



UNIVERSIDADE DO MINHO

PROCESSAMENTO DE LINGUAGEM NATURAL EM ENGENHARIA
BIOMÉDICA

Trabalho Prático Nº 1

Autores:

Ana Carolina Costa, A95694

Ana Margarida Sousa, A96095

Bernardo Moraes, PG53700

6 de abril de 2024

Conteúdo

1	Introdução	1
2	Pré-requisitos	1
3	Seleção de Ficheiros	1
3.1	Glossário do Ministério da Saúde	1
3.1.1	Introdução:	1
3.1.2	Estrutura de armazenamento definida:	2
3.1.3	Fase 1: Conversão	2
3.1.4	Fase 2: Divisão	2
3.1.5	Fase 3: Limpeza	3
3.1.6	Fase 4: Extração e Construção do dicionário final	3
3.2	Anatomia na Prática - Sistema Musculoesquelético	5
3.2.1	Introdução:	5
3.2.2	Estrutura de armazenamento definida:	6
3.2.3	Fase 1: Conversão	7
3.2.4	Fase 2: Limpeza	7
3.2.5	Fase 3: Correção de anomalias	8
3.2.6	Fase 4: Sintaxe de extração definida	10
3.2.7	Fase 5: Construção do dicionário final	11
3.3	Minidicionário de Cardiologista	13
3.3.1	Introdução:	13
3.3.2	Estrutura de armazenamento definida:	14
3.3.3	Fase 1: Conversão	14
3.3.4	Fase 2: Limpeza e Extração	14
3.3.5	Fase 3: Correção de anomalias	15
3.3.6	Fase 4: Construção do dicionário final	16
4	Conclusão	17

1 Introdução

O presente relatório tem como objetivo descrever todo o processo implementado durante a resolução do primeiro trabalho prático da UC de Processamento de Linguagem Natural.

O trabalho em causa foi proposto de maneira a serem aplicados, e aprimorados, todos os conhecimentos adquiridos durante as aulas no que se refere ao processamento de documentos PDF.

Com isto, pretende-se extrair a informação com maior relevância, presente nos documentos, de uma forma conveniente, de maneira a assegurar a sua possível utilização em projetos futuros. Para tal, esta informação deve ser armazenada em ficheiros de formato JSON.

Assim sendo, numa primeira parte, serão introduzidos os ficheiros selecionados para processamento, bem como a estrutura definida para os ficheiros JSON finais de cada um. Por forma a alcançar este objetivo final, serão também relatadas todas as etapas de manipulação, envolvendo expressões regulares, de maneira a assegurar uma extração bem sucedida da informação mais relevante.

2 Pré-requisitos

Antes de se proceder ao desenvolvimento do projeto em si, certos pré-requisitos, a considerar, consistem em:

- Realização do processamento de **pelo menos 3 ficheiros PDF**;
- Processamento **obrigatório** do ficheiro correspondente ao **Glossário do Ministério da Saúde**;
- Desenvolvimento do projeto em linguagem **Python**.

3 Seleção de Ficheiros

Tal como foi referido anteriormente, para a realização do projeto, devem ser processados, pelo menos, 3 documentos PDF. Tendo isto em conta, e analisando os ficheiros fornecidos para seleção, o grupo decidiu processar os seguintes documentos:

- **Glossário do Ministério da Saúde**, obrigatório;
- **Anatomia na Prática - Sistema Musculoesquelético**;
- **Minidicionário de Cardiologista**;

3.1 Glossário do Ministério da Saúde

3.1.1 Introdução:

O documento intitulado “Glossário Médico do Ministério da Saúde” é um documento elaborado pelo Ministério da Saúde do Brasil, no ano de 2004, com o objetivo de criar um glossário de vocabulário controlado e de qualidade.

No que se refere ao conteúdo do documento, e excluindo os segmentos referentes ao Sumário, Apresentação, Introdução, Bibliografia e VCMS, foi possível identificar quatro secções com estruturas diferentes:

- **Siglas**: secção com a lista de siglas utilizadas ao longo do documento e o seu significado;

-
- Glossário: secção majoritária do documento, com os termos e o seu significado, bem como a, ou as, categorias às quais pertence.
 - Áreas temáticas: secção com todas as categorias referidas no glossário, bem como a sua definição.
 - Termos organizados por categoria: nesta secção, como o nome indica, para cada categoria existe a lista de termos que nela estão incluídos.

3.1.2 Estrutura de armazenamento definida:

Por forma a guardar a informação que seria extraída, posteriormente, foram definidos três ficheiros JSON com a seguinte estrutura:

```
siglas = {"Sigla": "Significado"}

categorias = {"Categoria":{

    "Descrição": Descrição da categoria,

    "Termos": [Lista de termos]}

}

glossario = {"Termo":{

    "Categoria": [Lista de categorias],

    "Definição": Definição do termo}

}
```

3.1.3 Fase 1: Conversão

De modo a dar início ao processamento e extração da informação do documento, foi essencial convertê-lo de formato pdf para formato xml. Para tal, foi utilizado o comando **pdftohtml -xml -i**. Este formato foi selecionado devido à sua maior retenção de informação acerca da estrutura do documento, comparativamente ao formato txt.

3.1.4 Fase 2: Divisão

Como referido anteriormente, as secções referentes ao Sumário, Apresentação, Introdução, Bibliografia e VCMS foram consideradas desnecessárias para a produção dos ficheiros JSON, acima descritos, e, como tal, foram eliminados, manualmente, do ficheiro xml.

Após esta exclusão, os segmentos do ficheiro xml, que correspondem às secções descritas anteriormente, foram divididos em 4 ficheiros xml distintos: **glossario**, **misc**, **categorias** e **siglas**. Esta divisão adveio do desejo de simplificar o processamento, uma vez que o uso de expressões regulares para processar uma zona do documento poderia prejudicar o processamento de outras.

3.1.5 Fase 3: Limpeza

De modo geral, foi necessário utilizar expressões regulares para eliminar as tags **text**, **page**, **image**, **itálico** e **fontspec** dos quatro ficheiros, bem como o número da página.

Para além disso, e especificamente no documento referente ao glossário, foi necessário identificar as expressões regulares que permitem eliminar:

- A indicação do primeiro e último termo presentes no cabeçalho de cada página;

```
1 texto = re.sub(r"(</?text.*?></text>\n){2,4}</page>", r"", texto)
```

- A letra do alfabeto à qual se refere cada nova secção do dicionário:

```
1 texto = re.sub(r"</?text.*?>\w</text>", r"", texto)
```

Por outro lado, no que se refere ao ficheiro misc, os termos que ocupavam mais do que uma linha foram identificados pelo seu alinhamento, parâmetro **left** da tag **text**, associado a 4 valores diferentes. Como tal, foi fundamental processar estas situações antes da eliminação das tags text, e proceder à junção dos dois parágrafos que contêm o termo completo.

```
1 texto = re.sub(r'\n<text .*? left="500" .*?>', r'', texto)
2 texto = re.sub(r'\n<text .*? left="138" .*?>', r'', texto)
3 texto = re.sub(r'\n<text .*? left="193" .*?>', r'', texto)
4 texto = re.sub(r'\n<text .*? left="444" .*?>', r'', texto)
```

Além disso, o elemento **new line**, entre cada termo de uma mesma categoria, foi substituído por uma vírgula.

```
1 texto = re.sub(r"([>])\n([<])", r"\1,\2", texto)
```

