



Relatório - Trabalho Prático 2

Processamento de Linguagem Natural

Mestrado em Engenharia Biomédica - Informática Médica

Grupo :

PG57810, Dinis Mesquita

PG52290, Flávio Ribeiro

Docentes:

Luís Filipe Cunha

José João Almeida



Índice

1	Introdução	2
2	Scraper	2
2.1	Scraper significados	2
2.2	Scraper áreas	4
3	Estruturação do dataset final	7
4	Plataforma Web	8
4.1	Página Home	8
4.2	Resultados de pesquisa	8
4.3	Consulta dos conceitos	9
4.4	Página de um Conceito individual	9
4.5	Página das Áreas Médicas	11
4.6	Página de adicionar um novo conceito	12
5	Conclusão	13

1 Introdução

Este trabalho tem como objetivo enriquecer o dataset previamente desenvolvido, contendo uma coleção de conceitos médicos e fazer o desenvolvimento de uma aplicação web interativa dedicada à visualização, gestão e pesquisa dos tais.

O enriquecimento do dataset foi feito a partir da técnica de *web-scraping*, de modo a complementar os conceitos já obtidos na primeira parte deste projeto. A plataforma web serve para permitir ao utilizador explorar facilmente esta base de dados, através de funcionalidades como a consulta detalhada de cada conceito, a pesquisa por palavra-chave, a visualização por áreas médicas, bem como a adição, edição ou eliminação de conceitos, e foi desenvolvida utilizando Python com o framework Flask.

2 Scraper

Para enriquecer o dataset de conceitos médicos inicialmente recolhidos, foram desenvolvidos scrapers específicos para obter informações complementares automaticamente. Este processo permitiu atribuir uma área médica a cada conceito (caso esta ainda não estivesse presente) e recolher definições relevantes.

2.1 Scraper significados

Para o preenchimento automático dos significados dos conceitos em falta, foi utilizada a plataforma Wikipédia (ver Figura 1), como fonte principal de dados. Esta escolha justificou-se pela sua abrangência temática, incluindo conteúdos relevantes em áreas como Medicina, Química, Biotecnologia, Biologia, entre outras. A estrutura estável e acessível do seu HTML facilitou a aplicação de técnicas de *web scraping*.

Apesar da Wikipédia não ser a fonte mais fiável em contexto académico, a sua ampla cobertura temática e flexibilidade técnica tornaram-na a opção mais prática para a implementação deste módulo. A página de pesquisa da Wikipédia permite afunilar os resultados por categorias e tipos de conteúdo, como imagens, o que se revelou útil e foi considerado numa fase posterior do projeto.

A pesquisa avançada (ver Figura 2), utilizando expressões booleanas (operadores AND, OR e NOT), foi essencial para tentar manter o conteúdo no contexto médico.

O processo de extração funciona da seguinte forma:

1. Para cada conceito sem significado, o *scraper* realiza uma pesquisa na Wikipédia em português;
2. Se existirem resultados, é selecionado o primeiro link da lista (presumivelmente o mais relevante);
3. O *scraper* acede à página do artigo e extrai o primeiro parágrafo presente no bloco de conteúdo principal;
4. Apenas a primeira frase desse parágrafo é considerada como possível definição, e é automaticamente adicionada ao campo significado no ficheiro JSON atualizado.

Apesar da aplicação ter demonstrado eficácia geral e permitido automatizar um processo que seria manual e repetitivo, foram detetados casos em que o significado extraído estava incorreto ou fora de contexto. Isto deve-se ao facto de existirem múltiplos domínios para um mesmo termo, onde o conceito médico é apenas um dos possíveis significados.

Mesmo com a implementação de mecanismos de pesquisa baseados em expressões booleanas para melhorar a precisão da correspondência, os resultados não foram totalmente fiáveis.

Assim, conclui-se que, embora o *scraping* da Wikipédia seja uma solução eficaz para acelerar o preenchimento inicial do dicionário, a validação manual continua essencial para garantir a qualidade e o rigor dos conteúdos, sobretudo em áreas sensíveis como a terminologia médica.

