



Universidade do Minho
Escola de Engenharia

Trabalho Prático 2

Processamento de Linguagem Natural em Engenharia Biomédica

Membros do grupo:

Ana Beatriz Salgado Andrade PG56107

Filipa José Rodrigues de Araújo Costa PG56123

Leonor de Amorim Pereira PG57813

Professores:

Luís Filipe Cunha

José João Almeida

Resumo

Este trabalho descreve o desenvolvimento de um sistema terminológico médico baseado em técnicas de Processamento de Linguagem Natural. O sistema integra três glossários e enriquece-os com dados externos obtidos via web scraping do *Vocabulaire de la Médecine* (OQLF). Após a normalização e fusão dos dados, foram aplicadas técnicas de embeddings para análise semântica e identificação de relações entre termos. O resultado é um dicionário médico multilíngue, enriquecido e estruturado, acessível através de uma aplicação web desenvolvida com Flask e Bootstrap. A solução suporta funcionalidades como pesquisa semântica, filtragem por domínios e visualização detalhada dos conceitos, demonstrando o potencial de ferramentas computacionais na gestão de terminologia em saúde.

Conteúdo

1	Introdução	3
2	Descrição do Dataset Inicial	4
3	Enriquecimento do Dataset	5
3.1	Junção dos Datasets	5
3.2	Web Scraping	6
3.3	Integração dos Dados	7
3.4	Validação e Qualidade dos Dados	8
3.5	Exemplo Ilustrativo	8
4	Análise de Conceitos e Relações	9
4.1	Similaridade semântica	9
4.2	Agrupamento por domínios	9
5	Desenvolvimento da Ferramenta Web	10
5.1	Tecnologias utilizadas	10
5.2	Funcionalidades	11
6	Testes e Validação	14
7	Conclusão	20

1 Introdução

Este trabalho teve como principal objetivo o desenvolvimento de um sistema capaz de integrar, enriquecer e representar dados terminológicos médicos, previamente extraídos no Trabalho Prático 1. O ponto de partida consistiu na fusão de três dicionários distintos: o "Diccionari Multilingüe de la COVID-19", o "Glossário Temático de Monitoramento e Avaliação" e o "Glossário de Neologismos em Saúde". Estes conjuntos de dados foram inicialmente normalizados para uma estrutura comum, assegurando coerência interna e facilitando a sua manipulação e posterior enriquecimento.

Numa segunda fase, procedeu-se à expansão deste dicionário unificado através da integração de informação externa recolhida por técnicas de *web scraping*. A fonte escolhida para este enriquecimento foi o "Vocabulaire de la médecine", disponibilizado pelo *Office québécois de la langue française (OQLF)* [1]. Foram extraídos automaticamente conceitos médicos, definições, áreas temáticas, termos equivalentes em inglês e expressões utilizadas em contexto. Todos os termos foram traduzidos do francês para o português e integrados no dataset, respeitando a estrutura existente e evitando redundâncias ou conflitos com os dados previamente existentes.

Além da fusão e enriquecimento dos dados, o sistema incorpora um módulo avançado de análise de similaridade semântica, baseado em embeddings de texto. Utilizando o modelo *SentenceTransformer* 'pt-mteb/average_fasttext_cc.pt.300', são geradas representações vectoriais dos conceitos médicos que capturam o seu significado contextual. Esta abordagem permite identificar automaticamente relações semânticas entre termos, sugerir agrupamentos conceituais com base na proximidade vetorial, detetar sinónimos implícitos e potenciais duplicações, bem como realizar buscas por similaridade semântica para além da correspondência literal.

Este módulo inclui ainda um mecanismo de *boosting* que aumenta a relevância dos resultados quando o termo de pesquisa surge no nome do conceito ($\alpha = 0.2$) ou na sua descrição ($\beta = 0.1$), melhorando a qualidade da recuperação da informação.

Para operacionalizar estas funcionalidades, foi desenvolvida uma aplicação *web* completa utilizando o *framework Flask*. Esta plataforma disponibiliza uma interface de consulta com navegação alfabética e por categorias temáticas, visualização detalhada de todos os atributos (definições, sinónimos, traduções multilingues, códigos alternativos), e um sistema de pesquisa com filtros combináveis por dicionário de origem, categoria lexical e área temática.

No que respeita à manutenção dos dados, a aplicação suporta operações completas de

criação, com persistência imediata das alterações num ficheiro *JSON*, validação de campos obrigatórios e normalização automática da informação introduzida.

A arquitetura técnica assenta num *backend* em *Python* com *Flask* para a gestão da lógica de negócio e *routing*, um *frontend* responsivo baseado em *Bootstrap 5*, templates dinâmicos *Jinja2* e um sistema de paginação adaptado a grandes volumes de dados.

A aplicação constitui uma prova de conceito sólida das capacidades do sistema desenvolvido, combinando técnicas de processamento de linguagem natural com uma interface intuitiva e eficiente para a gestão de terminologia médica multilingue. A persistência em formato *JSON* garante a portabilidade dos dados, mantendo, ao mesmo tempo, a simplicidade da implementação.

O presente relatório descreve as várias etapas de construção do sistema, desde a recolha e fusão dos dados, passando pelo enriquecimento e análise semântica, até à implementação da plataforma interativa.

2 Descrição do Dataset Inicial

O conjunto de dados inicial utilizado neste trabalho é composto por três ficheiros no formato *JSON*: *bd.json*, *glossario.json* e *glossario_completo.json*, correspondentes à extração de dados efetuada a partir dos ficheiros *diccionari-multilinguee-de-la-covid-19.pdf*, *m_glossario-tematico-monitoramento-e-avaliacao.pdf* e *glossario_neologismos_sau-de.pdf*. Cada um destes ficheiros contém informações terminológicas relevantes para o domínio médico, embora com estruturas distintas, vocabulário variado e níveis de detalhe diferentes.

O ficheiro *bd.json* apresenta-se organizado por letras do alfabeto, onde cada letra agrupa vários termos numerados. Cada termo contém campos como “designação”, que representa o nome do termo; “categoria lexical da designação”, que indica a sua classe gramatical; e um conjunto de “complementos da designação”, que podem incluir siglas, sinónimos absolutos e complementares, denominação comercial e referências cruzadas para outras entradas. Este ficheiro também inclui traduções do termo para várias línguas, como espanhol, inglês, francês, português, neerlandês e árabe, bem como códigos alternativos (tais como símbolo, nome científico e número CAS), informações temáticas (área e descrição) e um campo de “notas” com texto livre.

