University of Minho
School of Engineering

Trabalho Prático 1
Extração de Termos de Documentos Médicos

Processamento de Linguagem Natural

Daniel Sá (PG56120)

João Cardoso (PG56135)

Jorge Costa (PG56136)

Pedro Flores (A100475)

april 2025

Índice

1 Contextualização

A necessidade de processamento e análise de grandes volumes de dados textuais é uma realidade crescente em diversas áreas, e a extração de informação relevante de documentos é um passo fundamental para permitir a utilização desses dados em aplicações como pesquisa, análise de tendências e suporte à decisão. Por conseguinte, no âmbito da disciplina de Processamento de Linguagem Natural da Engenharia Biomédica, foi desenvolvido um sistema de extração de informação de documentos em formato PDF pré-definidos.

O presente trabalho prático teve como objetivo a aplicação dos conhecimentos teóricos e práticos de Processamento de Linguagem Natural na construção de um conjunto de ferramentas (parsers) capazes de identificar e extrair informação relevante de documentos. A informação extraída foi organizada, produzindo, em última análise, uma coleção abrangente de recursos estruturados em JSON de terminologia médica.

Este trabalho pretende abordar vários desafios fundamentais no processamento e análise de grandes volumes de dados:

- **Acessibilidade da informação:** Ao converter glossários em formatos JSON estruturados, o conteúdo torna-se facilmente acessível para integração em várias aplicações, como registos de saúde eletrónicos, portais de pacientes, plataformas de educação médica e serviços de tradução.
- **Interoperabilidade:** A terminologia médica normalizada suporta uma melhor interoperabilidade entre diferentes sistemas e aplicações de cuidados de saúde, permitindo uma partilha perfeita de informações em todo o ecossistema de cuidados de saúde.
- **Suporte multilingue:** O formato estruturado facilita os futuros esforços de tradução e o suporte multilingue, particularmente importante em ambientes linguísticos diversos.

2 Sistema de Extração de informação

2.1 Glossário Neologismos Saúde

Para preparar o documento para processamento, foi realizada a sua conversão para formato de texto. Este procedimento foi efetuado através da execução do comando "pdftotext" na linha do terminal.

2.1.1 Limpeza do documento e marcação de conteúdo

Após a análise do conteúdo do ficheiro, foi definida a estrutura do ficheiro JSON para o armazenamento da informação extraída. Deste modo, foi determinado que o JSON deveria armazenar a meta-informação do documento, bem como os termos identificados e a informação associada a cada um.

Inicialmente, procedeu-se à limpeza do texto. As quebras de página foram substituídas por "\bn", e múltiplas ocorrências consecutivas desse caractere, "\n+" foram reduzidas a uma única ocorrência "\bn". Após esta etapa de limpeza, o conteúdo relevante foi marcado para facilitar a sua posterior extração. Para destacar a informação de interesse, foram inseridas marcas delimitadoras antes e depois de cada ocorrência dos seguintes elementos: o nome do autor (##AUTOR##), o título do documento (##TITULO##), o ano de publicação (##ANO_PUBLICACAO##), o resumo (##RESUMO##), as palavras-chave (##PALAVRAS-CHAVE##) e os termos e respetivas informações (##ENTRADA##).

2.1.2 Extração da metadata

Concluída a marcação do conteúdo, iniciou-se a fase de extração. A meta-informação foi o primeiro conjunto de dados a ser extraído. A extração desta informação revelou-se direta, dado que se encontrava delimitada por marcadores específicos, requerendo apenas a localização e obtenção do texto compreendido entre eles. Os dados extraídos foram armazenados num dicionário.

2.1.3 Extração dos termos e informação associada

A extração dos termos e da sua informação associada foi realizada seguindo um conjunto de etapas específicas.

Inicialmente, localizou-se o texto delimitado pelo marcador ##ENTRADA##. De cada segmento de texto assim identificado, extraíram-se primeiramente o nome do termo e a sua classe gramatical. Dado que estas informações se encontravam numa única linha, o nome do termo foi obtido desde o início da linha até ao ponto anterior à classe gramatical, enquanto a classe gramatical foi extraída desde o caractere seguinte ao último do nome até ao final da linha.

As traduções do termo foram extraídas da linha subsequente à que continha o nome e a classe gramatical. Após a extração, foi aplicada uma separação aos resultados para distinguir as traduções em inglês das traduções em espanhol.

A sigla, quando presente, é precedida pela sequência de caracteres "Sigla:". Consequentemente, a sigla foi extraída como o texto subsequente a esta sequência até à mudança de linha.