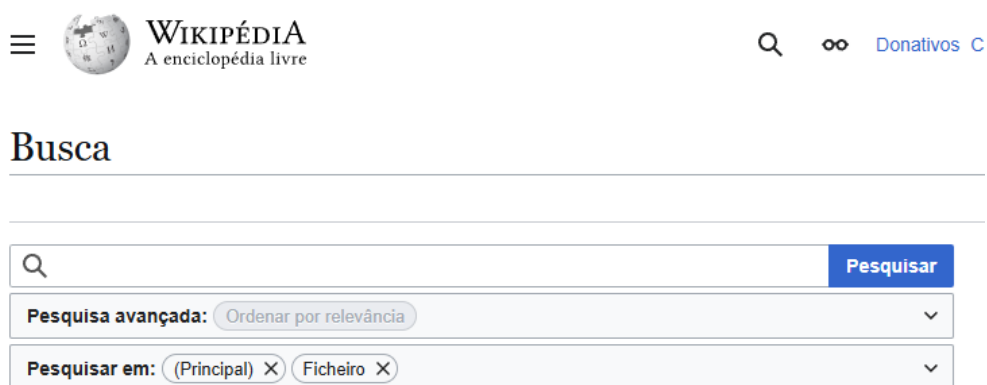
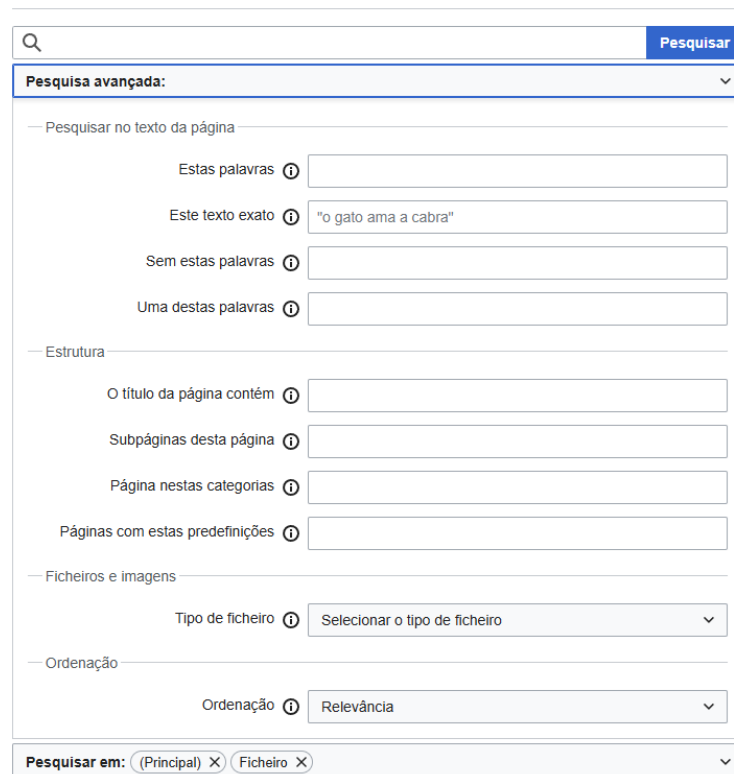


Figura 1: Página da Wikipédia utilizada como fonte de dados

Busca



The image shows a web search interface titled "Busca". It features a search bar at the top with a magnifying glass icon and a "Pesquisar" button. Below the search bar is a dropdown menu labeled "Pesquisa avançada:". The interface is organized into several sections, each with a minus sign icon to collapse it:

- Pesquisar no texto da página:** Contains four input fields with labels: "Estas palavras", "Este texto exato" (with a pre-filled example "o gato ama a cabra"), "Sem estas palavras", and "Uma destas palavras".
- Estrutura:** Contains four input fields with labels: "O título da página contém", "Subpáginas desta página", "Página nestas categorias", and "Páginas com estas predefinições".
- Ficheiros e imagens:** Contains a dropdown menu labeled "Tipo de ficheiro" with the option "Selecionar o tipo de ficheiro".
- Ordenação:** Contains a dropdown menu labeled "Ordenação" with the option "Relevância".

At the bottom, there is a section labeled "Pesquisar em:" with two buttons: "(Principal)" and "Ficheiro", and a dropdown arrow.

Figura 2: Exemplo de pesquisa avançada com expressões booleanas

2.2 Scraper áreas

Para atribuir áreas médicas específicas aos conceitos do glossário que não possuíam esta informação inicialmente, foi desenvolvido outro módulo de *web scraping* com recurso à National Library of Medicine (NLM). Ao contrário da Wikipédia, onde os conteúdos são mais abertos e interdisciplinares, a NLM apresenta um repositório de terminologia médica estruturada, o que permite garantir maior precisão e relevância clínica nas associações semânticas realizadas.

O scraping foi realizado sobre o sistema de classificação MeSH (*Medical Subject Headings*), utilizado mundialmente para organização de literatura biomédica. Cada conceito sem área médica associada foi enriquecido através de pesquisa no portal da NLM, usando o próprio nome do conceito como termo de consulta. Este portal e um exemplo de pesquisa estão ilustrados na Figura 3.

Ao contrário do *scraper* de significados (baseado na Wikipédia), não foi necessária a utilização de expressões booleanas nem operadores lógicos adicionais, uma vez que o domínio da informação está intrinsecamente limitado ao contexto médico, evitando ambiguidades significativas.

