UNIVERSIDADE DO SUL DE SANTA CATARINA

GIOVANNI SOARES LOBO

JOHN DE LARA

BUSCA SEMÂNTICA EM DOCUMENTOS DA ÁREA BIOMÉDICA RELACIONADOS COM COVID-19

Florianópolis

2021

GIOVANNI SOARES LOBO

JOHN DE LARA

BUSCA SEMÂNTICA EM DOCUMENTOS DA ÁREA BIOMÉDICA RELACIONADOS COM COVID-19

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Sistemas de Informação, da Universidade do Sul de Santa Catarina, como requisito parcial à obtenção do título de Bacharel em Sistemas de Informação.

Orientador: Prof. Flávio Ceci, Dr.

Florianópolis

2021

GIOVANNI SOARES LOBO

JOHN DE LARA

BUSCA SEMÂNTICA EM DOCUMENTOS DA ÁREA BIOMÉDICA RELACIONADOS COM COVID-19

Este Trabalho de Conclusão de Curso foi julgado adequado à obtenção do título de Bacharel em Sistemas de Informação e aprovado em sua forma final pelo Curso de Sistemas de Informação da Universidade do Sul de Santa Catarina.

Florianópolis, 04 de dezembro de 2021.

_____

Professor e orientador Flávio Ceci, Dr./
Universidade do Sul de Santa Catarina

_____

Professor Aran Bey Tcholakian Morales, Dr./
Universidade do Sul de Santa Catarina

_____

Professor Roberto Fabiano Fernandes, Dr./
Universidade do Sul de Santa Catarina

AGRADECIMENTOS

Primeiro agradecemos aos nossos familiares queridos que sempre nos deram muito apoio e tiveram muita paciência, permitindo-nos estar envolvidos por longas horas neste projeto.

Segundo agradecemos muito ao nosso orientador, professor Dr. Flávio Ceci, excelente mestre e modelo de sabedoria que nos serviu de inspiração para corrermos atrás de solução neste grande desafio, principalmente quando tínhamos que nos manter distanciados por conta da pandemia.

Finalmente a todos os nossos professores da Unisul, em especial à professora Dra. Vera Schuhmacher, que nos conduziram com dedicação por uma trilha de aprendizado durante todo nosso curso.

“A mente que se abre a uma nova idéia jamais volta ao seu tamanho original” (Oliver Wendell Holmes).

RESUMO

Sistemas de busca de informação tem grande importância na área biomédica, na qual os repositórios de publicações científicas armazenam volumes gigantescos de documentos. Junto aos desafios globais apresentados pelo Covid-19 somou-se a necessidade de criar alternativas eficientes para recuperar informação útil no universo de publicações geradas sobre esta doença. Entender a necessidade de informação de um usuário que realiza uma consulta em um sistema de recuperação de informação e selecionar, a partir de índices que tentam sintetizar o conteúdo de cada documento, aqueles documentos que têm mais relevância para a necessidade do usuário não é uma tarefa fácil. Diversos modelos matemáticos podem ser utilizados nas máquinas de busca. Entender o contexto dos termos nos documentos é muito importante para realizar uma busca por significado. Dentre as representações do conhecimento no domínio biomédico, as ontologias podem ser muito úteis para aumentar a eficiência de buscadores semânticos. Neste projeto desenvolvemos um buscador semântico para recuperar documentos em um corpus específico para Covid-19, o CORD-19, utilizando para isso o apoio de uma representação do conhecimento simplificada e a biblioteca python spaCy para recuperar entidades nomeadas nos documentos. Os índices invertidos foram criados a partir de uma imagem do Elasticsearch instanciada no Google Cloud Plataform. A fim de criar vetores densos de representação para análise de similaridade contextual foi utilizada o *framework* de modelo de linguagem pré-treinado sentence-transformers. A funcionalidade de busca semântica do projeto permitiu a recuperação de um número maior de documentos relevantes. Concluindo, a utilização de representações do conhecimento no domínio biomédico associada a modelos de linguagem pré-treinados que permitem associar a análise de contexto a nível de palavras e sentenças podem aumentar a relevância dos resultados de um sistema de recuperação de informação na área biomédica.

Palavras-chave: Busca semântica. Recuperação da informação. Biomédico. Covid-19. CORD-19. Índice invertido. Entidades nomeadas. SpaCy. Similaridade. Sentence-transformers.

ABSTRACT

Information search systems are of great importance in the biomedical area, in which the repositories of scientific publications store gigantic volumes of documents. In addition to the global challenges presented by Covid-19, there was also the need to create efficient alternatives to retrieve useful information from the universe of publications generated about this disease. Understanding the information need of a user who performs a query in an information retrieval system and selecting, from indexes that try to synthesize the content of each document, those documents that are most relevant to the user's need is not an easy task. Several mathematical models can be used in search engines. Understanding the context of terms in documents is very important to perform a search for meaning. Among the representations of knowledge in the biomedical domain, ontologies can be very useful to increase the efficiency of semantic searchers. In this project we developed a semantic search engine to retrieve documents in a specific corpus for Covid-19, CORD-19, using for this the support of a simplified knowledge representation and the python spaCy library to retrieve named entities in documents. The inverted indexes were created from an Elasticsearch image instantiated in Google Cloud Plataform. In order to create dense representation vectors for contextual similarity analysis, the sentence-transformers pre-trained language model framework was used. The project's semantic search functionality allowed the retrieval of a larger number of relevant documents. In conclusion, the use of knowledge representations in the biomedical domain associated with pre-trained language models that allow associating context analysis at the level of words and sentences can increase the relevance of the results of an information retrieval system in the biomedical area.

Keywords: Semantic search. Information retrieval. Biomedical. Covid-19. CORD-19. Inverted index. Named entities. SpaCy. Similarity. Sentence-transformers.

LISTA DE FIGURAS

Figura 1 – Componentes de um sistema de recuperação da informação.....19
Figura 2 – Exemplo de índice invertido gerado a partir de um documento.....20
Figura 3 – Modelos de recuperação de informação conforme tipo de arquivo.....21
Figura 4 – Pilha de tecnologias da *web* semântica.....27
Figura 5 – Exemplo de RDF no formato grados (esquerda) e em XML (direita).....28
Figura 6 – Página de busca do PubMed.....32
Figura 7 – Estruturas centrais do *framework* spaCy.....35
Figura 8 – Pipeline de processamento dos componentes da classe Language do spaCy.....35
Figura 9 – Arquitetura de um transformer.....38
Figura 10 – Fluxo de trabalho para extração conhecimento em *Big Data* sobre Covid-19.....39
Figura 11 – Visão de desenvolvimento do buscador semântico.....45
Figura 12 – Tela da API REST do Elasticsearch.....46
Figura 13 – Representação do conhecimento utilizada para Covid-19.....48
Figura 14 – Máquina virtual no GCP com imagem do Elasticsearch.....49
Figura 15 – Tela inicial do IRCO-19.....50
Figura 16 – Tipo de busca no IRCO-19.....50
Figura 17 – Entidades nomeadas anotadas em um documento.....51
Figura 18 – Quadros com nuvem de palavras para Labels e Entities.....52
Figura 19 – Página de resultados da busca clássica.....53
Figura 20 – Página de resultados da busca com recurso de sentence-transformer.....55
Figura 21 – Página de resultados da busca com recurso de sentence-transformer e filtragem.55

LISTA DE TABELAS

Tabela 1 – Score e validação por especialista dos 20 documentos mais relevantes para a busca no modelo clássico …….. 53

Tabela 2 – Score e validação por especialista dos 20 documentos mais relevantes para a consulta do buscador semântico……55

Tabela 3 - Distribuição das categorias de validação do documento para o modelo clássico e para o buscador semântico……..….56

LISTA DE SIGLAS

BERT - Bidirectional Encoder Representations from Transformers

GCP – Google Cloud Plataform

IRCO-19 – Information Retrieval in Covid-19

MeSH - Medical Subject Headings

NCBI - National Center for Biotechnology Information

RDF - Resource Description Framework

RDFS - Resource Description Framework Schema

SIR – Sistema de Recuperação da Informação

UMLS - Unified Medical Language System

URI – Uniform Resource Identifier

SUMÁRIO

1 INTRODUÇÃO.....13
1.1 PROBLEMÁTICA.....14
1.2 OBJETIVOS.....15
1.2.1 Objetivo geral.....15
1.2.2 Objetivos específicos.....15
1.3 JUSTIFICATIVA.....15
1.4 ESTRUTURA DA MONOGRAFIA.....16
2 REVISÃO DA LITERATURA.....18
2.1 COMPONENTES DE UM SISTEMA DE RECUPERAÇÃO.....18
2.2 MODELOS DE RECUPERAÇÃO DA INFORMAÇÃO.....20
2.2.1 Modelo booleano.....22
2.2.2 Modelo espaço-vetorial.....22
2.2.3 Modelo probabilístico.....23
2.3 RECURSOS COMPLEMENTARES PARA RECUPERAÇÃO DA INFORMAÇÃO.....23
2.3.1 Reformulação e expansão de consultas.....23
2.3.2 Filtragem de informação.....24
2.3.3 Categorização.....24
2.3.4 Extração de informação.....24
2.3.5 Pipelines de busca.....25
2.4 RECUPERAÇÃO DE INFORMAÇÃO NA INTERNET.....25
2.5 WEB SEMÂNTICA.....26
2.6 REPRESENTAÇÃO DE DOMÍNIOS.....29
2.6.1 Ontologias biomédicas.....30
2.7 REPOSITÓRIOS DE DOCUMENTOS BIOMÉDICOS.....32
2.8 CORD-19: REPOSITÓRIO BIOMÉDICO PARA COVID-19.....33
2.9 RECURSOS PARA MODELOS DE BUSCA SEMÂNTICA NO CORD-19.....33
2.9.1 Anotação e classificação de documentos.....33
2.9.2 Text embeddings.....36
2.9.3 Grafos de conhecimento.....36
2.9.4 Modelos de linguagem pré-treinados.....37
2.10 BUSCA SEMÂNTICA EM BIG DATA SOBRE COVID-19.....38
2.11 SOLUÇÕES BIOMÉDICAS PARA RECUPERAÇÃO DA INFORMAÇÃO.....40

3 METODOLOGIA.....42
3.1 CARACTERIZAÇÃO DO TIPO DE PESQUISA.....42
3.2 ETAPAS METODOLÓGICAS.....43
3.2.1 Estudo bibliográfico.....43
3.2.2 Modelagem do sistema.....43
3.2.3 Implementação.....43
3.2.4 Avaliação.....44
3.3 DELIMITAÇÕES.....44
4 PROPOSTA DE BUSCADOR SEMÂNTICO PARA DOCUMENTOS COVID-19….45
4.1 ARQUITETURA.....45
4.2 FERRAMENTAS UTILIZADAS.....46
4.2.1 Elasticsearch.....46
4.2.2 spaCy.....47
4.2.3 Google Cloud Plataform.....48
4.2.4 Streamlit.....49
4.3 INTERFACE DE USUÁRIO.....49
4.4 SIMULAÇÃO DE BUSCA.....52
4.4.1 Busca clássica.....53
4.4.2 Busca semântica.....54
4.5 AVALIAÇÃO DOS RESULTADOS.....57
4.6 LIMITAÇÕES DO SISTEMA.....58
5 CONCLUSÃO.....60
REFERÊNCIAS.....63

1 INTRODUÇÃO

A exponencial evolução tecnológica que vivemos, a ubiquidade na digitalização de conteúdos e a interconexão crescente de uma ampla variedade de serviços e dispositivos eletrônicos, principalmente nas últimas décadas, tem impulsionado a geração, a coleta e o armazenamento de um massivo volume de dados em diversos setores. Segundo Wu (2016), embora o termo *Big Data* seja muito genérico e tenha sua origem não muito certa, ele tem sido amplamente utilizado para definir este novo paradigma e engloba atributos que devem ser considerados no desenvolvimento de tecnologias para buscar valor nesta grande quantidade de dados, entre eles a velocidade, o volume, a variedade, a veracidade, a visibilidade e o valor dos dados. Segundo Luo et al. (2016, p.1, tradução nossa), "o ecossistema do *big data* biomédico abrange muitos níveis diferentes de fontes de dados para criar uma matriz rica de dados para pesquisadores." Várias são as bases de conhecimento na área biomédica, incluindo bases de literatura específicas da área, bases de imagens médicas, repositórios de genes humanos, bases de dados clínicos de pacientes obtidos a partir de prontuários eletrônicos, bases de proteínas humanas e bases bioquímicas. "Integrar diferentes fontes de informação permite aos clínicos descobrir uma nova visão no processo de cuidado dos pacientes que considera o estado holístico do paciente, do genoma ao comportamento" (LUO et al., 2016, p.9).

Conforme Andrade et al. (2009, p.2), "sistemas de recuperação da informação tem o propósito de recuperar documentos para satisfazer as demandas do usuário." Porém, segundo Koopman et al. (2012), buscas em bases médicas apresentam alguns desafios específicos, como a necessidade de inferência de alguns relacionamentos de contextos e o intervalo semântico. O torrencial volume de conteúdo desenvolvido para compreender e buscar soluções para a crise global causada pelo Covid-19, ao mesmo tempo que sobrecarregou ainda mais os recursos analíticos disponíveis, exigiu ainda mais urgência no desenvolvimento de sistemas eficientes de recuperação de informação que tratem adequadamente as questões semânticas deste domínio.

A combinação de diversos métodos de mineração e processamento de textos utilizando busca semântica com reconhecimento de entidades nomeadas biomédicas vem sendo utilizada como arquitetura em diversos *frameworks* para recuperação de informação em bases de conhecimento na área biomédica (ANDRADE et al.,2009).

1.1 PROBLEMÁTICA

Na área biomédica há diversas bases de conhecimento de texto completo que têm indexados milhões de artigos científicos publicados, estando estes subdivididos em diversas subcategorias relacionadas com subáreas do conhecimento biomédico. Conforme observado em outros contextos de *Big Data*, a maior parte de seus dados encontra-se em forma não estruturada ou está organizada em estruturas heterogêneas. Para Coneglian et al. (2018, p.143), "[…] alcançar a interoperabilidade semântica nestes casos pode ser considerado um grande problema[…]".

Desde o final de 2019 o mundo passou por mudanças e consequências extremas devido a uma nova doença viral que ficou conhecida como Covid-19. A urgência em entender o contexto do Covid-19 e de buscar formas de tratamento e prevenção impulsionou um número enorme de publicações científicas a taxas sem precedentes, estimada na ordem de centenas de novos artigos ciêntificos relacionados apenas a esta doença por dia, impondo, desta forma, um grande desafio para a comunidade científica em utilizar métodos tradicionais de leitura, fazendo-se mandatório desenvolver e incorporar métodos mais efetivos para mineração de conteúdo e recuperação de informação relevante (WANG; LO, 2021).

Segundo Bahammam (2020), uma simples pesquisa sobre a doença "Covid-19" coletou na base de artigos científicos médicos PubMed cerca de 28.000 artigos publicados em apenas cerca de 6 meses. Contudo, a maioria destas bases de conhecimento da área biomédica fornece um mecanismo próprio convencional de busca através de máquinas de busca que utilizam apenas a indexação de seus conteúdos a partir de palavras-chave. Conforme demonstrado por Sievert et al. (1991), a comparação de buscas que utilizam máquinas de busca convencionais baseadas em indexação por frequência de aparecimento de palavras-chave com outras metodologias de busca, como por exemplo, com aquelas que utilizam técnicas de processamento de linguagem natural para acrescentar uma busca semântica, há uma maior sensibilidade destes mecanismos que utilizam relacionamentos semânticos normalizados para ontologias especializadas. Para Beppler (2008 *apud* CECI,2015, p.89), "[…] uma ontologia deve representar um conhecimento comum (compartilhado) de maneira consistente e não ambígua". Dessa forma, a ontologia adotada para recuperação da informação poderá influenciar as inferências e os resultados solicitados pelas buscas.

O reconhecimento de entidades nomeadas representa uma das metodologias de pesquisa mais utilizadas atualmente no processamento automático de artigos científicos biomédicos (GOULART; STRUBE DE LIMA; XAVIER, 2011). Porém, segundo Sasaki et

al. (2008, p.1), “quando a ambiguidade e variabilidade de termos é muito alta, o reconhecimento de entidades nomeadas a partir de dicionários não é uma solução ideal até mesmo quando recursos de terminologia em larga escala estão disponíveis.” Por esta razão, nomes de proteínas são extremamente difíceis de serem reconhecidos (SASAKI et al., 2008).

Diante da necessidade de recuperar informação de forma efetiva em bases de conhecimento biomédico que crescem exponencialmente em volume de informação, principalmente no contexto do Covid-19, surge a dúvida de como pode ser estruturado um sistema de busca baseado em reconhecimento de entidades nomeadas biomédicas e que utilize recursos para encontrar relacionamentos semâticos?

1.2 OBJETIVOS

A possibilidade de gerar conhecimento valioso utilizando-se de diversas técnicas de inteligência artificial, incluindo a transferência de conhecimento de diferentes domínios, a partir do cruzamento de informações contidas em documentos de texto em diferentes bases de dados da área biomédica, sustenta a necessidade de criar uma forma eficiente de recuperar informação a partir de textos que possuam uma semântica normalizada por uma ontologia.

1.2.1 Objetivo geral

Desenvolver um protótipo funcional de busca semântica em documentos da área biomédica relacionados com o Covid-19.

1.2.2 Objetivos específicos

Revisar o conceito estrutural de um buscador semântico;

Compreender as taxonomias e ontologias da área médica;

Estudar métodos para anotação semântica de documentos;

Desenvolver uma proposta de solução para recuperação de documentos relacionados a Covid-19.

1.3 JUSTIFICATIVA

Segundo Luo *et al.* (2016), embora o *Big Data* represente uma promessa significante para a melhora global da qualidade dos cuidados em saúde, vários desafios se

impõem à utilização deste potencial, dentre eles, a dificuldade para integração de bases de conhecimento. A recuperação de informação de várias bases com correlação semântica e de forma desambiguada representa a primeira etapa deste processo de aproveitamento do conhecimento acumulado em diferentes bases de conhecimento biomédicas.