O segundo ficheiro, *glossario.json*, possui uma estrutura mais simples, com entradas indexadas numericamente (por exemplo, “1”, “2”, etc.). Cada entrada contém o campo

“termo” (nome do conceito), os campos “tipo” e “número” que indicam o género e número gramatical do termo, e a “descrição” que fornece uma definição textual. Adicionalmente, podem existir sinónimos, traduções para espanhol e inglês, e campos adicionais como “notas”, “remissiva” e “expandida”, que fornecem informações auxiliares ou indicam outros termos relacionados.

Por fim, o ficheiro `glossario_completo.json` é o mais denso e abrangente dos três. Nele, as entradas estão organizadas por chave lexicográfica, ou seja, o próprio termo serve como chave principal. Cada termo inclui campos como “referencia_gramatical”, que utiliza abreviaturas padrão para indicar a categoria lexical (por exemplo, “s.f.” para substantivo feminino); “Sigla”, que representa uma forma reduzida do termo, caso exista; “Sinonimos”; e uma “Definicao” detalhada. Este ficheiro destaca-se ainda por incluir um campo de “equivalencias”, que fornece traduções para espanhol e inglês, bem como uma secção de “notas” ricas com elementos como “Informacao_enciclopedica”, “Abonação”, “Numero_identificacao” e “Marcas_Tipograficas”.

Em conjunto, os três ficheiros oferecem uma ampla base de dados terminológicos, cobrindo diferentes aspetos linguísticos, semânticos e técnicos dos termos. No entanto, as estruturas heterogéneas, os nomes de campos distintos para a mesma informação e os formatos inconsistentes entre os ficheiros exigiram um processo rigoroso de normalização e unificação, descrito nas secções seguintes deste relatório.

3 Enriquecimento do Dataset

3.1 Junção dos Datasets

Como ponto de partida para o processo de enriquecimento terminológico, realizou-se a integração de três fontes principais previamente tratadas no Trabalho Prático 1: o *Diccionario Multilingüe de la COVID-19*, o *Glossário Temático de Monitoramento e Avaliação* e o *Glossário de Neologismos da Saúde*. Cada um destes dicionários apresentava características distintas em termos de formato, granularidade terminológica, estrutura de dados e escopo semântico, o que exigiu um processo detalhado de normalização e padronização antes de qualquer fusão.

Para atingir uma estrutura comum, foram desenvolvidos *scripts* em *Python* capazes de processar os ficheiros JSON correspondentes, reorganizando os dados em conformidade com uma estrutura-padrão previamente definida. Esta estrutura contemplava os seguintes campos

principais: `termo_principal`, `categoria_lexical`, `sinonimos`, `siglas`, `traducoes` (com subdivisões por idioma), `area_tematica` e `notas`.

A conversão individual de cada dicionário para esta estrutura uniforme foi essencial para garantir a consistência e a integridade semântica dos dados, bem como para facilitar a fusão posterior. A junção dos três ficheiros foi então realizada por meio de uma função de mesclagem ordenada, que agrupava os termos segundo a sua letra inicial e os ordenava alfabeticamente dentro de cada grupo. Esta organização favoreceu a deteção de termos redundantes e permitiu a consolidação de entradas que compartilhassem a mesma raiz lexical ou conceito-base, sem perda de informação.

O produto final dessa etapa foi o ficheiro `dicionario_unificado.json`, que consolidava o conteúdo dos três glossários iniciais, servindo como base sólida e extensível para a etapa seguinte de enriquecimento com dados externos.

3.2 Web Scraping

Com o objetivo de ampliar o escopo informacional do dicionário unificado e incluir novos termos especializados, foi desenvolvido um processo automatizado de extração de dados a partir do *Vocabulaire de la Médecine*, uma base terminológica disponibilizada pelo *Office Québécois de la Langue Française (OQLF)* [1]. Essa fonte foi escolhida pela sua riqueza terminológica, especialização na área da saúde e pela sua curadoria institucional de qualidade.

A extração de dados foi dividida em duas fases complementares:

1. **Extração de links de entrada:** nesta fase inicial, o *script* percorreu sistematicamente a página principal do OQLF, identificando e armazenando os URLs associados a cada termo listado em ordem alfabética. Para garantir cobertura completa, foram tratados casos de paginação e redirecionamentos automáticos.
2. **Extração de conteúdo terminológico:** cada URL identificado foi acessado individualmente, sendo então extraídas informações específicas do termo correspondente. Entre os campos capturados destacam-se:

- Termo privilegiado e variantes (sinónimos, termos não recomendados, formas alternativas);
- Traduções disponíveis para inglês (com distinção entre termo principal e variantes);
- Definições técnicas e detalhadas;

- Domínio médico ou área temática associada;
- Autoria da entrada e data da última atualização;
- Notas complementares e observações terminológicas.

A extração foi realizada com o auxílio das bibliotecas `requests` e `BeautifulSoup` para requisições HTTP e *parsing* de HTML, respetivamente. Para garantir a limpeza textual e extração eficiente de conteúdo relevante, foram empregadas expressões regulares (`re`) e mecanismos de verificação de padrões linguísticos. A tradução dos conteúdos em francês para português foi realizada através da biblioteca `deep_translator`, utilizando o motor de tradução do Google. Esse processo foi validado manualmente para um subconjunto de termos, a fim de garantir fidelidade semântica nas traduções.

3.3 Integração dos Dados

Com os dados externos extraídos e traduzidos, procedeu-se à sua integração com o dicionário previamente unificado. Este processo envolveu a transformação das entradas extraídas do OQLF para o mesmo modelo estrutural dos dados já existentes, por meio de funções de normalização. Essa padronização assegurou a compatibilidade e coerência dos dados, minimizando riscos de inconsistência.

Para cada entrada extraída do OQLF, os campos foram mapeados e convertidos para os seguintes elementos estruturais:

- **termo_principal**: termo privilegiado em português;
- **traducoes**: subdivididas em **fr** e **en**, conforme os dados originais;
- **definicao**: tradução da definição técnica;
- **area_tematica**: domínio médico identificado na entrada original;
- **notas**: notas explicativas ou metadados adicionais.