Por fim, foi necessário proceder à eliminação de duplos espaços, bem como situações de mudança de linha a meio de uma palavra.

```
1 texto = re.sub(r"- ", r"", texto)
2 texto = re.sub(r" ", r" ", texto)
```

3.1.6 Fase 4: Extração e Construção do dicionário final

Ficheiro siglas:

Após a limpeza do documento, a informação restante está limitada a cada sigla, inserida numa tag **bold**, seguida do seu significado. Como tal, para extrair a sigla e o seu significado, foi utilizada a expressão:

```
1 <b>(.*?)</b>([<]+)
```

O tuplo resultante foi convertido num dicionário, com a estrutura definida previamente. Este dicionário foi guardado no ficheiro JSON “siglas.json”.

```
{
  "AB": "Atenção Básica ",
  "ABEn": "Associação Brasileira de Enfermagem ",
  "ADT ": "Assistência Domiciliar Terapêutica ",
  "AFE ": "Autorização de Funcionamento de Empresa ",
  "AIDPI": "Atenção Integrada às Doenças Prevalentes na Infância ",
```

Figura 1: Dicionário siglas

Ficheiro misc:

De forma semelhante, o ficheiro misc, após a limpeza, possuía apenas o nome de cada categoria rodeado pela tag **bold**, b, e a lista de termos que contém, pelo que foi utilizada a mesma expressão para extrair a informação.

Para além disso, foi utilizado um ciclo **for** para construir o dicionário cujas chaves correspondem às categorias, primeiro elemento do tuplo resultante do findall. Por sua vez, os valores destas chaves correspondem às listas resultantes da utilização do método split, aplicado ao segundo elemento do tuplo, pelo separador “,”. Este dicionário irá ser utilizado como auxiliar na construção de um outro, numa fase posterior, não tendo sido, portanto, mencionado anteriormente.

O dicionário produzido foi guardado no ficheiro “misc.json”.

Ficheiro categorias:

Novamente, após a etapa anterior, o ficheiro categorias continha apenas o nome de cada categoria rodeado pela tag bold, b, e a sua definição, pelo que foi utilizada a expressão já mencionada para extrair a informação.

No entanto, para construir o dicionário com a estrutura desejada, foi necessário carregar o dicionário criado a partir do ficheiro misc, designado termos_categoria. Deste modo, e utilizando um ciclo for, fez-se a correspondência entre cada categoria, primeiro elemento do tuplo resultante da função findall, e um novo dicionário. Este dicionário tem como chaves as palavras “Definição” e “Termos”, cujos valores são, respetivamente, o segundo elemento do tuplo e o valor, com a mesma chave, do dicionário termos_categoria.

```
"Ambiente e Saúde": {
  "Descrição": " Refere-se ao estudo das interações entre os seres vivos e o meio, dedica-se a analisar as formas de vida, substâncias agressivas e condições adequadas ou inadequadas, produzidas pela ação humana, que podem exercer alguma influência sobre a sua saúde e sobre o meio em que vive. Inclui subtemas como: águas de abastecimento para consumo humano, águas residuais, resíduos sólidos, controle ambiental e poluição, desastres naturais, emergências ambientais, legislação e direito ambiental, educação ambiental, política, planeamento e gestão ambiental, qualidade ambiental (do ar, da água, do solo), saneamento ambiental, ecologia sanitária, saúde e trabalho, economia e meio ambiente, desenvolvimento sustentável, gestão de riscos e de impactos ambientais, indicadores de contaminação, psicologia ambiental, efeitos sobre o consumo e exposição a produtos tecnológicos que tragam danos à saúde, agentes tóxicos, luz fluorescentes, eletricidade estática, computador, telefonia celular, torres eletromagnéticas, efeito estufa, cidades saudáveis, entornos saudáveis, etc. ",
  "Termos": [
    "Benzeno ",
    " Chumbo ",
    " Controle químico ",
    " Desinfetante ",
    " Detergente ",
    " Estudos ecológicos ",
    " Explosão demográfica ",
    " Inseticidas piretóides ",
    " Riscos ocupacionais ",
    " Ruído ",
    " Salubridade ambiental ",
    " Solventes orgânicos"
  ]
},
```

Figura 2: Dicionário categorias

Ficheiro glossario:

Antes da limpeza do ficheiro, o documento foi analisado com o objetivo de encontrar padrões para facilitar a extração da informação, e apesar das designações estarem rodeadas pela tag bold, b, não existe distinção entre o texto referente às categorias e a definição do termo.

Como tal, após a limpeza, foram iteradas as chaves do dicionário construído a partir do ficheiro categorias, e em todas as ocorrências das mesmas, no ficheiro glossário, foi adicionado o elemento @ no início e no fim da categoria.

No entanto, a categoria “Doenças” não possui correspondência direta com as chaves do dicionário, uma vez que, e como é explicado no próprio documento, esta categoria agrega doenças para além das “Doenças Crônicas e Degenerativas” e “Doenças Infeciosas e Parasitárias”. Deste modo, foi necessário que a sua identificação e marcação fosse feita utilizando duas expressões regulares específicas.

```

1 texto = re.sub("Categoria: Doenças", "Categoria: @Doenças@", texto)
2 texto = re.sub("@ Doenças", "@ @Doenças@", texto)

```

Após a marcação das categorias, procedeu-se à marcação dos casos em que existiam mais de uma categoria, num mesmo termo. Para tal, substitui-se a ocorrências de dois “@” seguidos, que representam o fim de uma categoria e início de outra, por um elemento “&”.

Após estas adições foi possível utilizar a seguinte expressão regular:

```

1 <b>([\^<]+?)</b> Categoria: @(.+?)@([\^<]+)

```

Esta permite extrair os conceitos, com o primeiro elemento do tuplo a corresponder ao termo, o segundo à categoria ou categorias, e o último à definição do conceito.

Para a construção do dicionário desejado, `dic_termos`, foi utilizado um ciclo `for`, de forma a iterar pelas instâncias extraídas pela função `findall`, para fazer a correspondência entre cada valor do primeiro elemento do tuplo, e novo dicionário, cujas chaves são “Categoria” e “Definição”. Os valores que correspondem a estas chaves, são respetivamente, a lista resultante do método `split`, pelo separador “&”, aplicado ao segundo elemento do tuplo e o último elemento deste.

