



Universidade do Minho  
Escola de Engenharia

Processamento de Linguagem Natural em Engenharia  
Biomédica  
2024-2025

# **Trabalho Prático de Grupo**

## **Trabalho Prático 2**

Beatriz Amorim, PG56112  
Carolina Santos, PG56116  
Catarina Nunes, PG56117

Mestrado em Engenharia Biomédica | Informática Médica

---

# Conteúdo

<b>1</b>	<b>Introdução</b>	<b>3</b>
<b>2</b>	<b>Descrição do Dataset Inicial (TP1)</b>	<b>4</b>
<b>3</b>	<b>Enriquecimento do Dataset</b>	<b>5</b>
3.1	Fontes externas . . . . .	5
3.2	Web Scraping . . . . .	6
3.3	Integração dos Novos Dados . . . . .	6
<b>4</b>	<b>Transformação e Normalização dos Dados</b>	<b>7</b>
4.1	Similaridade . . . . .	8
4.2	Agrupamento por Categorias . . . . .	9
<b>5</b>	<b>Desenvolvimento da Plataforma Web</b>	<b>10</b>
<b>6</b>	<b>Conclusão</b>	<b>16</b>

# 1 Introdução

O principal objetivo deste projeto foi enriquecer e explorar um conjunto de dados de termos médicos, originalmente obtido através da extração de termos a partir de documentos em formato PDF, com o intuito de construir um recurso linguístico mais completo, útil e interativo, com foco na área da saúde.

Com o intuito de uniformizar e facilitar a manipulação dos dados, foi realizada uma transformação estrutural do *dataset*, adotando as traduções em português como chave principal dos termos que se encontravam noutras línguas.

Posteriormente, foi realizado um processo de enriquecimento semântico do dicionário, através da extração de novos termos e definições provenientes de fontes externas confiáveis, recorrendo a técnicas de *Web Scraping*. Esta etapa permitiu adicionar novas entradas ao *dataset*, enriquecer entradas já existentes e ampliar o conteúdo do conjunto de dados.

Por fim, foi desenvolvida uma plataforma *web* interativa, com recurso ao *framework Flask*, que permite visualizar, pesquisar, editar e navegar pelas entradas do dicionário enriquecido, tirando partido das relações entre os diferentes termos que constituem o dicionário. Esta aplicação visa, assim, facilitar o acesso e a exploração dos dados.

Deste modo, o presente relatório documenta todas as etapas do projeto, desde a preparação dos dados até ao desenvolvimento da aplicação final, detalhando as decisões técnicas tomadas, os desafios enfrentados e os resultados alcançados.

---

## 2 Descrição do Dataset Inicial (TP1)

O ponto de partida deste projeto foi o conjunto de dados reunido no trabalho prático anterior (TP1).

O ficheiro JSON final resultou da integração de múltiplos dicionários de termos médicos, provenientes de diferentes fontes, o que se refletiu na presença de campos distintos entre as entradas, consoante a origem e a natureza da informação disponível.

Abaixo apresentam-se exemplos representativos das diferentes estruturas do ficheiro final do TP1, ilustrando a variedade de formatos e a necessidade de uma posterior uniformização dos dados:

```
{
  "conceito": {
    "Categoria Lexical": "",
    "Traduções": { 'oc': [], 'eu': [], 'gl': [], 'es': [], 'en': [],
      ↪ 'fr': [], 'pt': [], 'pt_PT': [], 'pt_BR': [], 'nl': [], 'ar':
      ↪ [] },
    "Categoria": "",
    "Descrição": "",
    "Notas": [],
    "Sigla": [],
    "Símbolo": "",
    "Entrada Principal": "",
    "Sinónimo": "",
    "Sinónimo Complementar": [],
    "Denominação Comercial": "",
    "Nome Científico": "",
    "Número CAS": []
  },

  "conceito" : {
    "categoria_lexical": "",
    "traducoes": {
      "en": "",
      "es": ""
    },
    "descricao": "",
    "sigla": "",
    "inf_encicl": "",
    "citacoes": [""]
  },

  "conceito": {
    "Traduções": { "es": "tradução", "en": "translation" },
    "Descrição": "Texto descritivo.",
    "Género": "masc",
    "Número": "pl",
    "Sinónimos": ["sinónimo1", "sinónimo2"],
```

```
    "Notas": ["Nota 1", "Nota 2"],  
    "Sigla": "sigla",  
    "Remissivas": ["relacionado1", "relacionado2"]  
  },  
  
  "conceito": {  
    "termo_popular": ["Termo 1", "Termo 2"],  
  }  
}
```

### 3 Enriquecimento do Dataset

De forma a melhorar a qualidade e utilidade do *dataset* construído no TP1, procedeu-se ao seu enriquecimento semântico, através da consulta de fontes externas. Esta etapa teve como objetivo complementar os conceitos já existentes com informação adicional, assim como introduzir novos conceitos relevantes.