Após submeter o termo de interesse, a página de resultados apresenta uma lista de correspondências, da qual é selecionado apenas o primeiro resultado, assumindo-se a sua relevância. A página do conceito selecionado está geralmente organizada em quatro secções principais (ver Figura 4):

- **Details**

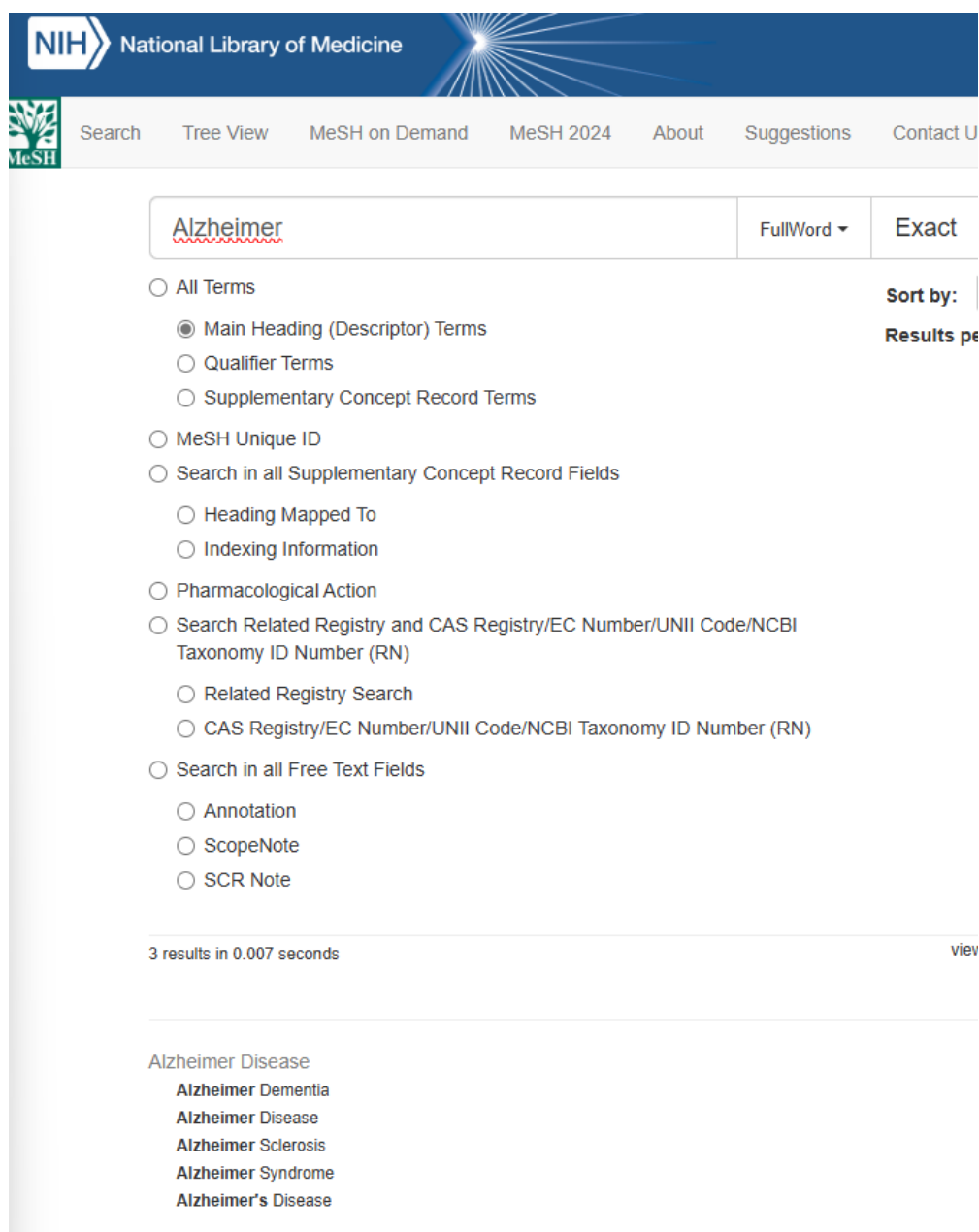
- **Qualifiers**
- **MeSH Tree Structures**
- **Concepts**

O foco do scraping incide sobre a secção **MeSH Tree Structures**, onde se encontra a categorização hierárquica do conceito. Este campo apresenta classificações como, por exemplo, **C10.228.140.617.738** (que pode corresponder a “Doenças do Sistema Nervoso” → “Transtornos Cognitivos” → “Doença de Alzheimer”), e foi considerada a melhor aproximação à ideia de ‘área médica’ dentro da estrutura do projeto.

A extração e associação automáticas destas classificações permitiram:

- Organizar os conceitos por área;
- Permitir filtragem posterior por domínio médico;
- Enriquecer o dataset com metadados estruturados.

Esta integração entre scraping e ontologias médicas contribuiu significativamente para a melhoria semântica do dicionário final, respeitando a hierarquia de classificações biomédicas reconhecidas internacionalmente.



