

UNIVERSIDADE DO MINHO

TP2

Realizado no âmbito da Unidade Curricular de
Processamento de Linguagem Natural em Engenharia Biomédica

Discentes

Mónica Carina Costa Martins (a95918)

Ricardo Lopes Araújo (a97311)

Docentes

José João Antunes Guimarães Dias Almeida

Luís Filipe Cunha

junho de 2024

Índice

1	Introdução	3
2	Enriquecimento dos dados	4
2.1	Web Scraping	4
2.2	Estrutura final dos dados	5
3	Desenvolvimento da interface	6
3.1	Funcionalidades implementadas	6
3.2	Rotas	7
3.3	Demonstração	8
4	Conclusão	17

1 Introdução

O presente relatório trabalho foi realizado no âmbito da Unidade Curricular de Processamento de Linguagem Natural em Engenharia Biomédica e tem como objetivo a criação de um *website* para visualização e manipulação de conceitos e definições médicas.

Durante a realização deste trabalho, foram utilizados dados em formato JSON extraídos de diversos PDFs e *websites* com informação médica, dados esses que foram extraídos tanto durante a realização deste trabalho como no trabalho prático anterior. Deste modo, um dos objetivos deste trabalho consistia no enriquecimento do conjunto de dados obtido anteriormente. Para tal, foram aplicadas técnicas de web scraping com recurso à biblioteca *Beautiful Soup* e expressões regulares.

Pretende-se, ainda, a realização de uma *interface* interativa utilizando Flask e Jinja, para uma melhor visualização dos dados. A plataforma desenvolvida também deve permitir a atualização, adição e remoção de novas informações, garantindo a consistência dos dados após as alterações efetuadas.

Assim, o objetivo principal deste trabalho consiste na aplicação e aprofundamento dos conhecimentos adquiridos durante a Unidade Curricular de Processamento de Linguagem Natural em Engenharia Biomédica.

2 Enriquecimento dos dados

O enriquecimento dos dados foi efetuado com recurso às bibliotecas *Beautiful Soup*, *requests* e *json* do Python, para que fosse possível extrair a informação dos websites. Para além do web scraping efetuado, também foi adicionada mais informação aos dados extraídos, como é o caso de "sinónimos" e "termos relacionados".

Após a extração e enriquecimento dos dados, foi definida uma estrutura final para que fosse possível juntar os dados provenientes de diferentes fontes e concatenar essas informações.

2.1 Web Scraping

No trabalho TP1 realizado anteriormente, foram extraídos termos, traduções e definições de 3 fontes diferentes:

- Minidicionário do Cardiologista [1]: com termos médicos traduzidos de inglês para português e vice-versa;
- Vocabulário de Medicina [2]: com termos médicos em galego, espanhol, inglês e português, tendo associado categoria gramatical e áreas de aplicação, por vezes sinónimos, formas variantes ou notas, tendo ainda entradas remissivas.
- Glossário do Ministério da Saúde [3]: este ficheiro continha termos médicos em português e a respetiva definição.

A extração de informação destes ficheiros foi efetuada com recurso ao comando `pdftohtml -xml` e, de seguida, expressões regulares para eliminar informação não relevante.

O enriquecimento do conjunto de dados, desta vez, foi efetuado (maioritariamente) com recurso ao *web scraping*. As fontes de onde foram retirados mais dados foram as seguintes:

- Honor Health [4];
- Health Careers [5];
- Harvard Medical School [6];
- Great Ormond Street Hospital (GOSH) [7].

Todos estes *websites* continham informação acerca de termos médicos e respetivas definições.

De modo a enriquecer os dados extraídos, para cada um dos termos extraídos, foi efetuado um pedido ao *website* Thesaurus [8] de modo a obter sinónimos para cada um dos termos extraídos.

Após a extração, foi necessário juntar os dados de forma a evitar termos repetidos na mesma base de dados. Assim, quando o mesmo termo existia em fontes diferentes, a informação era concatenada. Por exemplo, se o mesmo termo estava presente em duas fontes diferentes, as definições eram juntas.

Para além das fontes referidas anteriormente, foi extraída informação de um ficheiro do *Update em Medicina*[9] (através do mesmo comando utilizado no trabalho TP1) e expressões regulares. Esta informação, contrariamente à anterior, tem como objetivo o treino e criação de um modelo Word2Vec, cuja funcionalidade será explicado mais à frente.

2.2 Estrutura final dos dados

Para melhorar a eficiência da visualização dos dados, definiu-se que a estrutura final dos dados seria inspirada na estrutura de dados definida no Vocabulário de Medicina, isto é, identificando cada termo através de um índice. Esta decisão foi tomada uma vez que este ficheiro apresentava muitas entradas (termos), todas elas com tradução para Inglês, Português e Espanhol (tendo sido descartada a tradução galega, por ser próxima ao espanhol e introduzir uma componente de dificuldade acrescida na obtenção de traduções de outras fontes).

Assim, foram criados três ficheiros finais, um para cada idioma. A forma para identificar os termos entre os diferentes idiomas e ficheiros, seria pelo índice. Esta decisão permitiria uma melhor performance para a visualização dos dados, já que, caso contrário, o ficheiro final tornar-se-ia muito pesado e o carregamento dos dados seria lento.

Deve notar-se que se garantiu que a informação entre ficheiros era transversal e coerente, ou seja, os termos presentes no ficheiro com todos os dados em português, também estão presentes no ficheiro em inglês. Para além disso, efetuou-se a tradução de toda a informação antes da adição aos respetivos ficheiros (descrição, sinónimos, áreas de aplicação, etc). Estas traduções foram efetuadas com recurso à biblioteca `deep_translator`.

