Universidade do Minho
Mestrado em Informática Médica

Natural Language Processing
Assignment 2

Unidade Curricular: Processamento de Linguagem Natural em Eng. Biomédica

Ano Letivo: 2022/2023

Trabalho realizado por:

Maria da Conceição Vieira Mota, PG51210

Susana Isabel Pereira Martins, A93790

Índice

1. Introdução 2
2. Contexto e Requisitos 2
3. Abordagem e Implementação 3
3.1. Fase 1: Análise e extração de informação 3
3.2. Fase 2: Implementação da ferramenta 5
3. Conclusão 12

1. Introdução

Este relatório apresenta o trabalho realizado no desenvolvimento de uma aplicação utilizando técnicas de *web scraping* com a biblioteca *BeautifulSoup* e a *framework Flask* do *Python*.

O *web scraping* é uma técnica que permite extrair informações de sites de forma automatizada. Neste trabalho, foi utilizada a biblioteca BeautifulSoup para realizar o web scraping. Através dessa abordagem, foi possível a extração de dados essenciais como termos anatómicos e informações sobre ossos, de duas fontes online. A *BeautifulSoup* facilitou a extração de dados de forma eficiente, permitindo a obtenção de informações necessárias para a aplicação.

A *framework Flask*, por sua vez, é uma *framework* de desenvolvimento web em *Python*. Ao longo do desenvolvimento da aplicação, foram utilizados recursos como rotas, modelos de renderização e o mecanismo de template *Jinja* para criar páginas dinâmicas e interativas.

No decorrer deste relatório, serão detalhadas as etapas de extração de dados, a estrutura e os recursos da aplicação desenvolvida com o *Flask*.

2. Contexto e Requisitos

Este trabalho tem como objetivo enriquecer o conjunto de dados médicos gerado no trabalho anterior, procurando informações adicionais de fontes externas relevantes. Para isso, foram explorados sites online para ser realizado o *web scraping*.

Além disso, ao longo da realização deste trabalho, foi feita uma análise dos termos e as suas possíveis relações com o intuito de agrupá-los em domínios ou categorias específicas permitindo uma representação adequada das informações.

Com base nas informações enriquecidas, foi desenvolvida uma ferramenta que permite a manipulação eficiente do conjunto de dados. Esta ferramenta é capaz de atualizar as informações do conjunto de dados, garantido, assim, que este esteja sempre atualizado e completo.

Desta forma, o trabalho visa não apenas enriquecer o conjunto de dados, mas também fornecer uma ferramenta prática e eficaz para lidar com esses dados, permitindo uma análise mais abrangente e facilitando o acesso às informações necessárias para pesquisas médicas e análises posteriores.

3. Abordagem e Implementação

Dado o desafio, a implementação deste trabalho prático foi dividida em duas fases principais: análise e extração de informações através de técnicas de *web scraping* e implementação da ferramenta utilizando a *framework Flask*.

3.1. Fase 1: Análise e extração de informação

Nesta fase do desenvolvimento do trabalho, começou-se por enriquecer o conjunto de dados através do *web scraping* num site de referência médica. O objetivo era extrair termos, descrições e imagens para preencher a categoria de anatomia no ficheiro *JSON*, já existente, que estava incompleto.

Para realizar *o web scraping*, foi utilizada a biblioteca ***Requests*** para fazer uma requisição ao site e obter o HTML da página desejada. Em seguida, recorreu-se à biblioteca ***BeautifulSoup*** para analisar o HTML e extrair as informações desejadas.

O site escolhido para o *web scraping* foi o seguinte:

“https://reference.medscape.com/guide/anatomy"

A partir da página principal de anatomia do site referido, foi percorrida toda a sua estrutura HTML de forma a serem encontrados e extraídos os termos e as suas respetivas descrições e imagens.

As descrições e URLs das imagens foram armazenadas num dicionário, onde cada termo era a chave e as informações eram os valores correspondentes.

Na figura abaixo, encontra-se representado o ficheiro em formato JSON com os dados extraídos.


```
"Anal Canal Anatomy": {
    "description": "The anal canal is the most terminal part of the l
    "image_urls": [
        "https://img.medscapestatic.com/pi/meds/ckb/45/71345tn.jpg",
        "https://img.medscapestatic.com/pi/meds/ckb/65/13265tn.jpg",
        "https://img.medscapestatic.com/pi/meds/ckb/66/13266tn.jpg"
    ]
},
"Ankle Joint Anatomy": {
    "description": "The ankle joint is a hinged synovial joint with p
    "image_urls": [
        "https://img.medscapestatic.com/pi/meds/ckb/59/12459tn.jpg"
```


Figura 1. Ficheiro JSON

Foi ainda utilizada outra fonte online para extrair informações como definições, notas e links de imagens para vários ossos.