Várias técnicas de recuperação da informação podem ser utilizadas para ajudar os pesquisadores a localizarem informação relevante em grandes repositórios de documentos ou bases de conhecimento biomédicos. Segundo Wei, Barnaghi e Bargiela (2007), as técnicas tradicionais de recuperação de informação de documentos são baseadas sobretudo na computação de estatísticas das palavras e frequentemente negligenciam a semântica das *queries* realizadas nas buscas de uma determinada informação, não explorando assim, o significado e o contexto das necessidades dos usuários. Ainda para Wei, Barnaghi e Bargiela (2007), modelos de recuperação de informação que utilizam busca semântica em coleções de documentos são mais assertivos para a maior parte das aplicações. Segundo Bast, Buchhold e Haussmann (2016, p.119, tradução nossa), "[…] busca semântica é uma 'busca com significado'."

O reconhecimento de entidades nomeadas biomédicas é uma área de grande interesse do processamento de linguagem natural e que tem sido implementado em diversas arquiteturas de recuperação da informação na área biomédica (GOULART; STRUBE DE LIMA; XAVIER, 2011).

Dessa forma, faz-se necessário desenvolver um sistema que utilize busca semântica em documentos de bases biomédicas para posterior utilização em heurísticas de geração de valor para suporte à decisão clínica, proporcionando a integração do conhecimento obtido em documentos correlacionados.

1.4 ESTRUTURA DA MONOGRAFIA

Este trabalho é composto por cinco capítulos. Neste primeiro capítulo apresentamos, além da introdução do trabalho com uma abordagem inicial do tema, os objetivos gerais e específicos desta pesquisa, a problemática, a sua justificativa, além da estrutura da monografia.

No segundo capítulo abordamos a revisão bibliográfica, abrangendo todo o embasamento teórico necessário para o entendimento e desenvolvimento do presente projeto. Foram abordados conceitos fundamentais da recuperação da informação, uma breve síntese das bases de dados biomédicos disponíveis, incluindo o CORD-19, as bases para sistemas de

recuperação da informação que utilizam busca semântica, uma breve explicação sobre o papel da representação do conhecimento através de ontologias, além da apresentação de modelos de arquiteturas existentes para sistemas de busca semântica no domínio biomédico, mais especificamente para aqueles relacionados ao Covid-19.

No terceiro capítulo é apresentada a metodologia empregada neste projeto, desde as considerações práticas sobre as ferramentas apresentadas na literatura, a apresentação das etapas para estruturação e implementação deste projeto, até o desenvolvimento de um sistema de busca semântica no domínio biomédico relacionado ao Covid-19.

No quarto capítulo é apresentada uma solução que representa um produto de valor mínimo para um buscador semântico em documentos sobre Covid-19, utilizando uma ontologia específica.

Finalmente, no quinto capítulo, conclui-se o presente trabalho com as observações finais dos autores sobre os resultados obtidos e apresentados, além de ser apresentada uma breve análise sobre possíveis aperfeiçoamentos futuros deste projeto.

2 REVISÃO DA LITERATURA

Sistemas de recuperação de informação podem ser considerados os elementos intermediadores entre coleções de dados diversos armazenados e o usuário que almeja extrair ou derivar informações a partir dos dados, representando as informações através dos processos de indexação e descrição dos documentos bem como permitindo a recuperação direta de determinada informação dos próprios documentos ou a partir de outros tipos de bases de dados (SOUZA, 2006).

Para ser efetivo e eficiente no propósito de entregar ao usuário a informação que ele espera receber, um SRI deve processar o conteúdo de cada um dos documentos de uma coleção e apresentar ao usuário uma lista de documentos ordenados de acordo com a relevância, à consulta do usuário, enquanto deve ainda evitar recuperar documentos que não sejam relevantes e executar esta tarefa com o menor esforço computacional possível. (BAEZA-YATES; RIBEIRO-NETO, 2011).

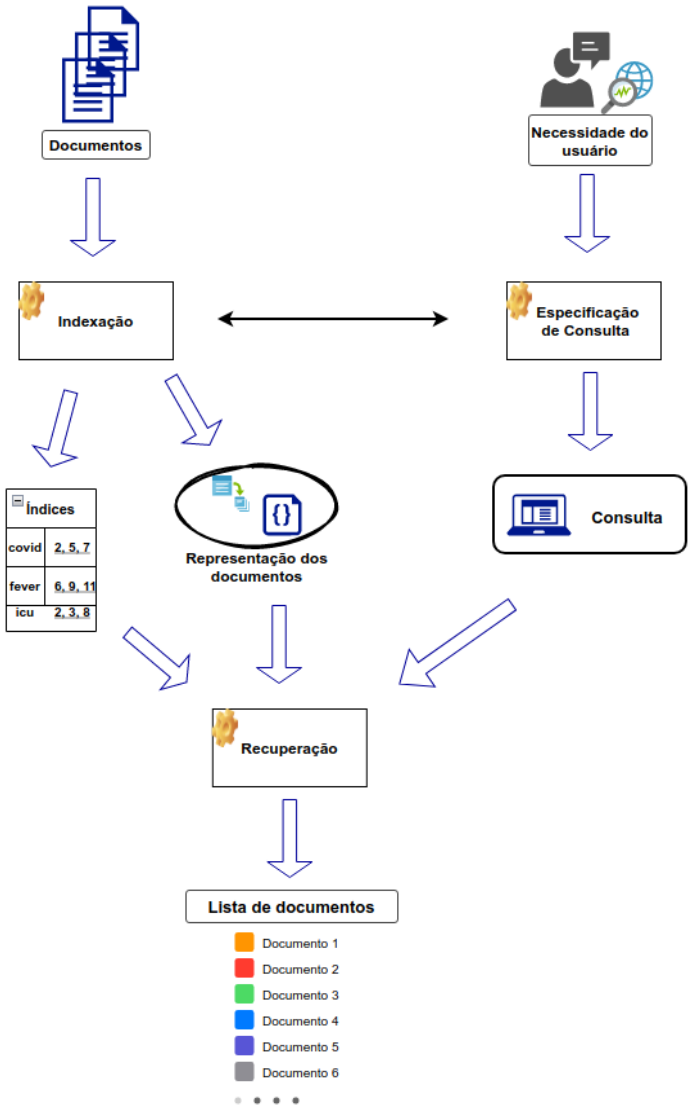
Contudo, a relevância atribuída aos documentos varia de usuário para usuário de acordo com a necessidade individual de informação, mesmo quando utilizam uma mesma especificação de consulta.

Outro ponto a ser levado em consideração é que um usuário de um SRI, na busca de resolver uma necessidade de informação, não busca recuperar apenas documentos que contenham os termos da busca, propósito este dos gerenciadores de banco de dados que funcionam com sistemas recuperadores de dados e trabalham sobre um esquema que possui uma estrutura e semântica bem definidas (BAEZA-YATES; RIBEIRO-NETO, 2011). O SRI, por sua vez, trabalhará com documentos não estruturados, representados por texto em linguagem natural, que carrega uma complexidade sintática e semântica intrínsecas.

2.1 COMPONENTES DE UM SISTEMA DE RECUPERAÇÃO

Basicamente um SRI é composto por uma base de dados (documentos), por uma pesquisa do usuário que representa sua necessidade e do processo de recuperação propriamente dito, que a partir da consulta formulada (*query*) e de uma estrutura de dados derivada dos documentos, recupera uma lista dos documentos classificada em ordem decrescente, de acordo com o nível de similaridade entre o documento e a consulta (CARDOSO, 2004). A figura 1 demonstra o relacionamento dos componentes de um SRI.

Figura 1 – Componentes de um sistema de recuperação da informação



Fonte: Adaptado de Cardoso (2004, p. 1).

De acordo com Frakes (1992 *apud* CARDOSO,2004, p.33) "a ferramenta mais importante para auxiliar o processo de recuperação é denominado índice, que é uma coleção de termos que indicam o local onde a informação desejada pode ser localizada", e estes devem ser adequadamente implementados para acelerar a busca sem ocupar muito espaço.

Um índice invertido é uma estrutura de dados composta de um vocabulário, que é o conjunto de todas as diferentes palavras do texto, e de suas ocorrências, permitindo que um texto seja reconstruído a partir deste índice, razão pela qual é chamado de índice invertido (BAEZA-YATES; RIBEIRO-NETO, 2011).

A representação de um documento a partir de um índice invertido (figura 2) pode ser realizada a partir de uma matriz cujas células demonstram para cada termo de toda a coleção o

número de vezes que aparece por documento (matriz termo-documento) ou por intermédio de uma lista invertida dos documentos para cada palavra (BAEZA-YATES; RIBEIRO-NETO, 2011). O problema da matriz termo-documento é que ela ocupa um espaço proporcional ao número de termos multiplicado pelo número de documentos.

Figura 2 – Exemplo de índice invertido gerado a partir de um documento

Vocabulário	n	d1	d2	d3	d4	d5	d6
covid-19	6	5	7	6	5	10	2
é	2	-	2	-	-	2	-
doença	5	3	5	4	1	6	-
perigosa	1	-	-	-	1	-	-
complica	3	-	2	1	-	2	-
penumonia	3	-	1	-	-	1	1
quando	2	-	-	1	-	-	1
tosse	3	1	2	-	-	3	-
frequente	1	-	-	-	1	1	-

Fonte: Criado pelos autores (2021).

Para poder realizar consultas de frases e poder identificar a proximidade dos termos de um documento, há necessidade de adicionar nos índices as posições de cada palavra nos textos, sendo estes chamados de índices invertidos totais (BAEZA-YATES; RIBEIRO-NETO, 2011).

O índice invertido pode ser representado por uma árvore binária para reduzir o tempo de pesquisa, entretanto, neste tipo de estrutura, há um aumento da necessidade de espaço em memória para armazenamento (AMAZONAS et al., 2008).

Segundo Souza (2006), a ordenação dos documentos em sistemas automatizados de recuperação da informação é executada por diversos tipos de algoritmos baseados em heurísticas previamente definidas que fazem uma ponderação da relevância presumida destes documentos.

2.2 MODELOS DE RECUPERAÇÃO DA INFORMAÇÃO

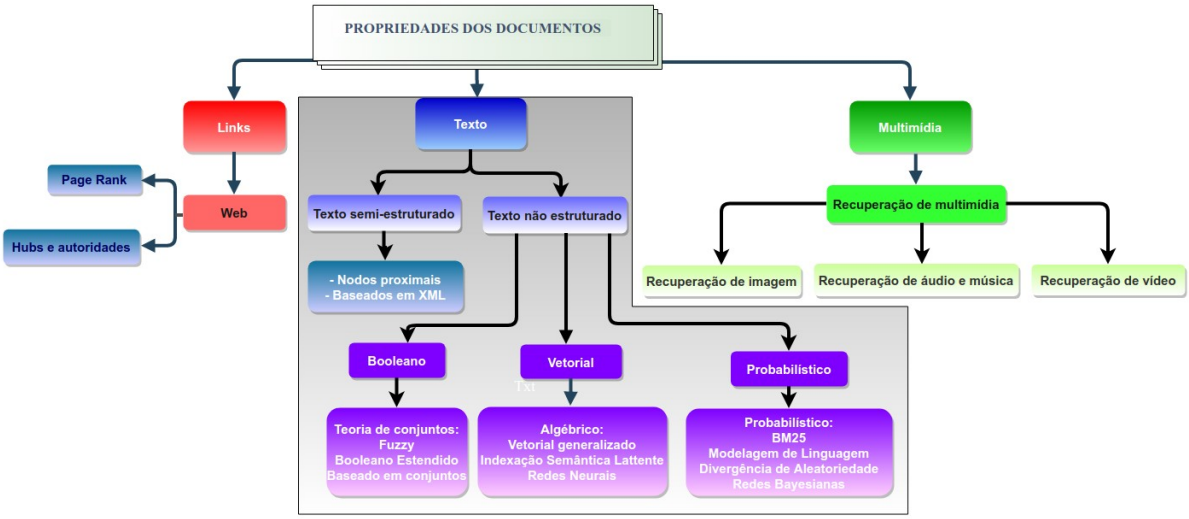
Um modelo de recuperação da informação pode ser definido como um *framework* matemático que define o princípio do processo de busca (SINGH et al., 2017).

Os sistemas de recuperação podem ser divididos em modelos clássicos e modelos estruturados. Sobre estes tipos de modelos Souza esclarece (2006, p.166):

> Nos modelos clássicos, cada documento é descrito por um conjunto de palavras-chave representativas – também chamadas de termos de indexação – que busca representar o assunto do documento e sumarizar seu conteúdo de forma significativa. Nos modelos estruturados, podem-se especificar, além das palavras-chave, algumas informações acerca da estrutura do texto (como seções a serem pesquisadas, fontes de letras, proximidade das palavras, entre outras informações.).

Segundo Baeza-Yates e Ribeiro-Neto (2011), os modelos clássicos de recuperação da informação para textos não-estruturados utilizam algoritmos booleanos, vetoriais e probabilísticos, sendo o texto modelado puramente como uma sequência de palavras, enquanto para textos semi-estruturados, que utilizam métodos de indexação baseados por exemplo em XML, os componentes do modelo incluem os seus títulos, seções e parágrafos. Para textos da web faz-se necessário utilizar informação contida nos links das páginas para ranquear uma consulta. Objetos multimídia (imagens, áudio, música, vídeo), que também podem ser pesquisados para recuperação, são codificados de forma diferente dos textos, exigindo assim modelos diferentes para recuperação, conforme pode ser visto na Figura 3 (BAEZA-YATES; RIBEIRO-NETO, 2011).

Figura 3 – Modelos de recuperação de informação conforme tipo de arquivo



Fonte: Adaptado de Baeza-Yates e Ribeiro-Neto (2011).

Contudo, para Gonzalez e Lima (2003), a classificação dos modelos de recuperação de informação é incerta e limitada, uma vez que, geralmente, diversos dispositivos e estratégias são combinados em um sistema.

Segundo Cardoso (2004), existem modelos mais avançados de recuperação de informação, dentre estes destacam-se os modelos fundamentados em bases de conhecimento, lógica *fuzzi* e redes neurais.

2.2.1 Modelo booleano

Neste modelo são recuperados todos os documentos que possuem as combinações de palavras que satisfazem uma consulta realizada a partir de uma expressão lógica booleana convencional contendo os conectivos lógicos AND, OR e NOT (CARDOSO, 2004). Ainda segundo Cardoso (2004), uma das vantagens deste modelo é a sua facilidade de implementação e entre as desvantagens estão a ausência de ordem na resposta, além do fato das respostas poderem ser nulas ou muito grandes. Outros problemas relacionados a este modelo são sua ineficiência para usuários que não sejam treinados a formular consultas utilizando operadores lógicos e o fato de ter como retorno basicamente uma partição da coleção de documentos em dois subconjuntos discretos: aqueles que satisfazem a consulta e os que não satisfazem (GONZALEZ; LIMA, 2003).

2.2.2 Modelo espaço-vetorial

No modelo espaço-vetorial os documentos devolvidos a partir de uma consulta são resultantes de um cálculo de similaridade aos termos da consulta, sendo estes cálculos representados por vetores (CARDOSO, 2004).

Para Gonzalez e Lima (2003, p.23), "o estabelecimento de pesos para os termos é um dos mais importantes fatores a determinar a efetividade de um sistema de RI […]". Os pesos quantificam a relevância de cada termo para as consultas e para os documentos (CARDOSO, 2004). Ainda segundo Cardoso (2004), o peso de um termo para um documento ($W_{t,d}$) pode ser calculado pela fórmula $W_{t,d} = tf * idf$, onde *tf* é a frequência do termo no documento e *idf* (*inverse document frequency*) é a razão entre o número de documentos no corpus e a quantidade de documentos que possuem o termo.

"Aos termos das consultas e documentos são atribuídos pesos que especificam o tamanho e a direção de seu vetor de representação. Ao ângulo formado por estes vetores dá-se o nome de θ. O cos θ determina a proximidade da ocorrência (CARDOSO, 2004, p.35)." Neste modelo, "a determinação do grau de relevância pode ser feita com base na distância vetorial entre os documentos e as consultas mapeadas no espaço *n*-dimensional" (AMAZONAS et al, 2008, p.198), sendo *n* o número de palavras contidas em todos os documentos que possuem representatividade no índice.

2.2.3 Modelo probabilístico

"O modelo probabilístico descreve documentos considerando pesos binários que representam a presença ou ausência de termos. O vetor resultado gerado pelo modelo tem como base o cálculo da probabilidade de que um documento seja relevante para uma consulta" (CARDOSO, 2004, p.35).

Para Souza (2006), o valor deste modelo está nas iterações contínuas do usuário com o modelo para buscar o conjunto ideal de atributos dos documentos recuperados, sendo a consulta inicial considerada uma estimativa inicial destas propriedades através de uma descrição probabilística. As desvantagens deste modelo é que ele depende da precisão das estimativas de probabilidade, além de não explorar a frequência dos termos nos documentos (CARDOSO, 2004).

2.3 RECURSOS COMPLEMENTARES PARA RECUPERAÇÃO DA INFORMAÇÃO

A especialização de um SRI para um buscador semântico envolve o entendimento de recursos complementares que podem ser utilizados na arquitetura de um SRI.

2.3.1 Reformulação e expansão de consultas

A dificuldade do usuário compreender a lógica de um SRI e por conseguinte em construir uma consulta assertiva prejudica muito o desempenho dos sistemas.

Após realizada uma consulta pelo usuário e um conjunto de documentos seja recuperado, existe uma variedade de técnicas para reformular a consulta ou dar um novo direcionamento para o processo de busca (BAEZA-YATES; RIBEIRO-NETO, 2011).

Duas das técnicas mais importantes de reformulação de consulta consistem em apresentar sugestões de termos com correções gramaticais na interface do usuário ou em fazer uma expansão dos termos por intermédio de sugestões contínuas de acréscimo de termos relacionados (BAEZA-YATES; RIBEIRO-NETO, 2011).

Para Cardoso (2004), uma alternativa seria agregar uma estratégia de realimentação de relevantes (*relevance feedback*), onde após ser retornado um vetor de documentos o usuário seleciona dentre estes os documentos de sua preferência e o sistema então refaz a consulta

utilizando uma fração dos termos mais relevantes contidos nestes documentos selecionados. A expectativa é que a consulta reformulada retorne um número maior de documentos relevantes.