Foi dado especial destaque ao enriquecimento dos termos provenientes do documento "Dicionário Multilingue da COVID-19", pois estes possuem os campos exclusivamente em catalão. Por isso, procurou-se complementar o *dataset* com fontes que contivessem conceitos relacionados com a COVID-19, de modo a garantir que, no final, pelo menos algumas descrições estivessem disponíveis em português.

#### 3.1 Fontes externas

Para este processo, foram selecionadas fontes *online* fidedignas e tematicamente adequadas ao contexto biomédico e terminológico do *dataset*. As principais fontes consultadas foram:

- Jornal Público de Notícias (JPN) [1]
- Hospital da Luz [2]
- Centro Hospitalar de Lisboa Ocidental (CHLO) [3]

Estas fontes foram escolhidas pela sua confiabilidade, atualidade da informação e adequação ao vocabulário biomédico que se pretende representar e enriquecer.

As duas primeiras fontes incidem especificamente sobre a temática da COVID-19, disponibilizando termos e definições que permitiram complementar os conteúdos do "Dicionário Multilingue da COVID-19". Já a terceira fonte apresenta um glossário mais generalista, abrangendo um leque mais amplo de termos médicos relevantes para o contexto do *dataset*.

---

## 3.2 Web Scraping

De modo a extrair os glossários das três fontes referidas, utilizou-se *web scraping*, fazendo uso das bibliotecas *python - requests* e *BeautifulSoup*.

Na fonte JPN, o glossário estava presente no código-fonte HTML da página, dentro de uma secção com a classe *post-content*. Neste caso, o processo consistiu em:

1. Localizar a secção do conteúdo principal;
2. Encontrar todas as listas `<ul>` dentro dessa secção;
3. Para cada item `<li>` nas listas, identificar o termo em *tags* `<strong>` e a definição em *tags* `<em>`;
4. Extrair e limpar o texto de cada termo e definição, guardando também o texto adicional relacionado.

Na segunda fonte, do Hospital da Luz, o glossário era carregado dinamicamente via *JavaScript*, pelo que não era possível aceder diretamente ao conteúdo no HTML da página. Para ultrapassar esta limitação, identificou-se a API que devolvia o glossário em formato JSON. Foram feitas requisições HTTP a esta API, com os *headers* necessários para simular um *browser* real e permitir o acesso ao conteúdo. O resultado da API continha o HTML, pelo que se recorreu ao *BeautifulSoup* para analisar esse HTML e extrair os termos, em *tags* `<strong>`, e as definições associadas, presentes dentro dos elementos `<li>`.

Na última fonte, do Centro Hospitalar de Lisboa Ocidental (CHLO), os termos estavam organizadas em tabelas HTML paginadas, onde cada página apresentava um conjunto limitado de entradas. Assim, o *web scraping* consistiu em automatizar a navegação pelas páginas, incrementando o parâmetro *start*, no URL, para aceder a todas as páginas sucessivas. Em cada página, extraíram-se as linhas da tabela onde cada linha (`<tr>`) continha um termo e a respetiva definição em células separadas (`<td>`). Foram ainda aplicados tratamentos para garantir a correta extração dos termos, tendo em conta variações na estrutura HTML entre as páginas.

Deste modo, o *web scraping* realizado nas três fontes distintas gerou três ficheiros JSON separados, que serão posteriormente combinados com o JSON obtido no TP1, enriquecendo assim o conjunto de dados original.

## 3.3 Integração dos Novos Dados

Os novos dados extraídos através de *web scraping* foram estruturados em formato JSON, compatível com o formato do *dataset* original. Durante a integração, procedeu-se à comparação entre os novos termos e os já existentes, recorrendo a uma abordagem insensível a maiúsculas/minúsculas para evitar duplicações desnecessárias. Em casos de sobreposição, os conteúdos foram fundidos, garantindo que nenhuma informação útil era descartada.

## 4 Transformação e Normalização dos Dados

Para garantir a consistência do *dataset* utilizado, foi necessário aplicar um conjunto de transformações e normalizações aos dados. Estas operações tiveram como objetivo uniformizar a estrutura dos termos e respetivas entradas, corrigir duplicações e consolidar campos semelhantes.

Inicialmente, foram identificadas entradas duplicadas que diferiam apenas na capitalização (por exemplo, “Anticorpo” e “anticorpo”). Esta situação já existia no ficheiro JSON criado no TP1, embora só tenha sido identificada nesta fase do trabalho. Para resolver este problema, todas as chaves do dicionário foram transformadas de forma a começarem com letra maiúscula. Sempre que essa transformação originava duplicados, as respetivas entradas eram fundidas, combinando os conteúdos sem existir perda de informação.

