



**Universidade do Minho**  
Escola de Engenharia

# Trabalho Prático 1

**Processamento de Linguagem Natural em Engenharia Biomédica**

**Mestrado em Informática Médica**

**Ano Letivo 2023/2024**

**Alunos:**

Madalena Passos, pg54023

Mariana Ribeiro, pg54061

Mariana Almeida, pg54062

**Docentes:** Luís Filipe da Costa Cunha e José João Almeida.

Braga, abril de 2024

## Introdução

No âmbito da Unidade Curricular de Processamento de Linguagem Natural em Engenharia Biomédica, foi desenvolvido o presente relatório técnico que visa a análise do primeiro trabalho prático realizado.

O objetivo principal do projeto consistiu na extração de informações de documentos em formato PDF e posterior preservação em ficheiros JSON. Desta forma, foram selecionados 4 dos documentos disponibilizados pelo docente, nomeadamente o documento obrigatório *Glossário Ministério da Saúde*, bem como os documentos *Ossos*, *Minidicionário de Cardiologia* e *Glossário de Termos Médicos Técnicos e Populares*.

Para atingir o objetivo proposto, foi seguido um processo metodológico que se iniciou com uma primeira análise dos documentos para a escolha do melhor tipo de conversão. Com efeito, foi decidido a conversão de três dos documentos (*Glossário Ministério da Saúde.pdf* , *Ossos.pdf* e *Minidicionário de Cardiologia.pdf*) para XML e do quarto para TXT, recorrendo, através do terminal, aos comandos **pdftohtml -xml** e **pdftotext**, respectivamente.

Posteriormente, foram utilizadas expressões regulares, nomeadamente os métodos *sub*, *search* e *findall*, para o tratamento e limpeza dos dados extraídos.

De seguida, procedeu-se à definição da estrutura de dados mais adequada para guardar a informação extraída de cada documento para posterior armazenamento num ficheiro em formato JSON.

Ao longo deste relatório, será apresentada uma visão detalhada das etapas realizadas durante o desenvolvimento do projeto para cada documento. Serão, ainda, discutidas as abordagens utilizadas, os desafios enfrentados e as soluções adotadas para superar as dificuldades encontradas.

# 1 Análise de Documentos

## 1.1 Documento 1: Glossário Ministério da Saúde

No documento *Glossário Ministério da Saúde.pdf* está presente vocabulário controlado próprio, composto quer por termos técnicos em Saúde quer por terminologias referentes aos atos normativos do Ministério e entidades vinculadas.

Para a realização do trabalho prático, foram consideradas três secções distintas: a primeira, onde estão presentes os termos, as respetivas categorias e as descrições dos mesmos; a segunda, onde se encontram apresentadas as categorias juntamente com as respetivas descrições; e a terceira, que contém os descritores organizados por categorias.

Para tratar cada uma destas secções, foram definidas três estruturas de dados diferentes, bem como diferentes técnicas de extração das informações relevantes. Para isso, foram desenvolvidos três programas distintos: *ministerio1.py*, *ministerio2.py* e *ministerio3.py*. Cada um dos programas gerou um JSON intermediário que correspondeu a cada secção analisada. Posteriormente, foi desenvolvido um novo programa, *juncoes\_ministerio.py*, cujo objetivo foi unir os três JSONs intermediários num único JSON consolidado.

Para a análise do processamento da primeira secção, numa fase inicial, através do método *sub* das expressões regulares, procedeu-se à eliminação de *tags* relacionadas a declarações de tipo de documento, de página, de texto, de fonte e de imagem, bem como elementos específicos de conversão de PDF para XML. De seguida, foi removido todo o texto até à terceira ocorrência do número 15 e todo o conteúdo após a primeira ocorrência do número 107, ficando apenas com o conteúdo relevante. Adicionalmente, foi também removida numeração de páginas e cabeçalhos que se encontravam no texto.

De forma a estruturar a informação contida no texto, definiu-se a seguinte estrutura:

```
{
  "Termo": {
    "categoria": "nome_categoria",
    "descricao": "descrição_termo"
  },
  ...
}
```

Atendendo à estrutura anterior, foi decidido que a informação deveria ser guardada num dicionário de dicionários, onde cada chave representaria um termo do glossário e cada valor um

outro dicionário que contivesse informações sobre a respetiva categoria e descrição. Posto isto, foram criadas marcas para destacar os campos relevantes e auxiliar no processo de extração. Assim, as tags `<b>` e `</b>` foram substituídas pela marca '@' de forma a delimitar cada termo. Para além disso, a marca anterior foi utilizada posteriormente num ciclo *while* para unir o conteúdo de um termo que ocupa mais do que uma linha. É ainda de salientar que foram efetuadas limpezas adicionais com o intuito de remover elementos desnecessários e efetuadas substituições para correções de casos particulares.

Posteriormente, foram usadas as marcas '£' e '#' para indicar o início e o fim da descrição de um termo, respetivamente, conforme ilustrado na Figura 1.

```
@Abordagem médica tradicional do adulto hospitalizado@
<i>Categoria: </i>
Atenção à Saúde
£Focada em uma queixa principal e o hábito
médico de tentar explicar todas as queixas
e os sinais por um único diagnóstico, que é
adequada no adulto jovem £ não se aplica em
relação ao idoso.
#
```

Figura 1: Marcas utilizadas na 1ª secção do texto xml do documento *Glossário Ministério Saúde*.