Posto isto, a estrutura definida para cada um dos ficheiros foi a seguinte:

```
1 {
2   "1": {
3     "Termo": "Conceito1",
4     "Fontes": ["Fonte1",
5               "Fonte2"],
6     "Definicao": "Descricao do termo1.",
7     "Area(s) de aplicacao": ["Area1",
8                              "Area2"],
9     "Sinonimos": ["sinonimo1",
10                  "sinonimo2"],
11    "Categoria Gramatical": "f",
12  },
13  "2": {
14    "Termo": "Conceito2",
15    "Fontes": ["Fonte1",
16              "Fonte2",
17              "Fonte3"],
18    "Definicao": "Descricao do termo2.",
19    "Area(s) de aplicacao": ["Area1"],
20    "Relacionado": "1",
21    "Variante": "variante1; variante2",
22    "Nota": "Observacao relevante ao termo",
23    "Index_Remissivo": "Indice de uma entrada para a qual a presente remete"
24  }
25 }
```

Deve notar-se que nem sempre os termos terão toda a informação representada. No entanto, no máximo, o mesmo termo poderá ter as seguintes informações:

- Fontes;
- Definição;
- Área(s) de aplicação;
- Sinónimos;
- Categoria gramatical;
- Relacionado;
- Variante;
- Nota;
- Index Remissivo;

3 Desenvolvimento da interface

A criação da interface para a visualização e manipulação dos dados foi efetuado, como já foi referido, com recurso às ferramentas Jinja e Flask. O foco do desenvolvimento desta interface foi permitir que o utilizador navegasse e interagisse com os dados de forma eficiente.

3.1 Funcionalidades implementadas

As funcionalidades implementadas consistem em:

- **Troca de dicionários/ficheiros com base no idioma selecionado pelo utilizador:** já que, uma vez que foram separados os dados extraídos, a troca entre idiomas deve ser suportada;
- **Implementação de datatables:** para permitir a visualização dos dados de forma mais agradável;
- **Pesquisa por conteúdo:** funcionalidade inerente à datatable criada;
- **Visualização detalhada de um conceito:** onde o utilizador pode ver, em detalhe, toda a informação de um conceito. Nesta página é dada a possibilidade de o utilizador traduzir toda a informação do conceito para um total de 24 idiomas diferentes (Árabe, Búlgaro, Dinamarquês, Alemão, Grego, Francês, etc);
- **Adição de conceitos:** com a verificação se o conceito já existe ou não na base de dados. Nesta funcionalidade é garantida a adição transversal aos 3 ficheiros;
- **Alteração de um conceito:** funcionalidade que permite ao utilizador alterar qualquer campo da entrada que esteja a analisar, desde o próprio termo (estando implementada uma verificação de unicidade antes da inserção), adicionar diferentes áreas, fontes e sinónimos, bem como remover campos. O utilizador pode ainda escolher se pretende alterar o dicionário relativo ao idioma da página ou se pretende alterar todos, sendo feita a tradução automaticamente neste último caso. O idioma da informação inserida irá sempre corresponder ao do dicionário destino, pois também esta é passada pelo tradutor com deteção automática;
- **Remoção de um conceito:** funcionalidade que permite ao utilizador remover entradas de todos os dicionários da base de dados, através do id único, fazendo primeiro uma verificação de existência, retornando um aviso caso contrário, e uma mensagem de sucesso caso esta remoção seja bem sucedida;
- **Pesquisa detalhada:** permitindo a procura avançada por conteúdo, permitindo que os utilizadores efetuem a procura por termos, sinónimos, definições e fontes, com a possibilidade de aplicar um operador lógico (AND ou OR) à pesquisa efetuada;

- **Q&A:** permite que o utilizador faça uma questão, utilizando um modelo de Hugging Face para responder à mesma;
- **Similaridades:** através de um modelo Word2Vec, o utilizador pode averiguar qual a palavra mais semelhante a outra ou, dado um conjunto de palavras, averiguar qual delas se adequa menos ao conjunto.

De um modo geral, as funcionalidades implementadas garantem a visualização dos dados de forma visualmente agradável e eficiente. Para além disso, são garantidas todas as operações *CRUD* (create, read, update e delete).

3.2 Rotas

Durante este trabalho, foram desenvolvidas 10 rotas, com recurso ao Flask. A rota `'/'` renderiza o template `'home.html'` e trata-se da página inicial da aplicação.

A rota `'/conceitos/<lang>'` permite que, dado um idioma (`lang`), seja renderizado o template `'conceitos.html'`. Este template apresenta ao utilizador uma tabela dos conceitos, termos relacionados e fontes, no idioma especificado (caso o idioma seja EN, ES ou PT).

Para visualizar cada detalhe com mais detalhe, desenvolveu-se a rota `'/conceitos/<lang>/<id_conc>'` que, dado um idioma e índice (`id_conc`) de um conceito válidos, renderiza o template `'conc.html'`. Caso o idioma não seja suportado (ou seja, não for português, inglês ou espanhol), os campos do conceito são traduzidos, com o `deep_translator`. Esta tradução acontece com o auxílio da rota `'/change_language'`.

A funcionalidade de adicionar um conceito é garantida pela rota `'/add_entrada'` que, utiliza o método `'POST'`. Após receber os dados de um formulário, são criadas novas entradas nos dicionários em inglês, espanhol e português e guardados os ficheiros JSON correspondentes.

Por outro lado, a rota `'/alterar_entrada/<lang>/<id_conc>'` permite alterar um conceito já existente. Nesta rota permite-se que o utilizador escolha se pretende que as alterações sejam feitas aos três dicionários ou apenas um deles.

Para concluir as operações *CRUD*, a rota `'/apagar_entrada'` recebe o índice de um conceito e apaga o conceito dos três dicionários. Conforme o sucesso (ou não) da operação, é enviada uma mensagem ao utilizador (`'Entry removed'` ou `'There is no entry with that key!'`, respetivamente).

A rota `'/table'`, renderiza o template `'table.html'` com os conceitos em cada um dos três idiomas principais abordados neste trabalho.

O utilizador pode ainda fazer uma pesquisa avançada, como já foi referido anteriormente, recorrendo aos operadores `"AND"` ou `"OR"` e a diferentes componentes de um termo (descrição, sinónimos, fontes, etc). Para tal, desenvolveu-se a rota `'/pesquisa_detalhada'`. Esta rota recebe os parâmetros da pesquisa e procura os *matches* nos dicionários adequados.