NIH National Library of Medicine

Search Tree View MeSH on Demand MeSH 2024 About Suggestions Contact Us

Alzheimer

FullWord Exact

☐ All Terms
☒ Main Heading (Descriptor) Terms
☐ Qualifier Terms
☐ Supplementary Concept Record Terms
☐ MeSH Unique ID
☐ Search in all Supplementary Concept Record Fields
☐ Heading Mapped To
☐ Indexing Information
☐ Pharmacological Action
☐ Search Related Registry and CAS Registry/EC Number/UNII Code/NCBI Taxonomy ID Number (RN)
☐ Related Registry Search
☐ CAS Registry/EC Number/UNII Code/NCBI Taxonomy ID Number (RN)
☐ Search in all Free Text Fields
☐ Annotation
☐ ScopeNote
☐ SCR Note

Sort by: Results per page

3 results in 0.007 seconds view

Alzheimer Disease

- Alzheimer Dementia
- Alzheimer Disease
- Alzheimer Sclerosis
- Alzheimer Syndrome
- Alzheimer's Disease

Figura 3: Exemplo de pesquisa no portal da NLM

Alzheimer Disease MeSH Descriptor Data 2025

[Details](#)[Qualifiers](#)[MeSH Tree Structures](#)[Concepts](#)

Nervous System Diseases [C10]
 Central Nervous System Diseases [C10.228]
 Brain Diseases [C10.228.140]
 Dementia [C10.228.140.380]
 AIDS Dementia Complex [C10.228.140.380.070]
 Alzheimer Disease [C10.228.140.380.100]
 Aphasia, Primary Progressive [C10.228.140.380.132] +
 Creutzfeldt-Jakob Syndrome [C10.228.140.380.165]
 Dementia, Vascular [C10.228.140.380.230] +
 Diffuse Neurofibrillary Tangles with Calcification [C10.228.140.380.254]
 Frontotemporal Lobar Degeneration [C10.228.140.380.266] +
 Huntington Disease [C10.228.140.380.278]
 Kluver-Bucy Syndrome [C10.228.140.380.326]
 Lewy Body Disease [C10.228.140.380.422]
 Mixed Dementias [C10.228.140.380.711]

Nervous System Diseases [C10]
 Neurodegenerative Diseases [C10.574]
 Tauopathies [C10.574.945]
 Alzheimer Disease [C10.574.945.249]
 Corticobasal Degeneration [C10.574.945.312]
 Diffuse Neurofibrillary Tangles with Calcification [C10.574.945.374]
 Supranuclear Palsy, Progressive [C10.574.945.500]

Figura 4: Estrutura das secções da página de um conceito

3 Estruturação do dataset final

Depois de enriquecer o dataset com informação de diferentes fontes, cada conceito do dataset ficou estruturado da seguinte forma:

"conceito": "neologismo", "sinónimos pt":

"*sinonimo_1*", "*sinonimo_2*", ...

Língua :

"*traducao_na_respeitva_Lingua_1*", "*traducao_na_respeitva_Lingua_2*", ...

"significado": "definição clara do conceito",

"definicao catalã": "definição do conceito em catalã",

"significado_encylopédico": "definição enciclopédica",

"contexto": "exemplo textual de uso do conceito",

"área médica": "área médica referente",

"outras associacoes a 'termo referente'":

"*termo_relacionado_1*", "*termo_relacionado_2*", ...

4 Plataforma Web

Após o enriquecimento do dataset, foi desenvolvido uma plataforma web, utilizando a framework *Flask* e o auxílio de *bootstrap* e *cdn.datatables*, com o objetivo de fornecer uma interface intuitiva e funcional, que permita ao utilizador poder consultar informação dos diferentes conceitos médicos, tendo a possibilidade de procurar por expressões específicas e conseguir atualizar, adicionar ou eliminar conceitos no dataset de forma permanente.

4.1 Página Home

A página principal (Figura 5) da plataforma web apresenta um interface simples e focada na usabilidade, servindo como ponto de entrada para os utilizadores interagirem com o dicionário médico.

Esta interface apresenta ao utilizador uma pequena introdução ao projeto sobre o projeto desenvolvido e incorpora uma barra de pesquisa destacada que permite procurar conceitos médicos diretamente

Durante a utilização da plataforma web é também disponibilizado uma barra de navegação no topo da página, que permite o acesso rápido aos diferentes conteúdos e funcionalidades.