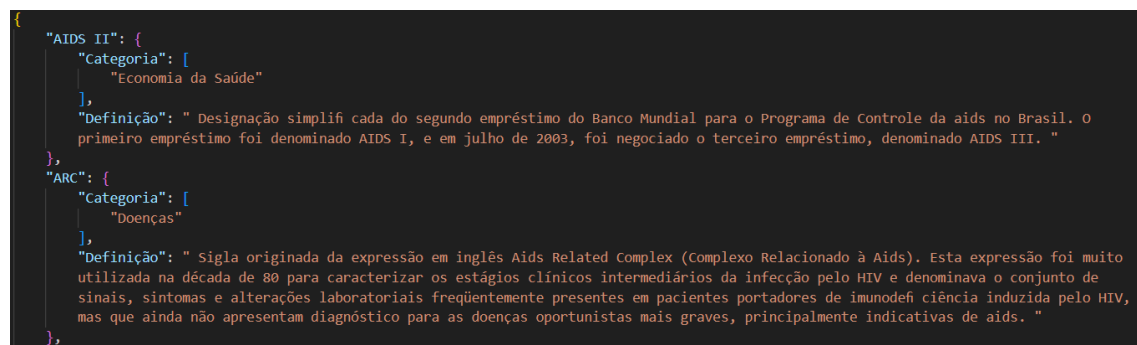
Por outro lado, também se verificou que nem todos os termos possuem categoria e descrição, alguns deles apenas fazem referência a um outro termo, por exemplo, **Didanosina Ver DDI**. Como tal, foi necessário definir uma expressão regular, para a extração da informação, distinta da utilizada para os termos com estrutura comum:

```

1 <b>([\^<]+)</b> (Ver [\^<]+)

```

Estes conceitos foram adicionados ao dicionário `dic_termos`, utilizando novamente um ciclo `for`, que permitiu criar uma entrada, cuja chave é o primeiro elemento do tuplo e o valor o segundo. Por fim, este dicionário foi ordenado de forma que as suas chaves estivessem organizadas alfabeticamente.



```

{
  "AIDS II": {
    "Categoria": [
      "Economia da Saúde"
    ],
    "Definição": " Designação simplifi cada do segundo empréstimo do Banco Mundial para o Programa de Controle da aids no Brasil. O primeiro empréstimo foi denominado AIDS I, e em julho de 2003, foi negociado o terceiro empréstimo, denominado AIDS III. "
  },
  "ARC": {
    "Categoria": [
      "Doenças"
    ],
    "Definição": " Sigla originada da expressão em inglês Aids Related Complex (Complexo Relacionado à Aids). Esta expressão foi muito utilizada na década de 80 para caracterizar os estágios clínicos intermediários da infecção pelo HIV e denominava o conjunto de sinais, sintomas e alterações laboratoriais frequentemente presentes em pacientes portadores de imunodeficiência induzida pelo HIV, mas que ainda não apresentam diagnóstico para as doenças oportunistas mais graves, principalmente indicativas de aids. "
  },
}

```

Figura 3: Dicionário glossario

3.2 Anatomia na Prática - Sistema Musculoesquelético

3.2.1 Introdução:

Tal como o nome indica, o ficheiro em causa trata-se de um documento que permite colocar em prática o conhecimento da anatomia humana, mais especificamente do sistema musculoesquelético, de estudantes de Medicina.

Em termos de conteúdo, numa parte inicial, é possível averiguar a presença de uma breve descrição dos autores, e tema do documento, bem como um índice/sumário em que é estabelecida uma enumeração das principais secções do documento.

Tendo em conta que o foco do ficheiro se trata do auxílio ao estudo da anatomia humana, verifica-se que a maioria deste é composto por imagens, em conjunto com alíneas, sendo estas usadas

para identificar determinadas partes do corpo humano. Assim, o aluno, durante a resolução dos exercícios, para cada alínea, toma nota da nomenclatura do componente anatómico em causa, podendo dirigir-se à secção final do documento para comparar as suas respostas.

É ainda importante realçar a divisão do documento em diversas secções e subsecções:

- Mais especificamente, verificam-se **dois grandes grupos**: o **Sistema Esquelético e Articular** e o **Sistema Muscular**.
- Seguidamente, para cada sistema global, averiguam-se várias secções devidamente numeradas, como **1. Crânio**, **2. Membro Superior**, **3. Membro Inferior**, etc.
- Por sua vez, cada secção em si, verifica múltiplas subsecções. Por exemplo, relativamente à secção **1. Crânio**, surgem as subsecções **1.1 Crânio: Vista Anterior - I**, **1.2 Crânio: Vista Anterior - II**, etc. Consequentemente, é cada subsecção que apresenta as mais diversas nomenclaturas associadas.
- No entanto, é de capital importância denotar que algumas subsecções poderão apresentar, também, outras secções, como é o caso de **4.4 Vértex Cervicais Atípicas: Áxis (C2) - Vista Pósterio-Superior** e **4.4.1 Vista Lateral**, sendo que é nesta última que se averiguam as diferentes terminologias.

1.1 CRÂNIO: VISTA ANTERIOR - I

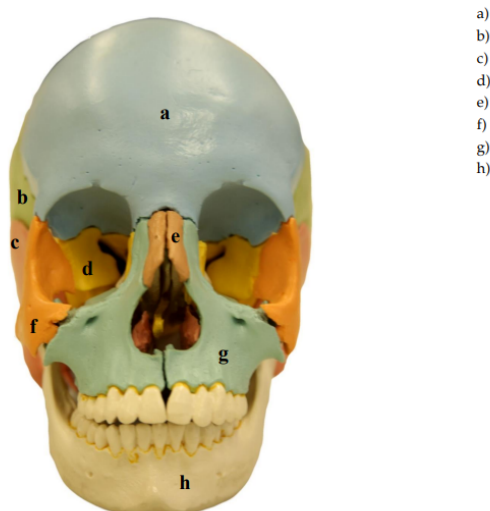


Figura 4: Exemplo de uma das páginas do documento

3.2.2 Estrutura de armazenamento definida:

Posto isto, uma vez que a informação mais relevante, no que diz respeito a este documento, se trata da nomenclatura dos mais diversos elementos anatómicos, decidiu-se implementar um dicionário, posteriormente armazenado no ficheiro JSON, com a seguinte estrutura:

dicionário =

{

 " SISTEMA ESQUELÉTICO E ARTICULAR ": { " SECÇÃO 01 ":

 { " SUBSECÇÃO 01 ": [terminologia1, terminologia2, terminologia3, ...],

```

    "SUBSECÇÃO2": {"SUB_SUBSECÇÃO1": [terminologia1, terminologia2, ...]}}
  }

  "SISTEMA MUSCULAR": {"SECÇÃO1":
    {"SUBSECÇÃO1": [terminologia1, terminologia2, terminologia3, ...],
    "SUBSECÇÃO2": {"SUB_SUBSECÇÃO1": [terminologia1, terminologia2, ...]}}}
}

```

Assim, basta apenas procurar pela subsecção desejada para averiguar as diferentes terminologias associadas.

3.2.3 Fase 1: Conversão

De maneira a dar início ao processo de preparação e limpeza do documento, por forma a dar origem ao ficheiro JSON final, o mesmo foi submetido a uma conversão para o formato **xml**.

Esta decisão reside no facto de a linguagem xml permitir uma maior facilidade de identificação de padrões uma vez que organiza os dados numa estrutura hierárquica, através do uso de tags, que, por sua vez, apresentam nomes descritivos, que auxiliam o ser humano na compreensão do conteúdo.

3.2.4 Fase 2: Limpeza

Eliminação Manual: Tal como foi referido anteriormente, a informação relevante, a ser extraída, diz respeito às respostas, presentes na secção final do documento, referentes às diferentes nomenclaturas dos componentes.

Esta secção diz respeito à secção **Gabarito**, que se inicia a partir da página 192 do documento. Posto isto, após abertura do documento, já em formato xml, no editor de texto, qualquer conteúdo presente acima desta secção foi eliminado.

Expressões Regulares:

No que diz respeito às expressões regulares, estas assumiram um papel bastante importante na fase de limpeza, uma vez que permitiram eliminar padrões específicos, que se repetiam continuamente ao longo do documento.