Ao longo do processo, foi detetado que existiam termos que não seguiam o padrão referido acima, isto é, não continham categoria nem descrição. No entanto, na linha imediatamente a seguir ao termo apresentavam uma expressão inicializada pela palavra *Ver*. Desta forma, foi utilizada a seguinte expressão regular para substituir estes casos pela estrutura anteriormente definida:

```
re.sub(r'@(.*)\n*(Ver.*)', r'@£1\n<i>Categoria:</i>\nVer descrição\n£2\n#', texto)
```

Na Figura 2, está representado o resultado do processo de substituição.

```
@Abuso sexual na adolescência@
<i>Categoria:</i>
Ver descrição
£Ver Abuso sexual na infância.
#
```

Figura 2: Substituição de casos particulares para uma estrutura padrão na 1ª secção texto xml do documento *Glossário Ministério Saúde*

Finalmente, foi desenvolvida uma expressão regular, utilizando o método *findall* com o intuito de extrair a informação mencionada, de onde resultaram 702 termos encontrados. De seguida, foi criado um dicionário onde cada termo corresponde a uma chave e os valores a um dicionário com a categoria e a descrição. Dado que algumas chaves podem ter mais de uma categoria e descrição associadas, é possível que alguns valores sejam representados como listas de dicionários. Posteriormente, criou-se um ficheiro json, *ministerio1.json*, com o conteúdo do dicionário, como se pode observar na Figura 3:

```
{
  "Abordagem médica tradicional do adulto hospitalizado": {
    "categoria": "Atenção à Saúde",
    "descricao": "Focada em uma queixa principal e o hábito médico de tentar explicar todas as queixas e os sinais por um ú"
  },
  "Abuso financeiro dos idosos": {
    "categoria": "Acidentes e Violência",
    "descricao": "Exploração imprópria ou ilegal e/ou uso não consentido de recursos financeiros dos idosos."
  },
  "Abuso incestuoso": {
    "categoria": "Acidentes e Violência",
    "descricao": "Consiste no abuso sexual envolvendo pais ou outro parente próximo, os quais se encontram em uma posição"
  },
  "Abuso sexual na adolescência": {
    "categoria": "Ver descrição",
    "descricao": "Ver Abuso sexual na infância."
  },
  "Abuso sexual na infância": {
    "categoria": "Acidentes e Violência",
    "descricao": "É todo ato ou jogo sexual, relação heterossexual ou homossexual, cujo agressor está em estágio de desenv"
  },
  "Ação racional": {
    "categoria": "Atenção à Saúde",
    "descricao": "Modelo de intervenção centrado no indivíduo no qual permite a relação entre a epidemiologia e a dimensão"
  }
}
```

Figura 3: *ministerio1.json*

Quanto ao processamento da secção 2, realizou-se, de forma semelhante, uma limpeza ao texto, removendo as *tags* e elementos irrelevantes do mesmo. De seguida, após uma análise do ficheiro, criou-se a seguinte sintaxe para representar a estrutura de dados a ser extraída:

```
{
  "categoria": "descricao",
  ...
}
```

Desta vez, atendendo à estrutura explicitada, foi decidido que a informação deveria ser guardada num dicionário, cujas chaves representam as categorias e os valores a descrição das mesmas.

Desta forma, foram aplicadas expressões regulares para encontrar as categorias e as respetivas descrições presentes na secção. Após a identificação das mesmas, o texto foi estruturado de acordo com o formato definido anteriormente. Para isso, foram utilizadas as marcas '@' e '§' para delimitar os termos e as descrições, respetivamente (Figura 4).

```
@Acidentes e Violência§
Refere-se ao conjunto de agravos à saúde que
pode levar a óbito ou seqüelas irreversíveis
que inclui as causas ditas acidentais: devidas
ao trânsito, trabalho, quedas, envenenamen-
tos, afogamentos e outros tipos de acidentes,
e as causas intencionais como agressões e le-
sões autoprovocadas.
Inclui subtemas como: abuso sexual, aciden-
tes com animais peçonhentos, acidentes de
trânsito, acidentes do trabalho, homicídios,
intoxicações e envenenamentos, maus-tratos
contra o idoso, maus tratos na infância, sui-
cídios, violência doméstica, violência contra
as mulheres, acidentes em ambientes domés-
ticos, etc.
@Administração e Planejamento em Saúde§
```

Figura 4: Marcas utilizadas na 2ª secção do texto xml do documento *Glossário Ministério Saúde*.

Assim como na primeira secção, também foram realizadas correções e limpezas adicionais para garantir a consistência dos dados. A expressão regular desenvolvida para extrair a informação da segunda secção resultou num total de 24 categorias encontradas. Posteriormente, foram inseridas no dicionário com o formato referido. Por fim, foi criado um ficheiro JSON, *ministerio2.json*, contendo o conteúdo do dicionário, conforme demonstrado na Figura 5:

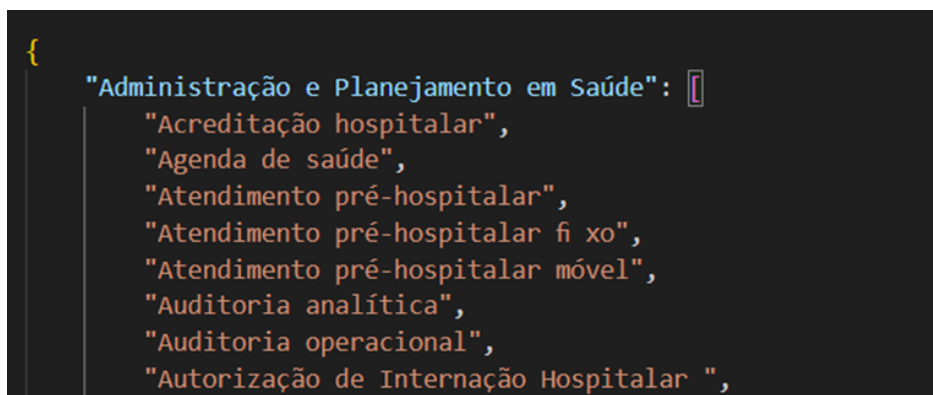
```
{
  "Acidentes e Violência": "Refere-se ao conjunto de agravos à saúde que pode levar a óbito ou seqüelas irreversíveis que incl",
  "Administração e Planejamento em Saúde": "Refere-se à organização, elaboração de planos e políticas públicas e governamentais",
  "Alimentação e Nutrição": "Refere-se a todos os tipos de substâncias que têm por função alimentar ou nutrir seres hu-manos e",
  "Ambiente e Saúde": "Refere-se ao estudo das interações entre os seres vivos e o meio, dedica-se a analisar as formas de ví",
  "Atenção à Saúde": "Refere-se à proteção e atenção à saúde dos di-versos grupos etários que correspondem aos ciclos vitais:",
  "Ciência e Tecnologia em Saúde": "Refere-se a investimentos públicos em ciência e tecnologia; desenvolvimento institucional",
  "Ciências Sociais em Saúde": "Refere-se aos estudos que se utilizam ou são elaborados pelas ciências sociais aplicados ao ca",
  "Comunicação em Saúde": "Refere-se ao conjunto dos meios de comu-nicação de massa voltados a divulgação de produtos, serviço",
  "Demografia": "Refere-se aos estudos das populações huma-nas, com o objetivo de caracterizá-las e anali-sar tendências popula",
  "Direito Sanitário": "Refere-se ao conjunto de leis e normas, na-cional e internacional, que compõe o sistema jurídico no qu",
  "Doenças Crônicas e Degenerativas": "Refere-se ao conjunto de doenças relaciona-das a múltiplos fatores de risco ambientais",
  "Doenças Infeciosas e Parasitárias": "Refere-se ao conjunto de infecções que po-dem ser adquiridas por contato direto de pe",
  "Drogas de Uso Terapêutico e Social": "Refere-se aos efeitos causados pelo consumo de substâncias químicas e seus desdobrame",
  "Economia da Saúde": "Refere-se aos estudos sobre gasto e fi nancia-mento em saúde, alocação e utilização de re-cursos no set",
  "Epidemiologia": "Refere-se aos estudos retrospectivos e pros-pectivos da distribuição e dos determinan-tes da prevalência d",
  "Equidade em Saúde e Social": "Refere-se à igualdade de recursos para neces-sidades iguais, de oportunidades de acesso para",
  "Ética e Bioética": "Refere-se ao conjunto de regras de conduta moral, deontológica e social que regulamen-tam as relações e",
  "História da Saúde Pública": "Refere-se às investigações voltadas a história das políticas, instituições e profi ssões de saú",
  "Medicamentos, Vacinas e Insumos": "Refere-se à produção cientifi ca e tecnológica referentes à biotecnologia e farmacologia",
  "Políticas Públicas e Saúde": "Refere-se à arena de interesses políticos, so-ciais e econômicos relativos ao domínio públic",
  "Promoção e Educação em Saúde": "Refere-se às diferentes formas de organização da sociedade no enfrentamento de seus proble",
  "Recursos Humanos em Saúde Pública": "Refere-se à formação e capacitação de pessoal técnico especializado, necessário ao fun",
  "Saúde Animal": "Refere-se aos cuidados e atenção à saúde dos animais, particularmente os de convívio hu-mano ou como fonte",
  "Vigilância em Saúde": "Refere-se à proteção e à promoção da saúde da população por meio da segurança sanitária de produtos,"
}
```

Figura 5: *ministerio2.json*

Relativamente à secção 3, o processo de limpeza do texto foi realizado da mesma forma que nas secções anteriores. Após uma análise do conteúdo do ficheiro, foi definida uma estrutura de dados para representar as informações a serem extraídas:

```
{  
  "categoria": ["descrição1", "descrição2", ...],...  
}
```

Neste caso, optou-se por guardar as descrições como uma lista associada a cada categoria. Desta forma, utilizou-se uma expressão regular que procura cada categoria entre `<b>` e `</b>` e a respetiva designação que se encontra entre `</b>` até ao próximo `<b>` consecutivo. Posteriormente, o texto estruturado foi convertido num dicionário, onde cada chave representa uma categoria e os valores são uma lista de descrições associadas a essa categoria. Este dicionário foi então convertido num ficheiro JSON, *ministerio3.json*.



```
{  
  "Administração e Planejamento em Saúde": [  
    "Acreditação hospitalar",  
    "Agenda de saúde",  
    "Atendimento pré-hospitalar",  
    "Atendimento pré-hospitalar fixo",  
    "Atendimento pré-hospitalar móvel",  
    "Auditoria analítica",  
    "Auditoria operacional",  
    "Autorização de Internação Hospitalar",  
  ]  
}
```

Figura 6: *ministerio3.json*

Para a criação de um único glossário abrangente, foi necessário combinar as informações contidas nos três JSONs: *ministerio1.json*, *ministerio2.json* e *ministerio3.json*.

Primeiramente, foram carregados os três ficheiros. De seguida, iterou-se sobre as chaves do primeiro JSON, *ministerio1.json*, verificando se os valores correspondentes eram listas ou dicionários. Para os valores que eram listas, foram extraídas para cada valor, a categoria e a descrição correspondentes e consultados os outros dois JSONs para obter informações sobre descrição e descritores associados das categorias.

