuação, formatação e resolução.
48	◆	≧	Correção de pontuação, formatação e resolução.

49	/	/	Correção de pontuação, formatação e resolução.
50	'		Correção de pontuação, formatação e resolução.
51	^(?!.*(?:..")\$). +	manual	Identifica sentenças completas.
52	.	'	Correção de pontuação, formatação e resolução.
53	."	"	Correção de pontuação, formatação e resolução.
54	않	≥	Correção de pontuação, formatação e resolução.
55	fi	fi	Correção de pontuação, formatação e resolução.
56	fl	fl	Correção de pontuação, formatação e resolução.
57	ff	ff	Correção de pontuação, formatação e resolução.
58	ffi	ffi	Correção de pontuação, formatação e resolução.
59	⊙	Δ	Correção de pontuação, formatação e resolução.
60	ⁿ _L	vazio	Correção de pontuação, formatação e resolução.
61	■	μ	Correção de pontuação, formatação e resolução.
62	↻	~	Correção de pontuação, formatação e resolução.
63	•	!	Correção de pontuação, formatação e resolução.
64	•	!	Correção de pontuação, formatação e resolução.
65	•	!	Correção de pontuação, formatação e resolução.

66	▶	vazio	Correção de pontuação, formatação e resolução.
67	◦	◦	Correção de pontuação, formatação e resolução.
68	⌞	<	Correção de pontuação, formatação e resolução.
69	~	~	Correção de pontuação, formatação e resolução.
70	U+2000 to U+2007	vazio	Remoção de espaços de todos os comprimentos
71	*	*	Correção de pontuação, formatação e resolução.
72	(?<!\\S)\\b\\w{1,15}, \\w+(?!\\S)	vazio	Localiza uma palavra de 1 a 15 caracteres, seguida de vírgula e outra palavra, delimitadas por "não-espços em branco". Usada para identificar palavras fundidas.
73	(?<=\\d), (?=\\d{3}(?!\\d))	vazio	Localiza uma vírgula precedida por um dígito e seguida por outros três dígitos que não sejam parte de um número maior. Usada para separar decimais, em números.
74	(\\d+)(\\d+)(\\d+)	vazio	Localiza três grupos de um ou mais dígitos. Usada para extrair partes separadas por dígitos.
75	¼	¼	Correção de pontuação, formatação e resolução.
78	°C	°C	Correção de pontuação e ortografia.
79	(\\d+)]	vazio	Localiza dígitos sucedidos por]. Usada para remoção de referências.
80	\\b[a-zA-Z]+\\s*\\d+, \\d+(?:, \\d+)*.	vazio	Localiza uma palavra (ou sequência de letras), seguida opcionalmente de espaços em branco, vírgula, um ou mais dígitos, outra vírgula e repetições opcionais de vírgula e dígitos, tudo delimitado por "fronteiras de palavra" e ponto final. Usada para identificar e remover referências.

81	<code>([^\w\d.])(\d{1,5})</code>	manual	Localiza caracteres diferentes de letra, dígito ou ponto, seguido de um a cinco dígitos. Usada para extrair valores numéricos com separadores não padrão.
82	<code>\w{4,6}(?! \d)\d{1,3}</code>	manual	Localiza 4 a 6 letras/números que não sejam precedidas por um dígito, seguido de 1 a 3 dígitos. Usada para remover referências.
83	<code>(?<=\d) days</code>	<code>__days</code>	Localiza `days` quando usados como unidade de medida. Usada para desambiguação.
84	<code>(?<=\d) years</code>	<code>__years</code>	Localiza `years` quando usados como unidade de medida. Usada para desambiguação.
85	<code>(?<=\d) months</code>	<code>__months</code>	Localiza `months` quando usados como unidade de medida. Usada para desambiguação.
86	<code>nmol / L</code>	nmol/L	Correção de pontuação, formatação e resolução. Usada na padronização da unidade de medida.
87	<code>nmol L 1</code>	nmol/L	Correção de pontuação, formatação e resolução. Usada na padronização da unidade de medida.
88	<code>nmol L-1</code>	nmol/L	Correção de pontuação, formatação e resolução. Usada na padronização da unidade de medida.
89	<code>nmol l-1</code>	nmol/L	Correção de pontuação, formatação e resolução. Usada a fim de padronizar a unidade de medida.

A Tabela 3.11 detalha as expressões regulares utilizadas nos processos de correção automatizados e semi-automatizados. Embora as expressões regulares tenham sido aplicadas de maneira sequencial, a ordem de aplicação no corpus não corresponde diretamente à ordem apresentada na tabela. Esta estratégia foi adotada para maximizar a precisão, aplicando-se as expressões regulares de forma iterativa e focada, de detecção em detecção, a fim de minimizar possíveis erros. Como resultado da execução dessas expressões regulares, foi gerado um novo arquivo no

formato CSV, que contém a versão revisada e corrigida do corpus. Este arquivo pode ser acessado através do seguinte link no Git: [GIT LINK].

Padronizações

Para assegurar a consistência nas quantidades, unidades de medida, intervalos e relações numéricas, recorreu-se à implementação de expressões regulares especificadas nas entradas 83–89 da Tabela 3.11, além de outras equivalentes para diferentes unidades de medida. Essa medida visou unificar o tratamento desses elementos, garantindo que as entidades numéricas incluíssem todas as informações essenciais. Essa abordagem também ajudou a superar as limitações da ontologia simples adotada neste estudo, baseada na estrutura sujeito-relação-objeto.

Adicionalmente, houve uma substituição de símbolos matemáticos por suas respectivas descrições em texto, convertendo "<", ">", "≤", "≥" em "**below**", "**above**", "**at most**", "**at least**", respectivamente. Tal estratégia teve como finalidade aproximar as expressões de relações matemáticas à linguagem natural, facilitando a compreensão.

Durante o processo manual de aplicação das expressões regulares, identificaram-se e corrigiram-se erros oriundos de fases anteriores do trabalho. Essas correções contribuíram significativamente para aprimorar a qualidade e a precisão do corpus desenvolvido.