Um exemplo prático desta situação consiste no menu vertical, vigente em todas as páginas, pelo que se revelou necessário utilizar a função **sub** das RegEx, de maneira a proceder à sua eliminação.

A identificação do seu respetivo padrão não foi difícil, uma vez que constituíam as únicas tags de texto com fontes correspondentes a 4 e a 5.

```

<page number="215" position="absolute" top="0" left="0" height="892" width="1262">
<text top="410" left="1227" width="0" height="22" font="4"><b>A</b></text>
<text top="403" left="1227" width="0" height="16" font="5"><b>n</b></text>
<text top="395" left="1227" width="0" height="16" font="5"><b>At</b></text>
<text top="381" left="1227" width="0" height="16" font="5"><b>omi</b></text>
<text top="358" left="1227" width="0" height="16" font="5"><b>A</b></text>
<text top="344" left="1227" width="0" height="22" font="4"><b> </b></text>
<text top="346" left="1227" width="0" height="16" font="5"><b>n</b></text>
<text top="337" left="1227" width="0" height="16" font="5"><b>A</b></text>
<text top="323" left="1227" width="0" height="22" font="4"><b> </b></text>
<text top="325" left="1227" width="0" height="16" font="5"><b>pr</b></text>
<text top="310" left="1227" width="0" height="16" font="5"><b>átic</b></text>
<text top="285" left="1227" width="0" height="16" font="5"><b>A</b></text>
<text top="271" left="1227" width="0" height="22" font="4"><b> S</b></text>
<text top="259" left="1227" width="0" height="16" font="5"><b>is</b></text>
<text top="248" left="1227" width="0" height="16" font="5"><b>tem</b></text>
<text top="224" left="1227" width="0" height="16" font="5"><b>A</b></text>
<text top="210" left="1227" width="0" height="22" font="4"><b> m</b></text>
<text top="197" left="1227" width="0" height="16" font="5"><b>u</b></text>
<text top="188" left="1227" width="0" height="16" font="5"><b>Scul</b></text>
<text top="159" left="1227" width="0" height="16" font="5"><b>oe</b></text>
<text top="144" left="1227" width="0" height="16" font="5"><b>Squelé</b></text>
<text top="99" left="1227" width="0" height="16" font="5"><b>tic</b></text>
<text top="82" left="1227" width="0" height="16" font="5"><b>o</b></text>

```

Figura 5: Menu Vertical - identificação

Para além disso, as expressões regulares foram também utilizadas para descartar outro tipo de conteúdo, como:

- As **page tags**;

```
1 text = re.sub(r"</page>\n<page.+>(\n)+", r"", text)
```

- As tags correspondentes a **elementos âncora**;

```
1 text = re.sub(r".+<a.+>\n.+>", r"", text)
```

- Os números, delimitados pelas tags bold, uma vez que se tratavam dos **números das páginas**;

```
1 text = re.sub(r".+<b>\d+</b>.+>\n+", r"", text)
```

- As **text tags**, juntamente com os mais diversos parâmetros, isto é font, width, height, etc.

```
1 text = re.sub(r"</?text.*?>", r"", text)
```

3.2.5 Fase 3: Correção de anomalias

Subsecções:

Uma vez obtida uma melhor identificação do conteúdo a ser extraído, foi importante proceder a uma análise de quaisquer anomalias a serem corrigidas.

Por exemplo, enquanto que determinadas subsecções se encontravam delimitadas pelas tags bold, numa única linha, outras, ou não se encontravam a bold, ou, então, a sua designação continuava numa linha posterior:

```

<b>4.6 VÉRTEBRAS CERVICAIS ATÍPICAS: VÉRTEBRA PROEMINENTE </b>
<b>(C7)</b>

```

Figura 6: Subsecções anómalas - continuação da descrição numa nova linha

```

<b>2.14 FALANGE E META</b>
<b>CARPO</b>
<b>: VISTA PALMAR</b>

```

Figura 7: Subsecções anómalas - continuação da descrição numa nova linha

```
3.2
<b>MÚSCULOS DO MEMBRO INFERIOR ESQUERDO: VISTA ANTERIOR</b>
```

Figura 8: Subsecções anómalas

É possível concluir, portanto, que o padrão, normal, identificador das **subsecções** corresponde a:

` n.n Designação ` sendo n um número inteiro

Por último, constatou-se a presença de mais uma anomalia, no que se refere às subsecções, mais concretamente das subsecções **1.25** e **1.26**, pertencentes ao Sistema Esquelético e Articular, dado que não apresentam quaisquer nomenclaturas associadas:

```
1.25 E 1.26 NÃO POSSUEM GABARITO
1.27 MANDÍBULA: VISTA ANTERIOR
a) Processo alveolar
b) Tubérculo mental
```

Figura 9: Subsecções sem terminologias

A decisão final consistiu na sua simples remoção, através da seguinte expressão regular:

```
1 text = re.sub(r"\n<b>1.25.+</b>", r"", text)
```

Terminologias:

No que se refere às terminologias dos respetivos elementos anatómicos, foi possível concluir que estas assumem a seguinte estrutura:

x) Nome do componente

No entanto, é importante notar que x pode corresponder a:

- uma letra **minúscula**, ou **maiúscula**, seguida de um **parênteses: a)**;
- uma letra **minúscula**, seguida de um **espaço** e **parênteses: f)**;
- um **par letra número**, seguido de **parênteses**, como por exemplo: **d1)**.

Em termos de anomalias, foi possível observar que uma das terminologias não apresentava um parênteses a seguir à respetiva letra:

```
i) Ligamento talofibular anterior
j Ligamento talonavicular dorsal
```

Figura 10: Terminologia anómala

Para tal, foi utilizada uma expressão regular específica por forma a gerar a estrutura desejada:

```
1 text = re.sub(r"\nj\s", r"\nj) ", text)
```

Seguidamente, averiguou-se a **presença de múltiplas terminologias numa única linha**:

```
j) Músculo Zigomático Maior k) Músculo Levantador do Lábio Superior
```

Figura 11: Terminologias numa única linha

Na imagem acima, as alíneas são constituídas por uma letra, seguida de um parênteses, contudo, foi também possível averiguar o mesmo caso para alíneas de formato letra+número, seguidos de parênteses, como também para alíneas de formato letra+espaço, seguidos de parênteses.

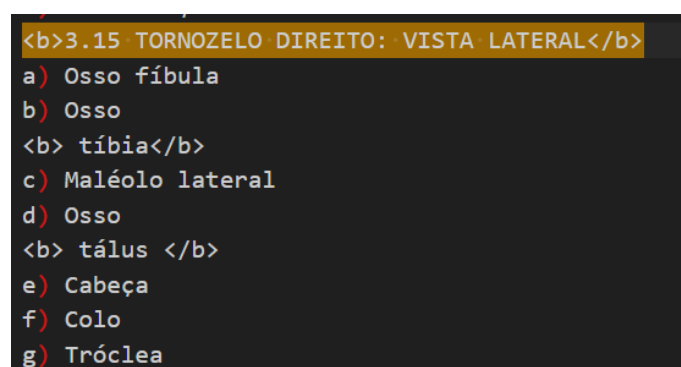
Por forma a contornar esta situação, foram utilizadas expressões regulares, para cada caso específico, uma vez que não foi possível estabelecer uma generalização. Tal decisão pode ser justificada pelo facto de determinadas alíneas poderem ser antecipadas, ou por letras, ou por espaços, o que complicava o processo de generalização.