”https://www.imaios.com/en/e-anatomy/anatomical-structure/frontal-bone-1536895744?from=2”

O processo de extração ocorre de forma semelhante ao descrito anteriormente para termos anatómicos, no entanto, surgiram algumas adversidades.

Foi necessária a utilização da biblioteca ***fake_useragent'*** para gerar um User-Agent aleatório para evitar bloqueios de requisições pelo site. Em seguida, foi necessário implementar a linha de código ***'time.sleep(1)'*** para criar um atraso de 1 segundo no programa. Este atraso foi importante, no sentido que evitou fazer um grande número de requisições em sequência, sem este tempo de espera estávamos a ser “banidos” no acesso ao site. De qualquer forma, este tempo também é importante para não sobrecarregar o servidor do site com vários pedidos.

Desta forma foi possível recolher alguma informação como é possível ver na figura abaixo, que representa o ficheiro JSON com os dados extraídos.

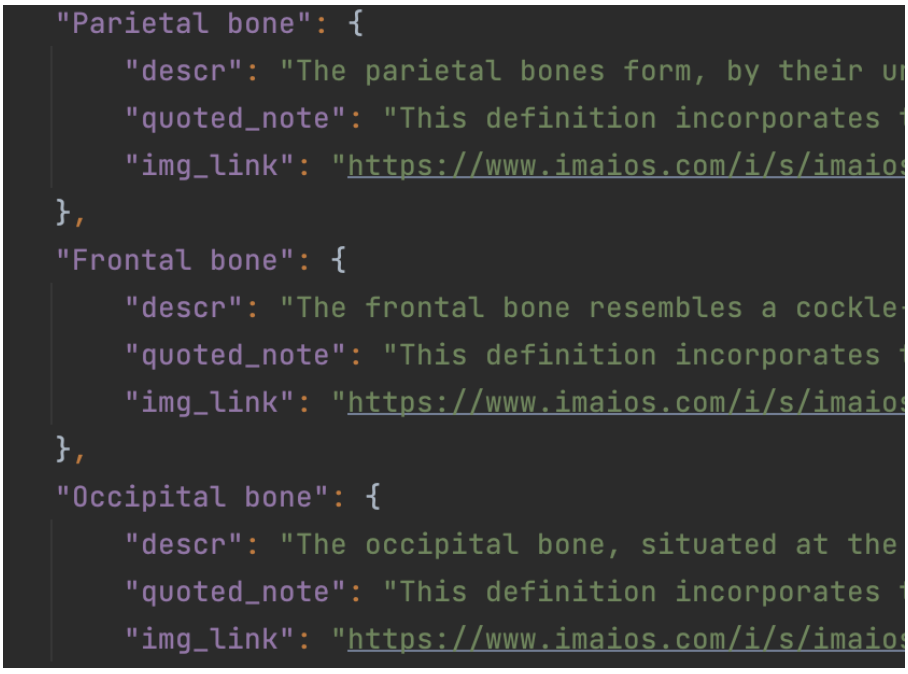


Figura 2. Ficheiro JSON

Além disso tivemos algumas limitações na recolha de informação recursivamente, dado que o HTML é adicionado dinamicamente à página, ou seja, seria necessário simular um clique para expandir a árvore da página para que o HTML fosse adicionado, para então conseguirmos navegar recursivamente nos vários links da página.

3.2. Fase 2: Implementação da Aplicação Web

3.2.1 Base de Dados

Além dos dados obtidos por meio do web scraping, também utilizamos o conteúdo de alguns ficheiros PDF que foram extraídos durante o trabalho prático anterior. Nomeadamente:

- Glossário Médico que é a junção de 3 glossários provenientes dos ficheiros:
  - WIPOPearl_COVID-19_Glossary.pdf
  - RU5HW615037.pdf
  - CIH Bilingual Medical Glossary English-Spanish.pdf
- Dicionários extraídos do ficheiro:
  - dicionario_termos_medicos_pt_es_en.pdf
    - Português-Inglês-Espanhol
    - Inglês-Espanhol-Português
    - Espanhol-Inglês-Português
- Ossos proveniente do ficheiro:
  - ossos.pdf

Optamos por utilizar apenas informação em inglês, a fim de manter a uniformidade da linguagem na aplicação.