Com todas as informações reunidas, foi criado um dicionário, que foi adicionado ao JSON combinado final, *ministério\_junto.json*, representado na Figura 7.

```
"Abordagem médica tradicional do adulto hospitalizado": {  
  "Categoria": {  
    "nome": "Atenção à Saúde",  
    "descrição categoria": "Refere-se à proteção e atenção à saúde dos di-versos grupos etários que correspondem aos c",  
    "descritores categoria": [  
      "Abordagem médica tradicional do adulto ",  
      "hospitalizado",  
      "Ação racional",  
      "Acidentes de trabalho",  
      "Ações estratégicas",  
      "Acompanhamento do crescimento e ",  
      "desenvolvimento infantil",  
      "Aconselhamento",  
      "Aconselhamento coletivo",  
      "Agentes comunitários da saúde",  
      "Alta complexidade",  
      "Amamentação exclusiva",  
      "Anticoncepção de emergência",  
      "Árvore de causas",  
    ]  
  },  
  "Descrição": "Focada em uma queixa principal e o hábito médico de tentar explicar todas as queixas e os sinais por um
```

Figura 7: *ministério\_junto.json*

## 1.2 Documento 2: Ossos.pdf

Inicialmente, foi feita a análise detalhada do documento *ossos.pdf*. Este documento está relacionado com a análise e legendagem anatômica de dois sistemas corporais: o sistema esquelético e articular e o sistema muscular.

Para uma melhor organização da informação o documento está organizado também pelas estruturas anatômicas legendadas e as diferentes vistas visualizadas. De forma mais visual, o ficheiro encontra-se estruturado da seguinte forma:



## SISTEMA ESQUELÉTICO E ARTICULAR

1. CRÂNIO
  - 1.1 CRÂNIO: VISTA ANTERIOR - I
  - 1.2 CRÂNIO: VISTA ANTERIOR - II
  - 1.3 ...
2. MEMBRO SUPERIOR
  - 2.1 CLAVÍCULA DIREITA: VISTA SUPERIOR
  - 2.2 CLAVÍCULA DIREITA: VISTA INFERIOR
  - 2.3 ...
4. TORSO MUSCULAR
  - 4.1 TORSO MUSCULAR: VISTA POSTERIOR DA CABEÇA E PESCOÇO
  - ...
  - 4.4 VÉRTEBRAS CERVICAIS ATÍPICAS: ÁXIS (C2)
    - 4.4.1 VISTA PÔSTERO-SUPERIOR
- ...

## SISTEMA MUSCULAR

1. CABEÇA E PESCOÇO
2. ...
3. ...

...

Este documento possui então várias páginas com imagens e as respectivas legendas agrupadas no final do documento, mais especificamente a partir da página 192. É de notar que entre estas páginas de, maioritariamente, imagens, encontram-se também textos introdutórios às estruturas anatómicas, como por exemplo, o crânio, membro superior e inferior e ainda o torso.

Como se pode observar pelo índice acima, cada secção (correspondente a cada sistema) tem diferentes títulos, correspondentes a estas estruturas anatómicas e, por fim, diferentes subtítulos correspondentes às vistas. Cada subtítulo tem as legendas respetivas no formato: “a) Osso frontal b) Osso parietal”, por exemplo.

De forma a armazenar a informação deste ficheiro em formato JSON, foi necessário desenvolver conceptualmente uma estrutura de dados capaz de responder aos requisitos do texto recolhido.

Assim, após análise decidiu utilizar-se vários dicionários pela maior facilidade de acesso e organização da informação. A legendagem de cada vista é formada por um dicionário com chaves as alíneas (“a”, “b”, “c”) e valores as respetivas legendas.

Consequentemente, estes vários dicionários inserem-se num outro dicionário com chave igual aos nomes das vistas e valores os dicionários referidos anteriormente. É de salientar que este dicionário tem também um par chave-valor correspondente à chave “Introdução” e valor correspondente a uma breve introdução à estrutura anatómica respetiva.

Seguidamente, este dicionário com as informações das vistas e dicionários com as suas legendas insere-se num outro dicionário com as estruturas anatómicas a legendar como chaves e o dicionário com informações das vistas como valor. Por fim, o primeiro dicionário engloba todos os anteriores tendo como chaves os títulos das 2 secções ("Sistema Esquelético e Articular" e "Sistema Muscular") e valores o dicionário geral mencionado anterior.

No caso particular, “4.4 VÉRTEBRAS CERVICAIS ATÍPICAS: ÁXIS (C2)” e outros casos semelhantes, é necessário definir mais um dicionário uma vez que existem dentro das estruturas anatómicas, outras sub-estruturas que têm as suas legendagens e vistas.

De forma mais visual, e para exemplificar, a estrutura de dados conceptualizada encontra-se abaixo:

```
{“SISTEMA ESQUELÉTICO E ARTICULAR” : {
  CRÂNIO: {“Introdução” : ”....” ,
    “Crânio lateral”: {"a":..., "b":...}, ...},
  COLUNAVERTEBRAL: {“Introdução” : ”....” ,
    “Vértebras Cervicais”: {"a":..., "b":....}, ...
    “Vértebras Atípicas”: {“Vista Anterior”: "a":...,“b”:...}
  }
}
“SISTEMA MUSCULAR” : {
...
}
}
```

Para então se passar à extração da informação para a estrutura definida acima, foi necessário converter o documento PDF num outro formato, neste caso, XML. Para isso, foi utilizado o comando **pdftohtml -xml**.