Abaixo encontram-se listadas as expressões regulares que permitiram contornar estas situações:

```
1 text = re.sub(r"\s([a-z]\))", r"\n1", text) #alíneas letra+parênteses
2 text = re.sub(r"([a-z]\))([a-z]\d)\)", r"\1\n2", text) #alíneas letra+número
3 text = re.sub(r"([a-z]\s)\)", r"\n1", text) #alíneas letra+espaço
```

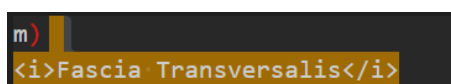
Em contrapartida, no que se refere à nomenclatura propriamente dita, determinados conceitos encontravam-se **separados por uma nova linha** e delimitados, ou pelas **tags bold**, ou pelas **tags itálico**.

Esta situação é passível de constatar relativamente a **tíbia**, **tálus** e **Fascia Transversalis** como é possível observar na imagem abaixo:



```
<b>3.15 TORNOZELO DIREITO: VISTA LATERAL</b>
a) Osso fíbula
b) Osso
<b> tíbia</b>
c) Maléolo lateral
d) Osso
<b> tálus </b>
e) Cabeça
f) Colo
g) Tróclea
```

Figura 12: Terminologias entre tags bold



```
m)
<i>Fascia Transversalis</i>
```

Figura 13: Terminologia entre tags itálico

De maneira a corrigir estas anomalias, procedeu-se à eliminação das tags, e posterior retorno para a linha anterior, através do uso da função **sub** das RegEx.

A linha abaixo demonstra a expressão regular empregue para o caso da terminologia **Fascia Transversalis**:

```
1 text = re.sub(r"\n<i>(.)</i>", r"\1", text)
```

3.2.6 Fase 4: Sintaxe de extração definida

Dada por concluída a fase de correção de anomalias, em que é possível observar os diferentes padrões distintivos de cada componente, isto é dos dois grupos globais, das secções e subsecções, e terminologias, a fase seguinte consiste na definição de uma sintaxe, própria para cada um, de maneira a facilitar o processo de extração pela função **findall** das RegEx.

Posto isto, foi definido que:

- Os dois grandes grupos, isto é, **Sistema Esquelético e Articular** e **Sistema Muscular** são antecipados pelo símbolo &;

```
1 text = re.sub(r"<b>([^\d]+)</b>", r"&\1", text)
```

- As **secções** são antecipadas pelo símbolo %;

```
1 text = re.sub(r"<b>\d\.\.?s?([A-Z\sÃÔÁÊËÕÚÇ]+)</b>", r"%\1", text)
```

- As **subsecções** são antecipadas pelo símbolo #;

```
1 text = re.sub(r"<b>\d\.\s?\d+\.\.?s?(.+)</b>", r"#\1", text)
```

- As **terminologias** são antecipadas pelo símbolo @;

```
1 text = re.sub(r"\n(([a-z]\s?\))|((([a-z][0-9])\s?\))|([A-Z]\s?\))", r"\n@" , text)
```

- As eventuais **secções** de **subsecções** são antecipadas pelo símbolo ;.

```
1 text = re.sub(r"<b>\d\.\d\.\d\.\.?s?(.+)</b>", r";\1", text)
```

Todo este processo facilitou a extração de todos estes diferentes elementos para as suas respetivas listas, de maneira a ser possível implementar o código de construção do dicionário final.

3.2.7 Fase 5: Construção do dicionário final

De um modo geral, o raciocínio de construção do dicionário final consistiu em:

- Escrita, num ficheiro **txt**, do conteúdo final do documento, já corrigido e com a sintaxe anteriormente definida;
- Definição de um dicionário, **dic**, vazio, para armazenar os dados analisados do ficheiro anterior;
- Definição de variáveis de controlo como **cur_titulo**, **cur_seccao**, **cur_subseccao** e **cur_sub_subseccao**, inicializadas como **None**;
- **Abertura** do ficheiro **txt** para leitura, sendo o seu conteúdo dividido em linhas;
- **Iteração** sobre cada linha;
- Para cada linha, **verificar** se o texto, sem o primeiro caractere, dado que não se pretende incluir os símbolos definidos anteriormente, está presente numa das listas resultantes do **findall**. Isto é, pretende-se averiguar se se trata de uma secção, subsecção, etc;
- Dependendo de onde a linha se encaixa, a **variável de controlo**, correspondente, é **atualizada**;
- Quando uma **linha começa com "@"** (indicando uma nomenclatura), esta é **adicionada ao dicionário**, no nível apropriado da hierarquia de secções;
- Após iterar por todas as linhas, o dicionário **dic** é **convertido em formato JSON**;
- O ficheiro final designa-se "**dicionario.json**" com codificação UTF-8 e formato de indentação de 4 espaços.

As imagens abaixo permitem observar partes do resultado final:

```
{
  "SISTEMA ESQUELÉTICO E ARTICULAR": {
    "CRÂNIO": {
      "CRÂNIO: VISTA ANTERIOR - II": [
        "Osso frontal",
        "Forame supraorbital",
        "Ossos nasais",
        "Lâmina perpendicular do osso etmoide",
        "Osso zigomático",
        "Osso vômer",
        "Osso maxila",
        "Processo estiloide do osso temporal",
        "Ângulo da mandíbula",
        "Mento",
        "Incisura frontal",
        "Glabela",
        "Asa maior do osso esfenóide",
        "Forame infraorbital",
        "Fissura orbital superior",
        "Osso nasal",
```

Figura 14: Parte inicial do dicionário final

```
    ],
    "MEMBRO SUPERIOR": {
      "CLAVÍCULA DIREITA: VISTA SUPERIOR": [
        "Extremidade acromial",
        "Corpo",
        "Face articular esternal",
        "Extremidade esternal",
        "Tubérculo conóide"
      ],
      "CLAVÍCULA DIREITA: VISTA INFERIOR": [
        "Face articular acromial",
        "Extremidade acromial",
        "Tubérculo conóide",
        "Sulco do músculo subclávio",
        "Impressão do ligamento costoclavicular",
        "Extremidade esternal"
      ]
    },
  ],
```

Figura 15: Parte intermédia do dicionário final

Exemplo de subsecções com secções:

```

"VÉRTEBRAS CERVICAIS ATÍPICAS: ÁXIS (C2)": {
  "VISTA INFERIOR DE C3": [
    "Corpo de C3",
    "Face articular inferior",
    "Processo articular inferior",
    "Forame transversário",
    "Processo transverso",
    "Lâmina",
    "Processo espinhoso bífido",
    "Forame vertebral"
  ]
},
"VÉRTEBRAS CERVICAIS ATÍPICAS: VÉRTEBRA PROEMINENTE (C7)": {
  "VISTA INFERIOR": [
    "Corpo de C7",
    "Forame vertebral",
  ]
}

```

Figura 16: Caso de subsecções com secções

3.3 Minidicionário de Cardiologista

3.3.1 Introdução:

O ficheiro em questão trata-se de um dicionário Inglês-Português e Português-Inglês das expressões, e termos mais frequentes, em escala mundial, no âmbito da Cardiologia.