No entanto, o ficheiro dos ossos está em português, mas consideramos que seria interessante acrescentar essa informação, e como as suas designações são pequenas, optamos por utilizar a biblioteca Python do GoogleTranslator, apresentada nas aulas práticas, para traduzir os termos para o inglês.

3.2.2 Aplicação Web

A aplicação Flask desenvolvida possui cinco páginas:

- **Home**

Representa a página inicial e é a página principal da aplicação, fornecendo uma visão geral do conteúdo disponível e convida os utilizadores a explorarem diferentes secções da aplicação.

A página é composta por quatro secções principais, cada uma apresentando um título, uma breve descrição e links relacionados aos respetivos recursos.

A página inicial está projetada para apresentar de forma clara e organizada as diferentes secções da aplicação, permitindo aos utilizadores que naveguem facilmente entre as secções que acharem relevantes, dependendo dos seus interesses e necessidades.

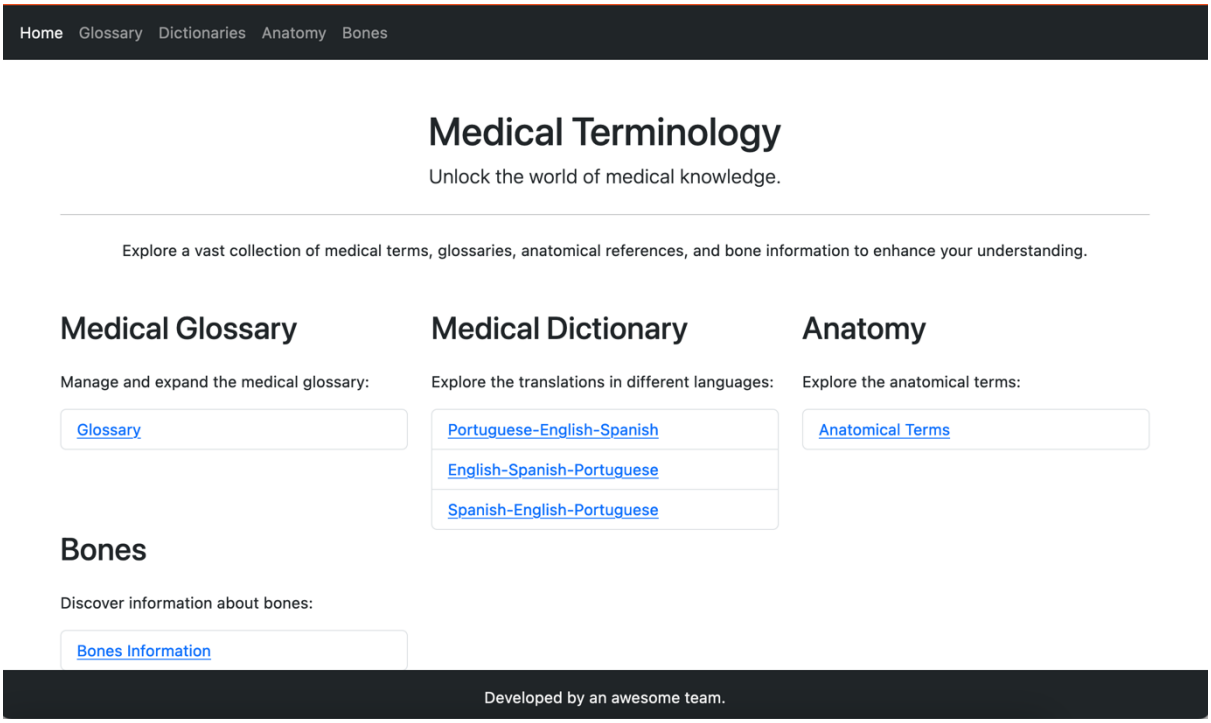


Figura 3. Página Home

- **Glossary**

Esta página mostra o glossário onde os termos podem ser pesquisados através de um formulário de pesquisa e os resultados correspondentes são exibidos na página.

No desenvolvimento, foram implementadas rotas e funcionalidades que fornecem uma interface interativa para manipular e visualizar os dados do conjunto de dados enriquecido.

Através da aplicação Flask, os utilizadores podem pesquisar, acrescentar, editar e eliminar termos, assim como toda a informação correspondente, neste caso, a descrição e a tradução para vários idiomas.