Em primeiro lugar, foi lido o ficheiro e feita uma limpeza ao texto, eliminando as tags como,

por exemplo ‘<>’, ‘</page>’, ‘</text>’, ‘<fontspec/>’ e ‘<image./>’. Foram ainda retirados cabeçalhos e removida a bibliografia. Esta limpeza foi conseguida através da função *sub*, substituindo estas expressões por uma expressão vazia, “ ”.

Foram também tratados alguns casos particulares de falta de numeração bold, <b>, em certos títulos, e a passagem para uma só linha dos mesmos, uma vez que alguns se encontravam divididos entre duas linhas.

Antes de passar à marcação do texto, este foi dividido em dois textos diferentes. O primeiro com as páginas com imagens e as introduções, que servirá, posteriormente, para a captação das introduções de cada título, e o segundo contendo as informações importantes de legenda-gem destas imagens.

De forma a facilitar a divisão e estruturação da informação, foram utilizadas marcações com caracteres especiais em locais do texto importantes, como por exemplo, o início e fim dos títulos e subtítulos.

Quanto à marcação, para sinalizar o início de cada uma das duas secções definidas acima, “Sistema Esquelético e Articular” e “Sistema Muscular”, foi utilizada a marcação “ℓℓℓℓℓℓℓℓℓℓ” no início e fim do seu título.

```
ℓℓℓℓℓℓℓℓℓℓ<b>SISTEMA ESQUELÉTICO E ARTICULAR</b>ℓℓℓℓℓℓℓℓℓℓ
```

Figura 8: Marcação de cada secção do documento *Ossos*

Seguidamente, para separar cada título do seu primeiro subtítulo, por exemplo, a separação de “1. CRÂNIO” e “1.1 1.1. CRÂNIO: VISTA ANTERIOR - I” foi utilizado o carater “@” repetidamente, da seguinte forma:

```
<b>1. CRÂNIO</b>
@@@@@@@@@@@@@@@@
#<b>1.1. CRÂNIO: VISTA ANTERIOR - I </b>§
```

Figura 9: Marcação da separação de cada título e o seu primeiro subtítulo e marcação de cada subtítulo do documento *Ossos*

Quanto à marcação de cada subtítulo esta foi feita através de dois caracteres diferentes: ‘#’ no início e ‘§’ no final, como se pode observar na figura anterior.

Por fim, uma vez que há existência de sub-subtítulos, foi também necessário proceder à sua marcação. Neste caso, foi utilizado “ℓ” no início e, novamente, o carater “ℓ” no final.

#### <b>4.5.1 VISTA SUPERIOR DE C3</b>£

Figura 10: Marcação da separação dos sub-subtítulos do documento *Ossos*

É importante realçar que após as marcações foi necessário tratar de alguns casos esporádicos de títulos e subtítulos mal marcados para permitir que o texto fosse bem dividido posteriormente. Por outro lado, ao longo de todo o processo foram feitos ajustes e tratados casos particulares encontrados.

Para passar à captação do texto, foram inicialmente guardados os textos correspondentes às introduções de cada título, através da função *findall* e da expressão '*Introdução</b>(.\*)<b>'*. A utilização desta função permite ter uma lista com todas estas introduções que serão adicionadas posteriormente ao dicionário de cada título.

Quanto ao texto constituído pelos títulos, subtítulos e legendas, este foi dividido em dois textos, *texto\_1* correspondente ao texto da secção 1, "Sistema Esquelético e Muscular", e *texto\_2*, correspondente ao texto da secção 2, "Sistema Muscular". Isto foi conseguido através da função *findall* e da expressão: '*££££££££££<b>(.+?)££££££££££<b>(.+)*'.

Assim, cada secção foi tratada separadamente, ainda que, de forma muito semelhante.

Na secção 1, o texto foi, inicialmente, dividido pelos títulos. Cada texto obtido foi então dividido pelos subtítulos e retiradas as informações importantes, nomeadamente, a alínea e descrição ("a) ..."). Esta divisão dos textos foi feita através de 3 ciclos *for*, utilizando três padrões principais e as marcações estabelecidas inicialmente, para a captação dos mesmos. O primeiro ciclo encontra o texto entre os títulos, por exemplo, o texto entre "1. CRÂNIO" e "2. MEMBRO SUPERIOR", através da função *findall* e a expressão: '*<b>titulos[j]/b>@@@@@ @@@@@@@@@@(.\*)s\*<b>titulos[j+1]*'. Esta ciclo recorre a uma lista *titulos*, construída previamente através de outro padrão de procura, percorrendo-a.

Após se obter o texto entre os títulos, é necessário obter o texto entre cada subtítulo. Para isso é utilizado outro ciclo *for* que percorre, neste caso, a lista de subtítulos, utilizando a expressão '*<b>subtitulos[i]/b>§(.\*)?<b>subtitulos[i+1]*'.

Por fim, com o texto de cada subtítulo separado, procedeu-se à leitura de cada linha alínea-legenda, como, por exemplo, "a) Osso frontal", separando a alínea e legenda através de grupos de captura na expressão de procura '*()(.+)?'* na função *findall*.

À medida que os títulos e subtítulos foram utilizados, bem como encontradas as legendas,

foram sido criadas as chaves e valores de cada dicionário. Foi também criada a chave “Introdução” e valor com o respetivo texto, adicionando este par ao dicionário correto.

É importante ressaltar que o facto de se utilizarem ciclos *for* comparando dois índices, não permitiu que o último índice tanto do título, como do subtítulo fosse tratado e captada o seu texto. Assim, foi necessário fazer esse tratamento fora do ciclo, ainda que, de forma equivalente.