Por fim, a rota `'/qa'` tem como objetivo responder a perguntas do utilizador e encontrar termos mais ou menos semelhantes. As respostas às perguntas são geradas a partir de um modelo Question Answering do HuggingFace (`lfc/bert-portuguese-squad`). Por outro lado, a similaridade entre termos é calculada através de um modelo Word2Vec, desenvolvido com o conteúdo do Livro de Resumos do congresso de 2024 do *UpdateMedicina*[9].

3.3 Demonstração

Ao entrar na aplicação, o utilizador depara-se com uma página como a da Figura 1.



Fig. 1: Página inicial.

Como se pode verificar, a *navbar* no topo da página apresenta 5 secções:

- **Home:** a página inicial;
- **Conceitos:** com uma tabela com informação de todos os conceitos, termos semelhantes e respetivas fontes;
- **Tabela de idiomas:** que contém um datatable com os termos dos 3 idiomas diferentes (inglês, espanhol e português);
- **Pesquisa detalhada:** com a possibilidade de o utilizador fazer uma pesquisa avançada;
- **Q&A:** como explicado na secção 3.1.

Na Figura 2 mostra-se a secção "Conceitos".

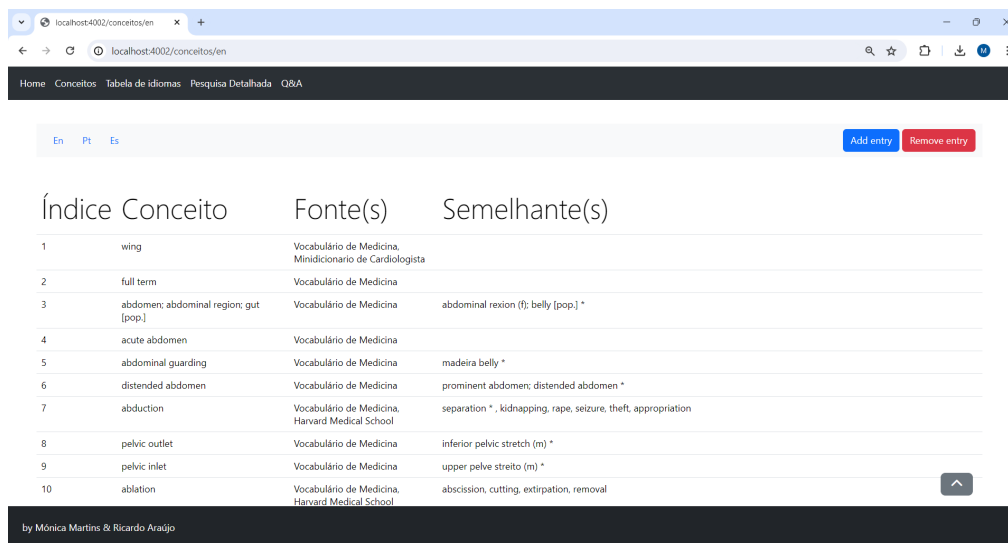


Fig. 2: Secção "Conceitos".

Uma vez que se tratam de muitos conceitos, foi adicionado um botão fixo no canto inferior direito que permite que o utilizador volte para o início da página diretamente. Para além disso, a *navbar* secundária contém 3 botões do lado esquerdo ('En', 'Pt' e 'Es') que permitem ao utilizador utilizar outro idioma. Na Figura 3, pode verificar-se a mudança do idioma para espanhol.

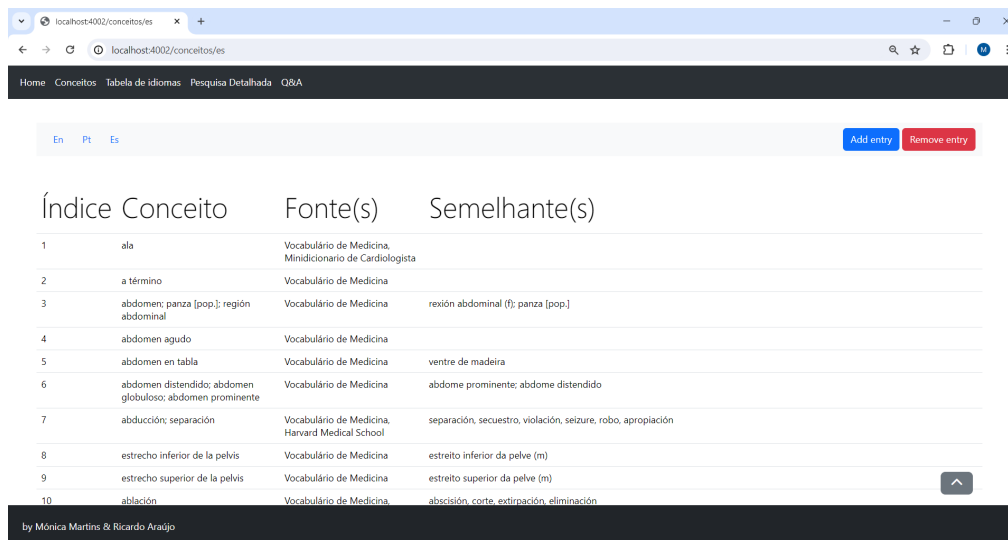


Fig. 3: Secção "Conceitos" em espanhol.

Do lado direito, o utilizador tem a possibilidade de inserir um novo termo ou de remover um existente. Na Figura 5 está apresentada a tentativa de inserção de um termo já existente na base de dados.