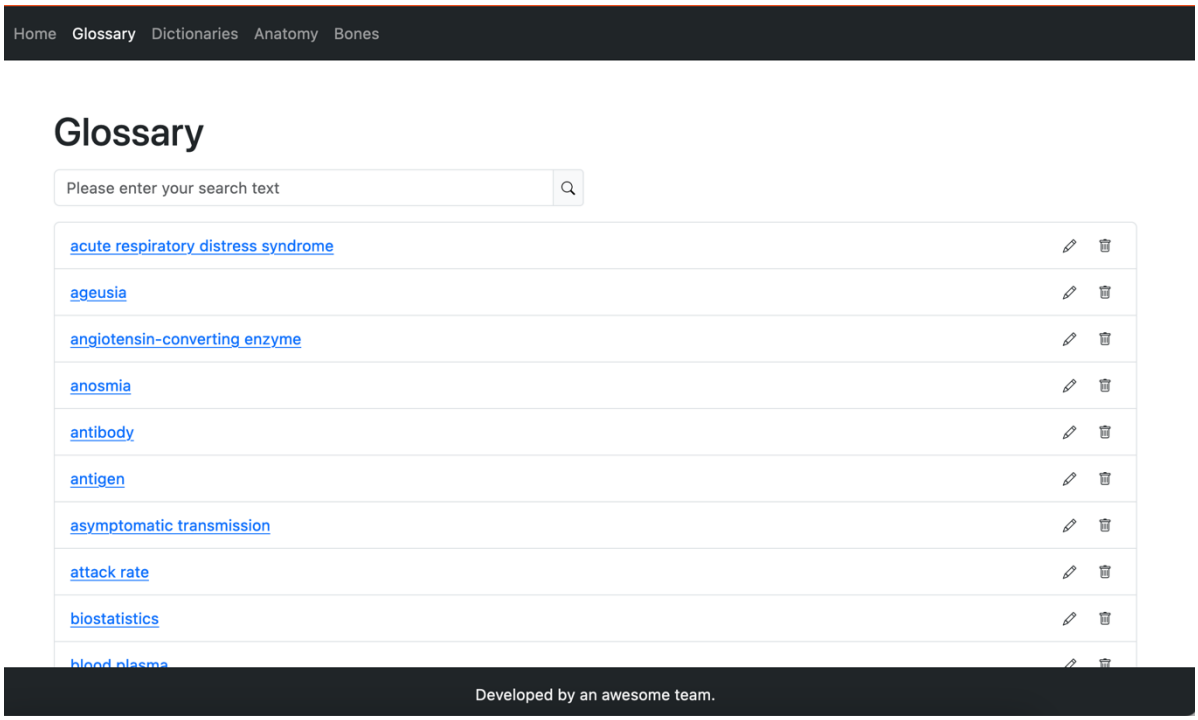


Figura 4. Página Glossary

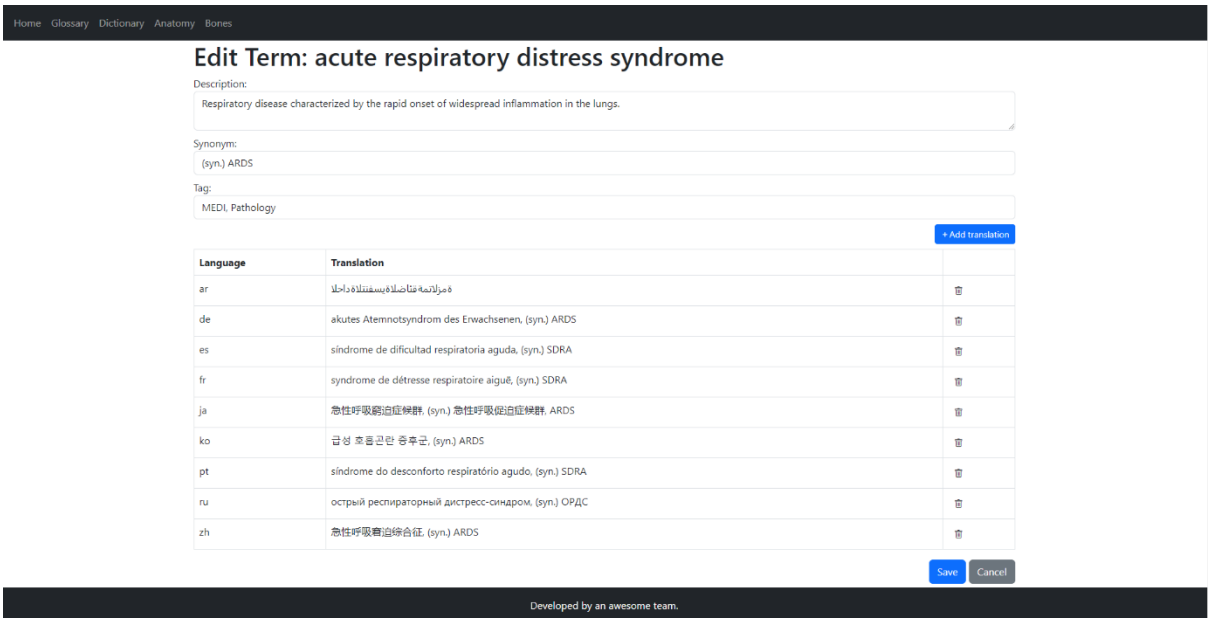


Figura 4.1 Edição de um termo do Glossário

- **Dictionary**

A página dictionary permite que os utilizadores acedam a uma lista de termos e as suas traduções em diferentes idiomas.

Ao aceder a página, o utilizador é recebido com uma visão geral dos idiomas disponibilizados, neste caso, PT, EN e ES. Os idiomas são exibidos como abas na parte superior da página, sendo que cada aba é um link que direciona o utilizador para a visualização dos termos e as suas traduções no idioma selecionado.