Como o dicionário multilingue da COVID-19 apresentava, originalmente, as chaves em catalão, de maneira a conseguir enriquecer os respetivos termos com a nova informação obtida, foi realizada uma transformação que substituiu essas chaves pelos termos equivalentes em português, utilizando as traduções disponíveis no campo *Traduções*. Priorizou-se a tradução em *pt\_PT*, quando presente, ou em *pt*, caso contrário, garantindo que cada termo em catalão fosse substituído pela sua correspondência em português. As traduções em português foram removidas do campo *Traduções*, e o termo original em catalão foi adicionado como tradução em *ca*, preservando a relação bidirecional. Em casos de chaves duplicadas após a conversão, as entradas foram combinadas, juntando listas e dicionários sem perda de informação. Este passo permitiu que, para além das descrições dos termos em catalão, armazenadas no campo *Descrição*, alguns dos termos deste dicionário passassem a apresentar também uma descrição em português, resultante do processo de *web scraping*, guardada no campo *Descricao\_pt*, de forma a distinguir as diferentes línguas, mantendo ambas as versões da informação e tornando a informação mais acessível. Após essa transformação, os *datasets* foram consolidados novamente num único ficheiro JSON.

Foi também detetada alguma inconsistência nos campos utilizados para representar sinónimos: alguns termos apresentavam campos como *Sinónimo*, *sinonimo*, ou *Sinónimos*, variando entre forma singular e plural, com ou sem acentuação. Para garantir uniformidade, todos esses campos foram normalizados para uma única chave: *Sinónimos*, contendo sempre uma lista de valores. Foram removidos valores vazios e unificados sinónimos repetidos.

Os novos dados provenientes de dois dos ficheiros obtidos por *web scraping*, estavam estruturados da seguinte forma:

```
{
  "Termo": {
    "Definição": ""
  },
}
```

---

No entanto, verificou-se que a Definição nesses casos continha o mesmo tipo de informação que os campos `Descrição` ou `Descricao_pt` presentes no dicionário. Para uniformizar a estrutura e evitar redundância, todas essas informações foram consolidadas numa única lista sob o campo `Descrição`. Esta lista passa a conter tanto definições como descrições, independentemente da sua origem.

Durante o processo de integração de dados provenientes de diferentes fontes, foram ainda identificadas mais situações de redundância. Um dos exemplos foi a coexistência simultânea dos campos `traducoes` (com inicial minúscula) e `Traduções` no mesmo termo. Para evitar duplicação de conteúdos e garantir a coerência estrutural, foi definido que, sempre que ambas as chaves estivessem presentes, o campo `traducoes` seria removido, mantendo-se apenas a versão normalizada `Traduções` que possui as informações mais completas (traduções para mais línguas). Adicionalmente, nos casos em que apenas a chave `traducoes` existia, esta foi renomeada para `Traduções`, garantindo assim a uniformidade em todo o dicionário.

Outro caso de sobreposição semântica foi a existência dos campos `Remissivas` e `Sinónimos Complementares`, ambos utilizados para indicar termos relacionados. Tendo em conta a sobreposição semântica entre ambos, procedeu-se à sua uniformização sob a chave `Remissivas`, garantindo, mais uma vez, consistência na estrutura dos dados.

De igual modo, procedeu-se à normalização do campo da sigla, uma vez que se verificou a existência de entradas com a chave escrita com capitalizações diferentes - `sigla` e `Sigla`. Assim, todas as ocorrências passaram a utilizar exclusivamente a forma com letra maiúscula.

## 4.1 Similaridade

De forma a enriquecer o dicionário com conexões semânticas relevantes e a potenciar a melhoria das consultas na aplicação *web* desenvolvida, foi realizada uma análise de similaridade entre os termos. Este processo permite identificar quais termos estão semanticamente relacionados, facilitando sugestões mais precisas e contextuais, além de possibilitar uma navegação mais intuitiva e informativa.

Para explorar a relação semântica entre os termos e as respetivas definições presentes no dicionário, foi utilizada uma abordagem baseada em *embeddings* gerados por modelos de linguagem pré-treinados. Em particular, optou-se por utilizar o modelo `bert-base-portuguese-cased`, que é uma versão do BERT treinada especificamente para a língua portuguesa.

A primeira etapa consistiu na extração dos termos e das suas descrições, a partir do campo `Descrição`, ou, caso este não estivesse presente, através dos termos populares associados (`termo_popular`). Em seguida, os textos das descrições e termos populares foram convertidos em vetores utilizando o modelo BERT, que transforma cada frase num *embedding* contextualizado. Para representar cada descrição de forma concisa, foi utilizado o vetor médio dos *embeddings* de cada *token* da frase.