Em relação ao conteúdo, numa parte inicial, são encontradas as secções referentes à apresentação, à breve descrição do autor, bem como aos agradecimentos. Em seguida, o documento é dividido em duas grandes secções: a primeira apresenta os termos e expressões em Inglês, com as suas respetivas traduções para Português; a segunda, o oposto. O ficheiro termina, por sua vez, com as referências bibliográficas utilizadas pelo autor durante o desenvolvimento do dicionário.

<p>CUFF</p> <p>CUFF (CUFFED HYPERTENSION) (AORTIC CUFF = BALÃO) – Dar um tapa com as mãos abertas / Punho de camisa / Parte do tecido que vai ao redor do braço num esfigmomanômetro (manguito)</p> <p>CULPRIT – Réu / Culpado / Acusado / Criminoso</p> <p>DISPOSABLE</p> <p>CULTURE MEDIA – Meios de cultura</p> <p>CURRENTLY – Vigente / Em andamento / Correntemente</p> <p>CUTOFF VALUE – Valor de corte</p> <p>CUTTING EDGE TECHNOLOGY – Tecnologia de ponta</p> <p>D</p> <p>DEADLINE – Prazo máximo</p> <p>DEAFNESS – Surdez</p> <p>DEEMED / DEEM (TO) – Julgar / Supor / Considerar como</p> <p>DELAY – Atraso / Movimento vagaroso / Adiar / Deter temporariamente</p> <p>DELIVERY – Entrega ou parto / Dar à luz</p> <p>DELIVERY CABLE – Cabo liberador</p> <p>DENGUE FEVER OR BREAKBONE FEVER – Dengue ou febre do quebra-ossos</p> <p>DE NOVO – Inédito / Inusitado</p> <p>DEPLOY – Implantar / Estender / Desenrolar / Abrir</p> <p>DEPLOYMENT / STENT DEPLOYMENT – Formação / Colocação frontal / Colocação estratégica / Posicionamento</p> <p>DEPTH / DEEP / DEEPEN (TO) – Profundidade / Fundo / Aprofundar</p> <p>DEVICE – Dispositivo / Instrumento / Algo construído com um "design"</p> <p>DIASTOLIC GRUNT – Estalido / Grunhido (emitir som gutural) / Som inarticulado (que expressa indiferença)</p> <p>DIASTOLIC THRILL – Frêmito diastólico</p> <p>DISABILITY – Incompetência / Falta de habilidade física ou mental</p> <p>DISARRAY – Desordem, desaranjo / Fora de disposição / Fora de conjunto</p> <p>DISCHARGE / DISMISSAL (TO DISMISS) – Dispensar / Liberar de obrigação / Dar alta médica</p> <p>DISCLOSURES – Confidencialidade (desvendar)</p> <p>DISCONTINUED – Interrompido / Descontinuado</p> <p>DISPENSARY – Um lugar onde os medicamentos estão dispostos; especialmente em uma instituição pública onde o serviço é oferecido gratuitamente</p> <p>DISPOSABLE – Descartável</p>	<p>AMIDO</p> <p>Amido / Fécula / Goma / Alimentos ricos em amido – STARCH</p> <p>Amígdalas – TONSILS</p> <p>Ampla margem – WIDE-RANGING (EFFECTS)</p> <p>Amplamente / Amplo / Mais amplo – BROADLY / BROAD / BROADER</p> <p>Amplamente espalhado / Largamente espalhado – WIDESPREAD</p> <p>Anestesiari – ANAESTHETIZE (TO)</p> <p>Anestesista (tomar cuidado, o final não é sist.); é igual a dentista – ANESTHETIST</p> <p>Angulação – TENTING</p> <p>Ano bissexto – LEAP YEAR</p> <p>Ansiedade – ANXIETY</p> <p>Apesar de / Não obstante / Contudo / Todavia – NEVERTHELESS</p> <p>Apoio / Colaboração – SUPPORT</p> <p>Aqueles que exercem arte / profissão etc. / Profissionais – PRACTITIONERS</p> <p>Aqui contido / Anexo / Incluso – HEREIN</p> <p>Aquilo que é posto ou aplicado / Intervir / Energia que é aplicada a algo – INPUT / INPUT (TO)</p> <p>Argola de guardanapo, que é um tipo de calcificação causada pelo estreitamento similar a uma argola de um guardanapo – NAPIKIN RING – CALCIFICATION / STENOSIS</p> <p>Armação / Moldura – FRAMEWORK</p> <p>Armadilha / Arapuca / Cilada – PITFALL</p> <p>Arrotar / Arroto – BELCH (TO)/BELCHING</p> <p>Artelhos / Dedos dos pés – TOES</p> <p>AZIA</p> <p>Assemelhar-se a / Parecer-se com – RESEMBLE (TO)</p> <p>Assentar / Colocar o stent – TO LAND THE STENT</p> <p>Assoberbante / Esmagador / Inesistível / Dominante – OVERWHELMING</p> <p>Assumir cargo / Prometer fazer se for possível / Prometer / Assumir / Tentar – UNDERTAKE (TO)</p> <p>Assunto / Fascículo de revista / Emitir um bilhete – ISSUE</p> <p>Atraso / Movimento vagaroso / Adiar / Deter temporariamente – DELAY</p> <p>Através – THROUGH</p> <p>Através de todo – THROUGHOUT</p> <p>Aumentar a altura (elevant) / Altura / Alto – HEIGHTEN (TO) / HEIGHT / HIGH</p> <p>Aumento / Intensificação / Crescimento (um toque a mais) – ENHANCEMENT</p> <p>Aumento de tamanho ou de capacidade / Engrandecer / (Estender / Amplificar / Incrementar / Aumentar de tamanho) – ENLARGEMENT / (TO ENLARGE)</p> <p>Auxiliar / Subserviente / subsidiário / subordinado – ANCILLARY</p> <p>Avaliação – ASSESSMENT</p> <p>Averiguar / Verificar / Determinar / Avaliar – ASCERTAIN</p> <p>Avermelhar (fluxo) – FLUSH (TO)</p> <p>Aves domésticas – POULTRY</p> <p>Aviso / Advertência / Coisas importantes / Um alerta / Precaução – CAVEAT</p> <p>Axilas – PITS / ARMPITS</p> <p>Azia / Pirose – HEARTBURN</p>
---	--

(a) Exemplo de uma página da secção Inglês-Português

(b) Exemplo de uma página da secção Português-Inglês

Figura 17: Secções do Minidicionário de Cardiologista

3.3.2 Estrutura de armazenamento definida:

De acordo com o reconhecimento das duas grandes secções no documento em estudo, decidiu-se implementar um dicionário, posteriormente armazenado num ficheiro JSON, com a seguinte estrutura:

```
dic_traduc = {  
  
    "ingles-portugues": {"termo1": "traducao1", "termo2": "traducao2", ...}  
  
    "portugues-ingles": {"termo1": "traducao1", "termo2": "traducao2", ...} }
```

3.3.3 Fase 1: Conversão

Para iniciar o processo de preparação e limpeza do documento, com o objetivo de gerar o ficheiro JSON final, optou-se por converter o documento para o formato **xml**.

Esta escolha baseia-se na capacidade da linguagem xml facilitar a identificação de padrões, uma vez que organiza os dados numa estrutura hierárquica por meio do uso de tags que, por sua vez, possuem nomes descritivos que auxiliam na compreensão do conteúdo.