Um ciclo de realimentação de relevância é composto de duas etapas. A primeira consiste em determinar quais termos ou informações dos textos recuperados podem estar relacionados com a consulta inicial e a segunda etapa trata em como reformular a nova consulta (BAEZA-YATES; RIBEIRO-NETO, 2011). A primeira etapa pode ser obtida de forma explícita a partir do próprio usuário ou de forma implícita pelo sistema, que pode utilizar um método automatizado de seleção de termos complementares a partir dos documentos considerados mais relevantes à primeira consulta (análise local) ou a partir de fontes externas, seja um tesauro ou por extração de relacionamentos entre termos nos próprios documentos (análise global).

A realimentação de relevância tem custo computacional alto e também pode apresentar dificuldades do ponto de vista de usabilidade em interfaces padrões (BAEZA-YATES; RIBEIRO-NETO, 2011). Uma alternativa a ela, que tem sido utilizada em sistemas de busca de literatura biomédica, é a geração automática de uma seção de "documentos relacionados".

2.3.2 Filtragem de informação

Estes sistemas agregam de forma estática em seu fluxo de informação tópicos pré-definidos de interesse ao usuário que serão utilizados como filtros na seleção de documentos recuperados de forma dinâmica (CARDOSO, 2004).

2.3.3 Categorização

Sua aplicação consiste em classificar documentos em categorias pré-definidas, que podem ser genéricas ou relacionadas a domínios de conhecimento, para em uma etapa subsequente servirem de parâmetro para filtragem e recuperação da informação.

2.3.4 Extração de informação

A extração de informação compreende um conjunto de técnicas de processamento de linguagem natural utilizadas com o fim de recuperar alguma informação específica dos documentos podendo ignorar o restante dos textos.

2.3.5 Pipelines de busca

"A fim de otimizar a busca de similaridade, um SRI trabalha com representações de documentos e de consultas através de linguagens de representação (de indexação e de consulta) […]" (GONZALEZ; LIMA, 2003).

A este propósito, Gonzalez e Lima (2003) ressaltam que o processo de busca envolve as operações de inspeção, verificação de similaridade, classificação (*scoring*) e saída.

Dentre os recursos que podem ser utilizados para aumentar a eficiência da busca em SRI, Gonzalez e Lima (2003) ressaltam a utilização de clusters de documentos (geralmente pré-classificados), a utilização de diversas estratégias de classificação, geralmente envolvendo mais de uma versão da mesma consulta, a utilização de indexação semântica latente que reduz o número de índices ao considerar interdependências e relacionamentos termo a termo, além de utilização de trechos de texto para busca em bloco nos documentos.

2.4 RECUPERAÇÃO DE INFORMAÇÃO NA INTERNET

Todo o processo de recuperação da informação a partir de textos inicia-se com a busca e aquisição dos documentos ou publicações que, no domínio biomédico, podem ser obtidos a partir de diversas bases bibliográficas disponíveis na *internet* (MIROŃCZUK; PROTASIEWICZ, 2018). Entretanto, no que tange à estrutura dos dados em documentos, estes encontram-se geralmente de forma não-estruturada e não são agrupados de acordo com as palavras-chaves da busca.

No contexto da *internet*, a organização do conhecimento na forma documental tradicional hierarquizada em categorias e subcategorias pré-estabelecidas tornou-se inexequível, dadas às características intrínsecas da *internet* de ausência de centro de significância estruturado e de instrumentos de organização do conhecimento (PICKLER, 2007). Segundo Pickler (2007, p.66), "[…] para recuperar informações disponíveis na Web, os mecanismos de busca valem-se da indexação de palavras (em linguagem natural) presentes nos próprios documentos […]".

No processamento de linguagem natural existe também a possibilidade de que uma palavra apresente diversos significados (polissemia), o que pode gerar ambiguidade de interpretação e, dessa forma, a recuperação de informação realizada pelos mecanismos de busca na internet pode ser prejudicada (PICKLER, 2007). Ainda no contexto da *internet*,

dependendo da base de conhecimento utilizada, uma mesma entidade pode ser representada de maneiras diferentes. A este propósito, Koopman et al. (2015) afirmam que existe naturalmente um desajuste entre o que os dados brutos nos dizem e a forma como os humanos os interpretam e que este intervalo semântico existe em todos os domínios, embora seja particularmente prevalente em pesquisas biomédicas.

Para Fatehi, Gray e Wootton (2013, p.479, tradução nossa) "[…] a busca ordinária na *internet* através do Google é insatisfatória para propósitos científicos porque os resultados incluem cada pedaço de informação digital que está disponível *online* […]."

A este propósito, Santarém Segundo, Coneglian e Lucas (2017, p.302) afirmam que "os motores de busca tradicionais utilizam habitualmente buscas heurísticas ou critérios de relevância durante a apresentação dos resultados aos usuários.

2.5 WEB SEMÂNTICA

A extração de uma informação específica dentre a imensa quantidade de informação da área biomédica disponível na *internet* não é muito efetiva através de buscadores genéricos que utilizam palavras-chave e metadados estruturais associados a heurísticas de categorização (SHEHNAZ; QAMAR; MUZAFFAR, 2016). Conforme Medeiros (2009, p.7), "os metadados estruturais descrevem a organização e estrutura dos dados como formato, tipo de dado e os relacionamentos sintáticos […]" enquanto os metadados semânticos descrevem os significados dos dados e as suas relações semânticas. Para resolver o problema dos buscadores genéricos perderem o relacionamento entre diferentes entidades, os pesquisadores tem utilizado uma metodologia conhecida como *web* semântica.

As tecnologias da *web* semântica podem ser utilizadas tanto na coleta, representação e armazenamento da informação quanto nos processos de inferência e consulta de uma busca semântica em uma base de dados (ROZSA; VIERA; DUTRA, 2019).

Fazzinga et al. (2011) descrevem a *web* semântica como estruturada em várias camadas hierárquicas, sendo uma delas formada por uma ontologia, permitindo com que as máquinas utilizem as informações disponíveis na *web* de forma mais eficiente ao tentar compreender o significado das consultas. Segundo Pickler (2007), por acrescentar a semântica ao atual formato de representação dos dados, gerando um sentido aos termos no contexto de um determinado documento, a *web* semântica permite trabalhar com associação e dedução.

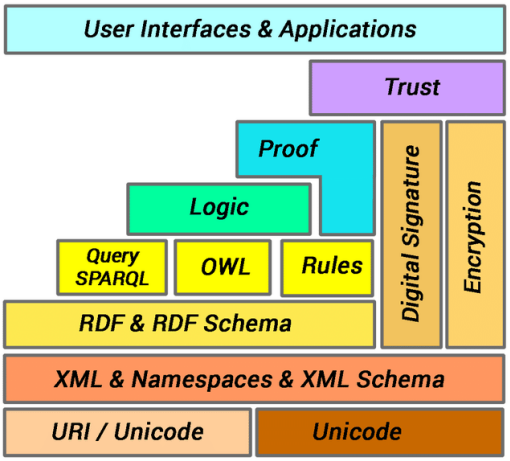
"No contexto da w*eb* semântica, o termo busca (*query*) é usado genericamente para se referir às tecnologias e aos protocolos desenvolvidos para a recuperação da informação […]" (CATARINO; SOUZA, 2012, p.80).

De acordo com Café e Brascher (2011, p.25), "relações semânticas são estabelecidas por meio da análise das características ou propriedades dos conceitos, as quais permitem identificar diferenças e semelhanças que evidenciam determinados tipos de relacionamentos […]". Para Messina et al. (2018, p.300, tradução nossa) "[…] toda solução deve abordar a questão de integridade semântica, ou seja, interpretar o real significado dos dados gerados a partir de múltiplas fontes ou manipulado por várias ferramentas."

Neste sentido, Nietto e Camargos (2015) definem a busca semântica como "forma de se consultar objetos em uma base de conhecimento e obter como resultado vários atributos e conceitos relacionados que são interligados".

Segundo Pasquier (2008), as principais tecnologias envolvidas nas camadas da *web* semântica incluem um modelo central RDF (*Resource Description Framework*), uma linguagem de esquema RDF, uma linguagem para definição de ontologias *web* OWL (*web ontology language*) e uma linguagem de consulta para RDF chamada SPARQL. A Figura 4 demonstra uma visão da pilha de tecnologias da *web* semântica proposta pela W3C (*World Wide Web Consortium*), uma comunidade internacional que desenvolve padrões para a *web*.

Figura 4 – Pilha de tecnologias da *web* semântica



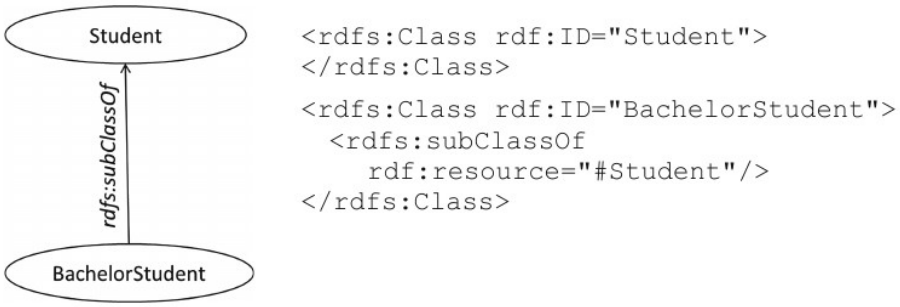
Fonte: Gezer e Bergweiler (2017, p.37)

Um modelo RDF é baseado em declarações de recursos representadas por uma asserção ou tripla (*triples*) na forma de "sujeito-predicado-objeto" (PASQUIER, 2008). Bases de dados em forma de grafo (*triplestores*) que permitem utilizar ontologias em sua

modelagem são utilizados para armazenamento de triplas RDF e seus relacionamentos de forma eficiente (ROZSA; VIERA; DUTRA, 2019). Ainda neste contexto, Medeiros (2009) considera o padrão RDF como sendo o melhor para representar o conhecimento através da modelagem dos metadados por ser de estrutura simples e pelo fato que cada conceito representado em RDF tem associado um URI único, o que resolve conflitos encontrados no compartilhamento de conhecimento. Um modelo RDF pode ser apresentado em vários formatos, entre eles XML, Turtle, N3 ou JSON (ALAOUI; BAHAJ, 2020).

Para Rozsa, Viera e Dutra (2019, p. 170), "[…] o RDFS trata-se de uma expansão semântica do RDF e fornece mecanismos para a descrição de grupos de recursos e o relacionamento entre estes recursos." A Figura 5 apresenta uma representação RDFS no formato XML e seus correspondente no formato de grafos.

Figura 5 – Exemplo de RDF no formato grados (esquerda) e em XML (direita)



Fonte: Alaoui e Bahaj (2020, p.413)

Se por um lado a ambiguação pode diminuir a especificidade na recuperação de informação na internet, os componentes semânticos de um sistema devem envolver todos os modelos de relacionamentos de determinado domínio, incluindo suas ambiguidades e variações, sob pena de provocar baixos índices de recuperação da informação (CAFE; BRASCHER, 2011).

A este propósito, Rozsa, Viera e Dutra (2019, p. 168) afirmam que "a busca sensível à intenção do usuário e seu contexto é conhecida como busca semântica e os buscadores que fornecem este tipo de busca são denominados "motores de busca semântica". Buscadores semânticos tem como principal objetivo melhorar a precisão de uma pesquisa de acordo com a intenção do usuário.

2.6 REPRESENTAÇÃO DE DOMÍNIOS

A implementação de sistemas computacionais inteligentes baseados em conhecimento biomédico extraído de diversas bases de dados especializadas no domínio tem sido motivo de pesquisa em inteligência artificial (CALLAHAN et al., 2020).

Segundo Callahan et al. (2020, p.24, tradução nossa), para gerar hipóteses e avaliar respostas:

> Os sistemas baseados em conhecimento especificam uma representação do conhecimento (como um sistema computacional representa o conhecimento internamente) e um ou mais métodos de inferência ou raciocínio para saber como o processamento sobre estas representações (talvez combinado com outras entradas) são usados para produzir saídas.

Davis, Shrobe e Szolovits (1993), ao compararem a representação de conhecimento como um substituto de algo que se conhece para ser utilizado com o reconhecimento das consequências de sua ação, dada a complexidade de representar o mundo, estabelecem a necessidade de um conjunto de 'comprometimentos ontológicos' que serviriam como filtros que definiriam melhor os componentes e conexões de uma realidade observada. Assim, em recuperação da informação, métodos baseados em bases de conhecimento têm sido propostos tanto para consultas de expansão quanto para realizar buscas semânticas (YEPES; BERLANGA, 2015, p.301).

Para Vanopstal et al. (2011, p.2, tradução nossa) "[…] termos como taxonomia, tesauro, ontologia, vocabulário controlado, etc, podem ser definidos de várias formas dependendo do domínio da aplicação." No domínio da recuperação da informação ou recuperação bibliográfica as estruturas subjacentes para buscas dos artigos nas bases de dados incluem taxonomias, tesauros, ontologias, vocabulários controlados e mapas de tópicos (VANOPSTAL et al., 2011).

De acordo com Garshol (2004 *apud* VANOPSTAL et al., 2011, p.8, tradução nossa), "uma ontologia geralmente consiste em conceitos, relacionamentos e propriedades […]". As ontologias são concebidas como representações formais do conhecimento por almejarem refletir a realidade através de muitos tipos de relações semânticas, estruturando os conceitos na forma de uma rede, conferindo-lhes assim muito mais versatilidade e robustez quando comparadas com tesauros ou taxonomias que tem seus conceitos estruturados hierarquicamente (VANOPSTAL et al., 2011).

Para Dupré (1993 *apud* PASQUIER, 2008, p.585, tradução nossa) "[…] não há uma ontologia única". "Ontologias de domínio devem também ser ancoradas à uma ontologia superior para permitir o compartilhamento e reuso do conhecimento" (PASQUIER, 2008, p.585, tradução nossa).

2.6.1 Ontologias biomédicas

No campo biomédico, apesar do uso de ontologias específicas do domínio facilitarem o reconhecimento de relacionamentos semânticos, ainda assim a correção do intervalo semântico exige mecanismos de inferências para outros relacionamentos e, neste sentido, Koopman et al. (2015, p.7, tradução nossa) afirmam que "existe valioso conhecimento de domínio explicitamente representado, ainda que preso, em recursos de conhecimento estruturado como as ontologias […]". Para Rozsa, Viera e Dutra (2019), as ontologias facilitam alguns processos de inferências sobre os dados pesquisados, incluindo as modalidades de herança, transitividade, simetria e equivalência, contudo, ontologias utilizadas em bases biomédicas possuem alta complexidade o que pode impactar na manutenção da base de dados e na eficiência das inferências sobre as informações anotadas.

O MeSH (*Medical Subject Headings*) é um tesauro produzido pela *National Library of Medicine* e é usado para indexar e buscar artigos na base de artigos científicos biomédicos do Pubmed/Medline de acordo com a consulta construída pelo usuário (VANOPSTAL et al., 2011). O Pubmed possui mais de 25 milhões de registros e considerando que cada um deles tem em média de 5 a 10 elementos de metadados calcula-se que, eliminando-se duplicatas e repetições, um simples processo de busca deva percorrer mais de 100 milhões de termos (SUBRAMANYAM; KRISHNAMURTHY; ASUNDI, 2017). Por outro lado, o MeSH tem cerca de 27.000 termos que em uma consulta são acrescidos aos termos em linguagem natural por intermédio de operadores booleanos integrados (SUBRAMANYAM; KRISHNAMURTHY; ASUNDI, 2017). Dado o número de termos do MeSH, Venkatesh et al. destacam o fato que para classificação de documentos os métodos de *machine learning* tendem a ser inefetivos devido ao grande número de categorias.

Segundo Bodenreider (2004), o *Unified Medical Language System* Metathesaurus é um repositório de vocabulários biomédicos que contém mais de 2 milhões de nomes, cerca de 900.000 conceitos e 12 milhões de relacionamentos entre os conceitos e tem a finalidade de ajudar a sobrepor questões como grande variedade de nomes e ausência de padronização. O UMLS inclui outros vários vocabulários controlados, dentre eles o SNOMED CT, o do NCBI

e o MeSH, o que lhe confere uma estrutura poli hierárquica para representação do conhecimento no domínio biomédico, sendo considerado assim uma ontologia (VANOPSTAL et al., 2011). Segundo Leão (2008 *apud* Medeiros, 2009, p.18) "[…] a UMLS é muito grande, complexa e na prática pouco útil".

Ontologias, tesauros e vocabulários controlados biomédicos são bases de conhecimento que tem contribuído para o desenvolvimento de métodos para realizar tarefas de mineração de texto como reconhecimento de entidades nomeadas e extração de relacionamentos. (YEPES; BERLANGA, 2015). A propósito, Pickler (2007, p.78) destaca:

> […] enquanto o tesauro é visto como uma ferramenta de organização do conhecimento, com a finalidade de facilitar a indexação e recuperação de informações, operando para a comunicação entre usuários e linguagens documentárias, a ontologia é tida como um instrumento de representação do conhecimento no ambiente computacional, voltada para o registro de conceitos de um domínio visando a inferência automatizada.

Lembrando Pickler (2007), as ontologias podem auxiliar na inserção e recuperação de documentos em sistemas computacionais, na mineração de textos em linguagem natural, na troca de informações entre agentes inteligentes e em outras tarefas que envolvam o uso e a representação do conhecimento. Contudo, Yepes e Berlanga (2015, p.300, tradução nossa) ressaltam sobre as bases de conhecimento:

> Infelizmente uma vez que estes recursos não foram desenvolvidos para realizar tarefas de mineração de texto, os métodos baseados em conhecimento geralmente exibem menor performance do que métodos supervisionados *ad hoc.* Apesar desta limitação, abordagens baseadas em conhecimento tornam-se cruciais quando há paralelamente uma escassez de dados rotulados para treinar métodos supervisionados. Devido à heterogeneidade e larga escala dos recursos biomédicos, os métodos baseados em conhecimento tem se tornado cada vez mais populares.