Quanto ao título “4. TORSO” da secção 1, “Sistema Esquelético E Articular”, é importante referir que este continha subtítulos com sub-subtítulos, como mostrado na Figura 10. Assim, aquando da captação do texto de cada subtítulo, foi criada uma condição capaz de verificar se o subtítulo continha sub-subtítulos. Em caso negativo, o procedimento será igual ao dos subtítulos anteriores. Em caso afirmativo, cada subtítulo é tratado como um título, e é feito um novo ciclo *for*.

Como referido anteriormente, o mesmo raciocínio foi utilizado para a secção 2.

Concluído todo o processo, a estrutura obtida foi a idealizada, e guardada em formato JSON, como se mostra nas Figuras 11, 12 e 13.

```
{
  "SISTEMA ESQUELÉTICO E ARTICULAR": {
    "1. CRÂNIO": {
      "1.1. CRÂNIO: VISTA ANTERIOR - I ": {
        "a": " Osso frontal",
        "b": " Osso parietal",
        "c": " Osso temporal",
        "d": " Osso esfenóide",
        "e": " Osso nasal",
        "f": " Osso zigomático",
        "g": " Osso maxila",
        "h": " Osso mandíbula"
      },
      "1.2 CRÂNIO: VISTA ANTERIOR - II": {
        "a": " Osso frontal",
        "b": " Forame supraorbital",
        "c": " Ossos nasais",
        "d": " Lâmina perpendicular do osso etmoide",
        "e": " Osso zigomático",
        "f": " Osso vômer",
        "g": " Osso maxila",
        "h": " Processo estiloide do osso temporal",
        "i": " Ângulo da mandíbula",
        "j": " Mento"
      }
    }
  }
}
```

Figura 11: Estrutura em JSON com informação do documento *Ossos*

```

"SISTEMA ESQUELÉTICO E ARTICULAR": {
  "4. COLUNA VERTEBRAL E TÓRAX": {
    "4.4 VÉRTEBRAS CERVICAIS ATÍPICAS: ÁXIS (C2)": {
      "4.4.1 VISTA PÓSTERO-SUPERIOR": {
        "a": " Processo odontoide (dente) de C2",
        "b": " Face articular para o arco anterior de C1",
        "c": " Corpo de C2",
        "d": " Face articular superior",
        "e": " Processo transverso",
        "f": " Processo articular inferior",
        "g": " Processo espinhoso bifido",
        "h": " Lâmina",
        "i": " Pedículo",
        "j": " Forame vertebral"
      },
      "4.4.2 VISTA LATERAL": {
        "a": " Processo odontoide (dente) de C2",
        "b": " Pedículo ",
        "c": " Face articular superior",
        "d": " Processo articular superior",
        "e": " Forame transversário ",
        "f": " Processo transverso",
        "g": " Face articular inferior",
        "h": " Processo articular inferior",
        "i": " Lâmina",
        "j": " Processo espinhoso bifido"
      }
    }
  },
}

```

Figura 12: Apresentação dos subtítulos e respectivos sub-subtítulos na estrutura JSON implementada

```

"SISTEMA MUSCULAR": {
  "2. MEMBRO SUPERIOR": {
    "Introdução": "A musculatura do membro superior é dividida em músculos que posicionam o cingulo superior, músculos que movimentam o braço, músculos que movimentam o antebraço e a mão e músculos que movimentam as mãos e dos dedos. Os músculos que posicionam o cingulo superior são: levantador da escápula, peitoral menor, romboides, serrátil anterior, subclávio e trapézio. Estes trabalham em conjunto com a musculatura que movimentam o braço que é formada pelos músculos: coracobraquial, deltoide, supraespal, infraespal, subescapular, redondo maior, redondo menor, latíssimo do dorso e peitoral maior. Os tendões dos músculos supraespal, infraespal, subescapular e redondo menor unem-se à cápsula da articulação do ombro, formando o chamado "manguito rotador", que reforça e sustenta a articulação. A musculatura que movimentam o antebraço e a mão geralmente tem origem no úmero, sendo formada pelos seguintes músculos: braquial, braquiorradial, biceps braquial, tríceps braquial, anconeio, flexor ulnar do carpo, flexor radial do carpo, palmar longo, pronador quadrado, pronador redondo, supinador, extensor ulnar do carpo, extensor radial longo do carpo e extensor radial curto do carpo. Em geral, a musculatura extensora está localizada na superfície posterior e lateral do antebraço, enquanto que a musculatura flexora está localizada nas porções anterior e medial do antebraço. Os músculos que movimentam a mão e os dedos são divididos em músculos intrínsecos e extrínsecos da mão. Os músculos extrínsecos que movimentam os dedos estão localizados no antebraço: abductor longo do polegar, extensor dos dedos, extensor curto do polegar, extensor longo do polegar, extensor do indicador, extensor do dedo mínimo, flexor superficial dos dedos, flexor profundo dos dedos e flexor longo do polegar. A musculatura intrínseca é formada pelos músculos da mão: adutor do polegar, oponente do polegar, palmar curto, abductor do dedo mínimo, abductor curto do polegar, flexor curto do dedo mínimo, flexor curto do polegar, oponente do dedo mínimo, lumbricais, interosseos dorsais e interosseos palmares."
  }
}

```

Figura 13: Apresentação da Introdução do título “2. MEMBRO SUPERIOR” da seção “Sistema Muscular” na estrutura JSON implementada

### 1.3 Documento 3: Minidicionário de Cardiologia

O documento *Minidicionário de Cardiologia* contém designações em inglês e português relacionadas a termos e conceitos da área de cardiologia, juntamente com as traduções correspondentes, em português e inglês.