Os termos e as traduções, nesta página, estão organizados no formato de tabela com três colunas. A primeira coluna corresponde ao termo no idioma selecionado e as duas restantes representam as respetivas traduções.

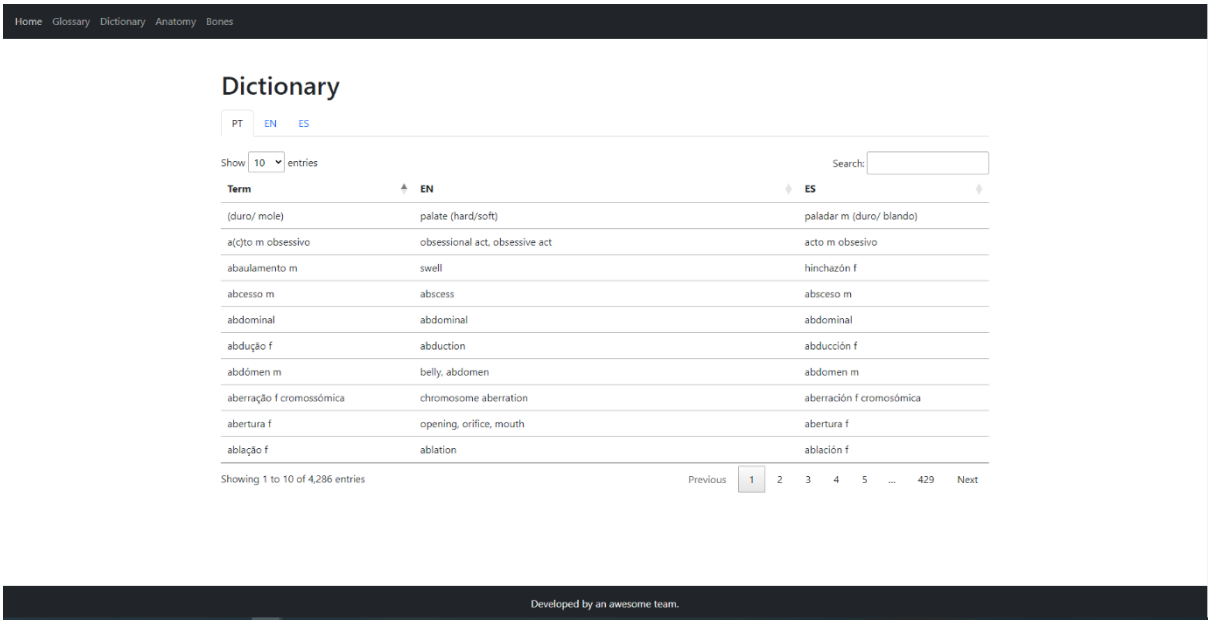


Figura 5. Página Dictionary

- **Anatomy**

Nesta página, os utilizadores podem explorar um dicionário anatómico com termos e descrições relacionadas.

Ao clicar num termo específico, os detalhes desse termo são exibidos numa página individual. A página mostra o termo como título e, em seguida, apresenta a sua descrição relacionada a ele.

Além da descrição, a página também pode exibir imagens relacionadas ao termo, caso estejam disponíveis.