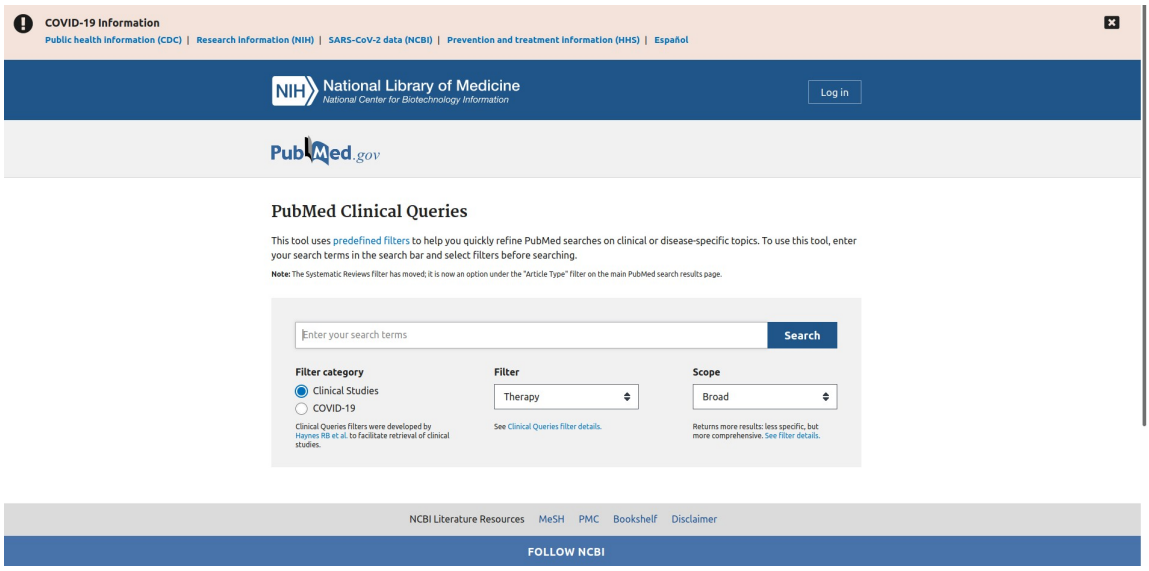
Segundo Pasquier (2008, p.586, tradução nossa) "existem múltiplas ontologias em que cada uma modela um domínio específico." Este autor também ressalta que muitas bases de dados utilizam identificadores próprios para suas entidades e embora exista um trabalho concentrado de padronização da nomenclatura para ser utilizada em aplicações de *web* semântica na área biomédica (*Life Sciences Identifier*), o que facilitaria a interoperabilidade dos dados, esta iniciativa ainda não tem sido amplamente utilizada pelos curadores das inúmeras bases de dados biomédicos.

2.7 REPOSITÓRIOS DE DOCUMENTOS BIOMÉDICOS

Atualmente existe uma quantidade enorme de recursos biomédicos disponíveis para pesquisadores incluindo, por exemplo, bases de dados de sequências proteicas e genéticas, bases de fármacos e bases bibliográficas biomédicas (BODENREIDER, 2004). Diante desse pressuposto, Messina et al. (2018) afirmam que problemas comuns enfrentados ao tentar integrar os dados e informações oriundos destas diversas fontes independentes de recursos incluem a dificuldade para localização dos dados relevantes, os diferentes formatos de dados, a identificação dos relacionamentos entre eles, a ambiguidade e duplicação, a granularidade da informação e ainda o fato de não existir um formato padrão de consultas para estas várias fontes de dados. Entretanto, um denominador comum entre elas é o seu vocabulário controlado que, quando compatível, permite a sua interoperabilidade (BODENREIDER, 2004).

O PubMed é um sistema de recuperação de informação baseado na *web*, desenvolvido pelo NCBI nos Estados Unidos e que possibilita pesquisas em bases de dados biomédicas, possuindo atualmente mais de 23 milhões de arquivos textos, sendo a busca gerada a partir de palavras-chave que são digitadas pelos usuários (FATEHI; GRAY; WOOTTON, 2013). A Figura 6 mostra a página de busca do PubMed.

Figura 6 – Página de busca do PubMed



Fonte: NCBI (2021).

O Google Scholar é uma ferramenta gratuita de busca em bases de dados que pode ser utilizado para pesquisas de documentos na área biomédica. Tem a vantagem de indexar textos

completos o que aumenta sua sensibilidade quando comparado com outros buscadores, encontrando artigos no qual a entidade pesquisada não pode não ser mencionada no título, no resumo ou nos tesauros do documento (BRAMER et al., 2017).

2.8 CORD-19: REPOSITÓRIO BIOMÉDICO PARA COVID-19

O CORD-19 é um acrônimo para *Covid-19 Open Research Dataset*, o maior corpora de literatura biomédica criado para facilitar a pesquisa e o desenvolvimento de sistemas de recuperação da informação e de mineração de texto em artigos relacionados ao Covid-19 (WANG et al., 2020). Atualmente compreende mais do que 267.000 mil documentos publicados e outros ainda na fase de pré-publicação, coletados a partir de várias fontes, incluindo o Pubmed, Worl Health Organization e arXiv.org (WANG et al., 2020). Neste corpora, os documentos em PDF foram convertidos para o formato JSON, de mais fácil processamento automático, e seus metadados foram normalizados para eliminar redundâncias (WANG; LO, 2021).

Segundo Wang e Lo (2021), o CORD-19-on-FHIR (https://github.com/fhircat/CORD-19-on-FHIR) é uma modificação do CORD-19 que pode facilitar a construção de SRIs que utilizam o CORD-19 porque ele já vem com entidades nomeadas anotadas em seus documentos para 3 categorias (*condition*, *medication*, *procedure*).

2.9 RECURSOS PARA MODELOS DE BUSCA SEMÂNTICA NO CORD-19

Diversas tecnologias de extração de conhecimento podem ser utilizadas para aumentar a assertividade de sistemas de recuperação de informação por auxiliarem a construção de uma base conceitual de conhecimento para apoio nas pesquisas.

2.9.1 Anotação e classificação de documentos

Uma alternativa aos modelos tradicionais para recuperação de informação de documentos obtidos a partir de uma base bibliográfica é a anotação semântica dos documentos (CECI; WOSZEZENKI; GONÇALVES, 2014). De acordo com Popov (2003 *apud* CECI, WOSZEZENKI e GONÇALVES 2014, p. 5) “ […] a anotação semântica é um processo de geração de metadados específicos para possibilitar novos métodos de acesso à informação e mesmo estender métodos existentes”, permitindo o reconhecimento de algum

significado a partir dos metadados e possibilitando a inferência de relações semânticas pelos motores de busca, que por sua vez facilitam a recuperação do documento.

Conforme foi abordado no capítulo anterior, as ontologias podem ser utilizadas em tecnologias de extração do conhecimento, especialmente para agregação de conhecimento a partir de documentos não estruturados, por permitirem a associação de palavras ou frases com entidades definidas pela própria ontologia, o que é conhecido como associação semântica (PECH et al., 2017).

Anotações semânticas em documentos podem ser realizadas de forma manual por especialistas no domínio, semiautomática, automática ou de forma híbrida, com ou sem o emprego de ontologias ou outras representações de domínios (CECI; WOSZEZENKI; GONÇALVES, 2014). Segundo Wang e Lo (2021), as anotações automáticas podem usar modelos automáticos para reconhecimento de entidades nomeadas e conexão de entidades de bases de conhecimento, com ferramentas como o MetaMap Lite ou o *framework* ScispaCy. As anotações podem ser realizadas no próprio documento recuperado (intrusivas) quanto em um documento à parte (não-intrusivas) (ROCHA; SALVADOR; SOUZA, 2017). Ainda a respeito deste processo, a "[…] anotação semântica recebe como entrada o índice invertido e uma ontologia de domínio e produz uma estrutura semântica estendida de forma a incorporar conceitos e propriedades da ontologia ao índice invertido" (CECI; WOSZEZENKI; GONÇALVES, 2014, p.6).

Para Pech et al. (2017, p.1, tradução nossa), "a anotação tem um importante papel para uma variedade de aplicações semânticas, como a geração de dados conectados, extração de informação aberta, alinhamento de ontologias, e a busca semântica."

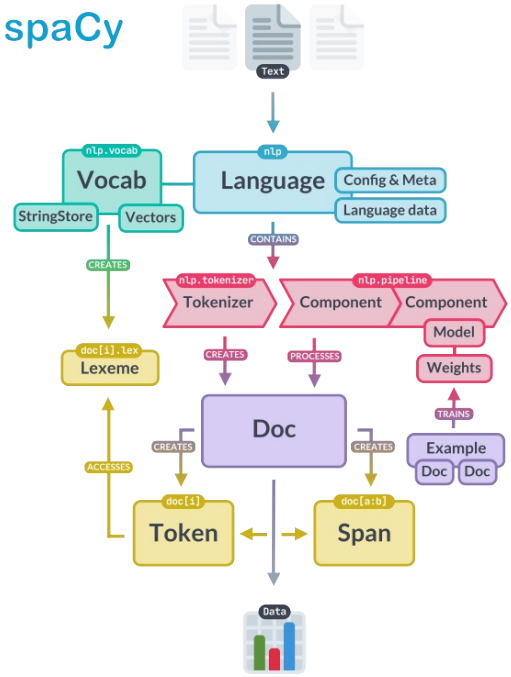
Entidade nomeada biomédica em um documento pode ser definida como uma frase ou mais que possui uma relação com um determinado objeto da literatura biomédica, podendo ser um diagnóstico, sintomas, genes, drogas, proteínas ou mais particularmente qualquer outro objeto definido em uma ontologia ou base de conhecimento biomédica (GOULART; STRUBE DE LIMA; XAVIER, 2011). Dado um documento tipo texto e um conjunto de palavras definidas como entidades, a técnica de reconhecimento de entidades nomeadas (NER) retornará uma lista de sequências de palavras cada uma associada a um tipo de entidade representado por uma tripla (índice de início, índice de término, tipo da entidade) (BAST; BUCHHOLD; HAUSSMANN, 2016).

A biblioteca baseada em python chamada spaCy fornece uma variedade de ferramentas para processamento de linguagem natural, incluindo o reconhecimento de

entidades nomeadas e a partir desta foi criada a biblioteca scispaCy que é especializada para o domínio biomédico (NEUMANN et al., 2019).

As estruturas centrais no spaCy são a classe Language, a classe Vocab e o objeto Doc, conforme está demonstrado na Figura 7.

Figura 7 – Estruturas centrais do *framework* spaCy

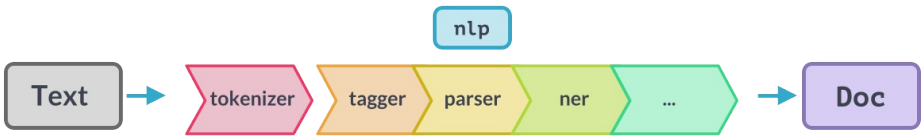


Fonte: Spacy (2021).

A classe Language é usada para processar o texto por intermédio de um pipeline (Figura 8), gerando um objeto Doc, e tipicamente é armazenada em uma variável chamada nlp.

O objeto Doc armazena a sequência de *tokens* gerados e suas anotações (SPACY, 2021).

Figura 8 – Pipeline de processamento dos componentes da classe Language do spaCy



Fonte: Spacy (2021).

A classificação de documentos é um processo sofisticado que envolve o pré-processamento dos textos, transformações e redução de dimensionalidade para o treinamento de um modelo que possa classificar novos documentos em classes pré-definidas (MIROŃCZUK; PROTASIEWICZ, 2018). Neste sentido, Venkatesh et al. (2019, p.629, tradução nossa) destacam que "a classificação automática de documentos médicos em classes pré-definidas está crescendo rapidamente em repositórios *online* […]".

2.9.2 Text embeddings

Text embeddings são representações vetoriais computadas de extensões de texto que capturam semelhanças semânticas e sintáticas entre esses textos (WANG; LO, 2021). Ainda segundo Wang e Lo (2021), estas podem ser realizadas em diferentes níveis de granularidade, seja para palavras ou *tokens* (*word embeddings*), entidades nomeadas, sentenças, parágrafos, documentos ou além, e têm sido utilizados em vários sistemas de busca e recuperação de informação em corpora relacionados ao Covid-19.

Segundo Bast, Buchhold e Haussmann (2016), *word embeddings* são aprendidos de forma não-supervisionada, sem entrada adicional de conhecimento, a partir de volumosos corpus de textos.

Wang e Lo (2021) relatam que um método de *embedding* de documentos (SPECTER) utiliza o modelo pré-treinado SciBERT para capturar a similaridade entre documentos em sistemas de recuperação relacionados a Covid-19. Também estão disponíveis para o CORD-19 *embeddings* de conceito (algoritmo JET), de relacionamento (SeVeN) e de co-ocorrência para entidades nomeadas utilizando o CORD-19-on-FHIR (WANG; LO, 2021).

Conforme Awatramani e Gupta (2020, p.733), "medir a similaridade textual semântica envolve a estimativa do grau de equivalência semântica entre duas sentenças através de um escore numérico", o que é primordial em questões complexas como recuperação e extração da informação.

2.9.3 Grafos de conhecimento

Wang e Lo (2021) afirmam que os grafos de conhecimento oferecem uma estrutura composta de entidades e seus relacionamentos de um determinado domínio, úteis para inferir ou descobrir novos relacionamentos além de servirem como uma excelente forma para visualizar os dados e informações complexas.

Segundo Callahan et al. (2020), muitos desafios são impostos para construção e utilização de grafos de conhecimento no domínio biomédico, incluindo a necessidade de envolvimento de especialista no domínio para curadoria, a dificuldade de integração de diferentes bases de dados, conflitos das dependências de diferentes bases e a alta performance computacional por utilizar algoritmos complexos em volume expressivo de dados. Ainda neste contexto, Callahan et al. (2020) citam que questões relacionadas com licença para utilização de dados diminuem a disponibilidade destes.

Abordagens automáticas de construção de grafos de conhecimento têm sido aprimoradas por métodos de extração de informação baseadas em processamento de linguagem natural.

Wise et al. (2020) construíram um sistema de recuperação de informação chamado de *Covid-19 Knowledge Graph* (CKG) com o objetivo de utilizar grafos de conhecimento a fim de extrair e visualizar relações complexas entre artigos relacionados à Covid-19..

2.9.4 Modelos de linguagem pré-treinados

O treinamento de modelos de inteligência artificial, na área de *deep learning,* que têm sido extensivamente utilizados para processamento de linguagem natural, exige volumosa quantidade de dados para que estes modelos adquiram performance eficiente. Em domínios que não possuem tantos dados disponíveis para treinamento cada vez mais tem sido utilizada a estratégia de *transfer learning*, que consiste em utilizar modelos treinados para uma tarefa geral como ponto inicial para o treinamento de outros modelos que serão utilizados para outros fins (AWATRAMANI; GUPTA, 2020).

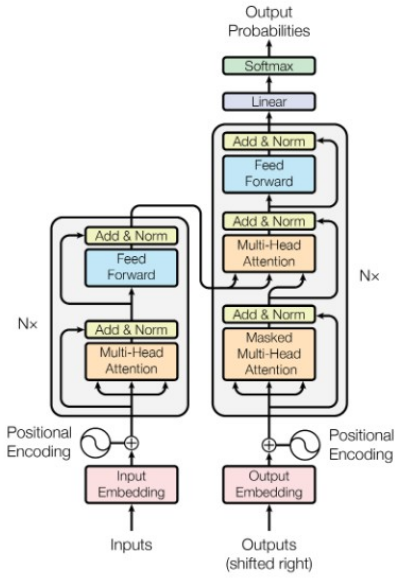
Modelos de linguagem pré-treinados, incluindo ELMo, BERT e RoBERTa, tem revolucionado a performance de *deep learning* para processamento de linguagem natural (AWATRAMANI; GUPTA, 2020). Melhora na eficiência tem sido relatada tanto para tarefas a nível de sentenças, incluindo inferência de linguagem natural e parafraseamento, quanto para tarefas a nível de *tokens*, entre elas o reconhecimento de entidades nomeadas e resposta a questões (Devlin et al., 2019).

BERT é um modelo de representação de linguagem que trabalha com *transformers*, um mecanismo de atenção que aprende associações contextuais entre palavras em um texto (Devlin *et al*., 2019). Segundo Vaswani et al. (2017), uma função de atenção mapeia uma consulta e um conjunto de pares chave-valor para uma saída, onde consulta, chaves, valores e saída são todos vetores. A saída é computada como uma soma ponderada dos valores, onde o

peso de cada valor é computado por uma função de compatibilidade entre a consulta e a chave correspondente (VASWANI et al., 2017).

A Figura 9 apresenta a estrutura de um *transformer*, que inclui um *encoder*, sua estrutura mais importante, que lê toda a sequência de palavras do texto de uma vez (bidirecional) para criar uma representação intermediária, e o *decoder*, que faz uma predição utilizando os *embeddings* (AWATRAMANI; GUPTA, 2020).

Figura 9 – Arquitetura de um transformer



Fonte: Vaswani *et al*. (2017, p. 3).

Modelos adaptados para domínios incluindo o SciBERT e BioBERT foram treinados especificamente para textos científicos e biomédicos respectivamente (WANG; LO, 2021).