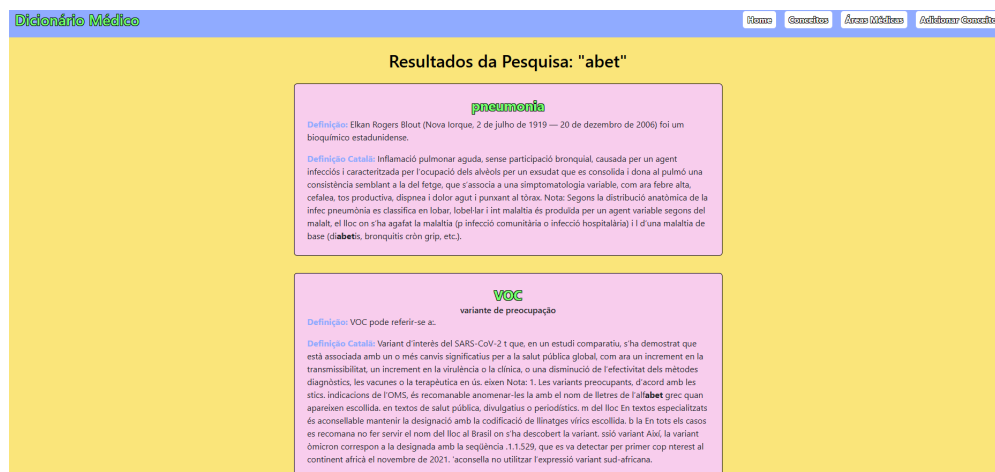


Figura 5: Página Principal da plataforma desenvolvida

4.2 Resultados de pesquisa

A página de Resultados da Pesquisa (Figura 6) apresenta os conceitos que correspondem à expressão procurada pelo utilizador na página principal. Os resultados são organizados em cartões individuais, onde cada conceito é identificado pelo seu nome, que aparece como um link clicável que redireciona para a página detalhada do conceito correspondente.

Cada cartão pode incluir informações adicionais, como sinónimos em português, definição principal, definição enciclopédica, contextualização e definição em catalão, caso esses dados estejam disponíveis. Esta estrutura permite ao utilizador consultar rapidamente os principais dados de cada conceito de forma resumida, antes de aceder detalhadamente ao conceito.



Resultados da Pesquisa: "abet"

pneumonia
Definição: Elkan Rogers Blout (Nova Iorque, 2 de julho de 1919 — 20 de dezembro de 2006) foi um bioquímico estadunidense.
Definição Catalã: Inflamació pulmonar aguda, sense participació bronquial, causada per un agent infeccios i caracteritzada per l'ocupació dels alveòls per un exsudat que es consolida i dona al pulmó una consistència semblant a la del fegat; que s'associa a una simptomatologia variable, com ara febre alta, cefalea, tos productiva, dispnea i dolor agut i punxant al tòrax. Nota: Segons la distribució anatómica de la infec pneumònia es classifica en lobar, lobel·lar i int malaltia és produïda per un agent variable segons del malalt, el lloc on s'ha agafat la malaltia (p infecció comunitària o infecció hospitalària) i d'una malaltia de base (diabetis, bronquitis cròn grip, etc.).

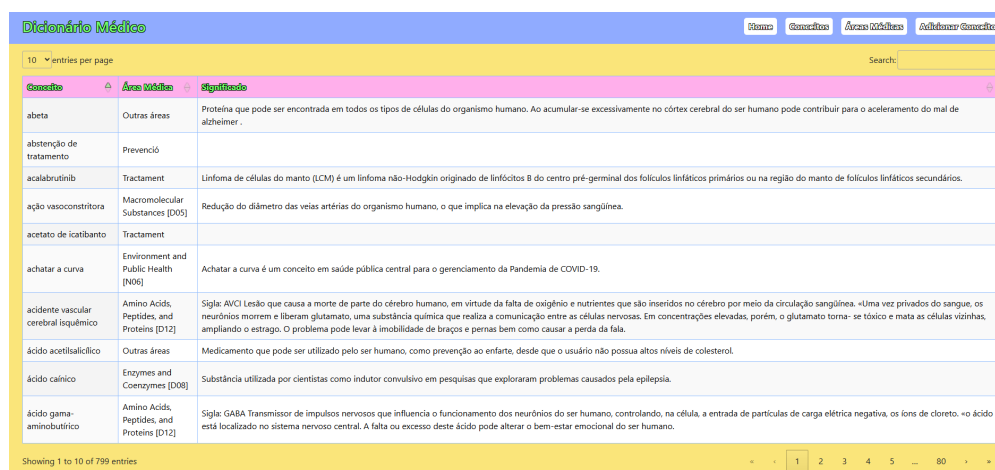
VOC
variante de preocupação
Definição: VOC pode referir-se a:
Definição Catalã: Variant d'interès del SARS-CoV-2 i que, en un estudi comparatiu, s'ha demostrat que està associada amb un o més canvis significatius per a la salut pública global, amb ara un increment en la transmissibilitat, un increment en la virulència o la clínica, o una disminució de l'efectivitat dels mètodes diagnòstics, les vacunes o la terapèutica en ús. eiven Nota: 1. Les variants preocupants, d'acord amb les stics, indicacions de l'OMS, és recomanable anomenar-les la amb el nom de lletres de l'al·labet grec quan apareixen escollida, en textos de salut pública, divulgatius o periodístics. m del lloc En textos especialitzats és aconsellable mantenir la designació amb la codificació de linatges vírics escollida. b la En tots els casos es recomana no fer servir el nom del lloc al Brasil on s'ha descobert la variant, ssó variant Alvi, la variant òmicron correspon a la designada amb la seqüència .1.1.529, que es va detectar per primer cop interest al continent africà el novembre de 2021. s'aconsetla no utilitzar l'expressió variant sud-africana.