Desambiguação

Um procedimento através do qual se pode tornar as relações de um GC mais precisas é a ambiguação, que consiste em .

3.8 Ingestão dos dados

A Langchain, a qual foi inicialmente utilizada, tendo os documentos sido carregados com o carregador padrão disponível *PyPDFDirectoryLoader*, que internamente utiliza a biblioteca *PyPDF*, a mesma que no benchmark apresentado

na Tabela 3.10. Como todos os carregadores próprios utilizam alguma das bibliotecas do Benchmark, utilizou-se a metodologia descrita nas seção 3.4-3.6.

A ingestão de dados são feitas através de objetos da classe *Node* do módulo *llama_index.core.schema*, que são instanciados a partir de:

```
metadata = {
    "source": f_name,
    "block_size": Settings.chunk_size,
    "size": size,
    "start": i + 1,
    "end": i,
}
node = Node(text=content, metadata=metadata)
```

F(9)

Onde, *f_name* é o nome do arquivo, que representa a origem do bloco de texto. O *block_size* é o tamanho do bloco de texto que o modelo está preparado para processar, enquanto que o *size* é o comprimento do bloco de texto que está de fato sendo enviado para o modelo. O *start* e o *finish* são os índices, no CSV, do início e final do bloco de texto que compõem o *Node*, respectivamente.

Na biblioteca *LlamaIndex*, há duas classes: *Node* e *Document*. Um objeto *Document* pode ser instanciado como os mesmos parâmetros indicados no F(9), *text* e *metadata*. Compreender a diferença entre eles é crucial para trabalhar e entender suas informações indexadas.

- **Document:** Representam toda a peça de informação que você está indexando - uma página da web, um parágrafo de um artigo de pesquisa, um e-mail ou até mesmo uma transcrição de áudio. Um documento também pode conter metadados específicos de toda a fonte, como seu URL ou título.
- **Node:** Esses são os blocos de construção. São partes menores e mais granulares de informação extraídas do documento original. Imagine dividir um livro em capítulos ou parágrafos. Cada capítulo ou parágrafo se torna um *Node* no *LlamaIndex*. *Nodes* podem ser de vários tipos, dependendo de como você processa o documento - trechos de texto, imagens, fragmentos de código e assim por diante.

Principais Diferenças:

- **Tamanho:** *Documents* são as entidades maiores, enquanto *Nodes* são os trechos extraídos.
- **Conteúdo:** *Documents* contêm metadados gerais da fonte, enquanto nós possuem conteúdo e metadados específicos de sua informação extraída.
- **Relacionamentos:** *Documents* geralmente não possuem conexões internas, enquanto *Nodes* podem descrever como se relacionam com outros nós dentro do documento (por exemplo, frase anterior, parágrafo seguinte).

Assim, o objeto mostrado no F(9) se presta a alimentar o *LLM* com blocos de texto, de forma que, através dos metadados, seja possível rastrear o texto a sua origem no *corpus.csv*. Isso se faz necessário para que, em casos de eventuais falhas ao longo da indexação dos conjunto de dados, seja

Sem superposição

Concatena-se o texto de cada uma das linhas do bloco designado, de forma que o conteúdo de cada bloco esteja única e exclusivamente presente em si.

Com superposição

Concatena-se o texto de cada uma das linhas do bloco designado, de forma que o conteúdo de um dado bloco contenha as *N* últimas linhas do bloco anterior, onde *N* é um parâmetro ajustável.

Ao contrário dos carregadores de documentos padrão fornecidos pela *Langchain* (Document Loaders) e *LLamaIndex* (Understanding Loading), o sistema aqui construído requer uma abordagem mais personalizada para a ingestão de dados para atender à forma como os dados aqui foram extraídos, tratados e com base no que se quer extrair deles. Os carregadores integrados, embora eficientes para propósitos gerais, não oferecem a flexibilidade necessária para nosso gerenciamento específico de tamanhos de blocos e enriquecimento de metadados.

Essa abordagem personalizada permite controlar precisamente o fluxo de dados para o *LLM*, garantindo que cada pedaço de texto seja apropriadamente

processado. Ao contornar as limitações dos carregadores integrados, podemos otimizar nosso processo de ingestão de dados para um melhor desempenho e análises mais perspicazes.

Uma inconveniência da biblioteca *LLamaIndex* é a de não permitir a ingestão dos dados de forma paralela, o que reduziria em muitas vezes o consumo e processamento dos dados.

3.9 Construção do GC

Prompt

O fator que inviabilizou esta abordagem foi a dificuldade de construir um único prompt que fosse capaz de processar todo o conteúdo aqui analisado.

NaLLM

O que impossibilitou seu uso foi o preço da API para os modelos *gpt-3.5-turbo* e *gpt-4* e a imprevisibilidade do custo total do processamento de todo o conjunto de dados. Isto se deve a dificuldade de tratar possíveis erros que podem ocorrer ao longo do processamento e ser necessário reiniciar todo o procedimento repetidamente.

spaCy

O exemplo apresentado na Figura 2.6 e 2.7 evidencia a dificuldade da biblioteca *spaCy* em processar os dados deste trabalho.

Na Figura 2.6, tem-se que no contexto dado, ***enveloped*** e ***single-stranded*** são empregados como adjetivos, porém classificados como verbos. Este é apenas um exemplo não complexo, dentre outros, presente no conjunto de dados, cujo *spaCy* não conseguiu processar satisfatoriamente.

Com dependências e NERs satisfatoriamente determinadas, Figura 2.7, pode-se derivar o GC, com a ontologia sujeito-verbo-objeto, a partir do objeto derivado. Este GC ainda poderia ter sua semântica enriquecida através da adição de informações aos sujeitos, verbos e objetos; tudo isso com base no nas classes e dependências agnósticas até palavras.

KnowledgeGraphIndex

O *KnowledgeGraphIndex* é o objeto que gerencia as duas estruturas índice e o GC. Nesta seção fala-se do GC criado através de sua inicialização.

O modo mais simples para a sua inicialização é mostrada em (Knowledge Graph Demo).