Para proceder à limpeza de elementos irrelevantes no documento, foram utilizadas expressões regulares para removê-los. Desta forma, foi realizada a remoção de *tags* XML, como `<?xml>`, `</page>`, `</text>`, `<fontspec>` e `<image>`, entre outras. Além disso, foram removidas quebras de linha e espaços em branco desnecessários.

Posteriormente, para uma melhor organização da informação no documento, foi definida a seguinte sintaxe para representar a estrutura de dados pretendida:

```

{
  "en:pt": {
    "Designação em Inglês": " Tradução em Português",
    ...
  },
  "pt:en": {
    "Designação em Português ": " Tradução em Inglês ",
    ...
  }
}

```

Após uma análise do documento, foi descoberto que texto dentro das *tags* `<font="7">` correspondia a uma designação em inglês, texto dentro das *tags* `<font="13">` indicava uma designação em inglês e texto entre as *tags* `<font="8">` se tratava de uma tradução em português ou inglês. Assim, de forma a facilitar a extração e manipulação da informação, foram criadas marcas para delimitar cada tipo de elemento:

1. **Designação em inglês:** O texto dentro das *tags* `<font="7">` foi inicializado pela marca '£'.
2. **Designação em português:** O texto dentro das *tags* `<font="13">` foi inicializado pela marca '€'.
3. **Traduções:** O texto dentro das *tags* `<font="8">` foi inicializado pela marca '&'.

Para o tratamento de possíveis situações em que os termos possuem mais de uma linha, foi desenvolvido um ciclo *while* que itera sobre a lista de marcas '£', '€' e '&'. Para cada marca, é definido um padrão de expressão regular que captura o conteúdo entre duas ocorrências da mesma marca, considerando que haja uma ou mais quebras de linha entre elas.

Na Figura 14, é possível observar o resultado da substituição e do tratamento mencionados anteriormente.

```
£A SURMISE (A CONJECTURE SUSP[C]ION) (TO ASSUME ON SMALL EVI[D]DENCE)
& 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 CON[D]DUCTION)
&Não significa Acidente Vascular Cerebral, um AVC em inglês é um "stroke"
€Imagem por ressonância magnética
&M.R.I. MAGNETIC RESONANCE IMAGING
€Implantar / Estender / Desenrolar / Abrir
& DEPLOY
€Impressão digital
&FINGER PRINT
€Impulso Simpático Central
&CENTRAL SYMPATHETIC DRIVE
€Inalar
&INHALE (TO)
```

Figura 14: Marcas utilizadas no texto xml do documento *Minidicionário de Cardiologia* .

Após o processamento do texto, foram utilizadas 2 expressões regulares, juntamente com o método *findall* para extrair as informações correspondentes às designações e traduções, resultando em 2 listas de tuplos.

Após a conversão de cada uma delas para um dicionário, os dicionários resultantes foram combinados num dicionário final com as chaves 'en:pt' e 'pt:en', representando as traduções bidirecionais entre inglês e português. No total, foram extraídas 1184 designações.

Finalmente, o dicionário final foi adicionado num ficheiro JSON, *card.json*, obtendo-se o ficheiro representado na Figura 15.



```

{
  "en:pt": {
    "A SURMISE (A CONJECTURE SUSP[ ] CION) (TO ASSUME ON SMALL EV[ ] DENCE) ": " 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 CON[ ] DUTION) ": "Não significa Acidente Vascular Cerebral, um AVC em inglês é um " stroke",
    "ABDOMINAL FEVER / TYPHOID FEVER ": "Febre tifoide",
    "ABSENTMINDED ": "Distraído / Desatento",
    "ACCEPTANCE ": "Novos pacientes / Aco- lhimento",
    "ACCRUING ": "Vir de maneira natural / Incre mentado de maneira natural / Devido a / Proveniente de",
    "ACHE ": "Termo para designar casos particulares de dor (dor contínua, que não passa, mas não é severa) / Dor localizada",
    "pt:en": {
      "1ª / 2ª / 3ª / 4ª / Bulha ": " 1 / 2 / 3 / 4 / SOUND",
      "A esquerda / Ao contrário dos ponteiros do relógio / Sentido anti[ ] horário ": "COUNTERCLOCKWISE",
      "A faculdade de fazer descobertas in[ ] portantes e valiosas de maneira ines[ ] perada 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 / Preme[ ] dita damente ": "OVERTLY",
      "Abordagem / uma aproximação / aproxi[ ] marse ": "APPROACH",
      "Abrangente / Extensivo / Que engloba ": "COMPREHENSIVE",
      "Acocoramento / Agachamento ": " SQUAT- TING (POSITION)",
      "Acoramento / Agachamento ": " SQUAT- TING (POSITION)",
    }
  }
}

```

Figura 15: *card.json*

#### 1.4 Documento 4: Glossário Termos Médicos Técnicos e Populares

O quarto documento analisado e processado consiste num glossário de termos médicos técnicos e populares em português.

Após a análise detalhada da sua estrutura em pdf foi concluído que existiam termos seguidos da sua respetiva definição, assim como o inverso, definições seguidas do termo a elas associado, e um pequeno texto inicial contendo informações relativas à fonte deste documento.

Tendo em conta que este texto inicial era um caso excecional, pois só se verificava uma vez no ficheiro, procedeu-se à sua remoção manual, reduzindo assim a carga computacional associada à inserção de comandos para a sua extração recorrente.