Com os vetores calculados, foi possível determinar a similaridade semântica entre as diferentes descrições, utilizando a métrica de similaridade do cosseno.

De seguida, para cada termo do dicionário, selecionaram-se os cinco termos semanticamente mais próximos, excluindo o próprio termo, com base nos valores da matriz de similaridade calculada. Esta lista de termos similares foi adicionada ao dicionário final, criando assim uma nova chave, *Similares*, que associa a cada termo os seus cinco termos mais relacionados semanticamente.

## 4.2 Agrupamento por Categorias

Para complementar a organização do dicionário e facilitar a navegação temática dos termos, foi implementado um processo de agrupamento automático por categorias. Esta abordagem visa atribuir a cada termo uma categoria conceptual que represente da forma mais adequada o seu contexto semântico.

Apenas a fonte relativa à COVID-19 possuía categorias previamente atribuídas aos seus termos. No entanto, essas categorias encontravam-se originalmente em catalão, pelo que foram traduzidas para português do seguinte modo:

- **Principis actius** → **Princípios activos**
- **Etiopatogènia** → **Etiopatogénese**
- **Tractament** → **Tratamento**
- **Clínica** → **Clínica**
- **Epidemiologia** → **Epidemiologia**
- **Prevenció** → **Prevenção**
- **Entorn social** → **Contexto social**
- **Conceptes generals** → **Conceitos gerais**
- **Diagnòstic** → **Diagnóstico**

Estas categorias serviram de base para a categorização dos restantes termos do dicionário, garantindo assim uma coerência semântica inicial no agrupamento dos conceitos. Para além destas, foram pré-definidas outras categorias, determinadas manualmente a partir da análise dos termos recolhidos, com o objetivo de abranger uma maior diversidade temática e conceptual. Assim, a lista final de categorias inclui:

- **Governança e administração pública;**
- **Políticas e intervenções na saúde pública;**

- 
- **Monitoramento e avaliação;**
  - **Gestão de dados e informação;**
  - **Planeamento e estratégia;**
  - **Metodologias e ferramentas;**
  - **Componentes e conceitos farmacológicos;**
  - **Anatomia.**

Para a categorização dos termos que não possuíam uma categoria previamente definida, foi implementada uma abordagem de agrupamento semântico automático. Para tal, foi definida uma estrutura de exemplos representativos por categoria. Cada categoria foi associada a um conjunto de termos (escolhidos manualmente do dicionário final) que refletem o seu significado semântico, funcionando como pontos de referência.

Utilizando, mais uma vez, o modelo pré-treinado de linguagem BERT, foram calculados *embeddings* para estes exemplos, e obtido um vetor médio que representa semanticamente cada categoria.

A categorização dos restantes termos foi realizada com base na semelhança do cosseno entre o *embedding* de cada termo e os vetores médios das categorias, sendo cada termo atribuído à categoria mais próxima.

Usaram-se duas formas de incluir as categorias nos ficheiros JSON, uma agrupando os termos por categoria (com categorias como chaves principais) e outra atribuindo a categoria diretamente a cada termo no seu próprio dicionário. A primeira abordagem facilita a análise e exploração temática, permitindo um acesso rápido a todos os termos associados a uma categoria, enquanto a segunda é mais prática para a pesquisa individual de termos. Estas duas metodologias resultaram na criação dos ficheiros `dicionario_final_com_tudo.json` e `dicionario_final_categorizado.json`.

## 5 Desenvolvimento da Plataforma Web

A plataforma *web* desenvolvida tem como objetivo principal disponibilizar o conteúdo do dicionário de termos médicos de forma interativa e acessível, com foco em termos em português, o seu significado e as suas traduções. A aplicação permite consultar, adicionar, editar e apagar termos médicos, organizados por categorias, além de oferecer funcionalidades de pesquisa avançada e navegação por índices alfabéticos. A plataforma foi projetada para ser intuitiva, facilitando o acesso a informações detalhadas sobre cada termo, incluindo descrições, sinónimos, remissivas e entradas principais, com redirecionamento automático para termos equivalentes em português quando aplicável.

Foram utilizadas as seguintes tecnologias no desenvolvimento da plataforma:

- **Python e Flask:** O *framework Flask* foi utilizado para criar a aplicação *web*, gerindo rotas, requisições HTTP e integração com *templates HTML*. O *Python* foi a linguagem principal para manipulação de dados e lógica do servidor.
- **Jinja2:** Motor de *templates* utilizado pelo *Flask* para gerar páginas HTML dinamicamente. Permite incluir lógica de programação (como condições e ciclos) diretamente nos ficheiros HTML, facilitando a exibição de dados enviados a partir do *backend* em *Python*.
- **HTML e Bootstrap:** Os *templates HTML*, estilizados com recurso ao *framework Bootstrap*, garantem uma interface visualmente consistente e esteticamente agradável.
- **JSON:** Os dados do dicionário são armazenados em ficheiros JSON, permitindo uma fácil manutenção e acesso estruturado.