O modo mais simples consiste alimentar o *KnowledgeGraphIndex* com documentos previamente carregados, como descrito em (Knowledge Graph Demo), de onde se tira o Fragmento 3:

```
from llama_index.core import StorageContext

graph_store = SimpleGraphStore()
storage_context =
StorageContext.from_defaults(graph_store=graph_store)

# NOTE: can take a while!
index = KnowledgeGraphIndex.from_documents(
    documents,
    max_triplets_per_chunk=2,
    storage_context=storage_context,
)
```

F(10)

Dado que o modelo aqui utilizado foi o *gemini-1.0-pro*, esta etapa foi limitada pelos chamados filtros de segurança da Google, o que pôde ser parcialmente pela abordagem alternativa.

Filtros de Segurança

Os filtros de segurança impedem o processamento de texto que usam palavras que a Google proíbe. Estes contemplam quatro categorias, as quais, segundo o próprio Gemini (<https://gemini.google.com/u/1/app/b9a57638035257bd>) são:

- **Discurso de ódio:** É qualquer expressão que ataque uma pessoa ou grupo com base em características como raça, religião, etnia, orientação sexual, deficiência ou gênero. Exemplos: Insultos racistas, homofóbicos, misogínicos ou propaganda discriminatória.
- **Conteúdo sexualmente explícito:** É qualquer conteúdo que retrate atividades sexuais de forma crua ou gráfica. Isso inclui imagens, vídeos e textos com descrições detalhadas de atos sexuais.

- **Assédio:** É qualquer comportamento que vise intimidar, atacar ou constranger outra pessoa. Pode incluir xingamentos, ameaças, perseguição virtual e a disseminação de boatos maliciosos.
- **Conteúdo perigoso:** É qualquer conteúdo que possa causar danos físicos ou emocionais a si mesmo ou a terceiros. Exemplos: Instruções para fabricar armas ou bombas, incentivo ao suicídio ou automutilação, glorificação da violência e promoção de atividades ilegais.

Para exemplificar e esclarecer como e porquê destes erros, apresenta-se algumas ocorrências encontradas ao longo da construção do GC e as respectivas justificativas segundo o próprio *gemini*.

category: HARM_CATEGORY_SEXUALLY_EXPLICIT	probability: HIGH	
innate and adaptive immunity are being influenced by sex hormones, which may explain the observed interaction of sex and 25ohd levels with respiratory disease mortality. according to data from the us-american national health and nutrition survey, women have a higher inflammation burden than men (the age range was 40-90+ years).		S(7)

Justificativa, dada pelo próprio modelo:

O filtro de segurança provavelmente classificou o texto como "categoria: HARM_CATEGORY_SEXUALLY_EXPLICIT" porque ele menciona hormônios sexuais e carga inflamatória, o que pode ser mal interpretado como conteúdo sexualmente sugestivo. Esses filtros geralmente dependem de palavras-chave e frases para identificar conteúdo potencialmente prejudicial e, neste caso, o filtro pode não ser sofisticado o suficiente para entender o contexto científico.

Veja porque o filtro pode estar interpretando mal o texto:

- **Hormônios sexuais:** Essa frase é um gatilho para o filtro porque os hormônios sexuais estão associados à reprodução e desenvolvimento sexual.
- **Carga inflamatória:** Embora a inflamação seja um processo biológico, o filtro pode interpretar mal "carga" como um termo sugestivo.

É importante lembrar que os filtros de segurança não são perfeitos e às vezes podem cometer erros. Nesse caso, o filtro está priorizando a segurança em vez da precisão, e é melhor prevenir do que remediar quando se trata de conteúdo potencialmente prejudicial.

Aqui estão alguns pontos adicionais a serem considerados:

- O texto que você forneceu discute um estudo científico sobre como os hormônios sexuais influenciam a imunidade. Esse contexto está totalmente ausente da análise do filtro.
- Os filtros de segurança estão constantemente sendo aprimorados, mas ainda possuem limitações.

Se você acredita que o conteúdo é benigno e foi sinalizado incorretamente, normalmente você pode denunciar a classificação incorreta para a plataforma onde o encontrou.

category: HARM_CATEGORY_HATE_SPEECH	probability: HIGH
table 1. general, physical and health characteristics of older adults with vitamin d sufficiency and insufficiency. limeira, piracicaba, and campinas, state of são paulo, 2018-2019. regarding skin color, reduced efficiency of vitamin d synthesis in people with dark skin is well described in the literature, to the greater amount of melanin, which acts as a natural sunscreen. it explains the higher prevalence of vitamin d insufficiency found among individuals reporting more pigmented skin color ethnicities.	

S(8)

Justificativa, dada pelo próprio modelo:

O filtro de segurança provavelmente classificou o texto como "categoria: HARM_CATEGORY_HATE_SPEECH" (discurso de ódio) devido à menção de "cor da pele" e "etnias." Infelizmente, alguns filtros de segurança básicos podem interpretar erroneamente discussões sobre raça e etnia como discurso de ódio.

No entanto, o contexto científico da passagem é totalmente ignorado pelo filtro. O texto está simplesmente explicando a relação entre a cor da pele, a síntese de vitamina D e a deficiência de vitamina D.

É importante ressaltar que o discurso de ódio se concentra em atacar e discriminar grupos de pessoas. A passagem fornecida está discutindo uma questão biológica de forma factual.

category: HARM_CATEGORY_HARASSMENT	probability: MEDIUM
in general, there are marked health disparities between white and african americans, where the mortality rate of multiple etiologies (cardiovascular and renal disease, diabetes, cancer, and others) for african americans is significantly higher than for white americans. some explanations for such disparities generally include social environment, lifestyle behaviors, socioeconomic status, and access to preventive health care services.	

S(9)

Justificativa, dada pelo próprio modelo:

No contexto de filtros de segurança, é improvável que um filtro bem-funcionamento categorize o texto fornecido como assédio (HARM_CATEGORY_HARASSMENT) com uma probabilidade MÉDIA. Aqui porque:

- **Ausência de Ataque Direcionado:** Assédio normalmente envolve atacar uma pessoa ou grupo específico com linguagem negativa ou abusiva. O texto discute disparidades de saúde entre raças de forma geral, sem destacar nenhum indivíduo ou grupo.