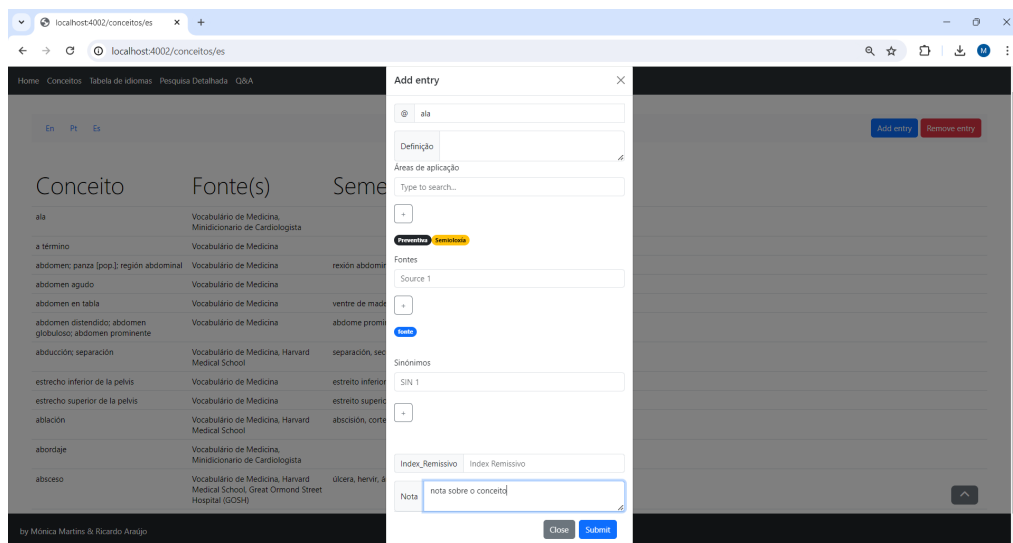


Fig. 4: Inserção de um termo existente.

Uma vez que o termo já existe, é enviada uma mensagem de erro ao utilizador, como mostra a Figura 5.

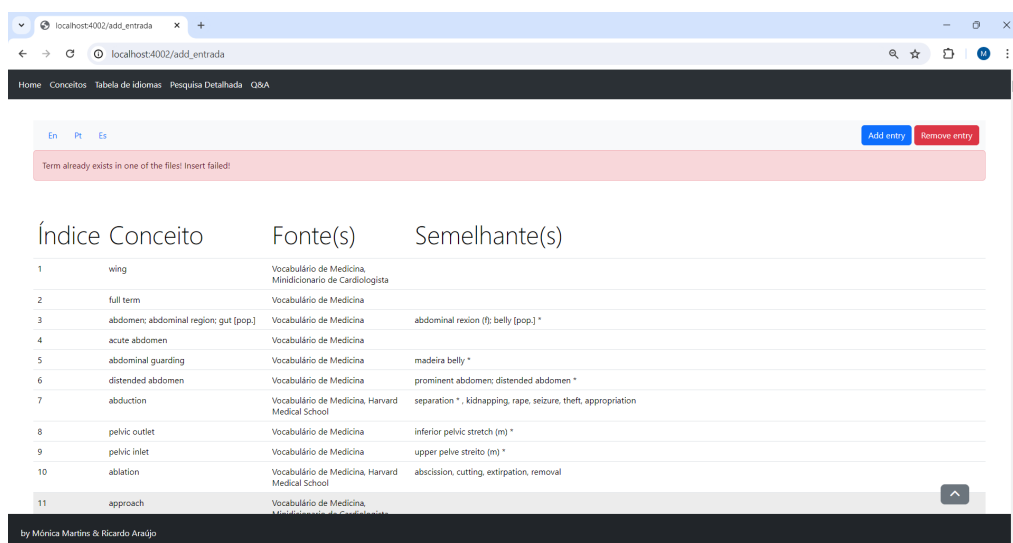


Fig. 5: Mensagem de erro ao inserir um termo já existente.

Por outro lado, quando o utilizador tenta inserir um termo que não existe na base de dados, é redirecionado para a página do conceito adicionado (Figura 6).

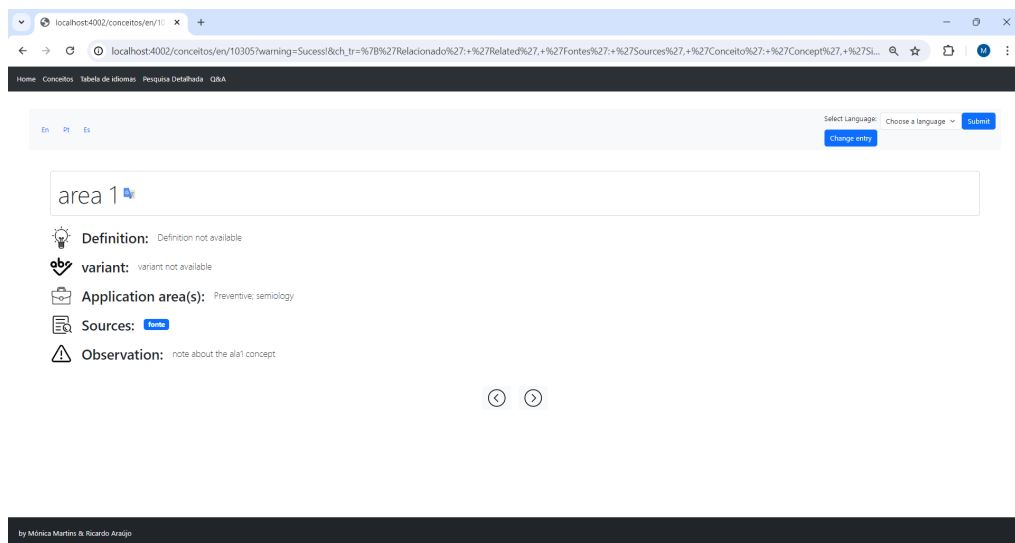


Fig. 6: Sucesso na inserção de um termo não existente.

Também é possível eliminar um conceito, sendo necessário, para isso, inserir o índice do termo a eliminar, para que seja possível eliminá-lo das 3 bases de dados diferentes (Figura 7).