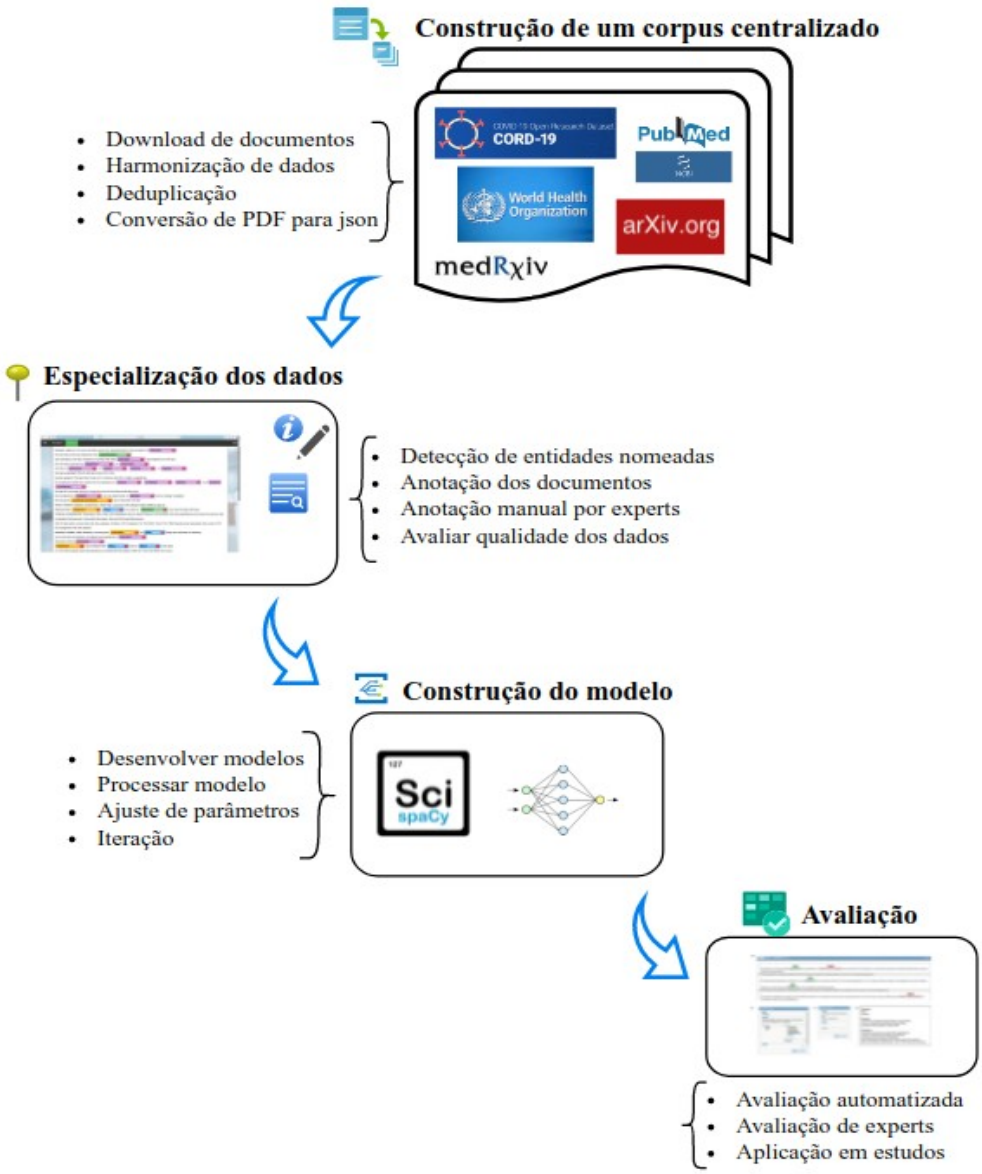
2.10 BUSCA SEMÂNTICA EM BIG DATA SOBRE COVID-19

São diversas as bases de dados e de documentos sobre Covid-19 disponíveis na *internet*. A imensa quantidade e variedade de dados disponibilizados em cada uma destas bases satisfaz plenamente todas as dimensões dos atributos de *Big Data*, tornando muito oneroso fazer o *deploy* de um sistema que faça todo o processo de busca semântica e extração de informação diretamente a partir destes repositórios biomédicos na *internet*.

Com este propósito, Wang e Lo (2021) sugerem um fluxo de trabalho (Figura 10) que pode ser utilizado como modelo de arquitetura de sistemas que trabalham com o enorme volume de literatura publicada referente ao Covid-19, mas que também pode ser generalizada

para outros contextos de *Big Data* na área biomédica ou mesmo de outros domínios de conhecimento.

Figura 10 – Fluxo de trabalho para extração conhecimento em *Big Data* sobre Covid-19



Fonte: Adaptado de Wang e Lo (2021, p. 783).

A primeira etapa do processo consiste em construir um corpus centralizado, que será a a base de dados sobre a qual o sistema trabalhará. Nesta etapa, além do *download* dos documentos, há necessidade de harmonização dos metadados, exclusão de documentos

duplicados e transformação dos documentos para um padrão de texto que seja legível por máquinas.

Conforme apresentado em seção anterior, o CORD-19 simplifica a primeira etapa do processo por ser um corpus centralizado de acesso público, que já foi otimizado para utilização em pesquisas e para o desenvolvimento de sistemas relacionados com recuperação de informação sobre Covid-19.

A segunda etapa do processo proposto consiste em entender os dados contidos em cada documento do CORD-19. Para isso, faz-se necessário utilizar técnicas de mineração de texto que anotem nos documentos as entidades nomeadas biomédicas bem como identifiquem suas classes. Algoritmos pré-treinados em bases biomédicas cumprem esta tarefa de anotação automática dos documentos, evitando assim a necessidade de consumir recursos para que especialistas treinem modelos baseados em documentos anotados manualmente.

A terceira etapa do modelo proposto por Wang e Luo (2021) consiste em implementar o que representa propriamente a máquina de busca de informação nos documentos, otimizada não apenas para recuperação de dados, mas para a recuperação de informação útil ao pesquisador de Covid-19. Esta etapa envolve iterações frequentes para ajustes da estrutura e dos parâmetros do modelo a fim de melhorar sua performance.

A etapa final da construção de qualquer sistema de recuperação de informação consiste em avaliar sua eficiência no que tange à relevância dos documentos recuperados. Esta etapa pode envolver também a validação por especialistas do domínio.

2.11 SOLUÇÕES BIOMÉDICAS PARA RECUPERAÇÃO DA INFORMAÇÃO

Para resumir a estrutura geral de um sistema de recuperação de informação eficiente, Monteiro et al. (2017, p.165) esclarecem:

> Os atuais SRI ou máquinas de busca possuem sofisticados processos de coleta, indexação, análise e interface de busca em que: linguagens e programas são utilizados para representar e descrever semanticamente a informação; ontologias para relações de domínio; bases de conhecimento (as chamadas bases secundárias com entidades descritas semanticamente) que fornecem dados estruturados para os agentes inteligentes; programas e tecnologias de visualização e apresentação dessa informação em contextos (Knowledge Graph ou mapa do conhecimento) nas interfaces de busca.

Soto, Przybyla e Ananiadou (2019) apresentam uma ferramenta de busca semântica para literatura biomédica chamado Thalia que realiza uma mineração automática de conceitos

em artigos indexados no PubMed e posteriormente faz sua normalização com base em uma ontologia especializada. O reconhecimento de entidades nomeadas do Thalia foi customizado para a mineração de entidades de 8 conceitos com as seguintes ontologias: químicos (ChEBI), drogas (DrugBank), metabólitos (HMDB), doenças (UMLS), proteínas (UniProt), genes (HGNC), espécies (NCBI) e entidades anatômicas (CARO). A busca semântica é disponibilizada ao expandir a área da *query* ou por interação com diferentes aspectos das entidades (SOTO; PRZYBYŁA; ANANIADOU, 2019).

Pasquier (2008) apresenta uma aplicação de *web* semântica (Thea-online) baseada nos padrões URIs, RDF, OWL, SPARQL para recuperar informação originada de fontes diferentes, mas que apresenta o contraponto de ser construída sobre uma arquitetura de *data warehouse*, limitando assim a capacidade de trabalho para volumes pequenos de dados.

MEDIE é uma ferramenta de busca semântica inteligente para ser usado no PubMed/Medline, provido de interface gráfica, que permite especificar quaisquer combinações de sujeito, verbo e objeto para a recuperação de documentos (OHTA et al.,2006).

Segundo Nietto e Camargos (2015, p.8), que fizeram uma análise de 22 artigos para avaliar características e limitações de interfaces de busca semântica utilizados em domínios diversos, “alguns dos maiores desafios das interfaces dos mecanismos de busca semântica são mensurar sua usabilidade, realizar os experimentos com um número grande de pessoas e tornar a interface do usuário mais simples e intuitiva”, mas finalizam afirmando que o campo da busca semântica tem futuro promissor.

Wang e Lo (2021) apresentam em seu artigo um compilado de estudos que demonstram diferentes formas de arquiteturas disponíveis atualmente, com utilização de recursos variados, para otimizar sistemas de extração de informação no contexto do Covid-19.

3 METODOLOGIA

De acordo com Trujillo Ferrari (1974, p. 248), conforme citado por Prodanov e Freitas (2013, p. 24) "O método científico é um traço característico da ciência", um mecanismo que organiza o pensamento em sistemas e arquiteta os passos do cientista durante o percurso até o objetivo final.

O capítulo atual propõe definir os tipos de pesquisa aplicada, bibliográfica e quantitativa. São definidas etapas do modelo de solução apresentada e as delimitações do trabalho.

Segundo Silva e Menezes (2005, p.20) "existem várias formas de classificar as pesquisas" dependendo do ponto de vista abordado. Quanto à sua natureza, as pesquisas podem ser classificadas em básica ou aplicada, do ponto de vista de abordagem do problema podem ser do tipo quantitativa ou qualitativa e quanto aos objetivos podem ser exploratória ou descritiva explicativa. As pesquisas podem ainda ser classificadas do ponto de vista técnico conforme Gil (1991 *apud* SILVA e MENEZES, 2005, p.20) que as dividem em pesquisa bibliográfica, documental, experimental, levantamento, estudo de caso, pesquisa *expost-facto,* pesquisa-ação e pesquisa participante.

3.1 CARACTERIZAÇÃO DO TIPO DE PESQUISA

A pesquisa para este trabalho será uma pesquisa do tipo aplicada quanto à natureza, uma vez que objetiva gerar conhecimentos novos que buscam trazer uma contribuição para o ramo das ciências da computação e do tipo qualitativa e quantitativa quanto à abordagem do problema, uma vez que abordará simultaneamente a quantificação da eficiência de métodos de buscas de informação e a qualidade dos relacionamentos semânticos dos objetos de busca. Do ponto de vista de seus objetivos, este trabalho envolve além de um caráter exploratório ao fazermos uma revisão bibliográfica sobre o assunto de busca semântica, haverá ainda o desenvolvimento de um sistema de busca utilizando um método experimental, caracterizando esta pesquisa também como explicativa. Finalmente, do ponto de vista técnico esta pesquisa de proposta de solução possui atributos de pesquisa-ação uma vez que se propõe ao desenvolvimento de um modelo de sistema.

3.2 ETAPAS METODOLÓGICAS

Este projeto está estruturado em basicamente quarta etapas:

3.2.1 Estudo bibliográfico

Nesta etapa propõem-se realizar uma ampla revisão de artigos científicos indexados em diferentes bases bibliográficas, com preferência para artigos de revistas *peer-reviewed* atuais identificadas por DOI (*digital object identifier*) e livros referenciados que abordem o assunto de busca semântica, com enfoque para a área biomédica, mais especificamente para o contexto da literatura sobre Covid-19. A estrutura analítica da pesquisa envolve progressivamente assuntos de recuperação da informação e seus principais componentes e métodos, anotação e classificação de documentos, representação de domínios com destaque para as ontologias e um abrangente enfoque no tema anotação semântica. Finalizamos esta etapa demonstrando algumas soluções já existentes aplicadas principalmente no domínio biomédico.

3.2.2 Modelagem do sistema

Ao longo da modelagem do sistema foram definidos as bibliotecas necessárias para o armazenamento e processamento dos documentos que utilizaremos, bem como das tecnologias necessárias para o cumprimento dos requisitos necessários para desenvolver um protótipo simplificado de buscador semântico. Em função da variedade e complexidade de todos os recursos que podem ser implementados em um motor de busca semântico no contexto do *Big Data*, este projeto seguirá seu objetivo específico de estruturar basicamente um produto de valor mínimo para um sistema de recuperação de informação.

3.2.3 Implementação

O protótipo de buscador semântico é codificado em um *script* python nesta etapa. Também foi instanciada uma imagem *docker* do Elasticsearch na nuvem para o armazenamento e processamento dos documentos. A interface da aplicação *web* foi criada através do *framework* Streamlit.

3.2.4 Avaliação

Na etapa de avaliação são realizados diversos testes a partir da digitação de termos ou frases no campo de consulta da interface *web* do usuário. As consultas serão realizadas a partir de uma funcionalidade que baseará a busca, sob a escolha do usuário, em uma máquina de busca baseada em um modelo clássico ou, alternativamente, a partir de uma análise semântica.

Serão avaliadas a qualidade da anotação de entidades nomeadas realizadas nos documentos armazenados no Elasticsearch, a relevância dos documentos recuperados em uma busca baseada no contexto e se haverá um incremento da relevância para a funcionalidade que permite uma busca semântica.

Ainda como parte da avaliação será realizada, por um conjunto de especialistas médicos, uma validação da relevância dos documentos recuperados, seguida de uma análise dos resultados.

3.3 DELIMITAÇÕES

A presente proposta de pesquisa não se propõe a descobrir informações novas do domínio biomédico a partir dos documentos pesquisados e trabalhará exclusivamente com referência de informação já disponível no meio científico biomédico relacionada à Covid-19. Também não será intuito da pesquisa desenvolver qualquer modelo de predição de informação. Finalmente não faz parte do escopo desta pesquisa fazer a integração de pesquisa em diferentes bases de conhecimento biomédico, embora esta pesquisa represente uma das etapas iniciais de um projeto futuro de integração de bases para extração de conhecimento.

4 PROPOSTA DE BUSCADOR SEMÂNTICO PARA DOCUMENTOS COVID-19

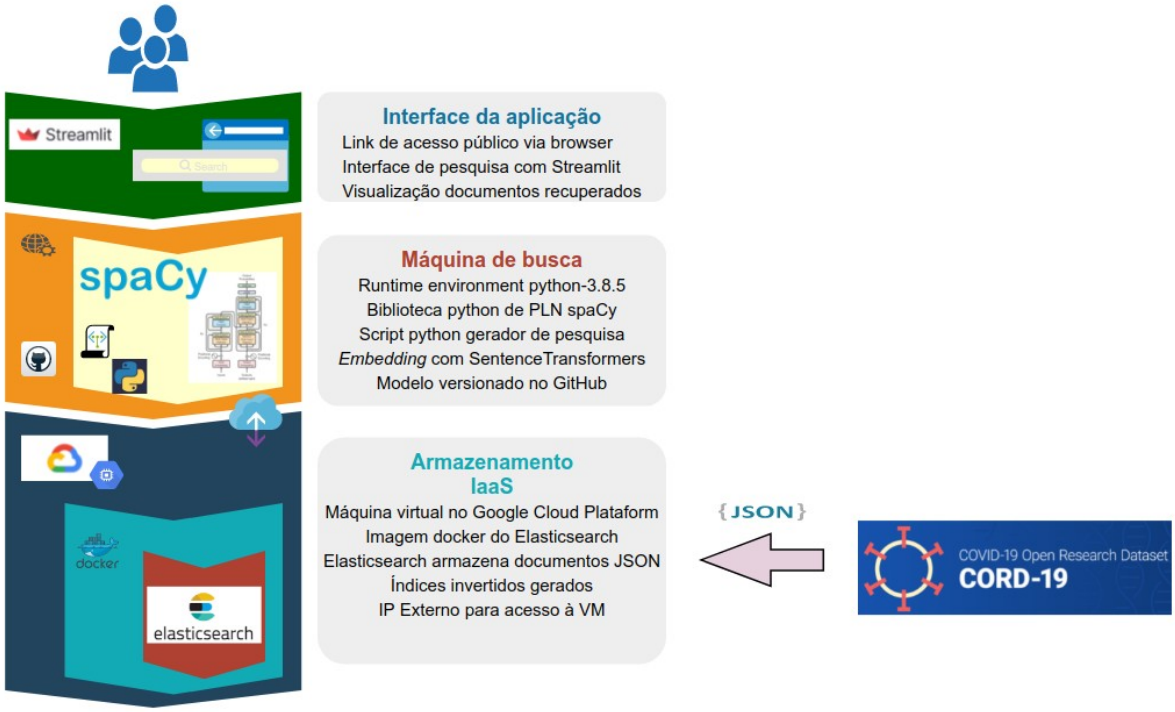
Neste capítulo apresentaremos uma proposta de um sistema que realizará um processo inicial de recuperação da informação baseado em uma arquitetura básica para busca semântica em um repositório de documentos específicos sobre Covid-19.

4.1 ARQUITETURA

Segundo Sommerville (2011, p.103), "o projeto de arquitetura está preocupado com a compreensão de como um sistema deve ser organizado e com a estrutura geral desse sistema." Ainda para este autor, esta é uma etapa crítica pois identifica os principais componentes de um sistema.

No que tange a visão de desenvolvimento da arquitetura, o buscador semântico para documentos sobre Covid-19 é organizado conforme a Figura 11.

Figura 11 – Visão de desenvolvimento do buscador semântico



Fonte: Criado pelos autores (2021).

Este projeto foi concebido com base na recomendação de arquitetura proposta por Wang e Lo (2021) para *Big Data* referente ao Covid-19, apresentada no capítulo anterior.

Para servir de base de documentos para o buscador semântico foram utilizados os textos sobre Covid-19 disponíveis no repositório CORD-19, disponíveis para *download* por intermédio de uma pasta compactada em https://ai2-semanticscholar-cord-19.s3-us-west-2.amazonaws.com/historical_releases.html. O arquivo utilizado nesta solução foi o *release* “cord-19_2021-09-27.tar.gz” de 27/01/2021.

Após descompactada a pasta localmente, todos os seus arquivos já em formato JSON foram transferidos para uma imagem *docker* do Elasticsearch instanciada em uma máquina virtual criada no Google Cloud Plataform.