- **Tom Informativo:** A linguagem é factual e objetiva, focando em apresentar informações sobre disparidades de saúde e potenciais causas. Não há hostilidade ou intenção de rebaixar nenhum grupo.

No entanto, pode haver alguns motivos pelos quais um filtro menos sofisticado possa classificá-lo incorretamente:

- **Tema Sensível:** Raça pode ser um tema sensível, e alguns filtros básicos podem sinalizar qualquer menção a questões raciais como potencial assédio, mesmo que o contexto seja neutro.
- **Palavras-chave:** O filtro pode interpretar mal palavras como "disparidades" ou "taxa de mortalidade mais alta" como indicativo de negatividade ou culpa, levando a um falso positivo.

Aqui estão alguns pontos adicionais a serem considerados:

- Os filtros de segurança estão constantemente evoluindo, e sua capacidade de entender o contexto está melhorando.
- Se você se deparar com uma classificação incorreta, denuncie-a para a plataforma. Você pode explicar porque o conteúdo é inofensivo e não deve ser sinalizado.

No geral, o texto fornecido tem maior probabilidade de ser categorizado como neutro ou informativo por um filtro de segurança bem projetado.

category: HARM_CATEGORY_DANGEROUS_CONTENT	probability: MEDIUM
<p>for all study subjects, peripheral venous blood samples were collected, by a qualified lab technician using a 5cc automated syringe with a needle, for measuring 25ohd levels. blood collected from the camp was transported to a single accredited laboratory in cool packs until they were separated and stored for further analysis.</p>	

S(10)

Justificativa, dada pelo próprio modelo:

No contexto de filtros de segurança, é bem improvável que o texto fornecido seja categorizado como conteúdo perigoso (HARM_CATEGORY_DANGEROUS_CONTENT) com uma probabilidade MÉDIA. Aqui porque:

- **Procedimento médico comum:** Coletar amostras de sangue é um procedimento médico rotineiro realizado por profissionais qualificados. O texto menciona especificamente um "técnico de laboratório qualificado".
- **Finalidade médica clara:** O texto explica que o sangue está sendo coletado para medir os níveis de 25ohd, o que indica um contexto de pesquisa médica.
- **Armazenamento seguro:** O texto menciona o transporte adequado do sangue em refrigeradores e armazenamento seguro para análise posterior.

Filtros de segurança bem elaborados reconheceriam esses detalhes e classificariam o texto como informativo ou neutro.

Agora, vejamos porque o filtro poderia ter te dado uma classificação média de conteúdo perigoso:

- **Menção de agulhas e sangue:** Embora comum na área médica, termos como "agulha" e "sangue" podem ser alarmantes para alguns filtros básicos.
- **Falta de contexto:** Sem entender o contexto da pesquisa, o filtro pode interpretar mal o procedimento como potencialmente perigoso.

Aqui estão alguns pontos adicionais para se atentar:

- Os filtros de segurança estão sempre aprendendo e melhorando na diferenciação de conteúdo médico de conteúdo perigoso.
- Se você se deparar com uma classificação equivocada, denuncie-a para a plataforma esclarecendo o contexto médico do texto.

A substituição das palavras que causam o erro (palavras problemáticas), por seus respectivos sinônimos aceitáveis, nas Saídas 1-3 tem como efeito o processamento do texto bem sucedido do texto. Contudo, esta estratégia não foi

utilizada pois sua aplicação exige um pós processamento não facilmente reprodutível.

Ainda que esta estratégia não tenha sido aqui empregada, convém mencioná-la, passo a passo:

1. Criar o GC através do F(3)
2. Fazer um levantamento do erros ocorridos
3. Para cada erro ocorrido, identificar a palavra problemática
4. Para cada palavra problemática, substitui-se-la e executar o passo 1 novamente
5. Repetir o passo 4 até que o erro deixe de ocorrer e o resultado seja satisfatório

Existem algumas heurísticas e ressalvas interessantes a respeito dos passos 3 e 4.

Heurísticas

1. Passo 3: Por vezes, é mais rápido obter dicas de sinônimos aceitáveis perguntando para o próprio *gemini*.
2. Passo 3: É possível substituir a palavra problemática, por uma sequência de caracteres não contida no corpus e posteriormente substituí-la (no GC criado) pela palavra original.

Ressalvas

Na Heurística 2, os resultados da extração das tríades pode se comportar de tal forma a ignorar as tríades esperadas; isto porque o contexto muda. Quanto ao Passo 4, substituir uma dada palavra ao longo de todo o corpus pode ser perigoso, não somente devido a eventuais detecções errôneas, mas principalmente à não adequação do contexto.

Voltando à construção do GC

O disparo dos erros acima mencionados exigiu que a abordagem alternativa fosse capaz de dividir os trecho alimentados ao modelo recursivamente para que

estes pudessem ser processados de forma satisfatória ou que, para casos de impossibilidade de processamento satisfatório do texto, a extração de tríades continuasse sem interferência até o final.

A implementação desta solução é técnica e consiste em utilizar o modelo escolhido para extrair as tríades do texto e inseri-las no GC a posteriori. Para tanto, foi necessário customizar a função extratora de tríades, que é utilizada internamente pelo *KnowledgeGraphIndex*.

É necessário mencionar que foi necessário fazer alterações ao código fonte da *LlamaIndex*. Estas alterações tiveram o objetivo de possibilitar o tratamento de diversos erros durante a construção do GC. Mais precisamente, esta alteração foi feita no método `_extract_triplets` da classe *KnowledgeGraphIndex*, e incluiu o parâmetro `metadata`, através do qual fez-se a redução recursiva dos *Documents* ingeridos.

3.4 Interação com a IA

A proposta do trabalho é compor um assistente, através do qual descobertas são feitas.