Figura 6: Página de Resultados da Pesquisa da plataforma desenvolvida

Caso não exista nenhum resultado correspondente à pesquisa, a página exibe uma mensagem clara ao utilizador, indicando que nenhum conceito foi encontrado.

4.3 Consulta dos conceitos

A página de de consulta de conceitos (Figura 7) tem como objetivo apresentar de forma clara e organizada todos os conceitos médicos incluídos na aplicação. Esta página disponibiliza uma tabela interativa que permite ao utilizador consultar rapidamente o nome do conceito, a sua área médica e respetivo significado. Cada nome de conceito está associado a um link que redireciona para uma página individual onde é possível visualizar mais detalhes sobre o mesmo. Da mesma forma, a área médica apresentada em cada linha funciona como um atalho para uma página que agrupa todos os conceitos pertencentes à mesma categoria.



Concepte	Àrea Mèdica	Significado
abet	Outras áreas	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 aceleramento do mal de alzheimer .
abstenção de tratamento	Prevenção	
acalabrutinib	Tractament	Linfoma de células do manto (LCM) é um linfoma não-Hodgkin originado de linfócitos B do centro pré-germinal dos folículos linfáticos primários ou na região do manto de folículos linfáticos secundários.
ação vasoconstritora	Macromolecular Substances [D05]	Redução do diâmetro das veias arteriais do organismo humano, o que implica na elevação da pressão sanguínea.
acetato de icatibanto	Tractament	
achatar a curva	Environment and Public Health [N06]	Achatar a curva é um conceito em saúde pública central para o gerenciamento da Pandemia de COVID-19.
acidente vascular cerebral isquêmico	Amino Acids, Peptides, and Proteins [D12]	Sígl: AVC Lesão que causa a morte de parte do cérebro humano, em virtude da falta de oxigênio e nutrientes que são inseridos no cérebro por meio da circulação sanguínea. «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.
ácido acetilsalicílico	Outras áreas	Medicamento que pode ser utilizado pelo ser humano, como prevenção ao enfarte, desde que o usuário não possua altos níveis de colesterol.
ácido cálcico	Enzymes and Coenzymes [D08]	Substância utilizada por cientistas como indutor convulsivo em pesquisas que exploraram problemas causados pela epilepsia.
ácido gama-aminobutírico	Amino Acids, Peptides, and Proteins [D12]	Sígl: GABA Transmissor de impulsos nervosos que influencia o funcionamento dos neurónios do ser humano, controlando, na célula, a entrada de partículas de carga elétrica negativa, os íons de cloreto. «o ácido está localizado no sistema nervoso central. A falta ou excesso deste ácido pode alterar o bem-estar emocional do ser humano.

Figura 7: Página de consulta de conceitos da plataforma desenvolvida

4.4 Página de um Conceito individual

Na página de visualização de um conceito individual (Figura 8), o utilizador tem acesso a todos os detalhes relacionados com um determinado termo médico. Para garantir a aces-

sibilidade e facilitar a compreensão, a informação proveniente do dataset é apresentada de forma clara e organizada, permitindo uma consulta eficiente e intuitiva.

No topo da página é apresentado o nome do conceito em destaque, seguido pelos seus sinónimos em português, se existirem. De forma organizada, são também apresentados os códigos linguísticos associados ao conceito (como "EN" para inglês, "FR" para francês, entre outros), bem como os respetivos termos traduzidos. Estes aparecem agrupados sob a forma de etiquetas, facilitando a identificação rápida das diferentes variantes linguísticas. A página inclui secções distintas para o significado principal do conceito, e para a respetiva área médica, que permite aceder diretamente à listagem completa de conceitos dessa mesma área. Se disponível, a definição enciclopédica, contextualização e a definição em catalão do respetivo conceito também são apresentados nesta página. Além disso, foram integrados tooltips nos diferentes vocabulários informativos, permitindo ao utilizador obter explicações adicionais ao passar o cursor sobre certas secções.



