Trabalho Prático 1

Processamento de Linguagem Natural em Engenharia Biomédica

Mestrado em Engenharia Biomédica - Informática Médica

2022/2023

Docentes:

Luís Filipe da Costa Cunha

José João Almeida

Diogo Guedes Lameira, PG50332

João Filipe Costa Alves, PG50471

José Miguel Moreira Santos, PG51190

Índice

Introdução

O presente relatório tem como objetivo fundamentar o primeiro trabalho prático realizado na unidade curricular de Processamento de Linguagem Natural em Engenharia Biomédica. De modo geral, o projeto passa pela extração de informação relevante de diferentes documentos médicos, fornecidos em formato PDF que é posteriormente preservada num ficheiro JSON.

De modo a manipular os textos e identificar correspondências de padrões, foram utilizadas expressões regulares. Estas permitem que a partir de um determinado padrão de pesquisa constituído por uma sequência de caracteres, possam ser efetuadas diversas operações como pesquisar, substituir e extrair texto a partir de grandes conjuntos de informação. É ainda possível efetuar a limpeza dos dados extraídos e remoção de elementos desnecessários como cabeçalhos, rodapés e números de página.

Assim, foram utilizados quatro documentos a partir dos quais foi efetuada a extração de informação, sendo eles o Dicionário de Termos Médicos Português-Inglês-Espanhol, o Glossário Médico, o CIH Bilingual Medical Glossary e o Dicionário de Termos Médicos e de Enfermagem. O primeiro documento constitui um dicionário de termos comuns na área médica com traduções em três línguas diferentes: português, inglês e espanhol. O segundo documento contém termos e a sua tradução em inglês. Quanto ao terceiro documento, este contém termos em inglês com a sua tradução em espanhol. Por fim, o quarto documento consiste num dicionário com definições para diversos termos médicos.

O objetivo do trabalho passou, portanto, por construir um ficheiro *JSON* com diversos termos médicos em português e a sua respetiva tradução em inglês e espanhol, assim como a sua definição. O procedimento passou por, em primeiro lugar, extrair os termos e respetiva tradução em inglês/espanhol do primeiro documento. O segundo passo foi extrair os termos e respetiva tradução do segundo e terceiro documentos e reunir a informação dos dois, ou seja, retirar apenas os termos presentes em ambos de modo a obter apenas aqueles que tivessem tradução em ambas as línguas. O terceiro passo na realização do trabalho prático foi a comparação da informação obtida no primeiro e segundo passo de modo a complementar as informações presentes em ambos. Por fim, o último passo foi acrescentar as descrições em comum contidas no quarto documento aos restantes, de modo a obter um ficheiro que reunisse toda a informação recolhida dos ficheiros selecionados e que, desta forma, contivesse apenas os termos que tivessem tradução em inglês e espanhol, assim como uma descrição. Para manusear os vários documentos efetuou-se a sua conversão de *PDF* para *XML* recorrendo ao comando `pdftohtml -c -xml` na *Bash*.

Documentos processados

Documento - Dicionário de Termos Médicos Português-Inglês-Espanhol

Neste documento está presente, para além do vocabulário técnico de medicina, também termos do dia-a-dia importantes, assim como expressões necessárias para diálogo com pacientes. Assim, este é constituído por três secções com traduções efetuadas em diferentes direções. Para a realização deste trabalho prático, utilizou-se unicamente a última secção do documento que é composta por termos em português e a sua respetiva tradução, em inglês e espanhol, com vista ao objetivo final. Assim, o processamento do ficheiro passou pelos seguintes passos:

- Leitura do ficheiro em formato *XML*.
- Limpeza das *tags <page>*, *<text>* e *<i>* e do seu respetivo conteúdo.
- Identificação do início da secção relevante para a extração de informação (dicionário português-inglês-espanhol) e remoção de todo o conteúdo anterior através de uma função *sub* e recorrendo às *flags re.DOTALL* e *re.MULTILINE*.
- Remoção de todas ocorrências de "português – inglês – espanhol" que surge nas laterais do documento *PDF*.
- Remoção do número da página e palavra que surge em destaque no cabeçalho para cada página do documento. Neste passo foi identificada uma exceção de palavra pertencente ao cabeçalho que teve de ser tratada separadamente às restantes por não seguir a mesma formatação (bloqueador-β).
- Remoção da linha com a *tag <fontspec>*.
- Remoção de cada letra do alfabeto que surge anteriormente à lista de termos começados por essa letra.
- Uniformizar os caracteres *newline* (/n) de modo a ficar com o mesmo espaçamento ao longo de todo o documento.
- Ao verificar a existência de termos que se encontravam separados em múltiplas linhas, começou-se por verificar no documento qual o máximo de linhas que um único termo ocupava. Assim, foram aplicadas uma série de expressões regulares para tornar os termos, que se encontram delimitados pela *tag <b>*, numa única linha. Para tal, foram utilizados vários grupos de captura para as várias situações presentes no documento, desde termos que ocupavam

desde duas linhas até um máximo que se verificou de seis linhas. Através destes grupos de captura, as várias linhas foram substituídas por uma só com os vários grupos nela, delimitados pela *tag*. Como resultado deste passo, obteve-se um documento com o termo numa única linha, de modo a poder ser extraído posteriormente.

- O mesmo raciocínio do passo anterior foi aplicado aqui de modo a apresentar toda a tradução em inglês do termo numa única linha. Assim, utilizaram-se grupos de captura para unificar o conteúdo presente entre as letras U e E que delimitam as traduções em inglês e espanhol, respetivamente.
- Neste passo foi efetuado o mesmo procedimento dos passos anteriores de modo a apresentar toda a tradução em espanhol numa única linha. O objetivo passa por obter a informação com a seguinte formatação:

termo
U
(tradução em inglês)
E
(tradução em espanhol)

- O passo seguinte foi juntar palavras que estavam separadas por "-", por estarem em linhas diferentes anteriormente, ou seja, remover este caracter quando este estava aplicado indevidamente. Para tal, utilizou-se uma expressão regular que juntasse partes de palavras que estavam separadas por este caracter e um espaço em branco de seguida, isto porque ao juntar as linhas quer para os termos, quer para as traduções, os grupos de captura foram separados por um espaço em branco.
- A etapa seguinte passou por tratar exceções que foram identificadas, como termos que não tinham ficado na mesma linha como pretendido pela diferente formatação (palato, cisto e bloqueador- β). Verificou-se ainda um termo que não apresentava traduções (ante-braço) e, por isso, foi removido. Finalmente, verificou-se dois termos que não apresentavam a *tag* final **, aos quais esta foi adicionada.
- Após esta sequência de passos, obteve-se o documento no formato pretendido para a extração da informação e para tal recorreu-se a funções *findall* para criar três listas diferentes com os termos, tradução em inglês e tradução em espanhol e construir um dicionário a partir destas listas, que foi posteriormente inserido num ficheiro *JSON*.

Documento - Dicionário de Termos Médicos e de Enfermagem

O documento além da explicação de milhares de termos da área da saúde, contém uma introdução sobre alimentação, hábitos saudáveis e outras temáticas que valorizam a qualidade de vida. Assim sendo e visto que o pretendido era um ficheiro *JSON* apenas com o termo e respetiva descrição, toda essa parte inicial desnecessária foi removida. De forma mais específica, o processamento do ficheiro é feito pelas seguintes etapas:

- Leitura do *'Dicionario_de_termos_medicos_e_de_enfermagem.xml'* e armazenamento do seu conteúdo na variável *lines* como uma lista;
- Criação de um novo arquivo chamado 'dicionario_simplificado.xml' e cópia de todo o conteúdo da lista *lines* a partir da 631ª linha, removendo desta forma as informações desnecessárias referidas acima;
- Leitura do arquivo 'dicionario_simplificado.xml' e armazenamento do seu conteúdo na variável *text;*
- Remoção do número da página que surge no rodapé e das inicias das palavras que surgem em destaque em cabeçalho para cada página do documento;
- Remoção das letras que surgem anteriormente à lista de termos começados por essa e que estão em itálico. Neste passo foi identificada uma exceção da letra *S* que teve de ser tratada separadamente às restantes pelo facto de existir uma letra *S* isolada mas que não servia de título de uma secção, mas sim parte da descrição de um termo;
- Limpeza através de expressões regulares das *tags <page>*, *<text> e <fontspec>*;
- Remoção do rodapé das páginas: "*Sou Enfermagem - Cadastre-se grátis em: https://souenfermagem.com.br*";
- Remoção das *tag <i>;*
- Uniformização dos caracteres *newline (/n)* de modo a ficar com o mesmo espaçamento ao longo de todo o documento.
- Substituição do "-" por " " de maneira a retirar o "-" que separava os termos das suas descrições;
- Junção das palavras que estavam separadas por "-" por estarem em linhas diferentes anteriormente, ou seja, remover este caracter quando este estava aplicado indevidamente. Para tal, utilizou-se uma expressão regular que juntasse partes de palavras que estavam separadas por este caracter e um parágrafo de seguida;

- Aplicação de uma uma série de expressões regulares para tornar os termos, que se encontram delimitados pela *tag *, numa única linha. Desta forma houve junção dos termos que estavam separadas por "-" por estarem em linhas diferentes anteriormente, bem como junção dos termos que estavam separados por "," *"\s"* e *"\("*. Desta forma juntou todos os termos numa única linha delimitada pelas *tags *;
- Dado que havia termos em negrito nas descrições dos termos, estas não podem ser esquecidas nem serem vistas como termos. Desta forma, foi utilizado um grupo de captura para que quando aparecesse uma frase delimitada pelas *tags * com letras minúsculas e com a possibilidade de ter até duas letras maiúsculas no início (visto que foi verificado exceções de frases que estavam a negrito mas que se iniciavam por letras maiúsculas mas faziam parte de descrições, esta removesse as *tags * e desta forma não haver confusão com os termos, que estes sim, têm de estar separados por essas *tags*;
- Remoção dos parágrafos para que algumas palavras que estavam separadas em linhas diferentes indevidamente, ficassem juntas e o texto desta forma ficasse seguido;
- Remoção de outra secção do cabeçalho;
- Através da função *findall* , extração dos termos médicos e as suas descrições usando expressões regulares para encontrar todas as ocorrências de texto entre as *tags * e **, juntamente com qualquer texto (descrição) que venha após estas até a próxima *tag *;
- Desta forma posterior criação de uma lista de listas em que cada sublista contém o termo médico, em minúsculas e a sua descrição limpa (sem espaços desnecessários);
- Conversão da lista de listas em um dicionário chamado *dicionario* onde as chaves são os termos médicos e os valores são suas descrições;
- Inserção do dicionário *dicionario* em um ficheiro JSON.

Posteriormente, após todo o processamento e criação das variáveis a trabalhar, foi executada a escrita do ficheiro intermédio JSON no formato "termo: descrição".

Documento - CIH Bilingual Medical Glossary English Spanish

O documento encontra-se elaborado através de tabelas sendo que na primeira coluna encontram-se os termos em inglês e na segunda coluna a respetiva tradução para espanhol, no formato *XML* as duas colunas encontram-se distanciadas por *tags <text>* vazias. Todas as partes desnecessárias

para o processamento encontram-se localizadas no início de cada *tag <page>*, desta forma, o processamento do ficheiro é feito pelas seguintes etapas:

- Remoção das *tags* iniciais desnecessárias.
- Remoção das *tags <text>* vazias entre cada termo-tradução.
- Remoção dos cabeçalhos do documento.
- Remoção da parte final do documento que contém "PREFIXES, ROOTS AND SUFFIXES".
- Remoção das *tags <i>* e *<b>.*
- Extração do texto necessário.
- Tratamento dos casos especiais.
- Alinhamento dos pares termo-tradução separadas por "\n".
- Extração dos grupos de captura referentes aos termos e às respetivas traduções.