O conjunto de dados obtido foi armazenado no ficheiro `oqlf_atualizado.json`. Em seguida, aplicou-se uma nova operação de junção com o `dicionario_unificado.json`, através de uma função de mesclagem com ordenação alfabética e verificação de duplicatas. O resultado foi o ficheiro final `dicionario_unificadofinal.json`, representando a versão enriquecida do dicionário terminológico.

3.4 Validação e Qualidade dos Dados

Para garantir a qualidade e robustez do dataset final, implementaram-se diversas medidas de validação:

- **Verificação de estrutura:** cada entrada foi validada para assegurar que continha os campos mínimos necessários e que estes estavam corretamente preenchidos;
- **Limpeza textual:** remoção de espaços em branco, normalização de pontuação e substituição de caracteres não padrão;
- **Deduplicação:** verificação de unicidade por termo e área temática, com fusão de informações complementares quando necessário;
- **Validação semântica de traduções:** análise amostral de traduções automáticas para garantir a preservação do sentido original.

Essas medidas aumentaram significativamente a confiança nos dados e a utilidade prática do recurso, tornando-o apropriado para uso em tarefas terminológicas, aplicações de PLN, sistemas de apoio à decisão clínica, entre outros.

3.5 Exemplo Ilustrativo

Um exemplo do impacto do processo de enriquecimento pode ser observado na entrada “**Acidente isquêmico de transição**”, termo inicialmente inexistente nos dicionários nacionais, mas extraído do OQLF. Após tradução, normalização e integração, a entrada resultante no ficheiro JSON assume a estrutura padronizada, conforme ilustrado na Figura 1.

```

"Acidente isquêmico de transição": {
  "dicionario": "vocabulaire de la medecine oqlf",
  "categoria_lexical_da_designação": "",
  "complementos_designação": {
    "siglas": [],
    "sinónimos_absolutos": "",
    "sinónimos_complementares": [],
    "denominação_comercial": "",
    "consultar_outra_entrada": ""
  },
  "traduções": {
    "castellà": "",
    "anglès": "transient ischemic attack, TIA",
    "anglès_associado": "mini-stroke",
    "francès": "accident ischémique transitoire n. m., AIT n. m., ischémie cérébrale transitoire n. f., ICT n. f.",
    "francès_contexte": "mini-accident vasculaire cérébral n. m., mini-AVC n. m."
  },
  "códigos_alternativos": {
    "símbolo": "",
    "nome_científico": "",
    "número_CAS": ""
  },
  "áreas_temáticas": {
    "área": "medicamento > Semiologia e Patologia, medicamento > neurologia",
    "descrição": "Isquemia temporária em uma parte do cérebro, que não causa infarto cerebral e resulta de uma obstrução arterial."
  },
  "notas": {
    "informacoes_notas": [
      "Os sinais e sintomas de um acidente isquêmico transitório aparecem repentinamente e desaparecem em menos de 24 horas. Pode, por exemplo, ser distúrbios visuais, parestia ou distúrbio da fala. [m acidente isq"
    ],
    "informacao_enciclopedia": "",
    "abonacao": "",
    "numero_identificacao": "",
    "marcas_tipograficas": [],
    "remissiva": "",
    "expandida": ""
  }
}

```

Figura 1: Exemplo do termo "Acidente Isquêmico de Transição".

Esse exemplo evidencia não apenas a ampliação lexical do *dataset*, mas também a contextualização técnica e multilingue do termo, que agora conta com tradução, definição especializada, domínio temático e notas explicativas. Assim, o dicionário enriquecido transforma-se numa ferramenta terminológica poderosa, com grande potencial para ser utilizada em textos profissionais e académicos na área da saúde.

4 Análise de Conceitos e Relações

4.1 Similaridade semântica

Para analisar a similaridade semântica entre os conceitos médicos do dicionário, foi utilizada a técnica de embeddings de frases através do modelo 'pt-mteb/average_fasttext_cc.pt.300', fornecido pela biblioteca *sentence-transformers*. Este modelo permite transformar texto em vetores numéricos que capturam o significado semântico dos termos.

A cada conceito foi atribuído um vetor de *embedding* calculado a partir de uma representação textual composta por várias componentes relevantes: o termo principal, a categoria lexical, os sinónimos, as denominações comerciais, as traduções e uma descrição temática, entre outros atributos. Cada uma destas partes foi ponderada de acordo com a sua relevância semântica, de forma a reforçar os elementos mais representativos do conceito.

Uma vez calculados os *embeddings* para todos os conceitos, a aplicação permite realizar buscas por similaridade. Quando o utilizador introduz um termo de pesquisa, este é convertido num vetor, e são calculadas as similaridades de cosseno entre esse vetor e os vetores dos conceitos existentes. Os resultados são ordenados por grau de similaridade, podendo ainda ser influenciados por fatores de reforço, como a correspondência direta com o termo ou com a descrição, permitindo destacar os conceitos mais semanticamente próximos da consulta.

Esta abordagem possibilita uma pesquisa robusta e inteligente, indo além da simples correspondência textual para identificar relações mais subtis entre conceitos do domínio médico.

4.2 Agrupamento por domínios

Para além da pesquisa por similaridade semântica, foi implementada uma funcionalidade de filtragem e agrupamento de conceitos com base em atributos específicos, nomeadamente o dicionário de origem, a categoria lexical e a área temática.

Cada conceito contém metadados que identificam a sua proveniência e o seu contexto

semântico. Estes dados são utilizados para permitir ao utilizador aplicar filtros personalizados, restringindo a pesquisa a subconjuntos específicos de conceitos que partilhem o mesmo domínio. Por exemplo, é possível pesquisar apenas conceitos pertencentes a uma determinada área temática (como cardiologia ou oncologia), ou que sejam classificados com uma determinada categoria lexical (como substantivos ou adjetivos).

Esta funcionalidade de agrupamento por domínios permite realizar análises mais dirigidas e contextuais, adaptando os resultados às necessidades do utilizador. Quando não é introduzida nenhuma consulta textual, o sistema pode apresentar todos os conceitos que respeitam os filtros definidos. Se for fornecida uma consulta textual, a filtragem atua como restrição adicional à pesquisa por similaridade.

Em conjunto, estas funcionalidades conferem à aplicação uma grande flexibilidade na exploração e análise dos conceitos médicos, permitindo tanto a descoberta de relações semânticas quanto a análise segmentada por domínios temáticos ou linguísticos.

5 Desenvolvimento da Ferramenta Web

5.1 Tecnologias utilizadas

A ferramenta *web* foi desenvolvida com o objetivo de proporcionar uma interface intuitiva e eficiente para a pesquisa e análise de conceitos médicos, explorando tanto a similaridade semântica como a filtragem por domínios.