Figura 8: Página de um Conceito individual da plataforma desenvolvida

Por fim, a página disponibiliza um botão para editar o conceito e outro para o eliminar. Ao pressionar no botão de eliminar o conceito e toda a informação respetiva são eliminados permanentemente do dataset, recebendo o alerta no *browser* caso a operação tenha sido, ou não, um sucesso. Ao pressionar no botão de editar, o utilizador é levado para uma página (Figura 9) que permite editar e adicionar novas informações ao respetivo conceito, podendo naturalmente cancelar esta operação. Caso pretenda guardar as alterações, o utilizador vai ser alertado sobre o sucesso desta operação e o conceito irá ficar permanentemente alterado no dataset.

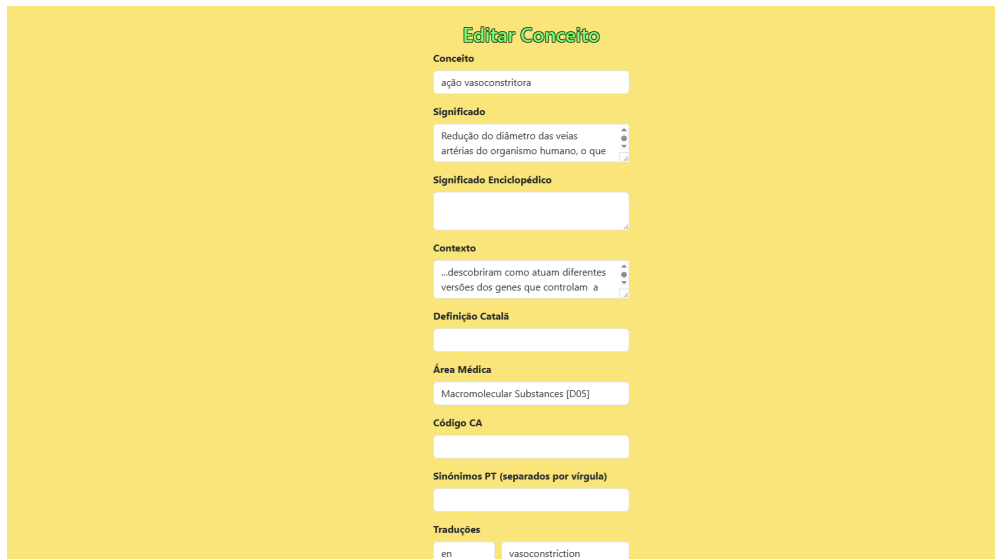


Figura 9: Página de edição dos conceitos da plataforma desenvolvida

4.5 Página das Áreas Médicas

A página das áreas médicas (Figura 10) apresenta todas as categorias médicas existentes no sistema, organizadas sob a forma de uma lista. Esta estrutura permite uma navegação temática, facilitando a consulta dos conceitos por especialidade. A organização em lista torna mais simples identificar e aceder rapidamente a grupos de termos relacionados, promovendo uma exploração eficiente da base de dados médica.

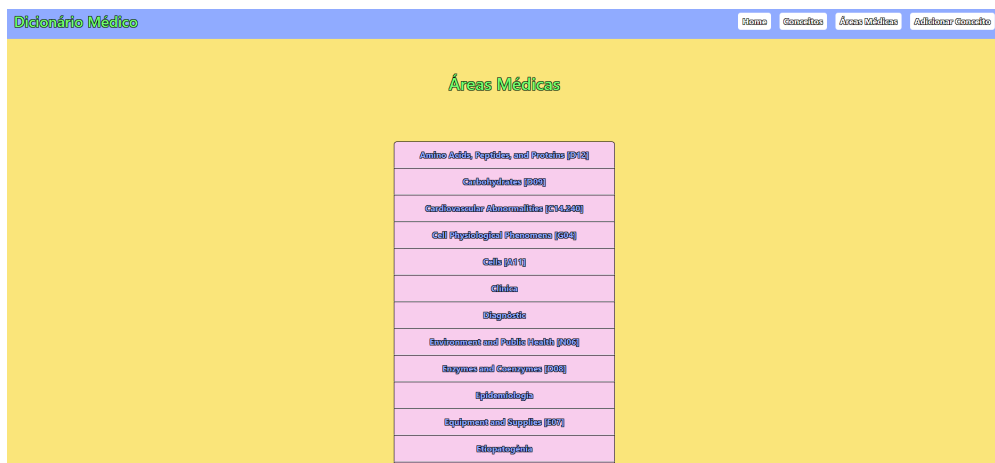


Figura 10: Página das áreas médicas da plataforma desenvolvida

Cada área pode ser clicada, levando o utilizador para uma página (Figura 11), similar à página de consulta dos conceitos, onde estão listados todos os conceitos pertencentes a essa área.