4.2 FERRAMENTAS UTILIZADAS

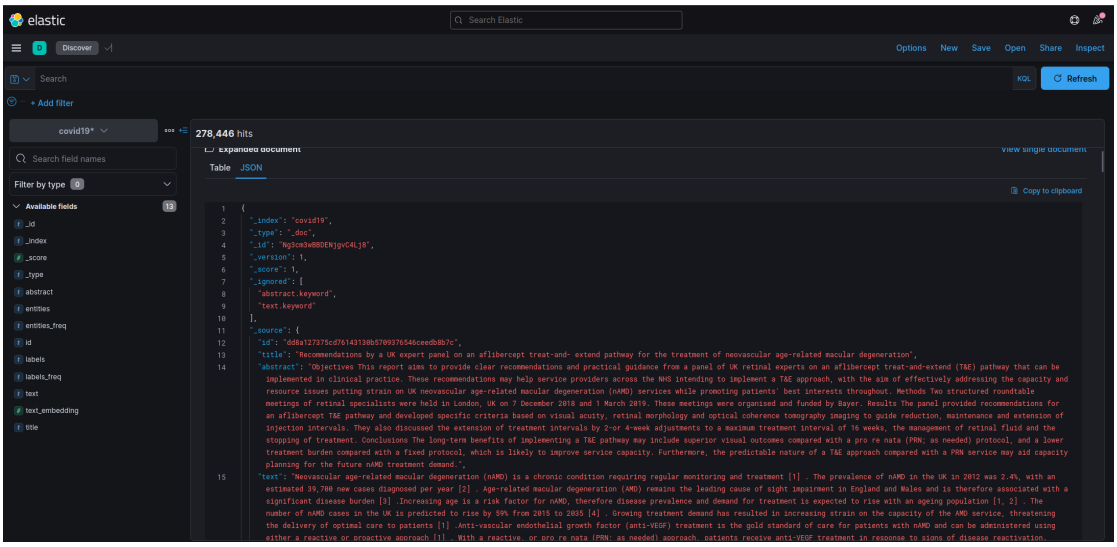
As ferramentas utilizadas para a construção do buscador semântico foram:

4.2.1 Elasticsearch

O Elasticsearch é uma ferramenta *open source,* baseado no Apache Lucene, que armazena dados estruturados e não-estruturados como documentos JSON.

O Elasticsearch possui uma API REST simples que permite, visualizar cada um dos documentos e a estrutura e conteúdo dos seus campos (Figura 12).

Figura 12 – Tela da API REST do Elasticsearch



Fonte: Criado pelos autores (2021).

A velocidade e escalabilidade do Elasticsearch, garantidos por sua natureza distribuída, aliados à sua capacidade de facilmente gerar índices invertidos para diversos tipos de conteúdo, tornam o Elasticsearch uma excelente opção para ser utilizada em máquinas de busca. Na máquina virtual do GCP deste projeto foi configurado um *script* em python (app.py) onde, a cada chamada realizada a partir de uma *query* na interface de usuário do buscador semântico, é gerada uma consulta parametrizada específica para as modalidades clássica ou semântica no Elasticsearch.

O Elasticsearch utiliza intrinsicamente dois tipos de função para calcular o escore de similaridade: TF-IDF para versões anteriores à versão 5.0 do Elasticserach e o Okapi BM25 para as versões posteriores. Conforme Baeza-Yates e Ribeiro-Neto (2011), BM25 é um modelo probabilístico alternativo que tem como vantagem, em relação ao modelo probabilístico clássico, de atribuir pesos não apenas ao componente IDF, mas também ao TF e à normalização do tamanho do documento, gerando melhores resultados do que modelos vetoriais clássicos.

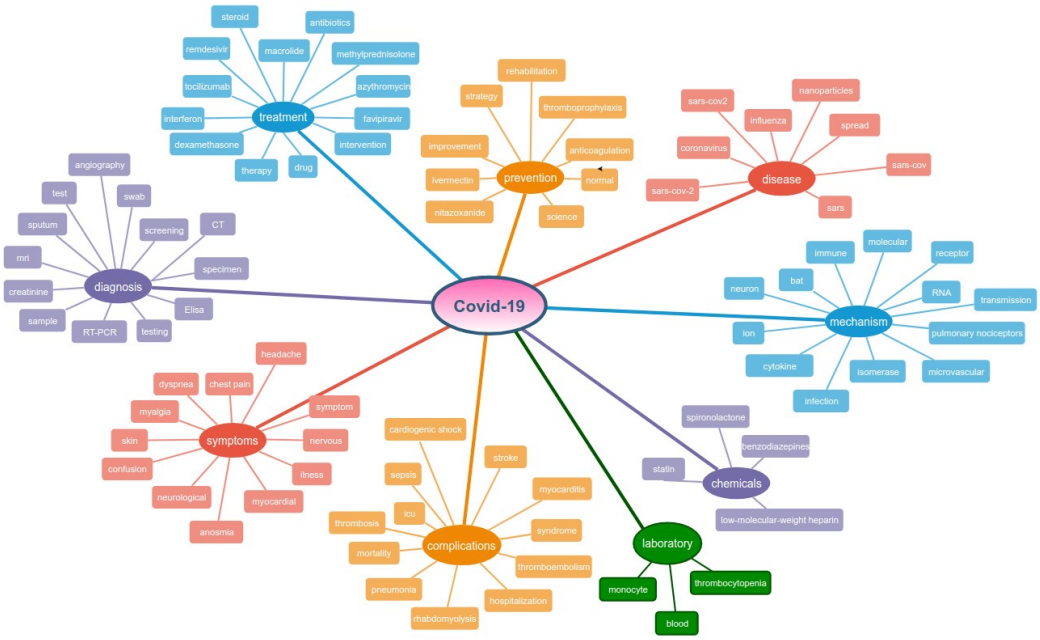
Na Figura 12 também podemos observar que, para cada um dos documentos do CORD-19 persistidos temos as seguintes chaves: _id, _index, _score, type, abstract, entities, entities_freq, id, labels, labels_freq, text, text_embedding e title. O campo ‘entities’ em cada um dos documentos no Elasticsearch recebe as entidades nomeadas biomédicas que foram anotadas em cada um dos respectivos documentos a partir da biblioteca spaCy.

Através da classe ElasticTransformers da biblioteca python sentence-transformers, e utilizando um modelo pré-treinado de propósito geral (all-mpnet-base-v2) inserido no *script* app-py, foi adicionada na busca do tipo ‘SEMANTIC’ no Elasticsearch um componente de análise vetorial de frases, identificada a partir do parâmetro ‘dense’. Na modalidade ‘CLASSIC’ de busca, o tipo de consulta no Elasticsearch foi parametrizada para ‘match’, que utiliza o princípio da frequência de palavras-chave.

4.2.2 spaCy

A biblioteca em python para processamento de linguagem natural spaCy foi utilizada no *script* app.py para *tokenização* e extração de entidades nomeadas biomédicas, que foram anotadas nos respectivos documentos de acordo com as classes representadas pela ontologia simplificada criada pelos autores e que foi representada na Figura 14. As classes definidas para este sistema foram ‘treatment’, ‘prevention’, ‘disease’, ‘mechanism’, ‘chemicals’, ‘laboratory’, ‘complications’, ‘symptoms’ e ‘diagnosis’.

Figura 13 – Representação do conhecimento utilizada para Covid-19



Fonte: Criado pelos autores (2021).

4.2.3 Google Cloud Plataform

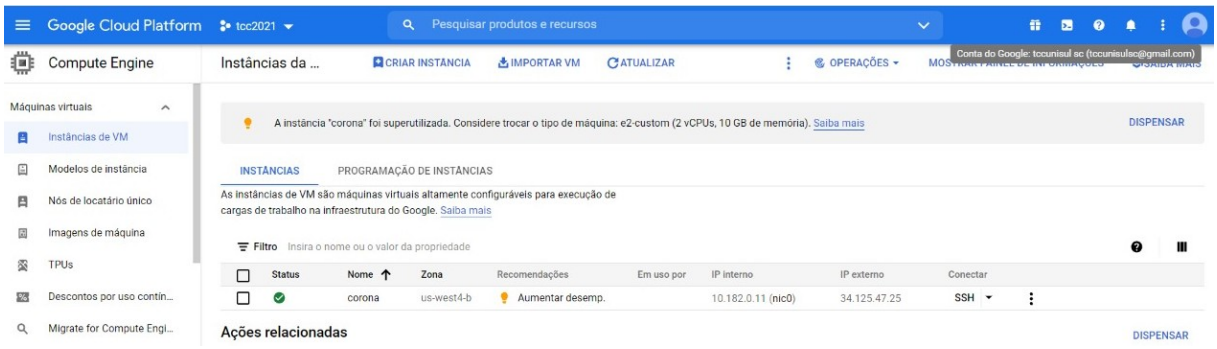
A opção por trabalhar com uma máquina virtual no GCP deve-se ao fato do buscador semântico trabalhar com um enorme volume de dados e exigir trabalho computacional intenso. Para este tipo de necessidade, o modelo de *Infrastructure as a Service* oferece recursos de infraestrutura computacional na nuvem conforme a demanda. Outras vantagens relacionadas com esta escolha estão a possibilidade de escalar dinamicamente a infraestrutura conforme a demanda e a segurança oferecida pelo GCP.

Um recurso do GCP utilizado em nosso sistema foi a disponibilização de um IP externo estático que serve de *link* para acesso do usuário à interface de busca de nossa aplicação.

No intuito de tornar custo-efetiva a utilização de recursos em nuvem do GCP, toda a etapa de desenvolvimento foi executada no Google Colaboratory (Colab), que é um serviço de nuvem gratuito da Google, e apenas quando em produção o sistema passou a ser rodado no GCP.

A Figura 13 mostra a máquina virtual instanciada no GCP, denominada ‘corona’ para este projeto, além de mostrar o seu IP externo, que é a única forma de um usuário externo acessar a máquina virtual a partir da *internet*.

Figura 14 – Máquina virtual no GCP com imagem do Elasticsearch



Fonte: Criado pelos autores (2021).

4.2.4 Streamlit

Streamlit é uma biblioteca python *open-source* que permite que *scripts* sejam facilmente convertidos em aplicações *web*, com poucas linhas de código e dispensando a necessidade de trabalhar com chamadas http, sendo muito utilizado para visualização de projetos relacionados a ciência de dados.

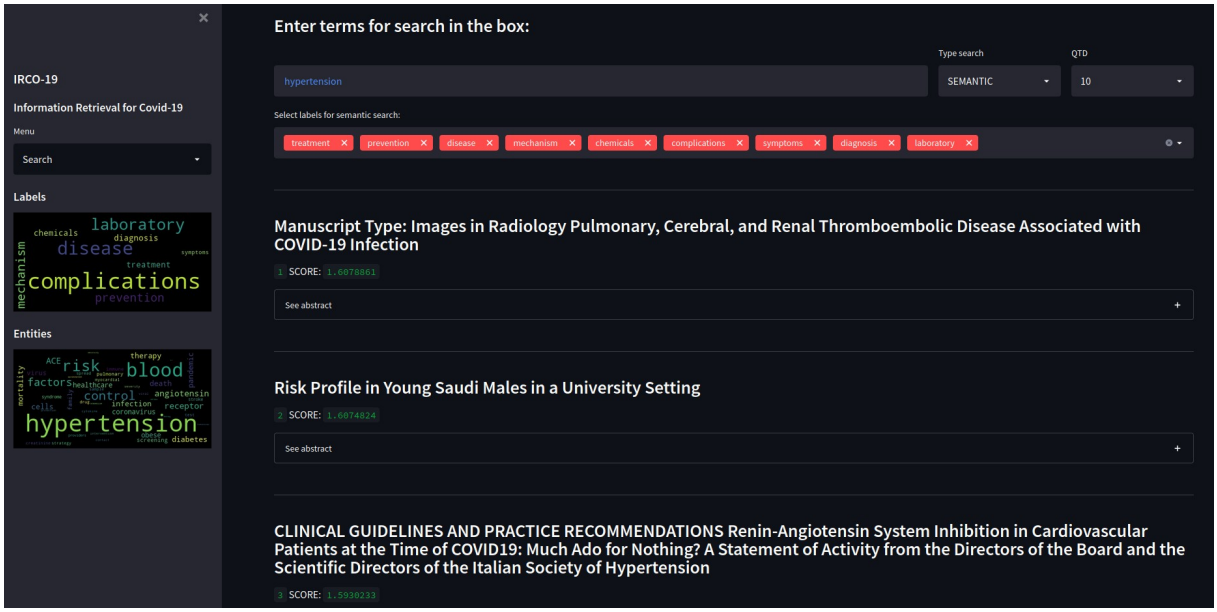
4.3 INTERFACE DE USUÁRIO

O URI para acesso à tela inicial do buscador semântico deste projeto é http://34.125.47.25:8501/.

Na Figura 15 podemos ver tela inicial para o usuário que acessa o *link* do buscador semântico para Covid-19, chamado de IRCO-19 (*Information Retrieval for Covid-19*).

Na caixa de entrada mais superior da tela aparece o termo ‘hypertension’, que foi aleatoriamente definido como *default* para consulta. A caixa de entrada permite a entrada de uma palavra até uma frase sem limite de tamanho.

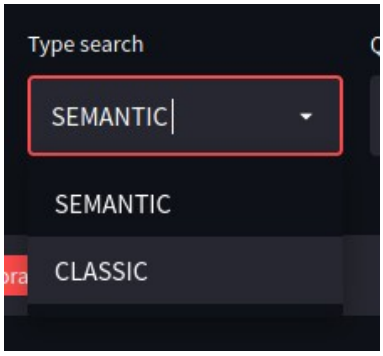
Figura 15 – Tela inicial do IRCO-19



Fonte: Criado pelos autores (2021).

A Figura 16 mostra o campo de escolha para o tipo de busca, com as opções ‘SEMANTIC’ ou ‘CLASSIC’.

Figura 16 – Tipo de busca no IRCO-19



Fonte: Criado pelos autores (2021).

O campo ‘QTD’ permite selecionar a quantidade de documentos mostrados por vez na tela rolável do sistema IRCO-19, permitindo 10, 20, 30, 40 ou 50 inserções.

No campo de recuperação dos documentos são visualizados o título do documento, o índice de relevância (SCORE) do documento comparado com a consulta realizada para o tipo de busca selecionada, seguido pelo texto do documento, no qual, quando expandido, permite visualizar as entidades nomeadas que foram anotadas nos documentos de acordo com a

ontologia Covid-19 simplificada utilizada. A Figura 17 mostra um trecho de um documento com entidades nomeadas anotadas.

Figura 17 – Entidades nomeadas anotadas em um documento

Fonte: Criado pelos autores (2021).

Na parte esquerda da tela do usuário são dispostos dois quadros com nuvens de palavras, conforme podem ser vistos em detalhe na Figura 18. A primeira nuvem de palavras (Labels) contém as classes relacionadas à ontologia simplificada que tiveram maior frequência de aparecimento entre os documentos anotados recuperados para a consulta em questão, em qualquer modalidade de busca. A nuvem de palavras inferior (Entities) refere-se aos termos mais frequentemente encontrados nos documentos recuperados para a consulta.

Figura 18 – Quadros com nuvem de palavras para Labels e Entities

Fonte: Criado pelos autores (2021).

4.4 SIMULAÇÃO DE BUSCA

Para fim de simulação de busca foi realizada uma consulta com a entrada de uma frase que buscava uma informação relacionada ao risco do sintoma tosse para a doença Covid-19: "*Is prolonged cough risky to covid-19 patients ?*".

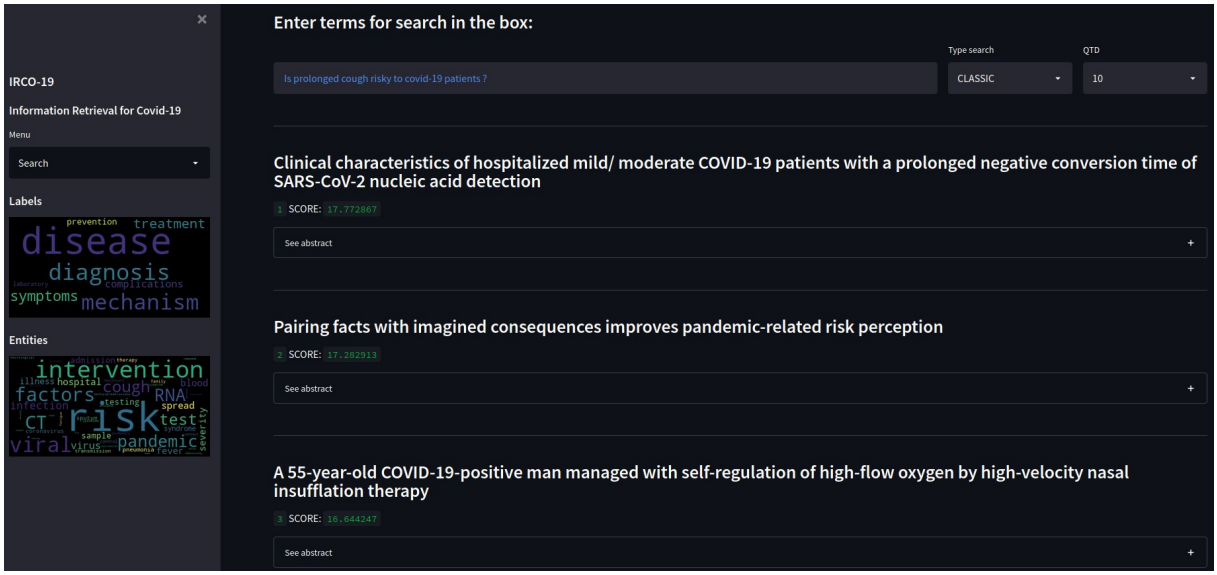
Foram analisados os 20 documentos classificados como mais relevantes pelo buscador para uma consulta baseada em um modelo clássico e os 20 documentos mais relevantes para uma consulta baseada em um modelo incrementado com a utilização do modelo de linguagem pré-treinado sentence-transformer associado ou não à filtragem por categorias de classes definidas pela ontologia simplificada de Covid-19.

Para cada um dos 20 primeiros documentos recuperados para cada um dos modelos foram realizadas validações de forma encoberta e *randomizada* por 10 médicos especialistas quanto à relevância de cada um dos documentos para à consulta, sendo atribuído pelo avaliador o sinal (+) para um documento considerado por ele nada ou pouco relevante em relação à *query*, (++) para os documentos com média relevância e (+++) para os documentos que pela análise do médico trazia informação pertinente e relevante quanto à consulta realizada. Para cada documento foi considerada o nível de relevância que obteve mais votos.

4.4.1 Busca clássica

A Figura 19 apresenta a página de resultados para a busca com um modelo clássico.

Figura 19 – Página de resultados da busca clássica



Fonte: Criado pelos autores (2021).

A Tabela 1 apresenta a lista de artigos recuperados pelo sistema e os índices de relevância calculados pelo sistema para cada um dos documentos em relação à consulta realizada, bem como a validação dos especialistas quanto à relevância.

A nuvem de palavras das classes de entidades nomeadas (Labels) mais relevantes nos documentos recuperados apontou ‘disease’, ‘diagnosis’ e ‘mechanism’ como sendo as classes mais frequentes.