Para a construção da interface do utilizador, optou-se pela utilização do *framework Flask*, uma biblioteca *Python* leve e flexível, que permite criar aplicações *web* de forma rápida e modular. O *Flask* facilita o desenvolvimento de rotas para servir páginas *HTML* dinâmicas, integrando-se perfeitamente com as funcionalidades de *backend* que processam as consultas e aplicam os algoritmos de similaridade.

No *frontend*, foi adotado o *Bootstrap 5*, um *framework CSS* amplamente utilizado para o design responsivo e estilização de componentes. O *Bootstrap* assegura que a aplicação seja acessível e visualmente agradável em diferentes dispositivos, desde computadores até dispositivos móveis. Este *framework* permitiu criar um *layout* organizado com filtros interativos para a pesquisa, utilizando componentes como botões, formulários, caixas de seleção e áreas de agrupamento com *scroll* vertical para melhorar a navegação.

O *backend* processa os pedidos *HTTP* recebidos, executa a lógica de consulta utilizando os *embeddings* semânticos previamente calculados e retorna os resultados ao *frontend* para

apresentação. A interação entre *frontend* e *backend* é gerida de forma simples e eficaz, utilizando métodos *GET* para a transmissão dos parâmetros da pesquisa, ou *POST* para inserção de novos dados.

Em suma, a combinação do *Flask* para o *backend*, com o *Bootstrap* no *HTML* no *frontend*, permitiu desenvolver uma aplicação *web* leve, responsiva e funcional, que suporta uma experiência de utilizador fluida e intuitiva para a exploração avançada de conceitos médicos.

5.2 Funcionalidades

O sistema desenvolvido oferece um conjunto abrangente de funcionalidades organizadas em quatro áreas principais: consulta, navegação, gestão de conteúdos e visualização. A interface *web*, construída com *Flask* e *Bootstrap*, proporciona uma experiência intuitiva e acessível, adequada tanto para utilizadores finais como para administradores.

No âmbito da consulta, o sistema disponibiliza um motor de busca avançado com três modos distintos de operação, Figura 2. Em primeiro lugar, permite a pesquisa por similaridade semântica, que possibilita encontrar conceitos relacionados mesmo na ausência de correspondência textual exata. Em segundo lugar, oferece a filtragem dos resultados com base em critérios específicos, como o dicionário de origem, a categoria lexical e a área temática. Por fim, incorpora um modo híbrido que aplica os filtros seleccionados antes do cálculo da similaridade semântica, combinando ambos os métodos para refinar os resultados.

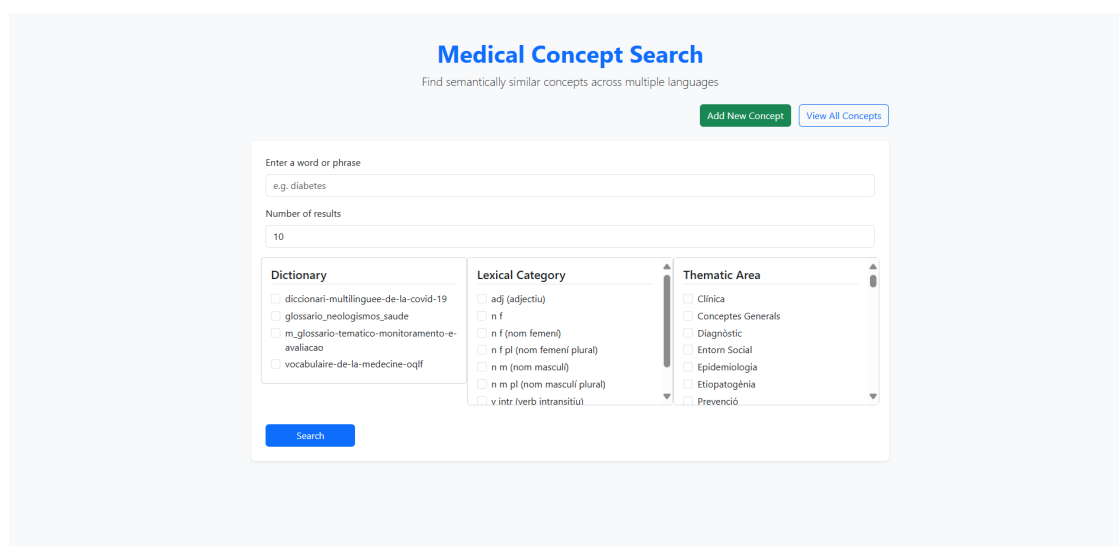


Figura 2: Página inicial com o sistema de busca por similaridade, aplicação de filtros ou sistema híbrido.

A página de resultados apresenta os conceitos ordenados por relevância, exibindo para cada entrada o termo principal, a descrição temática e o *score* de similaridade, Figura 3,

quando aplicável. O utilizador pode definir o número de resultados a apresentar, entre um e cinquenta, como é possível visualizar novamente na Figura 2, e um sistema de *boosting* é aplicado para aumentar a relevância quando o termo pesquisado surge no nome do conceito ou na sua descrição.

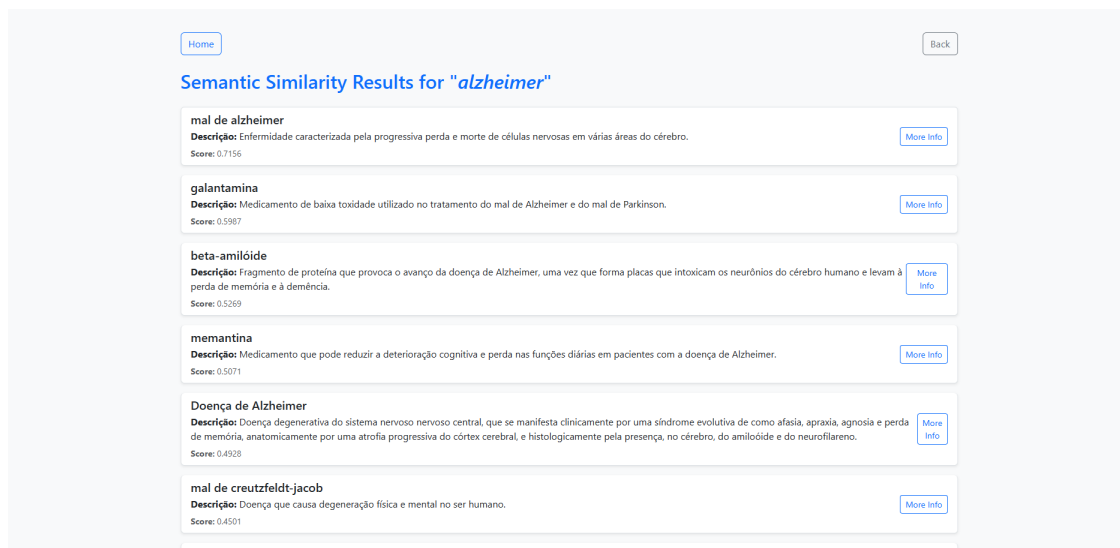


Figura 3: Página de resultados aplicando apenas a inserção de texto na página inicial, o que leva apenas a obtenção de resultados com base na similaridade.