A plataforma desenvolvida oferece um conjunto de funcionalidades principais voltadas para a navegação, consulta e gestão de termos no dicionário.

A navegação por termos está disponível na rota `/termos`, onde é apresentada uma lista de termos. Essa lista inclui um índice por letra inicial, permitindo ao utilizador filtrar os termos consoante a primeira letra, sendo os acentos ignorados nesse processo. Cada termo da lista pode ser clicado, redirecionando para uma página individual onde são apresentadas as suas informações detalhadas.

Também é possível navegar por categorias, através da rota `/categorias`. Nesta secção, o utilizador pode seleccionar uma categoria específica e visualizar todos os termos associados, os quais são apresentados na rota `/categorias/<categoria>`. Aqui, os termos de cada categoria são apresentados sob a forma de uma *DataTable*, que exhibe tanto o termo (que redireciona para a página individual de detalhes quando clicado) como a respetiva descrição, caso esta exista. Esta tabela permite ainda a pesquisa de termos recorrendo a expressões regulares.

A aplicação oferece ainda uma funcionalidade de pesquisa avançada, acessível pela rota `/pesquisa`. Esta funcionalidade permite ao utilizador procurar termos por palavra-chave, com a possibilidade de filtrar os resultados com base em critérios como campos de busca, palavras completas ou em categorias específicas. Os termos encontrados na pesquisa são destacados em **negrito** nos resultados para facilitar a visualização e interpretação dos mesmos.

Adicionalmente, cada termo possui uma página própria, onde são exibidas informações detalhadas como a categoria, a descrição, os sinónimos, as traduções, as remissivas, as siglas, a entrada principal, os termos similares, entre outros campos. Quando existem termos equivalentes em catalão, os sinónimos, remissivas, entradas principais e termos similares apontam para os seus correspondentes em português, sempre que disponíveis.

Para além da consulta, a plataforma permite ainda a adição de novos termos, através da rota `/novo_termo`. Nessa página, o utilizador preenche um formulário para criar um novo termo, sendo feita uma validação automática para evitar a criação de termos duplicados.

Também é possível editar termos já existentes. A rota `/editar/<termo>` é acessada a partir da página individual de cada termo e permite atualizar campos como a descrição e a categoria de um termo previamente inserido.

Por fim, os termos podem ser apagados diretamente através de uma requisição DELETE na rota `/termos/<termo>`. Esta ação é precedida por uma confirmação do utilizador no lado do cliente, implementada com *JavaScript*, para evitar eliminações acidentais.

Todas estas operações executadas sobre o *dataset* original são preservadas mesmo após reinicializações do sistema, sendo o mesmo também atualizado após qualquer um destes procedimentos, permitindo que o utilizador obtenha sempre a informação mais recente.

Abaixo, apresentam-se exemplos concretos de utilização da plataforma.

A Figura 1 apresenta a página que lista os termos médicos, exemplificando a funcionalidade de navegação alfabética da plataforma. O índice de letras permite filtrar os termos, enquanto o botão "Novo termo" destaca a opção de adicionar novas entradas, reforçando a interatividade do sistema.