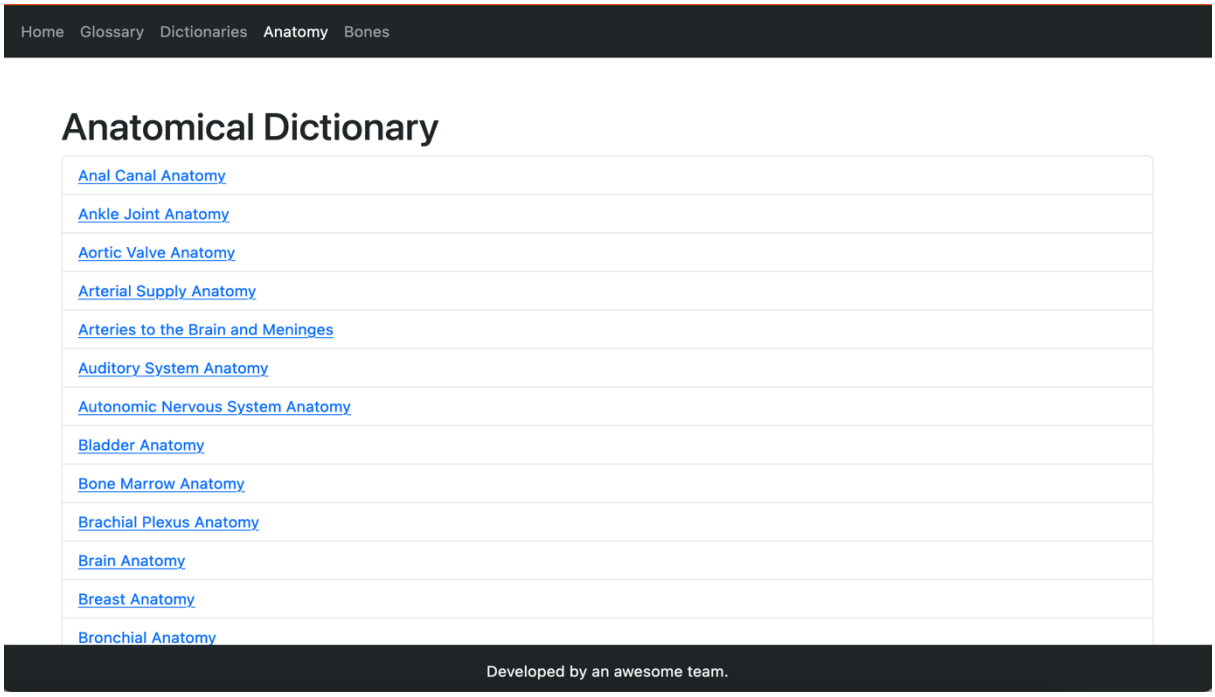


Figura 6. Página Anatomy

- **Bones**

A página relacionada aos ossos, permite aos utilizadores explorarem informações sobre diferentes categorias e grupos de ossos.

A página principal dos ossos exibe uma lista de categorias de ossos. Cada categoria é apresentada como um título e é seguida por uma lista de ossos relacionados a essa categoria.

Os ossos com imagens disponíveis são exibidos como links clicáveis que direcionam para a visualização das suas imagens.

Ao clicar num osso específico, o utilizador é redirecionado para uma página que inclui uma imagem do osso e uma lista de identificadores associados a ele. Cada identificador é clicável e, ao clicar nele, é exibida uma descrição correspondente.