3.3.4 Fase 2: Limpeza e Extração

Eliminação Manual: Conforme estabelecido anteriormente, as partes iniciais e finais do documento não apresentam informações relevantes sobre os termos e as suas respetivas traduções. Portanto, todo o conteúdo no ficheiro, em formato xml, acima da página 9 (início da secção Inglês-Portugal) e abaixo da página 51 (fim da secção Português-Inglês) foi eliminado.

Expressões Regulares: Através da estrutura do ficheiro em formato xml foi possível identificar padrões referentes aos termos e às suas traduções que facilitaram a limpeza e a extração.

Em relação à secção Inglês-Português, foi possível observar que os termos apresentavam, dentro da tag **<text>**, a propriedade **font="7"** e que, além disso, estavam dispostos entre tags ****. Logo, através do uso da função **sub** das RegEx foi possível sinalizar o início e o fim dos termos com o símbolo **@** a fim de facilitar a posterior extração.

```
1 line = re.sub(r'font="7"><b>(.*?)(</b>.*)', r'font="7"><b>@1\2', line)
```

De igual forma, para a secção Português-Inglês, o mesmo padrão foi observado, no entanto, verificou-se uma diferença sutil na função **<text>** em que a propriedade **font** apresentava um valor igual a 13.

```
1 line = re.sub(r'font="13"><b>(.*?)(</b>).*', r'font="13"><b>@1\2', line)
```

Após as devidas sinalizações, utilizou-se a função **findall** das RegEx para extrair todos os termos presentes no documento e armazená-los numa lista denominada **lista_termo**.

```
1 termos = re.findall(r'(<b>@[^\@]+@</b>)', text)
```

```
1 result = re.findall(r'<b>(.*?)</b>', t)
```

Em relação às traduções, o mesmo padrão foi observado nas duas secções: a tag **<text>** apresentava a propriedade **font="8"** e não estavam dispostas entre tags ****. Sendo assim, utilizou-se o mesmo mecanismo para extrair os termos, e respetivas traduções, pelo que foram armazenados numa lista chamada **lista_traducoes**.

```
1 line = re.sub(r'font="8">(.*?)(</text>)', r'font="8">1\2@</text>', line)
```



```
1 traducoes = re.findall(r'(\font="8">@[^@]+@</text>)', text)

1 result = re.findall(r'>(.*)</text>', trad)
```

```
<text top="95" left="34" width="97" height="12" font="7"><b>MEDICAL CLEARANCE </b></text>
<text top="95" left="131" width="76" height="12" font="8"> Liberação médica</text>
<text top="115" left="34" width="91" height="12" font="7"><b>MEDICAL ORDER </b></text>
<text top="115" left="125" width="87" height="12" font="8">Prescrição médica</text>
<text top="128" left="34" width="124" height="12" font="8">para paciente hospitalizado</text>
<text top="147" left="34" width="131" height="12" font="7"><b>MEDICAL PRESCRIPTION </b></text>
<text top="147" left="165" width="49" height="12" font="8"> Receita</text>
<text top="160" left="34" width="176" height="12" font="8">médica para adquirir medicamentos</text>
<text top="173" left="34" width="60" height="12" font="8">na farmácia</text>
<text top="193" left="34" width="47" height="12" font="7"><b>MEMBER </b></text>
<text top="193" left="81" width="126" height="12" font="8"> Sócio / Membro / Associado</text>
<text top="212" left="34" width="125" height="12" font="7"><b>MEMORY IMPAIRMENT </b></text>
<text top="212" left="158" width="57" height="12" font="8">Diminuir a</text>
<text top="225" left="34" width="176" height="12" font="8">qualidade de memorização / Enfraquecer</text>
<text top="239" left="34" width="136" height="12" font="8">a memória / Danos de memória</text>
<text top="258" left="34" width="38" height="12" font="7"><b>MESH </b></text>
<text top="258" left="72" width="27" height="12" font="8">Malha</text>
<text top="50" left="222" width="111" height="12" font="7"><b>MICROARRAY (STUDY) </b></text>
```

(a) Padrão dos termos na secção Inglês-Português

```
<text top="483" left="51" width="161" height="12" font="13"><b>Acontecer / Ocorrer / Realizar-se </b></text>
<text top="483" left="212" width="16" height="12" font="8">TO</text>
<text top="496" left="51" width="54" height="12" font="8">TAKE PLACE</text>
<text top="515" left="51" width="48" height="12" font="13"><b>Adesivos </b></text>
<text top="515" left="99" width="44" height="12" font="8"> PATCHES</text>
<text top="534" left="51" width="179" height="12" font="13"><b>Advertir / Chamar a atenção para </b></text>
<text top="546" left="51" width="50" height="12" font="8">WARN (TO)</text>
<text top="565" left="51" width="56" height="12" font="13"><b>AFFERENT </b></text>
<text top="565" left="107" width="41" height="12" font="8">Aferente</text>
<text top="111" left="239" width="113" height="12" font="13"><b>Afinar / Afinamento </b></text>
<text top="111" left="355" width="63" height="12" font="8">THIN (TO) /</text>
<text top="123" left="239" width="47" height="12" font="8">THINNING</text>
<text top="143" left="239" width="176" height="12" font="13"><b>Afinidade / Harmonia / Acordo / Afinidade </b></text>
<text top="156" left="239" width="73" height="12" font="13"><b>com o paciente </b></text>
<text top="156" left="312" width="83" height="12" font="8"> PATIENT RAPPORT</text>
<text top="175" left="239" width="85" height="12" font="13"><b>Afrouxadamente </b></text>
```

(b) Padrão dos termos na secção Português-Inglês

Figura 18: Padrões observados em relação aos termos

```
<text top="262" left="34" width="66" height="12" font="13"><b>Ano bissexto </b></text>
<text top="262" left="100" width="52" height="12" font="8"> LEAP YEAR</text>
<text top="281" left="34" width="59" height="12" font="13"><b>Ansiedade </b></text>
<text top="281" left="93" width="37" height="12" font="8">ANXIETY</text>
<text top="299" left="34" width="176" height="12" font="13"><b>Apesar de / Não obstante / Contudo /</b></text>
<text top="312" left="34" width="49" height="12" font="13"><b>Todavia </b></text>
<text top="312" left="83" width="71" height="12" font="8">NEVERTHELESS</text>
<text top="330" left="34" width="99" height="12" font="13"><b>Apoio / Colaboração </b></text>
<text top="330" left="133" width="46" height="12" font="8"> SUPPORT</text>
<text top="349" left="34" width="177" height="12" font="13"><b>Aqueles que exercem arte / profissão </b></text>
<text top="362" left="34" width="97" height="12" font="13"><b>etc. / Profissionais </b></text>
<text top="362" left="132" width="72" height="12" font="8">PRACTITIONERS</text>
<text top="380" left="34" width="138" height="12" font="13"><b>Aqui contido / Anexo / Incluso </b></text>
<text top="380" left="172" width="35" height="12" font="8"> HEREIN</text>
```

Figura 19: Padrão observado em relação às traduções

3.3.5 Fase 3: Correção de anomalias

Após a extração dos termos e traduções de acordo com os padrões identificados, foi possível observar que alguns elementos apresentavam erros devido à estruturação do ficheiro xml.