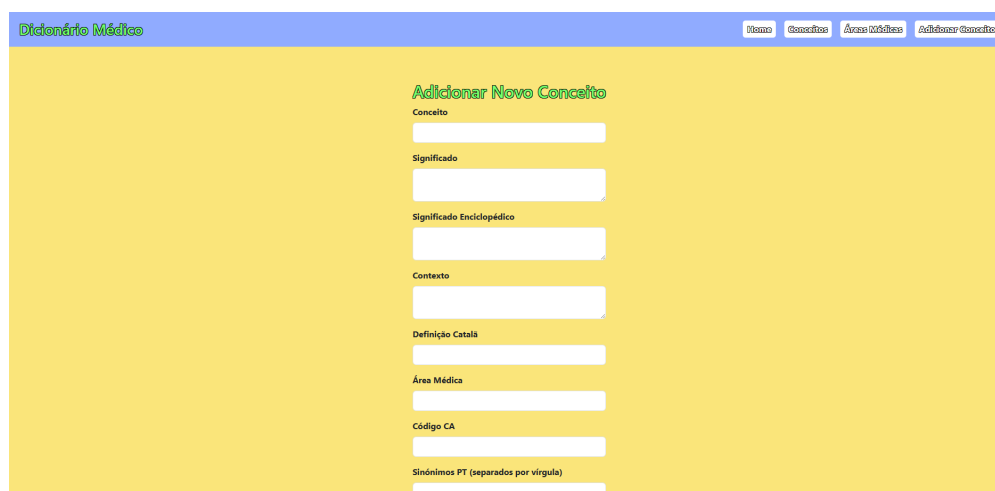


Conceito	Significado
calendário de vacinação	Calli Kairalla Farhat (Fernando Prestes, 7 de novembro de 1935 – São Paulo, 8 de setembro de 2010) foi um médico, pesquisador e professor universitário brasileiro.
capuz	Capuz Vermelho (Red Hood em inglês) é o codinome usado por alguns personagens fictícios da DC Comics, os principais sendo dois inimigos do Batman.
carga antígeno	O Glicocalix (também grafado glicocalice[1]) é uma matriz extracelular, uma camada externa à membrana, presente formada por glicolipídios, esfingolipídios, glicoproteínas e proteoglicanos.
cerca sanitária	Biomedicina (também conhecida como medicina ocidental, medicina tradicional ou medicina convencional[1]) é um ramo da ciência médica que aplica princípios biológicos e fisiológicos à prática clínica.
choque	Bioquímica (química aplicada à biologia) é a ciência e tecnologia que estuda e aplica as ciências químicas ao contexto da biologia, sendo portanto uma área interdisciplinar entre a química e a biologia.
cobertura vacinal	
colocado em quarentena	O 24.
Cominaty	Özlem Türeci (Siegen, Renânia do Norte-Vestfália; 1967) é uma médica, imunologista e empresária alemã de origem turca.
confinamento	-
contenção	Guerra Biológica é o conflito onde são usados microorganismos vivos como arma, dizimando vidas humanas, animais e vegetais.

Figura 11: Página dos conceitos pertencentes a uma área específica da plataforma desenvolvida

4.6 Página de adicionar um novo conceito

A página de adição de novos conceitos (Figura 12), permite ao utilizador introduzir manualmente um novo conceito no dataset. A informação é organizada por campos específicos, que orientam o preenchimento dos dados essenciais igual à estrutura dos restantes conceitos, sendo que o campo nome do conceito é o único obrigatório. Por fim, o utilizador dispõe de um botão para guardar o novo conceito e outro para cancelar a operação e regressar à página anterior, sendo que estes funcionam da mesma maneira que página de editar um conceito.



Adicionar Novo Conceito

Conceito

Significado

Significado Enciclopédico

Contexto

Definição Catalã

Área Médica

Código CA

Sinónimos PT (separados por vírgula)

Figura 12: Página de adicionar um novo conceito da plataforma desenvolvida

5 Conclusão

A utilização de scrapers foi fundamental para automatizar o enriquecimento dos conceitos médicos presentes no projeto, tanto em termos de atribuição da área médica (através da MeSH da National Library of Medicine) como na recolha de descrições introdutórias a partir da Wikipédia. No entanto, apesar da sua utilidade, os scrapers apresentam algumas limitações importantes. Entre os principais pontos negativos identificados, destacam-se a instabilidade da estrutura HTML das páginas, que pode mudar sem aviso e comprometer o funcionamento do scraper. Adicionalmente, como perspetiva futura, considera-se a substituição da Wikipédia por uma fonte mais estruturada e fiável, como a própria NLM ou outras bases biomédicas oficiais.

O desenvolvimento da plataforma web e a utilização do framework Flask no backend e HTML/CSS no frontend possibilitou a criação de uma plataforma acessível e funcional. A estrutura modular da aplicação permite ainda a sua futura expansão, como a integração de novos campos, dados ou até ligação a bases de dados externas.

Este trabalho demonstrou como é possível aplicar técnicas de Processamento de Linguagem Natural, promovendo a organização e difusão de conhecimento médico estruturado, com potencial utilidade em vários contextos.