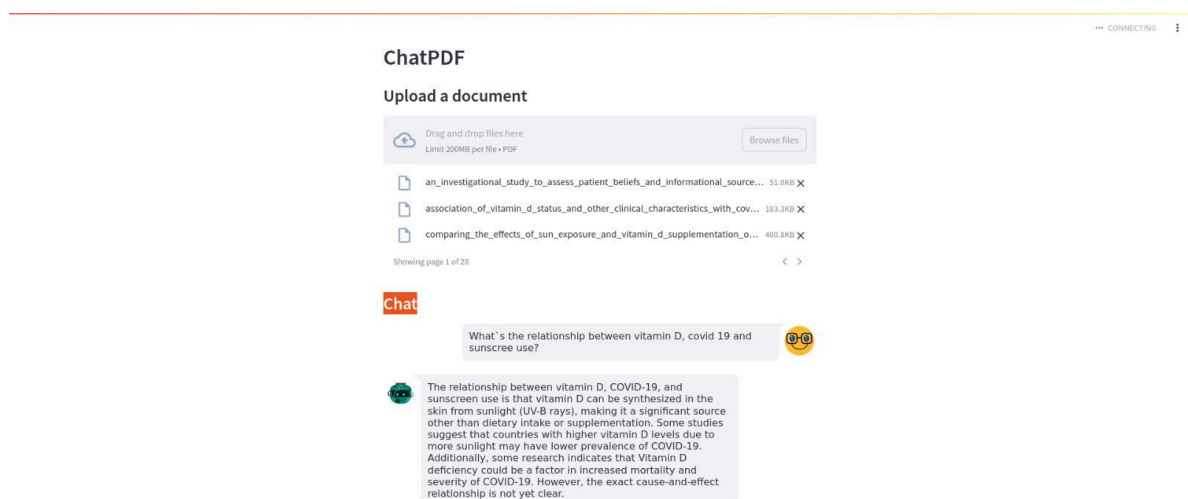


Figura 3.19: Exemplo da interação pergunta e resposta.

A Figura 3.9 mostra a interação com o assistente virtual, o ícone com *smiley* representa as perguntas do usuário e a IA, com ícone de Robô. Esta prova de

conceito foi feita utilizando o recurso 3 da Tabela 2.4 e alimentada apenas uma porção dos arquivos baixados, 28 no total. Este resultado foi obtido utilizando bases de dados vetoriais.

3.5 Metodologia

A metodologia funciona, com ressalvas nas capacidades computacionais e funcionais dos *LLMs* adotados. O GC torna a racionalização mais factual e verídica. O fator mais limitante desta abordagem reside na ontologia adotada e na qualidade da etapa de extração.

3.6 Empregabilidade

LLMs maiores e mais complexos, embora capazes de gerar textos mais influentes e criativos, não desempenham bem na tarefa de referenciar textos. Isso pode dificultar a tarefa de navegar entre todos os ensaios clínicos, de forma que selecionar dois ensaios clínicos específicos e compará-los é desafiador, posando como limitação dos *LLMs*.

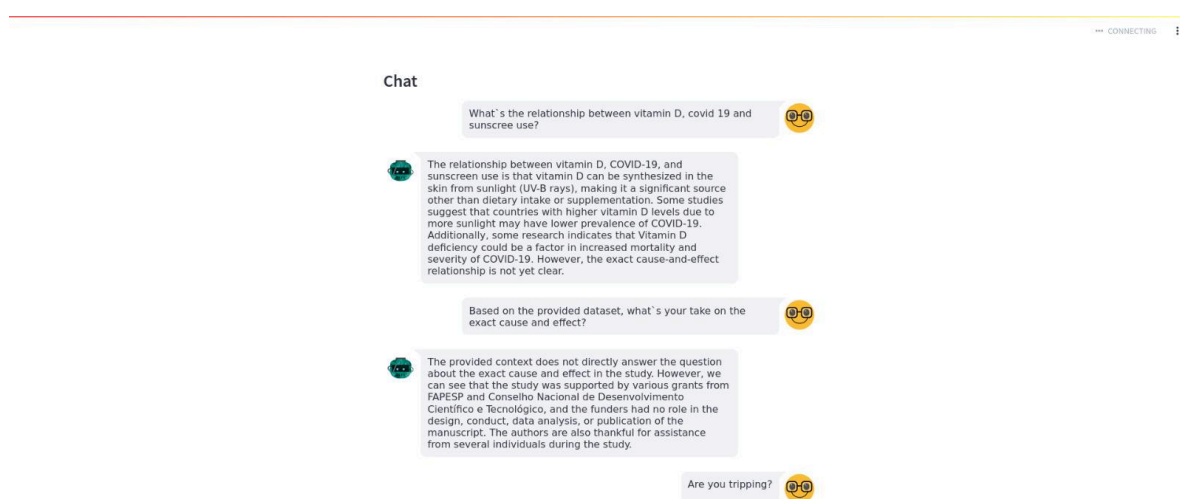


Figura 3.21: Performance em referenciamento de texto.

Como mostrado na Figura 3.21, o *LLM* não performou bem na referência do das evidências que suportam a argumentação de respostas. Este resultado foi obtido utilizando bases de dados vetoriais.

Esta limitação pode ser contornada através do tratamento das referências a outros trabalhos e a ensaios clínicos como conceitos ou criando uma estrutura apropriada específica para o entendimento de ensaios clínicos destacados do texto, isto é, com uma ontologia própria.

3.7 Modelos e Pipelines

Os *LLMs* se destacam em reconhecer padrões no texto, mas nem sempre compreendem o significado ou a lógica subjacente. Isso pode resultar em respostas aparentemente coerentes, mas semanticamente imprecisas ou implausíveis.

A forma como o *prompt* é formulado influencia significativamente a resposta do *LLM*. *Prompts* imprecisos, ambíguos ou excessivamente complexos podem desencadear alucinações de demorar excessivamente.

Os Pipelines da John Snow Labs performam no geral 3s por bloco enquanto que o *mistral* com o *Prompt* (1) leva 12s para processar um bloco de 512 caracteres sem sobreposição. Aumentando o tamanho do bloco, espera-se um aumento não linear do tempo de processamento.