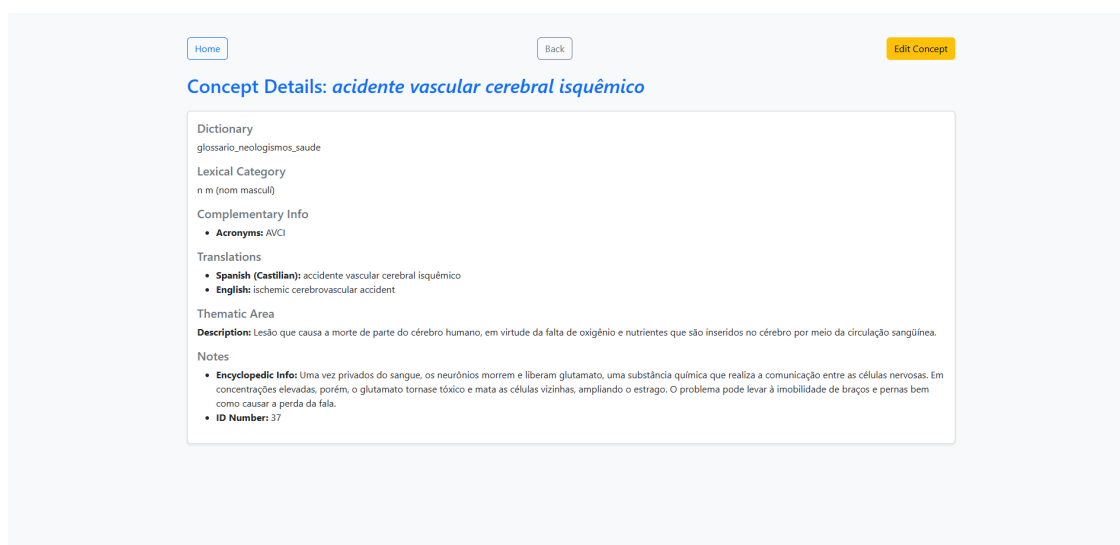
Em relação à navegação, o sistema permite explorar os conceitos através de uma listagem alfabética completa, Figura 4, com paginação que limita a vinte itens por página para melhor organização.

Concept	Description	More Info
abetaaaaaaa	Proteína que pode ser encontrada em todos os tipos de células do organismo humano. Ao acumular-se excessivamente no córtex cerebral do ser humano pode contribuir para o aceleração do mal de alzheimer.	Details
ACA		Details
acalabrutinib	Fármac: antineoplásico que bloca la tirosina-cinasa de Bruton i inhibeix la replicació dels limfócits T cancerosos.	Details
Accountability	Conjunto de mecanismos que permitem aos gestores da organização prestar contas dos planeamentos e execuções da área, bem como serem responsabilizados pelo resultado de suas ações.	Details
Acesso à informação	Direito de obter dados, processados ou não, contidos em qualquer meio, suporte ou formato.	Details
Acidente isquémico de transição	Isquemia temporária em uma parte do cérebro, que não causa infarto cerebral e resulta de uma obstrução arterial.	Details
acidente vascular cerebral isquémico	Lesão que causa a morte de parte do cérebro humano, em virtude da falta de oxigénio e nutrientes que são inseridos no cérebro por meio da circulação sanguínea.	Details
acidose metilmalónica	Doença que afeta o ser humano, causada em função de uma deficiência metabólica genética e que causa o retardamento do desenvolvimento infantil.	Details
Acompanhamento	Observação da evolução de um processo ou fenómeno, realizado por exame, medição e análise.	Details
AdS-nCoV		Details
adenomegalia	Hipertrofia de um ou mais nós linfáticos após uma infecção bacteriana ou viral ou ligados à presença de um tumor.	Details
adequació de les actuacions sanitàries	Decisió clínica que comporta l'aplicació de les actuacions sanitàries que afavoreixen el màxim benestar al pacient i l'abstenció de les que no li aporten prou beneficis, tenint en compte la seva situació concreta.	Details
adequació de l'esforç terapèutic		Details

Figura 4: Listagem alfabética completa de todos os conceitos.

Cada conceito pode ser visualizado detalhadamente numa página dedicada, onde a informação está organizada em secções lógicas, incluindo dados básicos como o dicionário e a categoria lexical, complementos como siglas, sinónimos e denominações comerciais, traduções

em onze línguas diferentes, códigos alternativos (como SBL, NC e CAS), área temática e descrição, bem como notas diversas que contêm informação enciclopédica e marcas tipográficas, Figura 5.