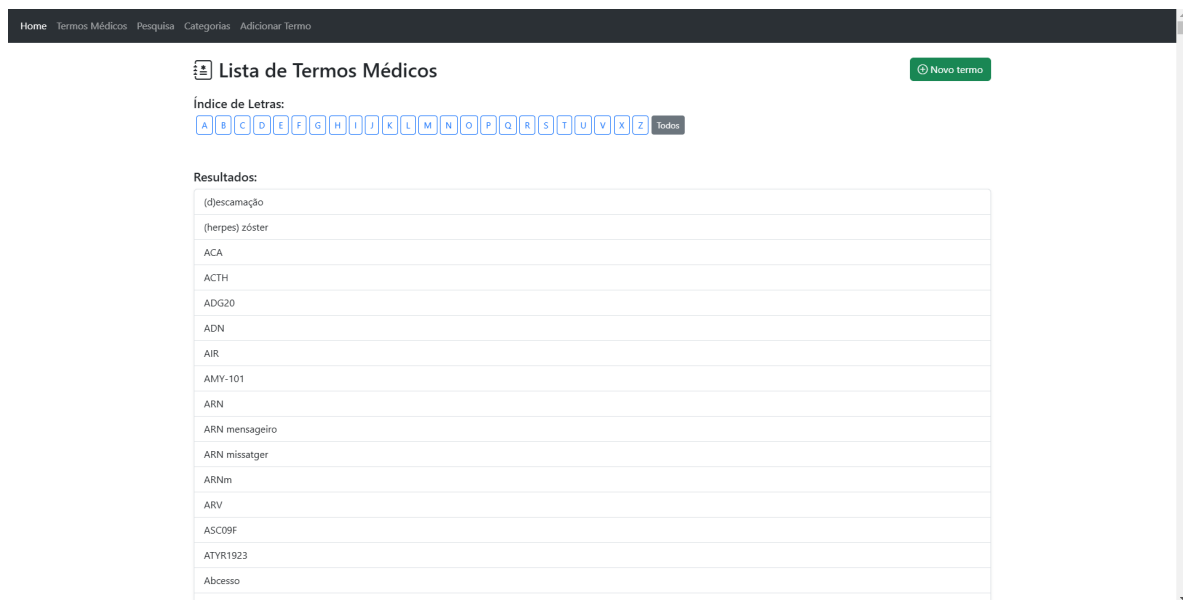


Figura 1: Página de lista de termos médicos, com índice alfabético e opção de adição de novos termos.

A Figura 2 apresenta a página detalhada do termo "Imunitário", exemplificando a funcionalidade de consulta de termos da plataforma. Esta interface reflete a integração das informações e a interatividade oferecida por opções de edição e exclusão do termo, além de permitir o redirecionamento automático de termos nas secções "Entrada Principal", "Remissivas" e "Termos Similares" para seus equivalentes em português, alinhando-se ao objetivo de fornecer um dicionário médico dinâmico e acessível.

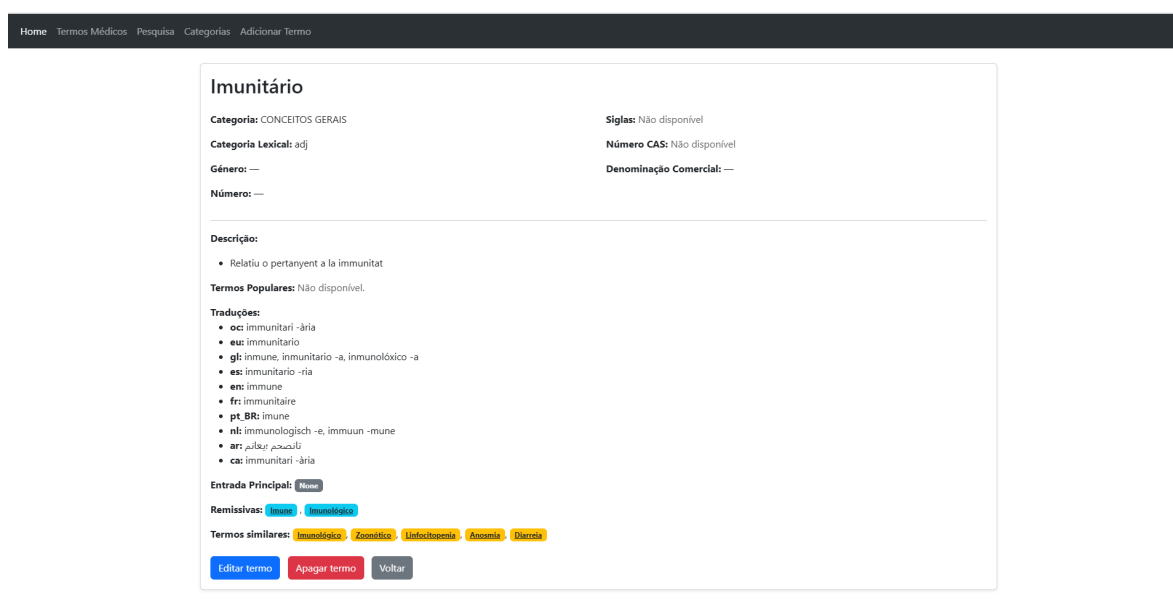


Figura 2: Página detalhada do termo 'Imunitário'

A Figura 3 apresenta a página detalhada do termo "Arn". Esta interface destaca o redirecionamento da "Entrada Principal" para "Ácido ribonucléico", a gestão de dados incompletos (como ausência de descrição) e as opções interativas de edição e exclusão do termo, refletindo a adaptabilidade da plataforma a diferentes entradas do dicionário.