Uma questão a se considerar é que os pipelines da John Snow Labs focam em tarefas específicas, já o *Prompt* (1) é geral apesar de emular um conhecimento nos domínios indicados para a interpretação do texto. Assim, os modelos do John Labs extraem menos informações.

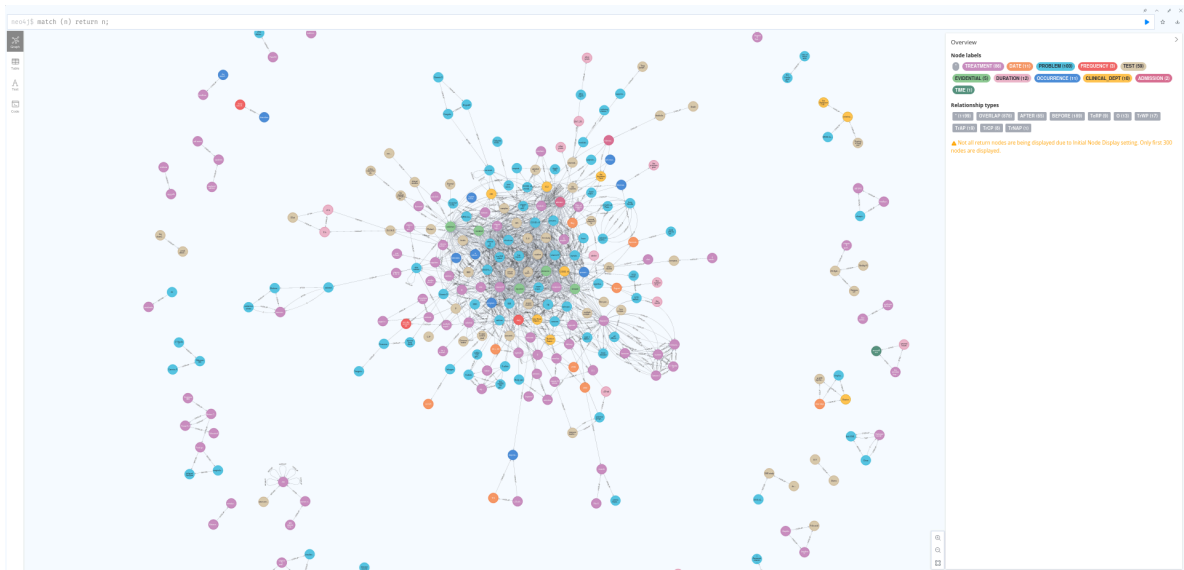


Figura 3.22: Grafo de Conhecimento Produzido pelos *Pipelines* da John Snow Labs.

A Figura 3.22 mostra uma imagem do GC criado apenas com as entidades e relações oriundas dos *pipelines* da John Snow Labs no Neo4j. Nota-se um número pequeno de tipos de relações, isso se deve às relações que o pipeline consegue entender.

3.8 Generalização

Esta abordagem agnóstica pode ser reutilizada para qualquer outro domínio. No entanto, como a etapa de extração de entidades e relações é específica ao domínio, se faz necessária a utilização de modelos específicos ou de *prompts* minuciosamente produzidos para a execução bem sucedida da tarefa proposta.

Capítulo 4 - Conclusões

Este estudo teve como objetivo investigar a relação entre COVID-19, vitamina D e o uso de protetor, que ainda não é totalmente compreendida. E para tanto, utilizou da inteligência artificial desde o pré-processamento dos dados até a extração da semântica de publicações científicas e obtenção de resultados de estudos que possam contribuir para o desenvolvimento de estratégias para o entendimento de relações diretas e indiretas dos elementos supracitados.

A implementação da abordagem *RAG* específica, personalizada com diferentes parâmetros e componentes, foi conduzida de maneira a descobrir a melhor configuração por meio de *benchmarks*. Os experimentos apontam para uma capacidade sem precedentes para a extração de conhecimento, consultas e perguntas e respostas em geral. O *KnowledgeGraphIndex*, foi capaz de buscar e extrair informações relevantes sobre a interação entre Vitamina D e a severidade do COVID-19.

A construção e validação de GCs são essenciais e foram rigorosamente analisados. De forma que as direções entre as entidades condizem com as relações presentes no texto. A performance dos motores de consulta foram avaliados para cada experimento e a conclusão é que a escolha do melhor motor de consulta vai depender da complexidade da pergunta. Para perguntas complexas, a recuperação de entidades baseada em vetores consegue responder perguntas complexas, porém perde precisão caso o não haja uma resposta clara no GC.

A implementação de uma *UI* interativa com um assistente virtual permite que os usuários explorem e compreendam as relações descobertas de maneira acessível e compreensível. O Assistente Virtual, coordenando todos os passos subjacentes à geração de respostas, fornece uma plataforma interativa e informativa para o público em geral.

Por fim, a compartilhamento do conhecimento, incluindo o código fonte em um repositório público no GitHub (https://github.com/vriez/doctor_rag), visa estimular a colaboração e permitir que outros pesquisadores e interessados contribuam para o avanço do entendimento sobre essas complexas interconexões entre o uso de protetor solar, a deficiência de vitamina D e a dinâmica do COVID-19.

Sendo assim, a metodologia utilizada mostrou-se bem sucedida e promissora, uma vez que proporciona uma compreensão aumentada do pesquisador, graças à inteligência artificial e à produção científica humana.

Perspectivas Futuras

Dadas as limitações das bibliotecas de manipulação de PDFs, especificamente no processo de extração do conteúdo de PDFs, trabalhos que visam interfacear componentes de detecção da ordem lógica de texto, utilizando redes neurais em grafos. Uma contribuição convergente é a compilação de um conjunto de dados contendo PDFs de publicações curados e anotados de forma a expressar a ordem natural do texto, especificamente.

Já existem conjuntos de dados para processamento de PDFs, como por exemplo o *PubLayNet* e *PubTabNet*, que contém anotações relativas ao layout e a tabelas respectivamente. No entanto, não se pode treinar um modelo para aprender a inferir a ordem lógica do texto a partir destes.