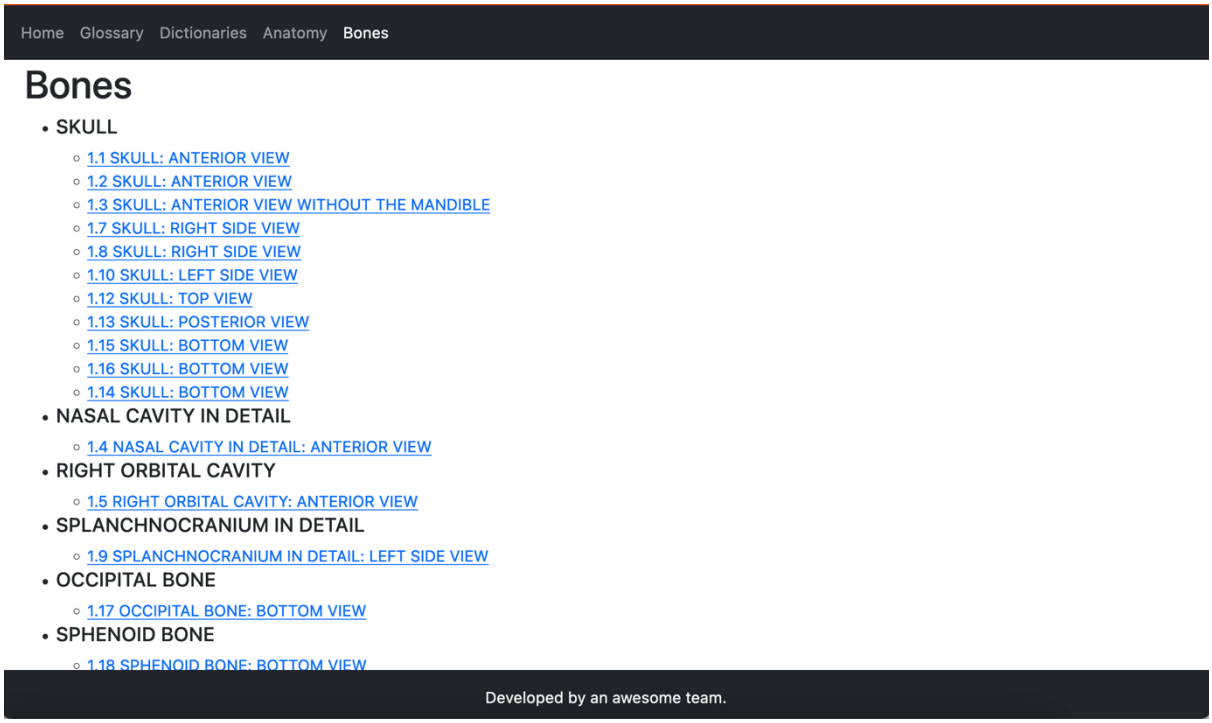
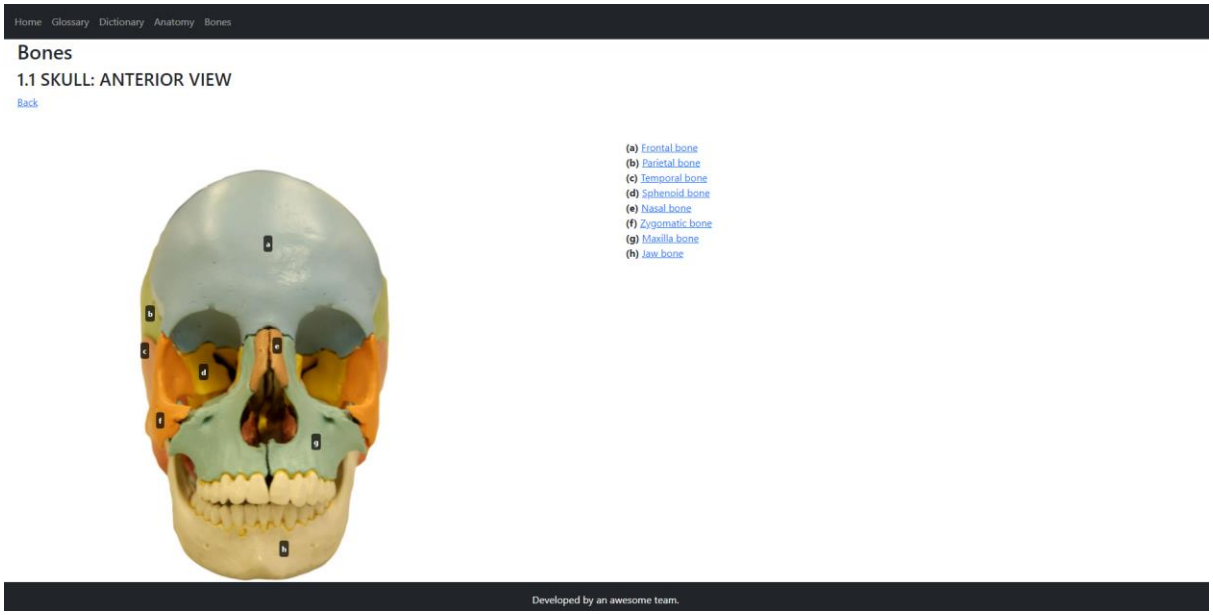