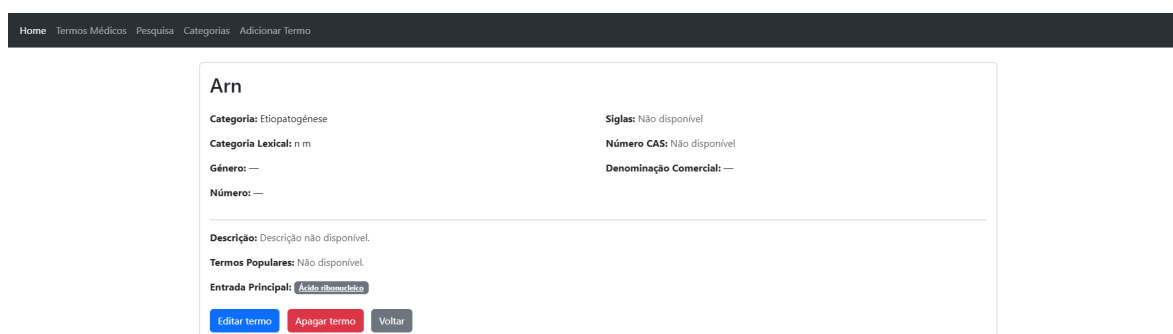


Figura 3: Página detalhada do termo 'Arn', mostrando redirecionamento na 'Entrada Principal'.

As Figuras 4 e 5 ilustram a funcionalidade de navegação por categorias da plataforma web, destacando a organização e acessibilidade dos termos médicos. A primeira imagem mostra a página de categorias, que lista as diferentes áreas do dicionário, permitindo ao utilizador selecionar uma categoria para explorar. A segunda imagem apresenta a visualização dos termos dentro da categoria "Diagnóstico", exibindo uma tabela com os termos e as respectivas descrições, tal como a possibilidade de pesquisa dentro da mesma tabela, demonstrando a estruturação dos dados e a facilidade de consulta. Adicionalmente, clicando num termo à escolha, é-se mais uma vez redirecionado para a página individual do termo que contém as suas informações mais pormenorizadas.

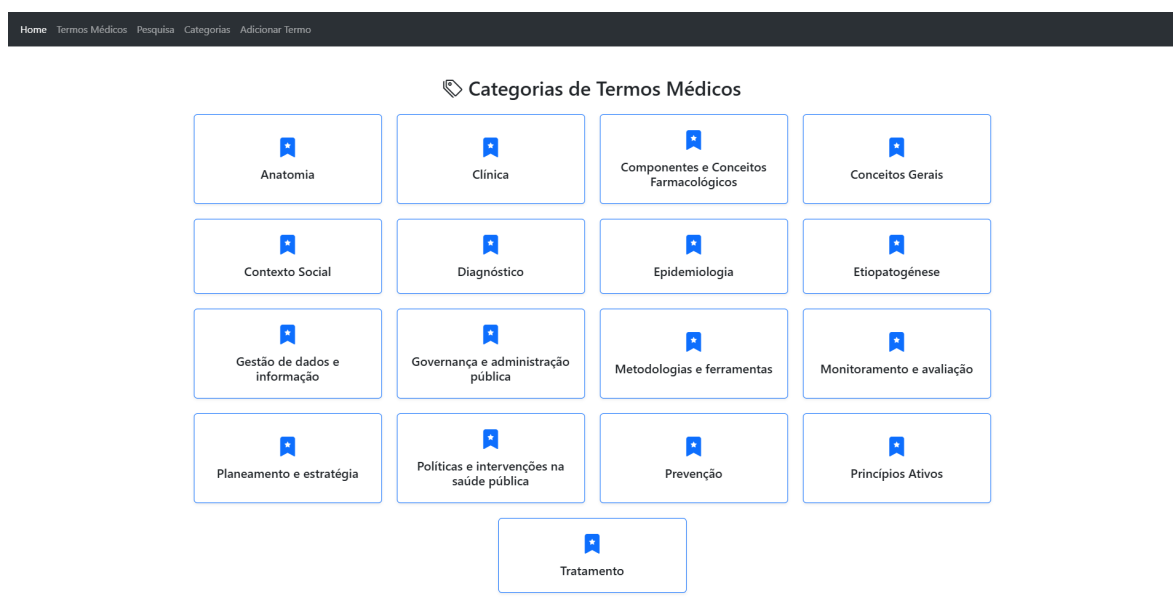


Figura 4: Página de categorias, exibindo as áreas do dicionário médico.

Termos na Categoria Diagnóstico	
10 entries per page	Search: <input type="text"/>
Termo	Descrição
Adenoidectomia	Operação para ablação das adenoides.
Amigdalectomia	Excisão de uma ou das duas amígdalas.
Análise biomecânica	Avaliação médica que pode ser feita em pessoas que estejam passando por uma reabilitação motora, através do uso de imagens de vídeo.
Análise de anticorpos	Prova diagnóstica serológica baseada na detecção d'anticorpos en la sang
Anestesia	Supressão temporária da sensibilidade e da consciência, mediante técnicas utilizadas em cirurgia, para fins operatórios, exploratórios, terapêuticos. Ausência ou perda de vários ou de um dos tipos da sensibilidade, em todo ou parte do corpo.
Angiografia	Radiografia obtida após a injeção endovenosa de um produto rádio-opaco (substância de contraste). Denomina-se arteriografia quando aplicado no estudo das artérias, e venografia quando aplicado no estudo das veias.
Angiografia de retina	Técnica utilizada para tratar de doenças oculares, como a Degeneração Macular Relacionada à Idade.
Angioplastia	Operação cirúrgica cujo objetivo é reparar um vaso sanguíneo.
Aponevrectomia	Excisão de uma aponevrose.
Artrocentese	Punção de uma articulação.