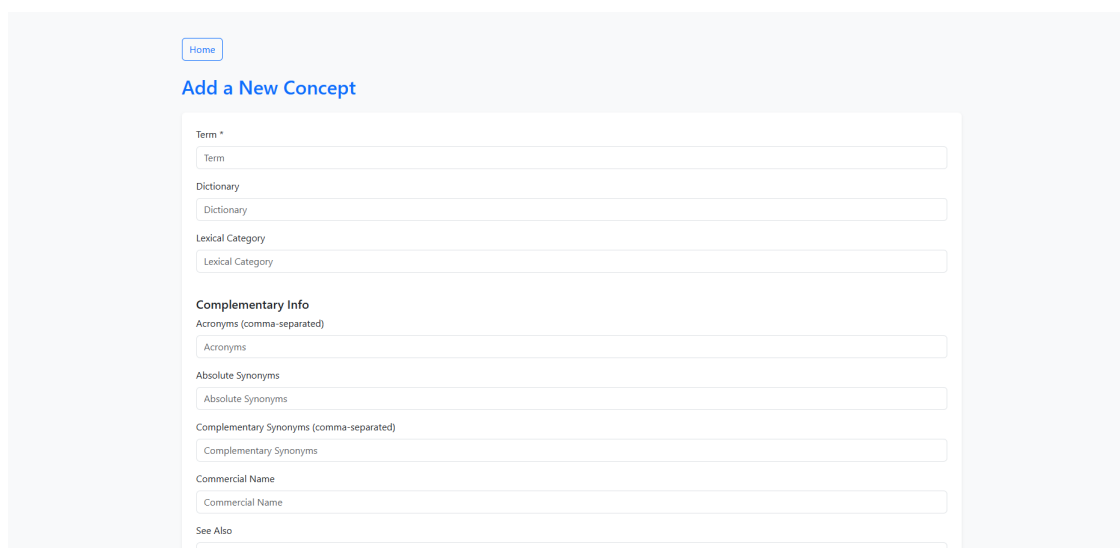


The screenshot shows a web interface for concept details. At the top, there are buttons for 'Home', 'Back', and 'Edit Concept'. The main heading is 'Concept Details: acidente vascular cerebral isquêmico'. Below this, a white box contains the following information:

- Dictionary:** glossario_neologismos_saude
- Lexical Category:** n m (nom mascul)
- Complementary Info:**
 - Acronyms:** AVC
- Translations:**
 - Spanish (Castilian):** accidente vascular cerebral isquémico
 - English:** ischemic cerebrovascular accident
- Thematic Area:**
- Description:** Lesão que causa a morte de parte do cérebro humano, em virtude da falta de oxigénio e nutrientes que são inseridos no cérebro por meio da circulação sanguínea.
- Notes:**
 - Encyclopedic Info:** Uma vez privados do sangue, os neurónios morrem e liberam glutamato, uma substância química que realiza a comunicação entre as células nervosas. Em concentrações elevadas, porém, o glutamato torna-se tóxico e mata as células vizinhas, ampliando o estrago. O problema pode levar à imobilidade de braços e pernas bem como causar a perda da fala.
 - ID Number:** 37

Figura 5: Visualização de todas as informações relativas ao conceito.

No que diz respeito à gestão de conteúdos, a aplicação disponibiliza uma funcionalidade que permite adicionar novos conceitos médicos, preenchendo todos os metadados associados, Figura 6.

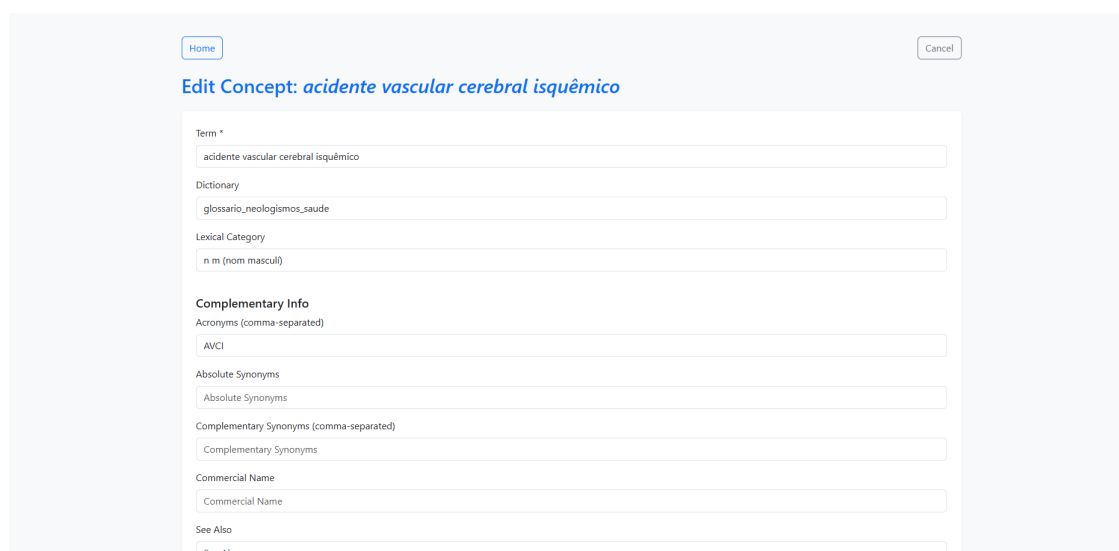


The screenshot shows a web interface for adding a new concept. At the top, there are buttons for 'Home' and 'Add a New Concept'. Below this, a white box contains the following form fields:

- Term *** (text input)
- Dictionary** (text input)
- Lexical Category** (text input)
- Complementary Info:**
 - Acronyms (comma-separated)** (text input)
 - Absolute Synonyms** (text input)
 - Complementary Synonyms (comma-separated)** (text input)
 - Commercial Name** (text input)
 - See Also** (text input)

Figura 6: Funcionalidade que permite a inserção de novos conceitos.

Esta interface inclui funcionalidades para editar entradas existentes, Figura 7, validação dos campos obrigatórios, como o termo principal, e processamento automático de listas separadas por vírgulas. Durante a criação dos conceitos, os *embeddings* vetoriais são gerados automaticamente para permitir a posterior pesquisa semântica.



Home Cancel

Edit Concept: *acidente vascular cerebral isquêmico*

Term *

acidente vascular cerebral isquêmico

Dictionary

glossario_neologismos_saude

Lexical Category

n m (nom mascul)

Complementary Info

Acronyms (comma-separated)

AVCI

Absolute Synonyms

Absolute Synonyms

Complementary Synonyms (comma-separated)

Complementary Synonyms

Commercial Name

Commercial Name

See Also

See Also

Figura 7: Funcionalidade que permite a edição de novos conceitos.

Por fim, a visualização e o acesso foram pensados com base em princípios de usabilidade e acessibilidade. A navegação é consistente, dispondo de botões para regressar e para aceder à página inicial, enquanto a paginação é inteligente, indicando a página atual e os limites da listagem. A aplicação oferece ainda feedback visual imediato quando não são encontrados resultados para uma pesquisa e organiza a informação de forma clara, recorrendo a tabelas para facilitar a leitura. Todas as alterações efetuadas são guardadas automaticamente num ficheiro JSON central, garantindo a consistência dos dados entre sessões.