Posteriormente, após todo o processamento e criação das variáveis a trabalhar, foi executada a escrita do ficheiro intermédio *JSON* no formato "termo: tradução"

Documento - Glossário Médico

O documento encontra-se elaborado por ordem alfabética, sendo que, dentro de cada letra encontramos o termo em inglês e a respetiva tradução para português. O documento no formato XML apresenta pouco texto desnecessário para o processamento, por outro lado, o texto necessário (termo-tradução) encontra-se elaborado dentro de *tags <text>*. Assim sendo, o processamento do comento no formato XML foi de encontro às seguintes etapas:

- Remoção das *tags* iniciais desnecessárias.
- Remoção das partes introdutórias de cada página que se encontravam dentro de *tags <i>* ou *<b>*, estas últimas dentro de *tags <text>*.
- Aquisição da informação do grupo de captura.

Ainda dentro do código foi realizado o processamento do texto adquirido, sendo que, havia ocorrências com termo e sem tradução, e outras em que o termo e a tradução foram separados por quebras, desta forma foram utilizados dois casos:

- Remoção dos casos em que o hífen era o último elemento.
- Junção dos termos e traduções separados por quebras.

Após a formulação do dicionário final foi escrito o ficheiro JSON no formato “termo: tradução”.

Junção de documentos

A junção dos documentos com traduções foi realizada sendo que, sempre que um dos termos de um ficheiro era equivalente a um dos termos encontrado noutro ficheiro era realizada a reunião da informação, desta forma, foi escrito um ficheiro *JSON* no formato:

Figura 1 - Formato da junção de documentos.

Junção final

O ficheiro proveniente do processamento do dicionário principal, ou seja, contendo termo em português e traduções em inglês e espanhol, foi atualizado com a junção dos documentos descritos anteriormente, desta forma, todas as ocorrências existentes na junção de documentos que não constavam no dicionário principal foram adicionadas.

Posteriormente, já com o dicionário intermédio com as definições dos termos em português, foi feita a interseção dos documentos para que o formato do ficheiro *JSON* final fosse:

Figura 2 - Formato do ficheiro final.

Através da utilização de uma expressão regular `\n\s\s\s\s".*"` foi possível perceber que o ficheiro final apresenta 1119 ocorrências que apresentam toda a informação completa.

Conclusão

O trabalho prático realizado na unidade curricular de Processamento de Linguagem Natural em Engenharia Biomédica permitiu a extração de informação relevante e limpeza de elementos desnecessários de diferentes documentos médicos, com a utilização de expressões regulares e manipulação de textos. O objetivo delineado que consistia na construção de um ficheiro JSON com diversos termos médicos em português e suas respetivas traduções em inglês e espanhol, assim como suas descrições foi alcançado.

Ao longo do trabalho foram surgindo muitas exceções aquando do processamento dos documentos, pelo que gerou dificuldades no manuseamento dos mesmos, assim sendo derivado da extensão dos documentos bem como das exceções, é possível que alguma informação não esteja corretamente manipulada e isso possa interferir com os resultados finais. Contudo, o objetivo do trabalho sendo um ficheiro *JSON* cujos termos são comuns nos quatro documentos faz com que possíveis erros sejam suprimidos do mesmo. Além disso, é possível que o código usado não seja o mais eficiente possível.

Para trabalho futuro, além da simplificação do código e consequente aumento da eficiência, seria possível também trabalhar ainda mais os documentos de modo a abranger todas as possíveis exceções existentes. Além disso, poder-se-ia utilizar mais documentos para analisar e, desta forma, ter um ficheiro *JSON* com mais termos traduzidos, mais designações e traduções. Bem como complementar a tradução dos termos utilizando outras ferramentas externas, como o *Google Tradutor*.

Por fim, destaca-se a importância do processamento de linguagem natural na área biomédica, pois permite o acesso mais rápido e preciso a informações relevantes para a tomada de decisões clínicas, contribuindo assim para a melhoria da qualidade do atendimento médico e, consequentemente, da saúde da população.