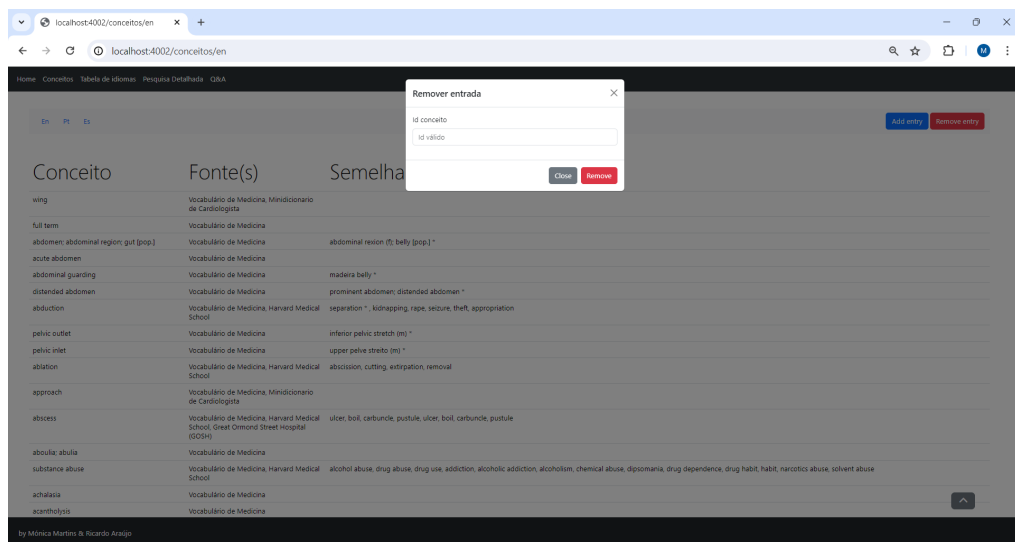


Fig. 7: Eliminação de um conceito.

Caso o termo seja eliminado com sucesso, é enviado um aviso ao utilizador, como mostra a Figura 8.

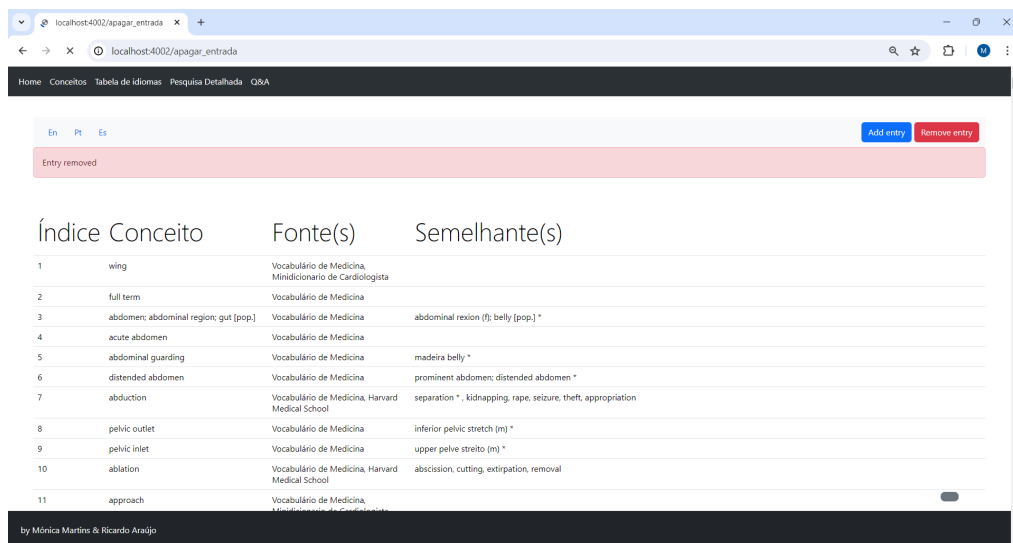


Fig. 8: Sucesso na eliminação de um conceito.

Na Figura 9, está representada a secção "Tabela de Idiomas", com todos os termos existentes na base de dados, nos 3 idiomas suportados. Note-se que, tanto nesta página como na da Figura 2, o utilizador pode carregar no conceito para ser direcionado para a página do conceito.

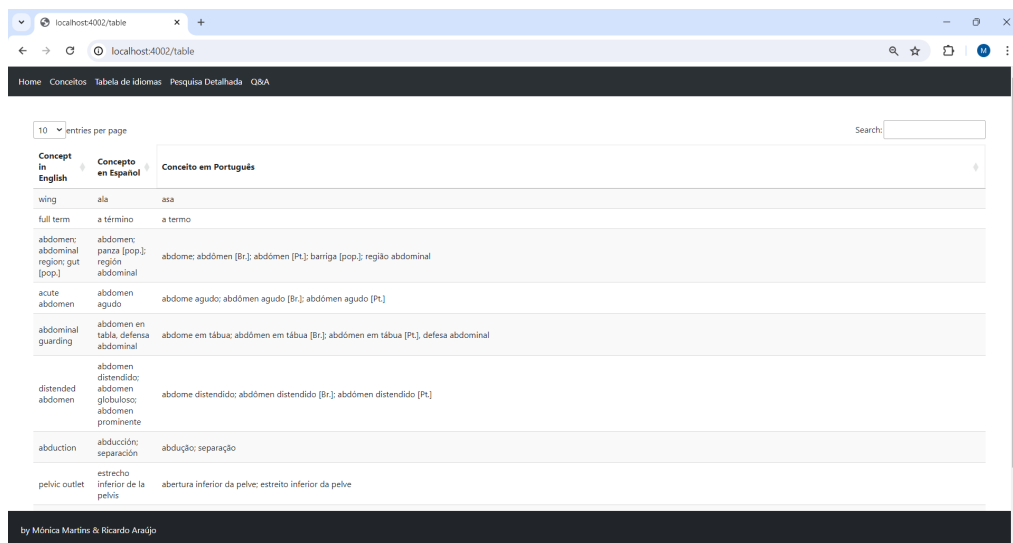


Fig. 9: Secção "Tabela de Idiomas".

Na Figura 10 está representada a página do conceito "reanimación cardiopulmonar".



Fig. 10: Página do conceito "reanimación cardiopulmonar".

Foram implementadas "setas" para que o utilizador pudesse avançar ou recuar nos termos. Como referido anteriormente, existe ainda a possibilidade de o utilizador mudar a página para outro idioma. Na Figura 11 mostra-se a tradução da página para alemão.