A nuvem de palavras dos termos mais frequentes (Entities) nos documentos recuperados trouxe ‘risk’, ‘RNA’, ‘viral’, ‘intervention’ e ‘pandemic’ como entidades mais relevantes nos documentos recuperados. Este padrão pode ser justificado pelo fato de a busca clássica ser baseada apenas na frequência de termos nos índices invertidos, sem considerar relacionamentos com significado.

Tabela 1 – Score e validação por especialistas dos 20 documentos mais relevantes para a busca no modelo clássico

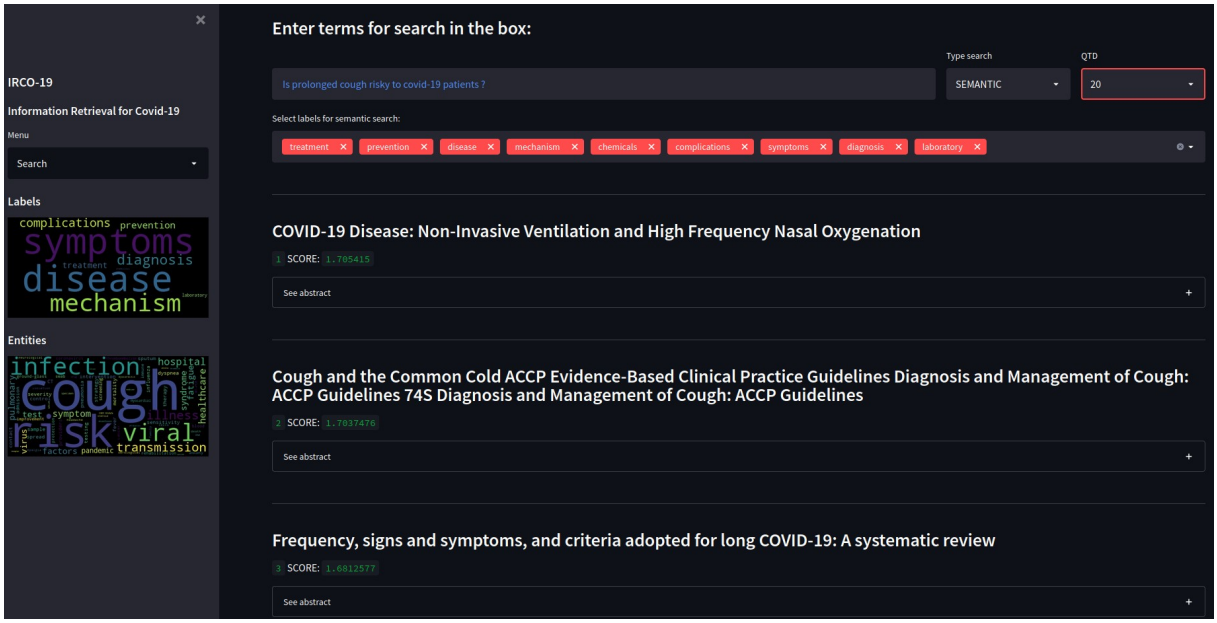
N	Documento	Score	Validação
1	Clinical characteristics of hospitalized mild/ moderate COVID-19 patients with a prolonged negative conversion time of SARS-CoV-2 nucleic acid detection	17.772	++
2	Pairing facts with imagined consequences improves pandemic-related risk perception	17.282	+
3	A 55-year-old COVID-19-positive man managed with self-regulation of high-flow oxygen by high-velocity nasal insufflation therapy	16.644	+
4	COVID-19 induced liver function abnormality associates with age	15.453	+
5	Impact of outpatient SARS-CoV-2 infections in minority children	15.286	++
6	Tracheostomy, ventilatory wean, and decannulation in COVID-19 patients	14.935	+
7	Emergency Conversion to Telehealth in Hospital-Based Psychiatric Outpatient Services: Strategy and Early Observations Emergency ...	14.867	+
8	Clinical relevance of serum Krebs von den Lungen-6 levels in patients with coronavirus disease 2019 Clinical relevance of serum Krebs von ...	14.850	+++
9	Sarcoidosis presenting with glazy mucoid sputum and dyspnea: a case report	14.740	+
10	Geographic Differences on Clinical Manifestations of COVID-19 Infection Between Overseas Chinese and Local Chinese Patients	14.729	+
11	Incidental CT Findings Suspicious for COVID-19-Associated Pneumonia on Nuclear Medicine Examinations Recognition and Management Plan	14.702	++
12	Risk-taking unmasked: Using risky choice and temporal discounting to explain COVID-19 preventative behaviors	14.701	+
13	Profiles of Social Distance Compliance: Psychological and Situational Predictors of Risky Behavior during COVID-19	14.530	+
14	BMC Public Health Gender difference in knowledge of tuberculosis and associated health-care seeking behaviors: a cross-sectional study in a rural area ...	14.424	+
15	Epidemiology and Infection Subcutaneous emphysema and pneumomediastinum in patients with COVID-19 disease; case series from a tertiary care hospital ...	14.111	+++
16	A COVID-19 Outbreak in a Rheumatology Department Upon the Early Days of the Pandemic	14.101	++
17	Love in Quarantine: Sexting, Stress, and Coping During the COVID-19 Lockdown	13.994	+
18	Altered mental status is a predictor of poor outcomes in COVID-19 patients: A cohort study	13.962	++
19	Prolonged Unconsciousness Following Severe COVID-19	13.875	+
20	Indomethacin: Can It Counteract Bradykinin Effects in COVID-19 Patients?	13.752	++

Fonte: Criado pelos autores (2021).

4.4.2 Busca semântica

A Figura 20 apresenta a página de resultados para a busca com um modelo incrementado com o recurso de modelo de linguagem pré-treinado sentence-transformer.

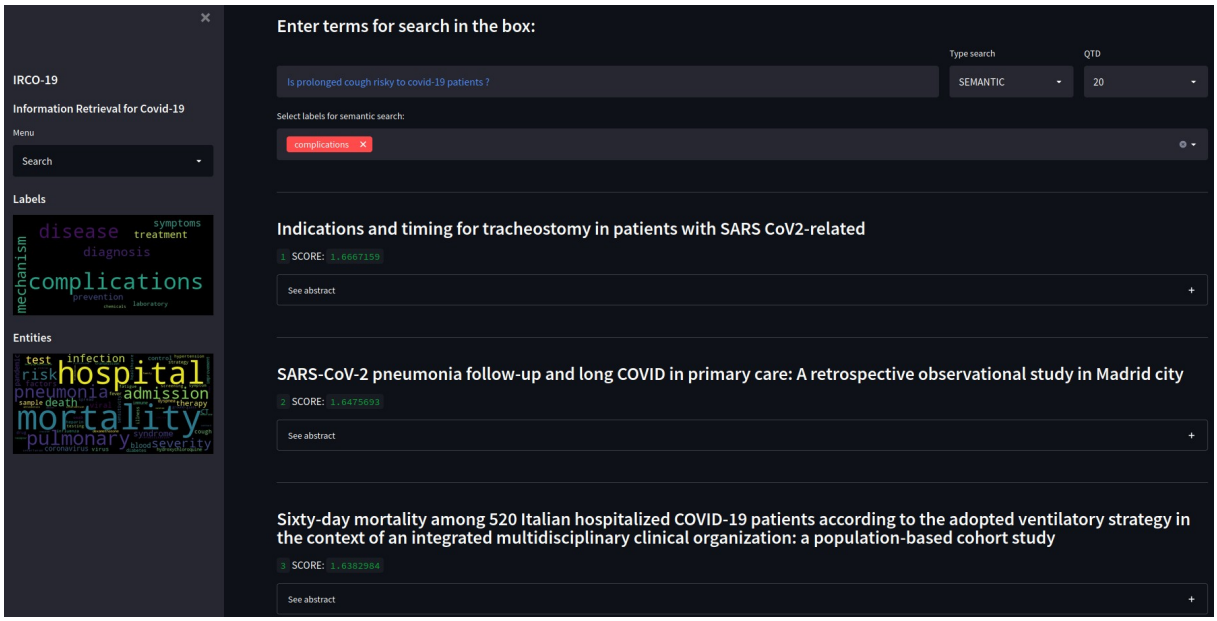
Figura 20 – Página de resultados da busca com recurso de sentence-transformer



Fonte: Criado pelos autores (2021).

A Figura 21 apresenta a página de resultados para a busca com um modelo incrementado com o recurso de modelo de linguagem pré-treinado sentence-transformer seguido de filtragem por categoria, tendo sido selecionada apenas a classe ‘complications’, que está mais relacionada com o contexto da consulta realizada.

Figura 21 – Página de resultados da busca com recurso de sentence-transformer e filtragem



Fonte: Criado pelos autores (2021).

A Tabela 2 apresenta a lista de artigos recuperados pelo sistema com buscador com recurso de sentence-transformer e filtragem por categoria. São demonstrados também os índices de relevância calculados pelo sistema para cada um dos documentos em relação à consulta realizada, bem como a validação dos médicos especialistas quanto à relevância.

Tabela 2 – Score e validação por especialista dos 20 documentos mais relevantes para a consulta do buscador semântico

N	Documento	Score	Validação
1	Indications and timing for tracheostomy in patients with SARS CoV2-related	1.666	+
2	SARS-CoV-2 pneumonia follow-up and long COVID in primary care: A retrospective observational study in Madrid city	1.647	+++
3	Sixty-day mortality among 520 Italian hospitalized COVID-19 patients according to the adopted ventilatory strategy in the context of an integrated multi ...	1.638	++
4	Tracheostomy in patients with COVID-19: predictors and clinical features	1.622	+++
5	Infectious Disease Initial assessment and management of respiratory infections in persons with spinal cord injuries and disorders in the COVID-19 era	1.622	+
6	Respiratory characteristics and related intraoperative ventilatory management for patients with COVID-19 pneumonia	1.621	+
7	Airway Hygiene in COVID-19 Pneumonia: Treatment Responses of 3 Critically Ill Cruise Ship Employees	1.620	++
8	Dysphagic disorder in a cohort of COVID-19 patients: evaluation and evolution	1.619	+
9	Mechanical ventilation utilization in COVID-19: A systematic review and meta-analysis	1.619	++
10	No Efficacy of the Combination of Lopinavir/Ritonavir Plus Hydroxychloroquine Versus Standard of Care in Patients Hospitalized With COVID-19: A Non- ...	1.618	++
11	Corticosteroids in the treatment of severe Covid-19 lung disease: the pulmonology perspective from the first United States epicenter	1.611	+
12	The value of tracheostomy of critically ill COVID-19 patients -A multicentral study	1.606	++
13	Clinical features, ventilatory management, and outcome of ARDS caused by COVID-19 are similar to other causes of ARDS	1.606	+++
14	Epidemiological Characteristics of COVID-19; a Systemic Review and Meta-Analysis 1	1.606	+++
15	Early vs. late tracheostomy in ventilated COVID-19 patients -A retrospective study	1.605	+++
16	Pulmonary function evaluation after hospital discharge of patients with severe COVID-19	1.605	+++
17	Therapeutic Advances in Respiratory Disease Early predictors for mechanical ventilation in COVID-19 patients	1.604	+++
18	Short-Term Consequences of SARS-CoV-2-Related Pneumonia: A Follow Up Study	1.604	++
19	Clinical review of COVID-19 patients presenting to a quaternary care private hospital in South India: A retrospective study	1.603	+++
20	Corticosteroid treatment in severe COVID-19 pneumonia: two cases and literature review	1.601	+++

Fonte: Criado pelos autores (2021).

A nuvem de palavras das classes de entidades nomeadas mais relevantes nos documentos recuperados no buscador semântico apontou ‘complications’, ‘disease’ e ‘mechanism’ como sendo as classes mais frequentes.

A nuvem de termos mais frequentes nos documentos recuperados trouxe ‘mortality’, ‘hospital’, ‘pulmonary’, ‘pneumonia’ e ‘admission’ como entidades mais relevantes nos documentos recuperados na busca na modalidade ‘SEMANTIC’.

4.5 AVALIAÇÃO DOS RESULTADOS

A Tabela 3 mostra a distribuição das categorias de validação dos documentos para o modelo clássico e para o buscador semântico em relação à relevância quanto à consulta.

Tabela 3 – Distribuição das categorias de validação do documento para o modelo clássico para o buscador semântico

Modelo \ Relevância	+	++	+++
Clássico	12 (60%)	6 (30%)	2 (10%)
Semântico	5 (25%)	6 (30%)	9 (45%)

Fonte: Criado pelos autores (2021).

Houve uma melhora na relevância dos documentos recuperados com o buscador semântico, que passou, entre os 20 primeiros documentos, de 2 documentos (10%) avaliados como muito relevantes (+++) com o modelo ‘CLASSIC’ para 9 documentos (45%) muito relevantes com o modelo ‘SEMANTIC’. Paralelamente, houve uma redução significativa dos documentos recuperados avaliados como sem relação com a *query* realizada (+) de 12 (60%) com o modelo ‘CLASSIC’ para 5 (25%) com o modelo ‘SEMANTIC’.

Vale ressaltar que a *query* realizada incluía o termo ‘risky’, que poderia em um processo de *tokenização* ser levado a orientar a busca para documentos que levassem mais ao termo ‘risk’, conforme foi observado na busca do modelo clássico, que teve esta como uma das principais entidades na nuvem de palavras. Contudo, na busca semântica, foram levantadas como entidades principais ‘mortality’ e ‘hospital’, termos muito relacionados com o contexto intrínseco da *query* que buscava saber dos riscos de tosse prolongada.

Outra observação importante refere-se aos escores gerados pelo Elasticsearch, que utilizam a função de similaridade BM25. Embora esta função leve em consideração tanto o TF, o IDF e o tamanho do documento normalizado no cálculo dos pesos dos termos, ainda assim a máquina de busca considerará fundamentalmente a presença de termos contidos nas *querys* e suas incidências nos documentos para calcular a relevância dos documentos. Dessa forma, o buscador clássico, que possui sua máquina no Elasticsearch com parâmetro de 'match' para termos, trará documentos com escores calculados muito maiores que na busca semântica, mesmo que o contexto do documento relacionado não esteja relacionado completamente com a necessidade de informação do usuário. Assim, a busca semântica trouxe documentos com escores menores, mas que após a validação demonstraram muito mais relevância dos documentos em relação à necessidade de informação expressa na *query*.

4.6 LIMITAÇÕES DO SISTEMA

A primeira etapa deste projeto, que foi baseado na recomendação de arquitetura proposta por Wang e Lo (2021), consistia na transferência dos artigos do CORD-19 para a imagem do Elaticsearch instanciada na GCP. Este processo foi realizado apenas uma vez durante o desenvolvimento deste projeto, utilizando o *release* do CORD-19 de 27/01/2021. Contudo, o repositório CORD-19 tem sido atualizado à uma cadência semanal e, portanto, diversos novos artigos foram integrados a este repositório. Para este SRI ser utilizado em produção para a comunidade científica que busca informação atualizada, faz-se necessário implementar uma rotina para atualização automática do repositório instanciado na GCP.

A utilização de uma ontologia simplificada criada pelos autores para definir as entidades nomeadas relacionadas ao Covid-19 a serem anotadas nos documentos buscava introduzir mais um componente para aumento da eficiência da busca semântica. A opção de criar uma estrutura super simplificada de classes e objetos para este complexo tema ocorreu em função da limitação de tempo para o desenvolvimento do projeto e da complexidade dos limites do conhecimento para o tema Covid-19. Uma ontologia específica para o Covid-19, criada no editor de ontologias Protégé sob os princípios definidos pela *Open Biological and Biomedical Ontology,* compreende 2270 classes, vem sendo aplicada com sucesso para auxiliar em processos de extração de conhecimento e no auxílio da interoperabilidade semântica no domínio do Covid-19. Embora esta ontologia esteja disponível para utilização já como um arquivo owl, e mesmo considerando que poderia enriquecer o poder do buscador semântico deste projeto, em acordo com a definição das tecnologias da *web* semântica

apresentadas anteriormente, embarcá-la no sistema representaria mais tempo para entender sua incorporação no código e para treinamento do reconhecimento de entidades nomeadas e anotações nos documentos. Vale ressaltar que a incorporação desta ontologia otimizada para o Covid-19 também teria um papel fundamental para a criação de grafos de conhecimento, o que agregaria a possibilidade de implementar modelos mais robustos de recuperação da informação.

5 CONCLUSÃO

Embora aparentemente "trivial" para qualquer ser humano moderno comunicar-se e compreender o mundo utilizando linguagem natural logo a partir dos primeiros anos de suas vidas, todas as etapas cognitivas cerebrais para construir significado a partir da linguagem foram desenvolvidas no ser humano com um esforço de adaptação de milhares de anos de evolução do órgão extremamente sofisticado, evoluído e complexo que é o cérebro humano.

Apesar de todo avanço tecnológico que dispomos hoje em dia, ainda temos muita dificuldade em fazer com que o computador trabalhe com a linguagem natural com tamanha simplicidade quanto fazemos em nosso dia a dia. A variabilidade das construções sintáticas das diferentes línguas e ambiguidade que pode estar carregada, mesmo a nível de palavra, tornam muito difícil criar regras automatizadas para processamento da linguagem natural.

A linguagem ainda permitiu o ser humano criar conhecimento e agora, no mundo contemporâneo digital, presenciamos uma explosão na quantidade de informação gerada a ponto de não conseguirmos mais extrair informação dos repositórios sem a ajuda de sistemas de recuperação de informação. A Biomedicina estuda toda a complexidade que envolve os organismos biológicos, mais especificamente os seres humanos. Paradoxalmente, já está "humanamente impossível" lidar com todo o conhecimento disponível sobre os seres humanos a nível biomédico. O Covid-19 pode ter catalisado mais um salto para a capacidade de produção exponencial de informação. Assim, criar sistemas eficientes para recuperação de informação útil passou a ser urgente.

O entendimento do princípio de um sistema de recuperação pode ser resumido, de forma muito simplificada, em 3 componentes principais.

O primeiro consiste na consulta, a famosa '*query*', mas traz na sua simplicidade a necessidade de compreensão que ela existe porque há uma necessidade de informação de alguém e, na grande maioria das vezes, a *query* representa em si toda a possibilidade de ambiguidade da linguagem natural, que mesmo para cérebros evoluídos, frequentemente há uma "mensagem de erro" do tipo "não entendi a pergunta!".