6 Testes e Validação

Para validar a aplicação desenvolvida, foram realizados diversos testes cobrindo as principais funcionalidades, como pesquisa por filtros, pesquisa por similaridade, combinação de filtros com similaridade, visualização e edição de termos, bem como adição de novos conceitos ao dicionário.

Inicialmente, testou-se a funcionalidade de pesquisa com filtros aplicados sem qualquer termo inserido na barra de pesquisa. Como mostra a Figura 8, ao selecionar um dicionário específico, os resultados são devidamente filtrados conforme esperado.

Medical Concept Search
Find semantically similar concepts across multiple languages

Enter a word or phrase
e.g. diabetes

Number of results
10

Dictionary

- ☒ dictionari-multilinguee-de-la-covid-19
- ☐ glossario_neologismos_saude
- ☐ m_glossario-tematico-monitoramento-e-avaliacao
- ☐ vocabulaire-de-la-medicine-ogif

Lexical Category

- ☐ adj
- ☐ n f
- ☐ n f pl
- ☐ n m
- ☐ n m pl
- ☐ v intr
- ☐ v tr

Thematic Area

- ☐ Clínica
- ☐ Conceptes Generals
- ☐ Diagnòstic
- ☐ Entorn Social
- ☐ Epidemiologia
- ☐ Etiopatogènia
- ☐ Prevenció

Search

Figura 8: Aplicação de um filtro referente a um dicionário sem introduzir texto na barra de pesquisa.

A Figura 9 demonstra que os resultados retornados condizem com os critérios estabelecidos, validando o comportamento da filtragem isolada e sem a apresentação de *scores*.

Home Back

Filtered Results

Active Filters

Dicionari: dictionari-multilinguee-de-la-covid-19

ACA
Descrição: [More Info](#)

acalabrutinib
Descrição: Farmac: antineoplàstic que bloca la tirosina-cinasa de Bruton i inhibeix la replicació dels limfòcits T cancerosos. [More Info](#)

AdS-nCoV
Descrição: [More Info](#)

adequació de les actuacions sanitàries
Descrição: Decisió clínica que comporta l'aplicació de les actuacions sanitàries que afavoreixen el màxim benestar al pacient i l'abstenció de les que no li aporten prou benefici, tenint en compte la seva situació concreta. [More Info](#)

adequació de l'esforç terapèutic
Descrição: [More Info](#)

Figura 9: Resultados da aplicação de apenas um filtro, sem *Scores*.

A funcionalidade de pesquisa por similaridade também foi avaliada. Na Figura 10, é apresentada a introdução do termo "Acidente vascular" na barra de pesquisa. A Figura 11 exibe os resultados obtidos, com a correspondente pontuação (*score*) de similaridade calculada, permitindo ao utilizador perceber o grau de correspondência entre os termos sugeridos e o termo pesquisado.

Medical Concept Search
Find semantically similar concepts across multiple languages

Enter a word or phrase
Acidente vascular

Number of results
10

Dictionary

- ☐ dictionari-multilinguee-de-la-covid-19
- ☐ glossario_neologismos_saude
- ☐ m_glossario-tematico-monitoramento-e-avaliacao
- ☐ vocabulaire-de-la-medicine-ogif

Lexical Category

- ☐ adj
- ☐ n f
- ☐ n f pl
- ☐ n m
- ☐ n m pl
- ☐ v intr
- ☐ v tr

Thematic Area

- ☐ Clínica
- ☐ Conceptes Generals
- ☐ Diagnòstic
- ☐ Entorn Social
- ☐ Epidemiologia
- ☐ Etiopatogènia
- ☐ Prevenció

Search

Figura 10: Pesquisa do termo "Acidente vascular" por similaridade.

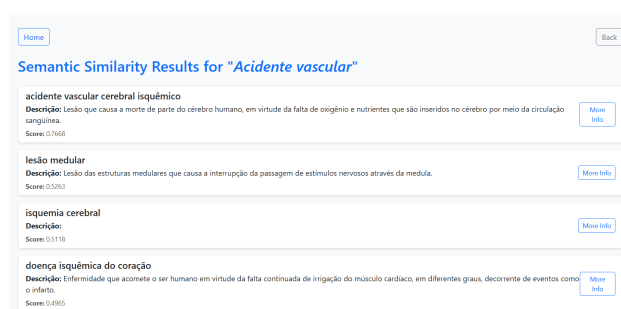


Figura 11: Resultado da pesquisa do termo "Acidente vascular" por similaridade, com verificação dos *scores*.

A pesquisa combinada, ou seja, por similaridade com aplicação simultânea de filtros, também foi verificada. A Figura 12 mostra essa configuração, e os resultados são apresentados na Figura 13, com os *scores* visíveis para análise detalhada.

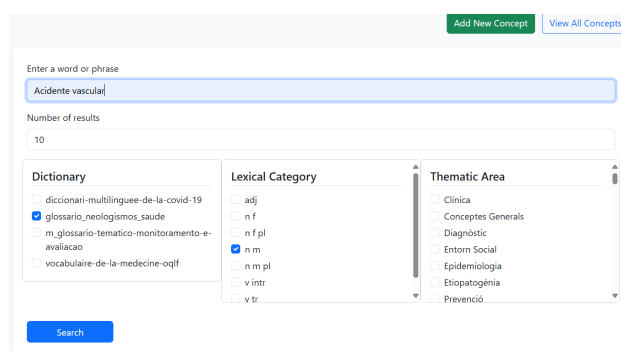


Figura 12: Pesquisa do termo "Acidente vascular" por similaridade e com aplicação de filtros.



Figura 13: Resultado da pesquisa do termo "Acidente vascular" por similaridade e aplicação de filtros, com verificação dos *scores*.

A seguir, foi testada a funcionalidade de visualização detalhada e edição de termos. A Figura 14 mostra os detalhes do termo "acidente vascular cerebral isquêmico", enquanto a Figura 15 exibe a informação completa antes da edição. Após a edição do termo, adicionando "1" ao final do nome (Figura 16), verificou-se a atualização correta tanto na interface da aplicação (Figura 17) quanto na base de dados (Figura 18).

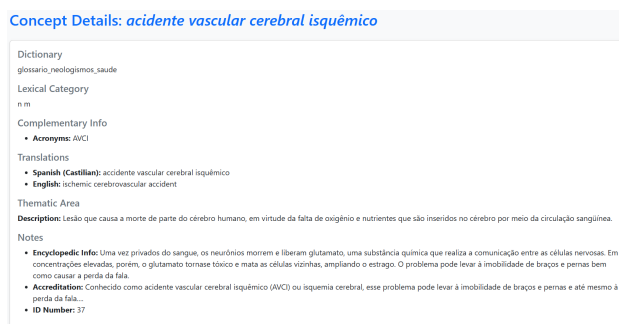


Figura 14: Visualização de detalhes associados ao termo "acidente vascular cerebral isquêmico".