O termo **1st / 2nd / 3rd / 4th / SOUND** na secção Inglês-Português apresenta um padrão diferente dos demais: os identificadores de números ordinais **st**, **nd**, **rd** e **th** apresentam tags **<text>** com a propriedade **font="11"**. Na secção Português-Inglês, a respetiva tradução deste termo, em relação à propriedade **font**, possui valor 14. Portanto, para realizar uma correta extração, foi necessário adicionar condições **if** ao código desenvolvido em Python:

```
1 if 'font="7"><b>' in line and ('font="7"><b>' not in previous_line and 'font
   = "11"><b>' not in previous_line)
```

```

1 if 'font="8">' in line and ('font="8">' not in previous_line and 'font="14">'
    not in previous_line)

1 if 'font="8"' not in next_line and 'font="14">' not in next_line

```

Além disso, foi possível verificar que alguns elementos apresentam, entre o termo e a tradução, tags <text> sem conteúdo, o que também gera erros no processo de extração. Logo, foi realizada a devida eliminação através de uma condição **if** juntamente com a função **search** das RegEx:

```

1 if re.search(r'font="8"> <.*', line)

```

3.3.6 Fase 4: Construção do dicionário final

Para a construção do dicionário final com a estrutura apresentada, foi necessário dividir as listas, citadas anteriormente, **lista_termos** e **lista_traducoes** em sublistas, porque, de acordo com a estruturação do documento, os primeiros 512 elementos de cada lista correspondem à secção Inglês-Português e, os demais, à secção Português-Inglês.

Após essa etapa, de maneira geral, o raciocínio consistiu em:

- Definição de um dicionário, **dic_traduc**, com as chaves **ingles-portugues** e **portugues-ingles**, para armazenar os dados extraídos em dicionários vazios;
- Para a chave **ingles-portugues**, foi adicionado ao seu dicionário vazio cada elemento da sublista **en_pt_termos** com a sua respetiva tradução presente na sublista **en_pt_traducoes**;
- Em relação à chave **portugues-ingles**, foram adicionados os elementos (termos e traduções) que estavam armazenados nas sublistas **pt_en_termos** e **pt_en_traducoes** respetivamente;
- Após a inserção de todos os termos e traduções, o dicionário **dic_traduc** é **convertido em formato JSON**;
- O ficheiro final designa-se "**dic_traduc.json**" com codificação UTF-8 e formato de indentação de 4 espaços.

```

"ingles-portugues": {
  "A SURMISE (A CONJECTURE - SUSPICION) (TO ASSUME ON SMALL EVIDENCE)": "Conjectura / Suposição",
  "A.C.L.S": "Advanced Cardiovascular Life Support",
  "A.E.D": "Automated External Defibrillator",
  "A.S.A.P. (AS SOON AS POSSIBLE)": "O mais rapidamente possível",
  "A.V.C. (ABERRANT VENTRICULAR CONDUCTION)": "Não significa Acidente Vascular Cerebral, um AVC em inglês é um",
  "ABDOMINAL FEVER / TYPHOID FEVER": "Febre tifoide",
  "ABSENT-MINDED": "Distraído / Desatento",
  "ACCEPTANCE": "Novos pacientes / Aco-lhimento",
  "ACCRUING": "Vir de maneira natural / Incrementado de maneira natural / Devido a / Proveniente de",
  "ACHE": "Termo para designar casos particulares de dor (dor contínua, que não passa, mas não é severa) / Do",
  "ACHIEVEMENT": "Realização / Conquistas / Feitos",
  "ADDICT": "Viciado (tóxicos)",
  "ADHESIVE TAPE": "Esparrapão",
  "AFFERENT": "Aferente",
  "AFFORD (TO)": "Dar-se ao luxo de / Ter recursos para",
  "AGE STRATA": "Faixas etárias",
  "AILMENT": "Uma desordem física ou mental, especialmente uma doença leve",
  "AIRBORNE": "Transportado via aérea / Que se pega pelo ar",
  "AKIN": "Consanguíneo / Aparentado",
  "AMBULATORY": "Passeio, galeria, cor-redor, ambulatorial, refere-se a paciente que pode caminhar/andar",
  "ANAESTHETIZE (TO)": "Anestesiar",
  "ANCILLARY": "Subserviente / Subsidiário / Auxiliar / Subordinado",
  "ANECDOTAL": "Pertencendo a / Caracterizado por ou repleto de histórias ou casos",

```

Figura 20: Secção Inglês-Português do dicionário final

```

"portugues-ingles": {
  "1ª / 2ª / 3ª / 4ª / Bulha": "1st / 2nd / 3rd / 4th / SOUND",
  "À esquerda / Ao contrário dos ponteiros do relógio / Sentido anti-horário": "COUNTERCLOCKWISE",
  "A facilidade de fazer descobertas importantes e valiosas de maneira inesperada ou por acaso": "SERENDIPITY",
  "A investigação profunda sobre um assunto / Exame minucioso": "SCRUTINY",
  "A.C.L.S": "Advanced Cardiovascular Life Support",
  "A.E.D": "Automated External Defibrillator",
  "Abertamente / Publicamente / Premeditadamente": "OVERTLY",
  "Abordagem / uma aproximação / aproximar-se": "APPROACH",
  "Abrangente / Extensivo / Que engloba": "COMPREHENSIVE",
  "Acocoramento / Agachamento": "SQUAT-TING (POSITION)",
  "Acomodar / Instalar (assentar pó ou sedimento) / Acalmar / Arranjar / Resolver": "SETTLE (TO)",
  "Acontecer / Ocorrer / Realizar-se": "TO TAKE PLACE",
  "Adesivos": "PATCHES",
  "Advertir / Chamar a atenção para": "WARN (TO)",
  "Afferent": "Aferente",
  "Afinar / Afinamento": "THIN (TO) / THINNING",
  "Afinidade / Harmonia / Acordo / Afinidade com o paciente": "PATIENT RAPPORT",
  "Afrouxadamente": "LOOSELY",
  "Aguilha": "NEEDLE",
  "Aguilha de buraco largo": "LARGE BORE NEEDLE",
  "Ala": "WING",

```

Figura 21: Secção Português-Inglês do dicionário final

4 Conclusão

Para a realização deste trabalho foi essencial estabelecer, primeiramente, uma análise cuidada da estrutura de todos os documentos, de maneira a tornar possível a identificação de padrões que permitem distinguir determinadas secções ou componentes.

Após esta fase de identificação, foi possível dar início à fase de limpeza, onde foi colocado em prática todo o conhecimento adquirido ao longo das aulas no que se refere à escrita de expressões regulares.

O trabalho em causa permitiu, efetivamente, aprimorar a capacidade de escrita destas expressões, uma vez que foram encontrados os mais diversos padrões, cada um com o seu grau de distinção. Além do mais, tornou-se evidente todo o poder e versatilidade das expressões regulares, uma vez que, com uma sintaxe adequada, e uso de caracteres, e metacaracteres especiais, é possível criar padrões de alta complexidade e efetuar operações em texto com um elevado grau de eficiência.

No entanto, uma das dificuldades encontradas consistiu no estabelecimento da generalidade das expressões regulares de forma a extrair apenas a informação necessária, sem envolver a escrita de expressões em demasia.

Assim sendo, as expressões regulares desenvolvidas ao longo do projeto foram estabelecidas de maneira a serem de fácil compreensão e implementação, evitando a possível influência sobre outros elementos. Com isto, procedeu-se à construção dos diferentes ficheiros JSON, podendo-se concluir, assim, que o objetivo do trabalho foi alcançado.