Adicionalmente, foi verificada a existência de termos repetidos ao longo do ficheiro, o que seria necessário remover.

Deste modo, investigou-se a estrutura do documento em xml e txt, chegando-se à conclusão que o meio mais apropriado para efetuar o processamento deste ficheiro seria recorrer à sua versão de texto (txt), sendo pretendido o armazenamento do resultado final no formato:

```

{
  "Designação" : "Definição",
  ...
}

```

Tendo escolhido a estrutura a utilizar, recorreu-se à análise e ao tratamento linha a linha do documento, visto que o formato dos pares designação, definição diferiam ao longo deste,

encontrando-se ora na ordem designação, definição, ora definição, designação.

Posteriormente, foi averiguado o padrão de separação da designação e da definição por “ , “, assim como a presença da expressão “(pop)” na última (Figura 16).

```
a milionésima parte de um grama (pop) , micrograma
à volta da boca (pop) , perioral
à volta da órbita (pop) , periorbital
à volta dos vasos sanguíneos (pop) , perivascular
abaixamento, abatimento, prostração (pop) , depressão
abcesso , abcesso, tumor (pop)
abcesso, tumor (pop) , abcesso
abcesso; acumulação de pus (pop) , empiema
abdômen , barriga, ventre (pop)
abdominal , ventral (pop)
aberrante , anormal (pop)
abertura; orifício (pop) , perfuração
ablação (pop) , extracção
ablação dos órgãos sexuais, capação, eviração, emasculação (pop) , castração
abocamento (pop) , anastomose
abortamento, desmancho (pop) , aborto
aborto , abortamento, desmancho (pop)
```

Figura 16: Marcação utilizada no documento *Glossário Termos Médicos Técnicos e Populares*.

Assim, para poder captar o conjunto na sua totalidade, sem que fossem desconsideradas linhas adicionais da definição, foi efetuada a procura por estas duas expressões na linha. Caso a linha captada não possuísse ambos os padrões reconhecidos, significaria que a definição teria linhas adicionais, procedendo-se à concatenação de linhas consecutivas até que se garantisse a condição anterior, guardando numa lista o resultado.

Durante este processo houve, também, a preocupação da remoção de quebras de linha “\n” e de quebras de página “\f” para que o texto ficasse corretamente formatado.

De forma a se obter o formato final desejado, seria ainda necessário separar a designação da definição, sendo que tal foi efetuado pela divisão de cada conjunto pela “ , “, recorrendo ao módulo de strings `split()`. Em seguida, procurou-se onde se encontraria o padrão “(pop)”, alocando-o ao valor e a parte restante à chave de um dicionário.

Por fim, este dicionário foi ordenado por ordem alfabética das chaves e inserido num ficheiro JSON (Figura 17).

```
{
  "blister": "frasco de X comprimidos recobertos de plástico (pop)",
  "(d)escamação": "formação excessiva de escamas na pele (pop)",
  "(herpes) zóster": "vírus instalado à volta das células sensitivas (pop)",
  "ACTH": "hormônio adreno-córticotrófico, corticotrofina (pop)",
  "Gram-negativo": "que não toma o corante de Gram (pop)",
  "Gram-positivo": "que toma o corante de Gram (pop)",
  "Petit mal; epilepsia menor": "pequeno mal, epilepsia com ataques pouco intensos (pop)",
  "abcesso": "abcesso, tumor (pop)",
  "abdominal": "ventral (pop)",
  "abdómen": "barriga, ventre (pop)",
  "aberrante": "anormal (pop)",
  "aborto": "abortamento, desmancho (pop)",
  "abrupto": "repentino, brusco (pop)",
  "absorção": "absorvimento, absorvência (pop)",
  "abstinência": "jejum (pop)",
  "acatisia": "incapacidade em permanecer sentado (pop)",
  "acidental": "por acaso, sem importância (pop)",
  "acidez": "acidade, azedume (pop)",
  "acidose": "alteração do equilíbrio ácido básico do sangue e líquidos teciduais (pop)",
  "acinesia": "ausência de movimento, acinese (pop)",
  "acne": "espinha (pop)",
  "acomodação": "adaptação (pop)",
  "acrocianose": "cor azulada das extremidades (mãos-pés) (pop)",
}
```

Figura 17: Estrutura final em JSON do *Glossário Termos Médicos Técnicos e Populares* .

## 2 Conclusão

Este trabalho explorou os fundamentos da unidade curricular Processamento de Linguagem Natural (PLN), destacando a aplicação de expressões regulares para extrair informações de documentos textuais e armazená-las em uma estrutura JSON. Durante o desenvolvimento, foi evidente o potencial das expressões regulares na identificação de padrões específicos, permitindo a extração precisa de informações relevantes. No entanto, é importante reconhecer que, apesar da eficácia das técnicas utilizadas, ainda há margem para melhorias na otimização dos algoritmos desenvolvidos para cada documento.

Um dos principais desafios encontrados foi garantir a robustez e generalização das expressões regulares diante de diferentes formatos e variações nos textos dos documentos analisados. Nesse sentido, estratégias para lidar com casos particulares e aprimorar a flexibilidade dos padrões definidos podem contribuir significativamente para a eficácia e precisão do processo de extração de informações.

Em suma, embora as expressões regulares tenham demonstrado ser uma ferramenta valiosa para a extração de informações em documentos textuais, é importante promover a continuidade da pesquisa e o desenvolvimento de técnicas mais eficazes de forma a melhorar a precisão, robustez e escalabilidade do sistema de processamento de linguagem natural em futuras aplicações.