Figura 15: Visualização da informação associada ao termo "acidente vascular cerebral isquêmico" antes de edição.

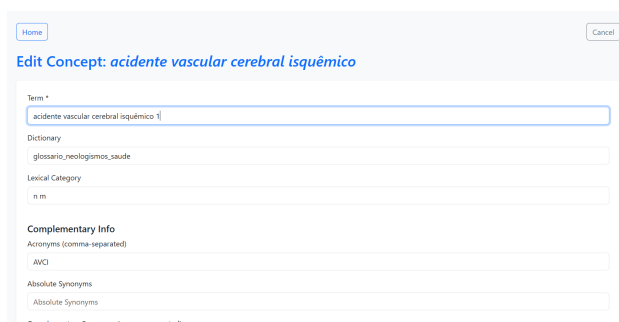


Figura 16: Edição do termo "acidente vascular cerebral isquêmico" adicionando "1" ao seu nome.

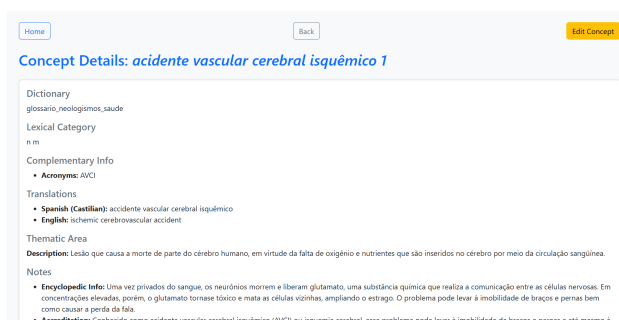


Figura 17: Verificação da edição do termo na aplicação.

[illegible]

Figura 18: Verificação de que as alterações ao termo foram guardadas na base de dados.

Por fim, foi realizada a adição de um novo termo ao dicionário. A Figura 19 apresenta o processo de adição através da interface, e a Figura 20 confirma o sucesso da operação. A Figura 21 valida que os dados foram devidamente armazenados na base de dados persistente.

[Home](#)

Add a New Concept

Term *

Agulheta

Dictionary

glossario_neologismos_saude

Lexical Category

Lexical Category

Complementary Info

Acronyms (comma-separated)

Acronyms

Absolute Synonyms

Absolute Synonyms

Figura 19: Adição de um novo termo ao dicionário.

[Home](#)[Back](#)[Edit Content](#)

Concept Details: *Aguilhet*

Dictionary

glossario, neologismos, saude

Thematic Area

Description: Pedaco de plastico ou metal na ponta de um atacante.

Figura 20: Verificação da adição do novo termo através da interface.

```
"Aguilheta": {
  "dicionario": "glossario_neologismos_saude",
  "categoria lexical da designação": "",
  "complementos designação": {
    "siglas": [],
    "sinónimos absolutos": "",
    "sinónimos complementares": [],
    "denominação comercial": "",
    "consultar outra entrada": ""
  },
  "traduções": {
    "occità": "",
    "basc": "",
    "gallec": "",
    "castellà": "",
    "anglès": "",
    "francès": "",
    "português": "",
    "[PT]português de Portugal": "",
    "[BR]português del Brasil": "",
    "neerlandê": "",
    "àrab": ""
  },
  "códigos alternativos": {
    "sbl": "",
    "nc": "",
    "CAS": ""
  },
  "áreas temáticas": {
    "área": "",
    "descrição": "Pedaço de plástico ou metal na ponta de um atacador."
  },
  "notas": {
    "informacoes_notas": [],
    "Informacao_encyclopedica": "",
    "Abonacao": "",
    "Numero_identificacao": "",
    "Marcas_Tipograficas": [],
    "remissiva": "",
    "expandida": ""
  },
}
```

Figura 21: Verificação da adição do novo termo à base de dados.

Como conclusão dos testes realizados, confirma-se que a aplicação cumpre com os requisitos funcionais definidos, proporcionando uma experiência de utilização fluida, intuitiva e robusta. As funcionalidades de pesquisa por filtros, similaridade e edição de termos demonstraram-se eficazes e confiáveis, assegurando a integridade e persistência dos dados manipulados. A capacidade de visualizar e validar as alterações diretamente na base de dados reforça a transparência e o controlo do utilizador sobre o conteúdo do dicionário, garantindo assim a qualidade e utilidade do sistema desenvolvido no apoio à terminologia médica multilingue.

7 Conclusão

O presente trabalho demonstrou a viabilidade e eficácia da integração de diversas técnicas de Processamento de Linguagem Natural na construção de um sistema terminológico médico robusto, multilingue e extensível. Através da junção de três glossários iniciais e do enriquecimento com dados externos extraídos do *Vocabulaire de la Médecine* (OQLF), foi possível criar um dicionário unificado e estruturado, com elevada riqueza semântica e cobertura lexical.

O processo de normalização, de duplicação e validação garantiu a coerência dos dados, permitindo a criação de um recurso fiável para futuras aplicações em contextos académicos e clínicos. A utilização de *embeddings* semânticos permitiu uma análise mais profunda das relações entre conceitos, possibilitando funcionalidades como a pesquisa por similaridade e o agrupamento temático.

A aplicação *web* desenvolvida evidencia o potencial prático da solução, oferecendo uma interface intuitiva para exploração terminológica, com suporte a filtros dinâmicos e visualização detalhada. A escolha de tecnologias leves e modulares (*Flask*, *Bootstrap*, *Sentence-Transformers*) garantiu flexibilidade e facilidade de manutenção.

Em suma, o sistema apresentado constitui uma base sólida para o desenvolvimento de ferramentas terminológicas inteligentes na área da saúde, podendo ser expandido futuramente com novos dicionários, integração com ontologias médicas ou inclusão de funcionalidades.

Referências

- [1] Office québécois de la langue française. *Vocabulaire de la médecine*. Disponível: <https://www.oqlf.gouv.qc.ca/ressources/bibliotheque/dictionnaires/vocabulaire-medecine.aspx>. Acesso a: 27 maio 2025.