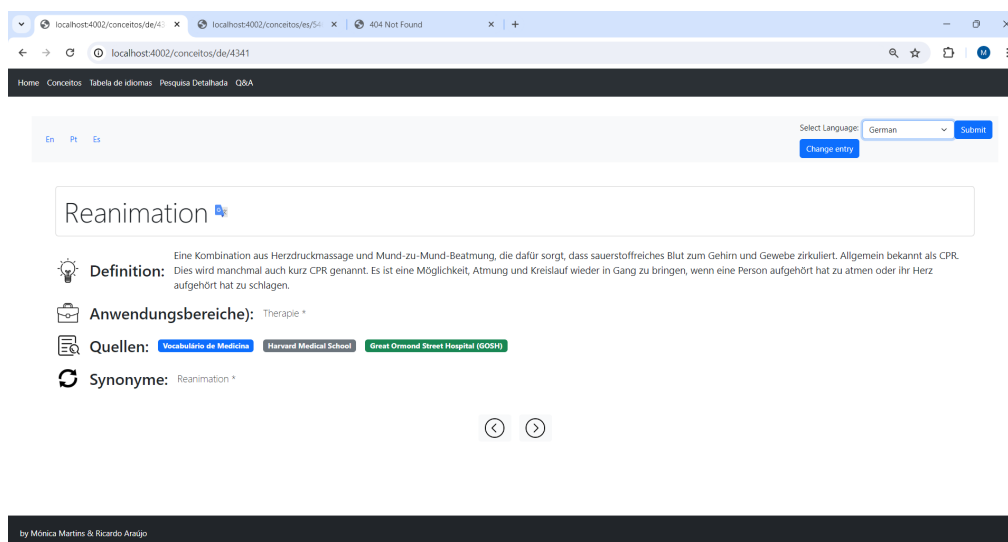


Fig. 11: Página do conceito "reanimación cardiopulmonar" em alemão.

Note-se que, no caso de existir um conceito na base de dados relacionado com o termo em questão, o utilizador pode navegar até à página desse "relacionado". Para além disso, o utilizador pode alterar a informação de um determinado termo, carregando no botão "change entry". Ao carregar, aparece um *pop-up* semelhante ao da Figura 12.

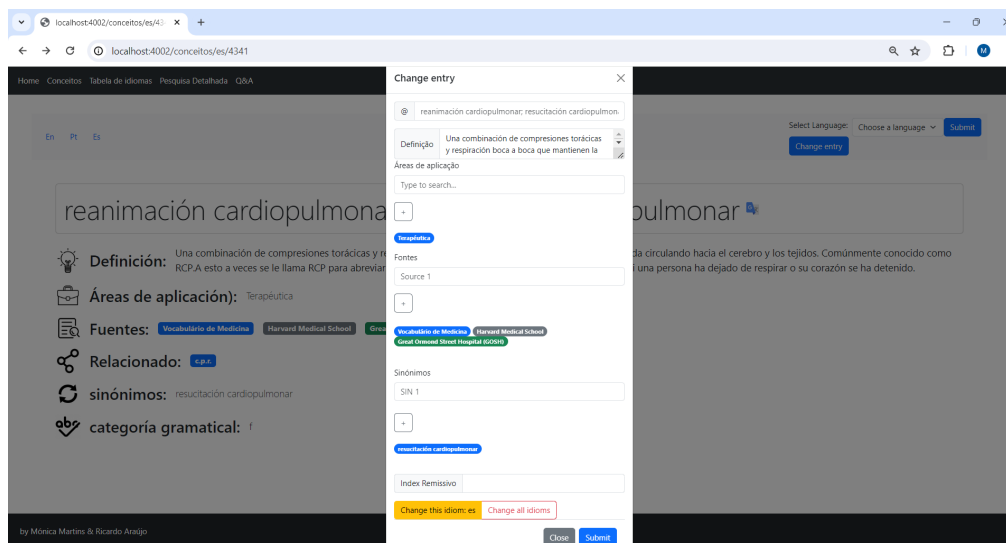


Fig. 12: Alteração da informação de um determinado termo.

Para facilitar a alteração, já é disponibilizada a informação atual do conceito, pelo que o utilizador só tem de modificar aquilo que pretende. Para além disso, é possível alterar o termo apenas no idioma selecionado, ou nos 3 idiomas (ou seja, ficheiros), com exceção das formas variantes, que devido à sua especificidade, apenas podem ser adicionadas a um documento de cada vez.

A secção "Pesquisa detalhada", permite, como o nome indica, fazer uma pesquisa avançada. Imagine-se um cenário em que se pretende procurar pelo termo "reanimación cardiopulmonar", apresentado anteriormente, na Figura 10. No entanto, temos algumas dúvidas em relação ao conteúdo. Lembrámo-nos que ou o conceito continha "pulmonar" ou que o descrição continha "combinación" ou que o termo estava em espanhol. Ora, devemos preencher a pesquisa da seguinte forma:

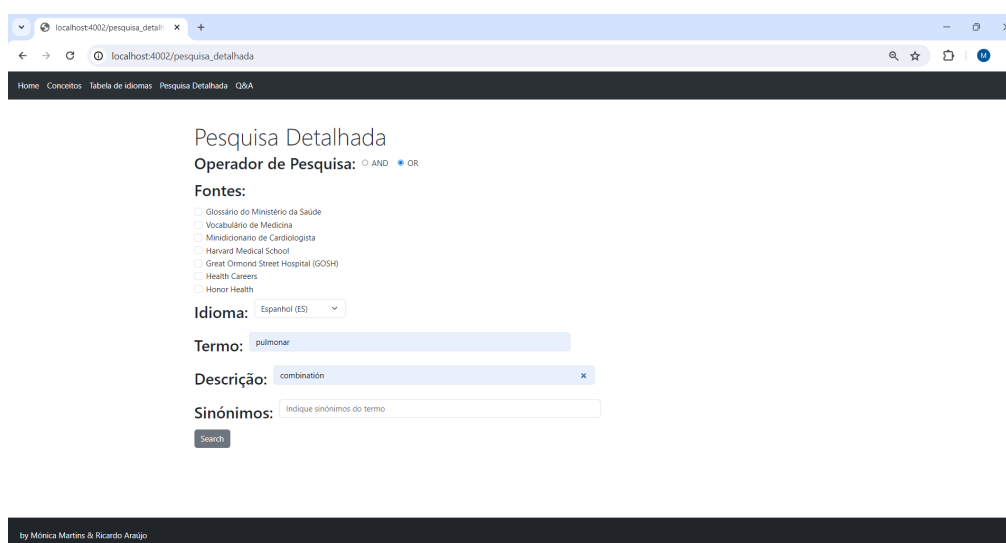
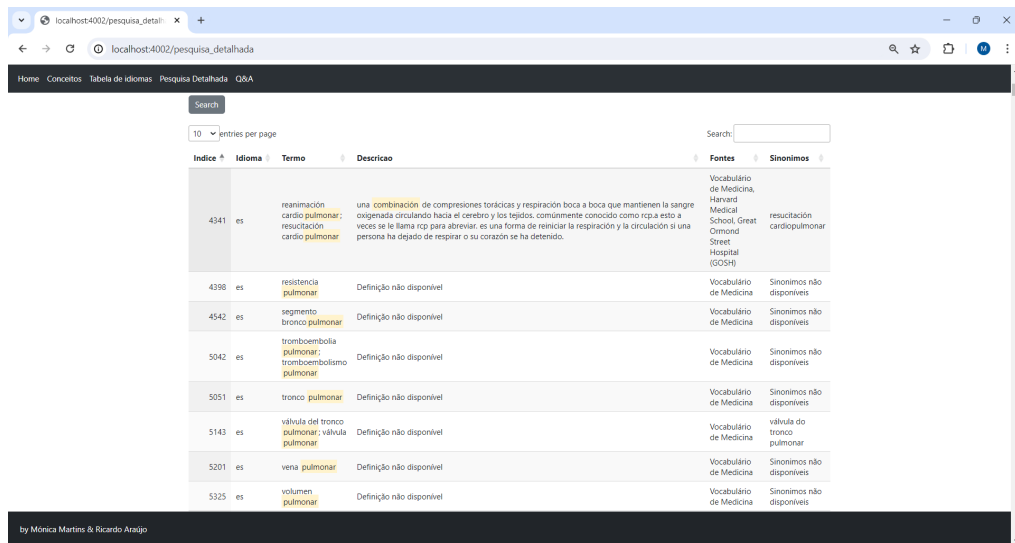


Fig. 13: Pesquisa detalhada.

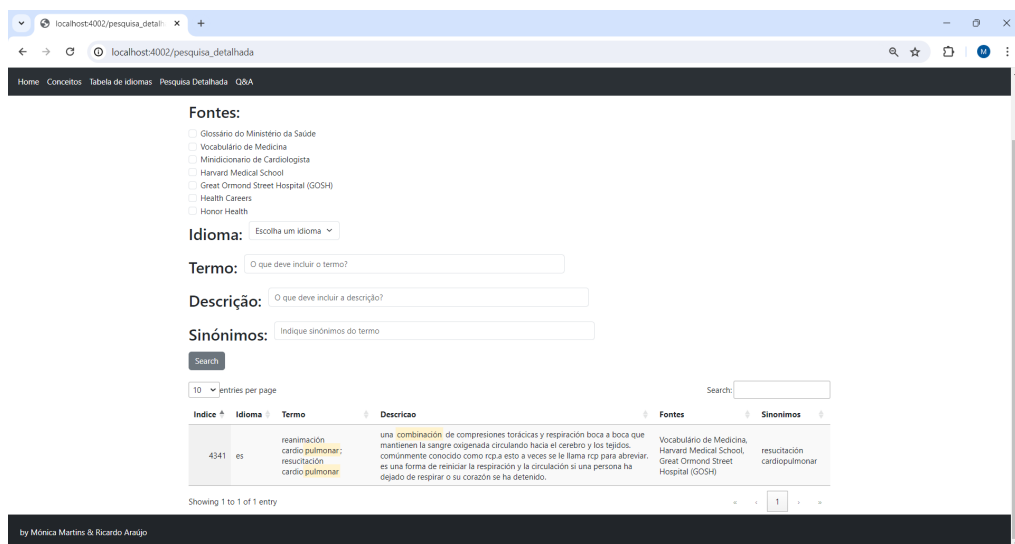
São-nos devolvidos 42 resultados, entre os quais, o pretendido, como mostra a Figura 14



Índice	Idioma	Termo	Descrição	Fontes	Sinónimos
4341	es	reanimación cardio pulmonar; resuscitación cardio pulmonar	una combinación de compresiones torácicas y respiración boca a boca que mantienen la sangre oxigenada circulando hacia el cerebro y los tejidos. comúnmente conocido como rcp a esto a veces se le llama rcp para abreviar. es una forma de reiniciar la respiración y la circulación si una persona ha dejado de respirar o su corazón se ha detenido.	Vocabulário de Medicina, Harvard Medical School, Great Ormond Street Hospital (GOSH)	resuscitación cardiopulmonar
4398	es	resistencia pulmonar	Definição não disponível	Vocabulário de Medicina	Sinónimos não disponíveis
4542	es	segmento bronco pulmonar	Definição não disponível	Vocabulário de Medicina	Sinónimos não disponíveis
5042	es	trombocitopenia pulmonar; tromboembolismo pulmonar	Definição não disponível	Vocabulário de Medicina	Sinónimos não disponíveis
5051	es	tronco pulmonar	Definição não disponível	Vocabulário de Medicina	Sinónimos não disponíveis
5143	es	válvula del tronco pulmonar; válvula pulmonar	Definição não disponível	Vocabulário de Medicina	válvula do tronco pulmonar
5201	es	vena pulmonar	Definição não disponível	Vocabulário de Medicina	Sinónimos não disponíveis
5325	es	volumen pulmonar	Definição não disponível	Vocabulário de Medicina	Sinónimos não disponíveis

Fig. 14: Resultados obtidos da pesquisa detalhada com o operador "OR".