O segundo componente consiste no processo de busca. Agora, se alguns poucos termos já podem gerar dúvida, a leitura por uma máquina de um universo de documentos em busca de uma informação tem inúmeros obstáculos, desde a usabilidade das interfaces de busca, o próprio acesso às bases dos documentos, o reconhecimento dos seus formatos, o manuseio dos metadados, os diferentes modelos matemáticos para cálculo da relevância dos documentos, a necessidade de um pré-processamento de textos antes da análise, a correta

utilização de métricas de avaliação, dentre outros tantos desafios. Mas talvez o maior dos desafios para ser programado em uma máquina seja o entendimento do contexto em que uma consulta está inserida. O que de fato é a necessidade de informação do usuário? Mesmo com um termo ou conjunto de termos inseridos idênticos, usuários diferentes podem estar esperando informações diferentes, e pior, às vezes completamente diferentes. Como fazer a máquina de busca extrair, antes de executar o seu trabalho, a direção correta da necessidade do usuário para que mais documentos importantes sejam recuperados de forma eficiente ao mesmo tempo que documentos não importantes não sejam apresentados ao usuário que não quer desperdiçar o seu bem mais precioso (o tempo)?

Com esta necessidade de trabalhar com contexto, muitas técnicas de processamento de linguagem natural têm feito verdadeiros milagres, simplificando o processo de extração de informação de textos. Neste projeto trabalhamos muito superficialmente com o spaCy, um fantástico *framework* em python para processamento de linguagem natural que oferece muitas possibilidades para facilitar a compreensão e a análise do conteúdo do corpus, incluindo o reconhecimento de entidades nomeadas.

À medida que se evolui no aprendizado destas técnicas evoluímos paralelamente na compreensão da imensa complexidade de nossa linguagem e nas inúmeras possibilidades de arquiteturas disponíveis para um sistema de recuperação da informação. Impressionante mesmo é observar o poder dos modelos de linguagem pré-treinados, que são *frameworks* de inteligência artificial, mais especificamente *deep learning*, que de forma não facilmente compreensível ensinam como encontrar contextos e relacionamentos entre palavras e sentenças contidos em textos de todo um corpus, o que parece fornecer um impulso nas aplicações que embarcam esta tecnologia. No sistema apresentado neste projeto embarcamos um *framework* muito poderoso chamado sentence-transformer, que com calcula o vetor denso de representações que nada mais é que um "nível de afinidade interna" entre as frases e palavras de um texto para depois comparar com as buscas a fim de inseri-las dentro de um contexto similar.

O terceiro componente do sistema de recuperação é o próprio retorno dos documentos ao usuário, e este pode incluir mecanismos para melhorar a relevância dos resultados. Neste projeto embarcamos a filtragem por categorias baseadas em uma ontologia simplificada para o Covid-19.

A utilização de representações do conhecimento biomédico, mais especificamente de ontologias, apesar de representar um desafio a mais na implementação e provável caminho para evolução deste projeto, acrescentam muito mais eficiência para um buscador semântico,

uma vez que são uma camada intermediária nos sistemas que buscam entender os relacionamentos semânticos não de palavras e frases gerais da nossa língua, mas permitindo revelar muitos relacionamentos escondidos na imensidão da informação do domínio biomédico, que podem gerar valor inestimável para o usuário e para toda a humanidade.

A funcionalidade de busca semântica do projeto permitiu a recuperação de um número maior de documentos relevantes no contexto do Covid-19.

Concluindo, a utilização de representações do conhecimento no domínio biomédico associada a modelos de linguagem pré-treinados que permitem associar a análise de contexto a nível de palavras e sentenças podem aumentar a relevância dos resultados de um sistema de recuperação de informação na área biomédica.

REFERÊNCIAS

ALAOUI, K.; BAHAJ, M. Semantic oriented data modeling based on RDF, RDFS and OWL. In: EZZIYYANI, M. (Ed.). . **Advanced intelligent systems for sustainable development (AI2SD'2019)**. Advances in Intelligent Systems and Computing. Cham: Springer International Publishing, 2020. v. 1105p. 411–421.

AMAZONAS, A. M. *et al*. Integrating data index engines in order to build information retrieval systems In Heterogeneous Environments. **JISTEM Journal of Information Systems and Technology Management**, v. 5, n. 2, p. 193–222, 1 ago. 2008.

ANDRADE, R. *et al*. **An Approach to Semantic Search in Medical Databases.** 8th International Information and Telecommunication Technologies Symposium. **Anais…** In: 8TH INTERNATIONAL INFORMATION AND TELECOMMUNICATION TECHNOLOGIES SYMPOSIUM. Florianópolis, Brasil: Fundação Bardal de Educação e Cultura, 2009. Disponível em: <https://www.researchgate.net/publication/235958741_An_Approach_to_Semantic_Search_in_Medical_Databases/>. Acesso em: 29 set. 2020.

AWATRAMANI, Vasudev; GUPTA, Pooja. Natural Language Transfer Learning for Physiological Textual Similarity. In: 10TH INTERNATIONAL CONFERENCE ON CLOUD COMPUTING, DATA SCIENCE & ENGINEERING (CONFLUENCE), 10., 2020, Noida. **10th International Conference on Cloud Computing, Data Science & Engineering (Confluence).** [S.L.]: IEEE, 2020. p. 733-737.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval: the concepts and technology behind search**. Second edition ed. New York: Addison Wesley, 2011.

BAHAMMAM, A. S.; JAHRAMI, H.; FARIS, M. A. Non-COVID-19 research and publications must never be at the mercy of COVID-19 research. **Journal of Nature and Science of Medicine**, v. 3, n. 3, p.143-145, 2020.

BAST, H.; BUCHHOLD, B.; HAUSSMANN, E. Semantic Search on Text and Knowledge Bases. **Foundations and Trends in Information Retrieval**, v. 10, n. 1, p. 119–271, 2016.

BODENREIDER, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. **Nucleic Acids Research**, v. 32, n. 90001, p. 267D – 270, 1 jan. 2004.

BOULOS, M. N. K. A first look at HealthCyberMap medical semantic subject search engine. **Technology and Health Care**, v. 12, n. 1, p. 33–41, 2004.

BRAMER, W. M. et al. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. **Systematic Reviews**, v. 6, n. 1, p. 245, dez. 2017.

CAFE, L.; BRASCHER, M. Organização do Conhecimento: teorias semânticas como base para estudo e representação de conceitos. **Informação & Informação**, v. 16, n. supl, p. 25–51, 23 dez. 2011.

CALLAHAN, T. J. et al. Knowledge-Based Biomedical Data Science. **Annual Review of Biomedical Data Science**, v. 3, n. 1, p. 23–41, 20 jul. 2020.

CARDOSO, O. N. P. Recuperação de Informação. **INFOCOMP Journal of Computer Science**, v. 2, n. 1, p. 33-38, 1 Nov. 2004.

CARROLL, J. J. et al. **Jena: implementing the semantic web recommendations**. Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters - WWW Alt. '04. **Anais**... In: THE 13TH INTERNATIONAL WORLD WIDE WEB CONFERENCE. New York, NY, USA: ACM Press, 2004Disponível em: <http://portal.acm.org/citation.cfm?doid=1013367.1013381>. Acesso em: 22 nov. 2020

CATARINO, M. E.; SOUZA, T. B. DE. A representação descritiva no contexto da web semântica. **Transinformação**, v. 24, n. 2, p. 77–90, ago. 2012.

CECI, Flávio. **Um modelo baseado em casos e ontologia para apoio à tarefa intensiva em conhecimento de classificação com foco na análise de sentimentos.** 2015. Tese (Doutorado) – Universidade Federal de Santa catarina, Florianópolis, 2015.

CECI, F.; WOSZEZENKI, C. R.; GONÇALVES, A.L. O uso de anotações semânticas e ontologias para a classificação de documentos. **International Journal of Knowledge Engineering and Management**, v.3, n. 5, p. 1-14, mar/jun. 2014.

CONEGLIAN, C. S. *et al*. O papel da web semântica nos processos do big data. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 23, n. 53, p. 137–146, 6 set. 2018.

DAVIS, R.; SHROBE, H.; SZOLOVITS, P. What Is a Knowledge Representation? **AI Magazine**, v.14, p.17-33, 1993.

DEVLIN, J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. **arXiv:1810.04805 [cs]**, 24 maio 2019.

FATEHI, F.; GRAY, L. C.; WOOTTON, R. How to improve your PubMed/MEDLINE searches: 1. background and basic searching. **Journal of Telemedicine and Telecare**, v. 19, n. 8, p. 479–486, dez. 2013.

FAZZINGA, B. et al. Semantic Web search based on ontological conjunctive queries. **Journal of Web Semantics**, v. 9, n. 4, p. 453–473, dez. 2011.

GEZER, Volkan; BERGWEILER, Simon. Cloud-based Infrastructure for Workflow and Service Engineering Using Semantic Web Technologies. **International Journal On Advances In Internet Technology.** Kaiserslautern, p. 36-45. jun. 2017.

GONZALEZ, Marco; LIMA, Vera L. S. de. **Recuperação de Informação e Processamento da Linguagem Natural**. 2003. PUCRS - Faculdade de Informática. Disponível em: https://www.marilia.unesp.br/Home/Instituicao/Docentes/EdbertoFerneda/mri-06---gonzales-e-lima-2003.pdf. Acesso em: 13 out. 2021.

GOULART, R. R. V.; STRUBE DE LIMA, V. L.; XAVIER, C. C. A systematic review of named entity recognition in biomedical texts. **Journal of the Brazilian Computer Society**, v. 17, n. 2, p. 103–116, jun. 2011.

KOOPMAN, B. et al. Information retrieval as semantic inference: a Graph Inference model applied to medical search. **Information Retrieval Journal**, v. 19, 20 nov. 2015.

KOOPMAN, B. et al. Towards semantic search and inference in electronic medical records: An approach using concept-based information retrieval. **The Australasian Medical Journal**, v. 5, n. 9, p. 482–488, 30 set. 2012.

LUO, J. *et al*. Big Data Application in Biomedical Research and Health Care: A Literature Review: **Biomedical Informatics Insights**, 19 jan. 2016.

MEDEIROS, W.M.C. **SISOnt: Sistema de Informação em Saúde Baseado em Ontologias**. 2009. Dissertação (Mestrado em Engenharia de Computação) – Universidade Federal do Rio Grande do Norte, Natal, 2009.

MESSINA, A. et al. BioGrakn: A Knowledge Graph-Based Semantic Database for Biomedical Sciences. In: BAROLLI, L.; TERZO, O. (Eds.). . **Complex, Intelligent, and Software Intensive Systems**. Advances in Intelligent Systems and Computing. Cham: Springer International Publishing, 2018. v. 611p. 299–309.

MIROŃCZUK, M. M.; PROTASIEWICZ, J. A recent overview of the state-of-the-art elements of text classification. **Expert Systems with Applications**, v. 106, p. 36–54, set. 2018.

MONTEIRO, S. D. et al. Sistemas de recuperação da informação e o conceito de relevância nos mecanismos de busca: semântica e significação. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 22, n. 50, p. 161, 6 set. 2017.

NCBI. **PubMed Clinical Queries**. Disponível em: https://pubmed.ncbi.nlm.nih.gov/clinical/. Acesso em: 24 out. 2021.

NEUMANN, M. et al. **ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing**. Proceedings of the 18th BioNLP Workshop and Shared Task. **Anais**... In: PROCEEDINGS OF THE 18TH BIONLP WORKSHOP AND SHARED TASK. Florence, Italy: Association for Computational Linguistics, 2019. Disponível em: <https://www.aclweb.org/anthology/W19-5034>. Acesso em: 13 out. 2021

NIETTO, P.; CAMARGOS, R. Interface para Busca Semântica: Um Estudo sobre o Estado da Arte. 2015. Disponível em: <https://www.researchgate.net/publication/285600076_Interface_para_Busca_Semantica_Um_Estudo_sobre_o_Estado_da_Arte>. Acesso em: 20 nov. 2020

OHTA *et al*. An intelligent search engine and GUI-based efficient MEDLINE search tool based on deep syntactic parsing. In: COLING/ACL Interactive presentation sessions, 2006, Sdney, Australia. **An intelligent search engine and GUI-based efficient MEDLINE search tool based on deep syntactic parsing**, p.17-20,2006.

PASQUIER, C. Biological data integration using Semantic Web technologies. **Biochimie**, v. 90, n. 4, p. 584–594, abr. 2008.

PECH, F. et al. Semantic Annotation of Unstructured Documents Using Concepts Similarity. **Scientific Programming**, v. 2017, p. 1–10, 2017.

PICKLER, M. E. V. Web Semântica: ontologias como ferramentas de representação do conhecimento. **Perspectivas em Ciência da Informação**, v. 12, n. 1, p. 65–83, abr. 2007.

PRODANOV, C. C.; FREITAS, E. C. **Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico**. 2. ed. Rio Grande do Sul: ASPEUR Universidade Feevale, 2013.

ROCHA, A. M.; SALVADOR, L. N.; SOUZA, M. V. S. Anotações Semânticas em Repositórios Acadêmicos. In: XXIII SIMPÓSIO BRASILEIRO DE SISTEMAS MULTIMÍDIA E WEB, 23., 2017, Porto Alegre. **Anotações Semânticas em Repositórios Acadêmicos.** Porto Alegre: SBC, p. 35-40, 2017.

ROZSA, V.; VIERA, A.F.G.; DUTRA, M. Aplicação de Tecnologias da Web Semântica em Motores de Busca na Internet. **Investigación bibliotecológica**, v. 33, n. 78, p. 165–191, mar. 2019.

RUSSEL, Stuart; NORVIG, Peter. Inteligência Artificial. 3ª edição (tradução). Rio de Janeiro. Elsevier, 2013.

SANTARÉM SEGUNDO, J. E.; CONEGLIAN, C. S.; LUCAS, E. R. DE O. Conceitos e tecnologias da Web semântica no contexto da colaboração acadêmico-científica: um estudo da plataforma Vivo. **Transinformação**, v. 29, n. 3, p. 297–309, dez. 2017.

SASAKI, Y. et al. How to make the most of NE dictionaries in statistical NER. **BMC Bioinformatics**, v. 9, n. S11, p. S5, dez. 2008.

SHEHNAZ, A.; QAMAR, U.; MUZAFFAR, A. W. **Building a Core Biomedical Knowledge Base**. 2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService). **Anais**... In: 2016 IEEE SECOND INTERNATIONAL CONFERENCE ON BIG DATA COMPUTING SERVICE AND APPLICATIONS (BIGDATASERVICE). Oxford, United Kingdom: IEEE, mar. 2016Disponível em: <http://ieeexplore.ieee.org/document/7474385/>. Acesso em: 2 nov. 2020

SIEVERT, M. C. *et al*. Retrieval from full-text medical literature: the dream & the reality. **Proceedings. Symposium on Computer Applications in Medical Care**, p. 348–352, 1991.

SILVA, E.L.; MENEZES. E.M. **Metodologia da Pesquisa e Elaboração da Dissertação.** 4. ed. Florianópolis: UFSC, 2005.

SINGH, A. et al. (EDS.). **Web Semantics for Textual and Visual Information Retrieval:** [s.l.] IGI Global, 2017.

SOMMERVILLE, Ian. Projeto de arquitetura. In: SOMMERVILLE, Ian. **Engenharia de Software**. 9. ed. São Paulo: Pearson, 2011. Cap. 6. p. 103-123. Tradução de Kalinka Oliveira e Ivan Bosnic.

SOTO, A. J.; PRZYBYŁA, P.; ANANIADOU, S. Thalia: semantic search engine for biomedical abstracts. **Bioinformatics**, v. 35, n. 10, p. 1799–1801, 15 maio 2019.

SOUZA, R. R. Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências. **Perspectivas em Ciência da Informação**, v. 11, n. 2, p. 161–173, ago. 2006.

SPACY: Library Architecture. Library Architecture. 2021. Disponível em: https://spacy.io/api. Acesso em: 22 out. 2021.

SUBRAMANYAM, N.; KRISHNAMURTHY,M.;ASUNDI,A.Y.. Developmental features of biomedical bibliographic databases: **Annals of Library and Information Studies,** p.16-20, mar. 2017.

VANOPSTAL, K. et al. Vocabularies and Retrieval Tools in Biomedicine: Disentangling the Terminological Knot. **Journal of Medical Systems**, v. 35, n. 4, p. 527–543, ago. 2011.

VASWANI, A. et al. Attention Is All You Need. **arXiv:1706.03762 [cs]**, 5 dez. 2017.

VENKATESH,R. et al. A Review on Biomedical Mining: **Journal of RNA and Genomics**, p.629-637, nov.2019.

WANG, L. L.; LO, K. Text mining approaches for dealing with the rapidly expanding literature on COVID-19. **Briefings in Bioinformatics**, v. 22, n. 2, p. 781–799, 22 mar. 2021.

WANG, L. L. et al. CORD-19: The COVID-19 Open Research Dataset. **arXiv:2004.10706 [cs]**, 10 jul. 2020.

WEI, W.; BARNAGHI, P. M.; BARGIELA, A. **Semantic-Enhanced Information Search and Retrieval**. Sixth International Conference on Advanced Language Processing and Web Information Technology. **Anais**... In: SIXTH INTERNATIONAL CONFERENCE ON ADVANCED LANGUAGE PROCESSING AND WEB INFORMATION TECHNOLOGY. Luoyang, Henan, China: IEEE, 2007. Disponível em: <http://ieeexplore.ieee.org/document/4460643/>. Acesso em: 28 set. 2020.

WISE, C. et al. COVID-19 Knowledge Graph: Accelerating Information Retrieval and Discovery for Scientific Literature. **arXiv:2007.12731 [cs]**, 24 jul. 2020.

WU, C.; BUYYA, R.; RAMAMOHANARAO, K. Big Data Analytics = Machine Learning + Cloud Computing. In: **Big Data**. [s.l.] Elsevier, 2016. p. 3–38.

YEPES, A.J.; BERLANGA, R. Knowledge based word-concept model estimation and refinement for biomedical text mining. **Journal of Biomedical Informatics**, v. 53, p. 300–307, fev. 2015.