Uma Ontologia mais apropriada para a representação de tabelas, ensaios clínicos e métricas estatísticas seria de grande valor, uma vez disponibilizadas os dados e metadados de tabelas no GC e no index, permitindo fazer perguntas, comparações e manipulações mais avançadas e específicas.

Referências

1. Acesso CAFe. Disponível em:
<https://www-periodicos-capes-gov-br.ez1.periodicos.capes.gov.br/index.php/acesso-cafe.html>. Acesso em: 16 jan. 2024.
2. American Academy of Dermatology. (2022). Estatísticas sobre vitamina D. Disponível em: <https://www.aad.org/media/stats-vitamin-d>. Acesso em: 27 jan. 2024.
3. American Academy of Dermatology. (2022). Estatísticas sobre protetor solar. Disponível em: <https://www.aad.org/media/stats-sunscreen>. Acesso: 27 jan. 2024.
4. Araújo, L. A. D., Veloso, C. F., Souza, M. D. C., Azevedo, J. M. C. D., & Tarro, G. (2021). The potential impact of the COVID-19 pandemic on child growth and development: a systematic review. *Jornal de pediatria*, 97, 369-377.
5. Bikle, D. (2015). Vitamin D: production, metabolism, and mechanisms of action.
6. Cambaza, E. M. (2022). COVID-19, SARS e MERS: semelhanças e diferenças. *Revista Internacional em Língua Portuguesa*, (42), 27-34.
7. Cao, T. T., Zhang, G. Q., Pellegrini, E., Zhao, Q., Li, J., Luo, L. J., & Pan, H. Q. (2021). COVID-19 and its effects on the digestive system. *World Journal of Gastroenterology*, 27(24), 3502.
8. DATANAMI. What's the Vector, Victor? Disponível em:
<https://www.datanami.com/2023/12/15/whats-the-vector-victor/>. Acesso em: 25 jan. 2024.
9. DataVersity. (2022). A Brief History of Natural Language Processing (NLP). Disponível em:
<https://www.dataversity.net/a-brief-history-of-natural-language-processing-nlp/>. Acesso em: 28 jan. 2024.
10. DataVersity. (2023). A Brief History of Large Language Models. Disponível em:
<https://www.dataversity.net/a-brief-history-of-large-language-models/>. Acesso em: 28 jan. 2024.
11. Fensel D, Şimşek U, Angele K, Huaman E, Kärle E, Panasiuk O, et al. Introduction: what is a knowledge graph? *Knowledge graphs: Methodology, tools and selected use cases*. 1-10; 2020.

12. Greve, J. M. D. A., Brech, G. C., Quintana, M., Soares, A. L. D. S., & Alonso, A. C. (2020). Impacts of COVID-19 on the immune, neuromuscular, and musculoskeletal systems and rehabilitation. *Revista Brasileira de Medicina do Esporte*, 26, 285-288.
13. Gupta A, et al. Vitamina D na COVID-19: Uma revisão sistemática e meta-análise. *Nutrients*. 2021;13(11):3736.
14. Harvard T.H. Chan School of Public Health. (2022). Vitamin D. Disponível em: <https://www.hsph.harvard.edu/nutritionsource/vitamin-d/>. Acesso em: 28 jan. 2024.
15. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*; 2023.
16. Latha, M. S., Martis, J., Shobha, V., Shinde, R. S., Bangera, S., Krishnankutty, B., ... & Kumar, B. N. (2013). Sunscreening agents: a review. *The Journal of clinical and aesthetic dermatology*, 6(1), 16.
17. Lee, D. (2021). The impact of COVID-19 on human reproduction and directions for fertility treatment during the pandemic. *Clinical and Experimental Reproductive Medicine*, 48(4), 273.
18. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*. 33:9459-9474; 2020.
19. MedlinePlus. (2022). Vitamin D deficiency. Disponível em: <https://medlineplus.gov/vitamindeficiency.html>. Acesso em: 03 jan. 2024.
20. Ministério da Saúde. Coronavírus (COVID-19): Sintomas. Disponível em: <https://www.gov.br/saude/pt-br/assuntos/coronavirus/sintomas>. Acesso em: 28 jan. 2024.
21. Ministério da Saúde. (Ano não disponível). Coronavírus (COVID-19): Transmissão. Disponível em: <https://www.gov.br/saude/pt-br/assuntos/coronavirus/transmissao>. Acesso em: 28 jan. 2023.
22. National Institutes of Health (NIH). (2022). Vitamin D - Consumer Fact Sheet. Disponível em: <https://ods.od.nih.gov/factsheets/VitaminD-Consumer/>. Acesso em: 03 jan. 2024.

23. Our World in Data. (2022). A brief history of AI. Disponível em: <https://ourworldindata.org/brief-history-of-ai>. Acesso em: 28 jan. 2024.
24. PAHO - Pan American Health Organization. (2022). Histórico da pandemia COVID-19. Disponível em: <https://www.paho.org/pt/covid19/historico-da-pandemia-covid-19>. Acesso em: 28 jan. 2024
25. Ponsonby AL, et al. Exposição solar e status de vitamina D: Uma revisão sistemática. *Am J Clin Nutr.* 2006;84(3):678-688.
26. Roberts, E. (s.d.). A Brief History of Neural Networks. Disponível em: <https://cs.stanford.edu/people/eroberts/courses/soco/projects/neural-networks/History/history1.html>. Acesso em: 28 jan. 2024.
27. Sander, M., Sander, M., Burbidge, T., & Beecker, J. (2020). The efficacy and safety of sunscreen use for the prevention of skin cancer. *Cmaj*, 192(50), E1802-E1808.
28. Santalucia. 2020. Quanto tempo leva para curar um paciente com COVID-19? Disponível em: <http://www.santalucia.com.br/noticias/quanto-tempo-leva-para-curar-um-paciente-com-covid-19/>. Acesso em: 28 jan. 2024
29. Shen, T., & Wang, T. (2021). Metabolic reprogramming in COVID-19. *International journal of molecular sciences*, 22(21), 11475.
30. Szczerbiński, Ł., Okruszko, M. A., Szablowski, M., Solomacha, S., Sowa, P., Kiszkiel, Ł., ... & Kamiński, K. (2023). Long-term effects of COVID-19 on the endocrine system—a pilot case-control study. *Frontiers in Endocrinology*, 14.
31. Typology. (2022). SPF 20, 30, 50: What Does This Mean? Disponível em: <https://uk.typology.com/library/spf-20-30-50-what-does-this-mean>. Acesso em: 28 jan. 2024.
32. UNICEF. Duas em cada cinco escolas em todo o mundo não tinham instalações básicas para lavar as mãos antes da pandemia. Disponível em: <https://www.unicef.org/brazil/comunicados-de-imprensa/duas-em-cada-cinco-e-scolas-em-todo-o-mundo-nao-tinham-instalacoes-basicas-para-lavar-maos-antes-pandemia>. Acesso em: 28 jan. 2024.
33. Valentine, Christina J., Aiman Q. Khan, Alexandra R. Brown, Scott A. Sands, Emily A. Defranco, Byron J. Gajewski, Susan E. Carlson, Kristina M. Reber,