Figura 5: Tabela de termos na categoria 'Diagnóstico', com as suas descrições.

Adicionalmente, a Figura 6 representa o sistema de pesquisa implementado, dando destaque para as várias opções de afinamento e personalização da pesquisa, nomeadamente seleccionando em que campos ou categorias se pretende fazer a pesquisa.

Figura 6: Menu de pesquisa e respectivas opções de configuração.

Um excerto dos resultados obtidos para uma pesquisa exemplo do termo "Nasal", sem nenhuma das opções anteriormente mencionadas selecionadas, pode ser consultado na Figura 7.

#### Resultados para "nasal":

<b>Termo:</b> Cânula <b>nasal</b> <b>Descrição:</b> Descrição não disponível. <a href="#">Consultar Informação Completa</a>
<b>Termo:</b> Congestão <b>nasal</b> <b>Descrição:</b> Sensació de disminució del flux d'aire que entra per les fosses <b>nasals</b> deguda a la inflamació dels vasos sanguinis de la mucosa <b>nasal</b> per agents exògens nocius <a href="#">Consultar Informação Completa</a>
<b>Termo:</b> Exsudado nasofaríngeo <b>Descrição:</b> Fluid corporal extret de la part de la faringe de darrere les fosses <b>nasals</b> , per damunt del vel del paladar, amb l'ajuda d'un escovilló, amb finalitats diagnòstiques <a href="#">Consultar Informação Completa</a>
<b>Termo:</b> Cânula <b>nasal</b> <b>Descrição:</b> Instrument utilitzat per a fluxos d'oxigen reduïts que consta de dos tubs petits flexibles que s'introdueixen a l'entrada de les fosses <b>nasals</b> <a href="#">Consultar Informação Completa</a>

Figura 7: Excerto dos resultados obtidos na pesquisa para o termo "Nasal".

Por fim, as Figuras 8 e 9 apresentam os formulários de adição e edição de termos no dicionário, sendo de destacar que, no caso da edição, o formulário vem já pré-preenchido com as informações atuais do termo. Adicionalmente, após a adição de um novo termo ou da edição de um já existente, é exibido o termo e os respectivos campos submetidos.

Figura 8: Formulário para a adição de um novo termo.

Figura 9: Formulário para a edição de um termo.

## 6 Conclusão

Este projeto alcançou o objetivo de enriquecer e explorar um conjunto de dados de termos médicos, inicialmente extraído de documentos PDF, transformando-o num recurso linguístico completo, interativo e voltado para a área da saúde. A reestruturação dos dados, centrando as traduções em português como chaves principais, garantiu maior uniformidade e simplificou o processamento do *dataset*.

Durante o processo de enriquecimento semântico, enfrentaram-se desafios significativos, como a duplicação de entradas, variações na capitalização e dados inconsistentes provenientes de múltiplas fontes. Foram aplicadas técnicas de *web scraping* sobre fontes externas confiáveis para complementar os dados com definições, categorias e traduções, exigindo a fusão de campos, além da reconciliação de formatos heterogêneos. A integração do dicionário da COVID-19 em catalão apresentou ainda o desafio de tradução bidirecional e normalização de termos, exigindo mapeamentos cuidadosos para evitar redundâncias.

A plataforma *web*, desenvolvida com o *framework Flask*, oferece uma interface intuitiva que permite visualizar, pesquisar, editar e navegar facilmente pelos termos. A implementação de funcionalidades como redirecionamentos automáticos e atualização de dados reforçou a robustez da aplicação, superando obstáculos técnicos relacionados à gestão dinâmica de rotas e à integridade dos dados exibidos ao utilizador.



## Referências

- [1] Sofia Frias. COVID-19: As palavras que entraram no nosso vocabulário - JPN, 6 2020. <https://www.jpn.up.pt/2020/04/07/covid-19-as-palavras-que-entraram-no-nosso-vocabulario/>.
- [2] Glossário para Covid-19 | Hospital da Luz. <https://www.hospitaldaluz.pt/pt/saude-e-bem-estar/glossario-para-covid-19>.
- [3] Glossário. <https://www.chlo.min-saude.pt/index.php/component/seoglossary/1-glossario>.