Figura 7. Página Bones

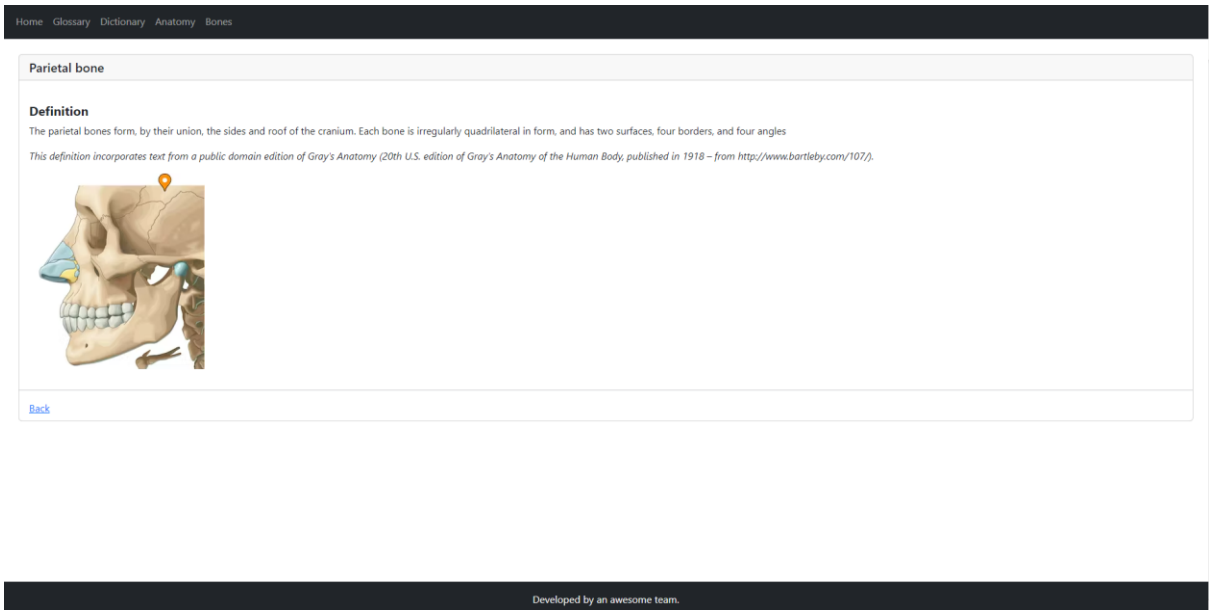
Durante o processo de tradução, aproveitamos para organizar esta informação de forma mais prática. Ao analisar todas as páginas do ficheiro de ossos, notamos que o título contém uma parte "comum" seguida por ":" e mais informações. Com base nisso, aplicamos um pré-processamento para agrupar os ossos pela primeira parte comum. Por exemplo, na lista abaixo, todos têm a parte comum "crânio (SKULL)", então consideramos "crânio" como uma categoria, seguida de todos os itens com seus títulos originais. Não removemos a numeração original, pois muitos têm o mesmo título para imagens diferentes.

- 1.1 SKULL: ANTERIOR VIEW
- 1.2 SKULL: ANTERIOR VIEW
- 1.3 SKULL: ANTERIOR VIEW WITHOUT THE MANDIBLE
- 1.7 SKULL: RIGHT SIDE VIEW
- 1.8 SKULL: RIGHT SIDE VIEW
- 1.10 SKULL: LEFT SIDE VIEW
- 1.12 SKULL: TOP VIEW
- 1.13 SKULL: POSTERIOR VIEW
- 1.15 SKULL: BOTTOM VIEW
- 1.16 SKULL: BOTTOM VIEW
- 1.14 SKULL: BOTTOM VIEW

Em cada um destes links é possível navegar para a respetiva imagem onde é possível ver também as respetivas designações dos identificadores.



Para cada um dos identificadores é possível navegar para a respetiva definição, caso exista.



A informação da definição é relativa à informação recolhia pelo web scraping no site e-anatomy. No entanto, devido às limitações mencionadas anteriormente, como a adição dinâmica de HTML na página, não foi possível obter uma quantidade significativa de informações através do web scraping recursivo.

3. Conclusão

A abordagem de web scraping permitiu-nos enriquecer o nosso ficheiro JSON com informações detalhadas que se tornaram valiosas para a manipulação e representação adequada dos dados na ferramenta desenvolvida.

A combinação dessas técnicas de web scraping com a biblioteca BeautifulSoup e ferramentas como o Flask e Jinja, resultou numa aplicação capaz de fornecer recursos abrangentes e centralizados para, por exemplo, profissionais de saúde e estudantes. Através da nossa aplicação, eles podem facilmente encontrar muitas informações detalhadas sobre termos, traduções e conceitos relacionados com a anatomia do corpo humano.

Ao longo deste trabalho, destacamos as capacidades e benefícios dessas ferramentas no desenvolvimento de aplicações web com funcionalidades complexas. Esperamos ter oferecido uma visão abrangente do trabalho realizado e demonstrado como essas tecnologias podem ser aproveitadas para criar soluções práticas e úteis no campo da terminologia médica.