- and Lynette K. Rogers. 2021. "Higher-Dose DHA Supplementation Modulates Immune Responses in Pregnancy and Is Associated with Decreased Preterm Birth" *Nutrients* 13, no. 12: 4248. <https://doi.org/10.3390/nu13124248>.
34. Zapier. (2023). AI Hallucinations: What They Are and Why They Happen. Disponível em: <https://zapier.com/blog/ai-hallucinations/>. Acesso em: 28 jan. 2024.
35. Williams CY, Miao BY, Butte AJ. Evaluating the use of GPT-3.5-turbo to provide clinical recommendations in the Emergency Department. *medRxiv*. 2023-10.
36. "Knowledge Graph Demo," LlamaIndex Documentation. Disponível em: https://docs.llamaindex.ai/en/stable/examples/index_structs/knowledge_graph/KnowledgeGraphDemo/. Acesso em: 26 mar. 2024.
37. "Knowledge Graph RAG Query Engine," LlamaIndex Documentation. Disponível em: https://docs.llamaindex.ai/en/stable/examples/query_engine/knowledge_graph_rag_query_engine/. Acesso em: 26 mar. 2024.
38. "Knowledge Graph Query Engine," LlamaIndex Documentation. Disponível em: https://docs.llamaindex.ai/en/stable/examples/query_engine/knowledge_graph_query_engine/. Acesso em: 26 mar. 2024.
39. PyPDF2 Documentation
PYPDF2. PyPDF2 Documentation. 3.0.0. Disponível em: <https://pypdf2.readthedocs.io/en/3.0.0/>. Acesso em: 8 fev. 2024.
40. PDFMiner.six Documentation
PDFMINER.SIX. Documentation. [S.l.: s.n.], [s.d.]. Disponível em: <https://pdfminersix.readthedocs.io/en/latest/>. Acesso em: 8 fev. 2024.
41. Apache Tika
APACHE TIKA. Apache Tika. [S.l.: s.n.], [s.d.]. Disponível em: <https://tika.apache.org/>. Acesso em: 8 fev. 2024

42. PyMuPDF Documentation
PYMUPDF. Documentation. [S.l.: s.n.], [s.d.]. Disponível em:
<https://pymupdf.readthedocs.io/en/latest/>. Acesso em: 8 fev. 2024.
43. pdfplumber on PyPI
PDFPLUMBER. pdfplumber. [S.l.: s.n.], [s.d.]. Disponível em:
<https://pypi.org/project/pdfplumber/>. Acesso em: 8 fev. 2024.
44. Tabula-py Documentation PDF
TABULA-PY. Tabula-py documentation. [S.l.: s.n.], [s.d.]. Disponível em:
<https://readthedocs.org/projects/tabula-py/downloads/pdf/latest/>. Acesso em: 8 fev. 2024.
45. PDF Text Extraction by G. Stavrakis on GitHub
STAVRAKIS, G. PDF Text Extraction. [S.l.: s.n.], [s.d.]. Disponível em:
https://github.com/g-stavrakis/PDF_Text_Extraction. Acesso em: 8 fev. 2024.
46. pdftotext on PyPI
PDFTOTEXT. pdftotext. [S.l.: s.n.], [s.d.]. Disponível em:
<https://pypi.org/project/pdftotext/>. Acesso em: 8 fev. 2024.
47. GOOGLE. **Gemini API for developers in Cloud**. *Google Blog*, Dec 13, 2023.
Disponível em: <https://blog.google/technology/ai/gemini-api-developers-cloud/>.
Acesso em: 12 fev. 2024.
48. SEO, Sumin et al. Structural Quality Metrics to Evaluate Knowledge Graph Quality.
49. PRADEEP. A comparison of Python libraries for PDF data extraction for text, images, and tables. Medium, year of publication. Disponível em:
<https://pradeepundefned.medium.com/a-comparison-of-python-libraries-for-pdf-data-extraction-for-text-images-and-tables-c75e5dbcfef8>. Acesso em: 26 mar. 2024.
50. STATISTICS CANADA. PDF extraction. Government of Canada, year of publication. Disponível em:
<https://www.statcan.gc.ca/en/data-science/network/pdf-extraction>. Acesso em: 26 mar. 2024.
51. SAVE EMAILS AS PDF. PDF file format: internal document structure explained. Year of publication. Disponível em:
<https://www.save-emails-as-pdf.com/news/pdf-file-format-internal-document-structure-explained/>. Acesso em: 26 mar. 2024.
52. INFOSEC INSTITUTE. Hacking PDF file format: basic structure. Year of publication. Disponível em:

<https://www.infosecinstitute.com/resources/hacking/pdf-file-format-basic-structure/>. Acesso em: 26 mar. 2024.

53. Document Loaders - Data Connection. Disponível em:

https://python.langchain.com/docs/modules/data_connection/document_loaders. Acesso em: 1 apr. 2024.

54. Understanding Loading. Disponível em:

<https://docs.llamaindex.ai/en/stable/understanding/loading/loading/>. Acesso em: 1 apr. 2024.