Imagine-se, agora, outro cenário, no qual tínhamos a certeza do que foi referido em cima, ou seja, tínhamos a certeza que: o conceito continha "pulmonar" e que o descrição continha "combinación" e que o termo estava em espanhol. Nesse caso, utilizávamos o operador "AND". Neste caso já é devolvido apenas um resultado, como mostra a Figura 15.



Índice	Idioma	Termo	Descrição	Fontes	Sinónimos
4341	es	reanimación cardio pulmonar; resuscitación cardio pulmonar	una combinación de compresiones torácicas y respiración boca a boca que mantienen la sangre oxigenada circulando hacia el cerebro y los tejidos. comúnmente conocido como rcp a esto a veces se le llama rcp para abreviar. es una forma de reiniciar la respiración y la circulación si una persona ha dejado de respirar o su corazón se ha detenido.	Vocabulário de Medicina, Harvard Medical School, Great Ormond Street Hospital (GOSH)	resuscitación cardiopulmonar

Fig. 15: Resultados obtidos da pesquisa detalhada com o operador "AND".

Por fim, resta apenas demonstrar a utilização da secção "Q&A", que tem como objetivo, como já foi referido anteriormente, a resposta a perguntas, a identificação do termo mais semelhante a outro e identificação do termo que não se enquadra dentro de um conjunto dado.

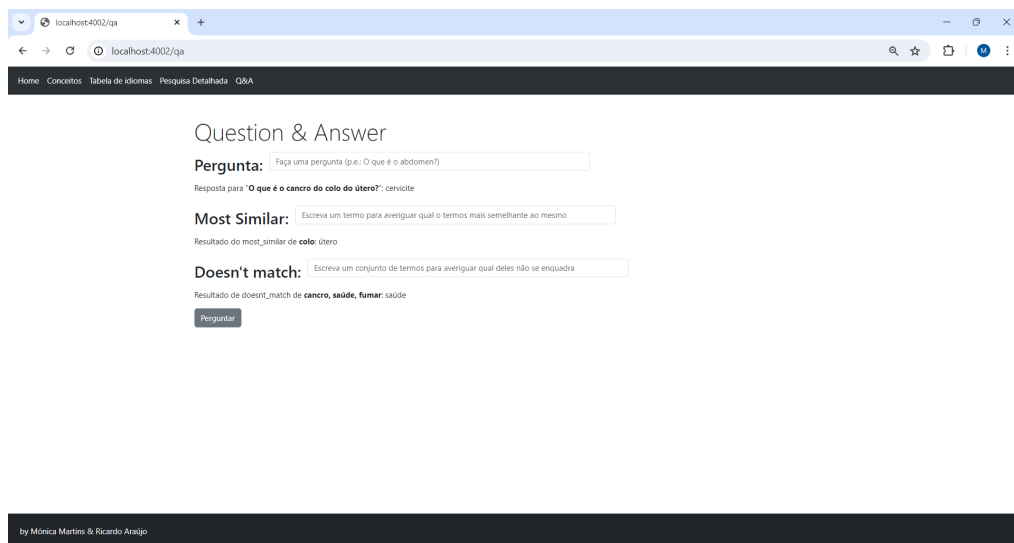


Fig. 16: Secção "Q&A".

4 Conclusão

O trabalho desenvolvido permitiu aplicar e aprofundar os conhecimentos adquiridos na Unidade Curricular de Processamento e Linguagem Natural em Engenharia Biomédica.

Considera-se que foram cumpridos os objetivos propostos inicialmente já que, foi possível enriquecer os dados anteriormente extraídos. Para além disso, criou-se uma interface gráfica que permite a visualização, manipulação e pesquisa dos dados.

As principais funcionalidades da aplicação desenvolvida consistem na visualização dos dados através de tabelas (com um sistema de pesquisa eficiente implementado) e a tradução automática da página do conceito para 24 idiomas diferentes. Para além disso, desenvolveu-se um sistema de pesquisa avançada que permite efetuar uniões ou interseções entre diferentes parâmetros de pesquisa.

Posto isto, o presente trabalho permitiu aprimorar a utilização de bibliotecas do Python como o *BeautifulSoup*, *deep_translator* e *Word2Vec*. Concomitantemente, permitiu também adquirir mais conhecimentos acerca das ferramentas Flask e Jinja, previamente estudadas.

Futuramente, poderiam ser incluídos mais dados acerca dos termos, nomeadamente a inclusão de grupos gramaticais (adjetivos, provérbios, nomes, etc). Para além disso, poderiam ser utilizadas mais fontes, de modo a incluir mais definições ou sinónimos, para os termos que não possuem esta informação.

Referências

- [1] Minidicionario Cardiologista, (2014). Sociedade brasileira de Cradiologia. Acedido em 10 jun. 2024. <https://pt.scribd.com/document/518804383/minidicionario-cardiologista-2014>
- [2] Vocabulario de Medicina, (2008). Servicio de Normalización Lingüística
- [3] Glossário do Ministério da Saúde, (2004). Ministério da Saúde
- [4] Honor Health: Glossary of Medical Terms: Common Procedures and Tests. Acedido em 10 jun. 2024. <https://www.honorhealth.com/patients-visitors/average-pricing/medical-glossary>
- [5] Health Careers (NHS): Glossary. Acedido em 10 jun. 2024. <https://www.healthcareers.nhs.uk/glossary>
- [6] Harvard Medical School: Medical Dictionary of Health Terms: A-C. Acedido em 10 jun. 2024. <https://www.health.harvard.edu/a-through-c>
- [7] Great Ormond Street Hospital for Children: Health Dictionary. Acedido em 10 jun. 2024. <https://www.gosh.nhs.uk/conditions-and-treatments/health-dictionary/>
- [8] Thesaurus Acedido em 10 jun. 2024. <https://www.thesaurus.com/>
- [9] Update em Medicina: Livro de Resumos. Acedido em 10 jun. 2024. <https://updatemedicina.pt/wp-content/uploads/2024/03/Livro-de-Resumos-2024.pdf>