A extração da definição incluiu a recuperação da sigla, caso esta esteja presente e por vezes parte da tradução. Posteriormente, foi realizado um processamento do texto para remover as partes irrelevantes para o caso. Nos casos em que o campo da definição continha uma referência para consultar outro termo, a informação extraída foi processada de modo a indicar a necessidade de consulta do termo referenciado.

A informação da enciclopédia, quando presente, é antecedida por "Inf encicl". O conteúdo a extrair foi, portanto, o texto que se seguia a esta sequência.

As abonações constituíram a última informação a ser extraída. Observou-se que as abonações apresentavam inconsistências na delimitação, utilizando diferentes tipos de aspas e pontuação.

Para assegurar a qualidade dos dados armazenados, toda a informação extraída foi processada utilizando as funções de *replace* e *strip*, de modo a remover caracteres indesejáveis.

A informação extraída foi organizada numa lista de dicionários, onde cada dicionário representava um termo individual.

2.2 Diccionari Multilinguee de la Covid 19

O documento em análise apresenta a informação disposta em duas colunas, o que inviabiliza a sua conversão direta para texto simples e dificulta o seu processamento automático. Para contornar esta limitação, procedeu-se à conversão do ficheiro PDF para o formato XML. Esta transformação permitiu o aproveitamento do atributo left das tags XML — indicador da posição horizontal do texto — para definir um ponto médio (valor numérico correspondente aproximadamente a metade da largura da página). Com base nesse ponto médio, foi possível determinar a que coluna pertencia cada linha de texto e, assim, reordenar o conteúdo do documento segundo a ordem natural de leitura.

2.2.1 Pré-processamento

Numa fase inicial, o processo de pré-processamento centrou-se na extração exclusiva do conteúdo relativo às páginas do dicionário multilingue, descartando outras secções irrelevantes para os objetivos do trabalho.

2.2.2 Limpeza do documento e marcação de conteúdos

Nesta etapa, o foco foi a reestruturação da informação e a introdução de marcas no texto, com o objetivo de facilitar a extração dos dados e distinguir os diferentes parâmetros associados a cada conceito do dicionário.

Foram utilizados caracteres especiais inexistentes no documento original (§§, ##, ▯ e @@) como marcadores auxiliares. Estes permitiram assinalar elementos como o nome do conceito, o identificador único, os diferentes tipos de traduções, classes gramaticais, definições, entre outros. A natureza semi-estruturada do conteúdo, após o parsing, possibilitou ainda o reaproveitamento de marcas linguísticas já presentes no texto para realizar uma separação mais granular da informação.

A limpeza consistiu na remoção de espaços em branco desnecessários e na exclusão das tags XML, já sem utilidade após a reorganização das linhas. Preservou-se apenas o conteúdo textual, devidamente marcado com os símbolos definidos.

2.2.3 Extração do conteúdo

A extração iniciou-se com a identificação dos diferentes conceitos presentes no documento e a separação dos respetivos conteúdos. Esta identificação baseou-se num padrão característico: uma linha composta exclusivamente por dígitos, seguida de zero ou uma linha vazia, uma ou duas linhas de texto, e a presença da marca '##', utilizada para assinalar classes gramaticais. Este padrão corresponde ao título do conceito (em catalão) e ao seu identificador único — ambos extraídos nesta fase.

Uma vez delimitado o conteúdo de cada conceito, procedeu-se à identificação de secções opcionais, como sinónimos, variantes ou referências a outros conceitos. Cada uma destas secções foi transformada numa chave distinta no dicionário final.

As traduções foram extraídas com base no marcador §§ e utilizando o carácter ";" como separador de múltiplas traduções dentro da mesma língua. A marca '##' foi novamente útil para associar cada tradução à respetiva classe gramatical, com exceção da língua árabe, cuja informação gramatical não

está presente no documento original.

As restantes secções seguiram estratégias de extração semelhantes. Para garantir que todos os tipos de parâmetros relevantes fossem reconhecidos, foram definidas constantes com os nomes possíveis para cada tipo de conteúdo, com base na secção de abreviações do dicionário e na árvore de campos do documento.

2.2.4 Organização da informação extraída

O resultado final da extração apresenta-se sob a forma de um dicionário por conceito, contendo:

- O identificador único do conceito sob a forma de número inteiro;
- As traduções em cada língua, incluindo o título (tratado como tradução em "ca"), organizadas como listas de dicionários com os pares "tradução" e "classe gramatical";
- Os parâmetros opcionais (ex. "CAS", "veg.", "sin."), incluídos apenas quando presentes, representados como chaves adicionais;
- Parâmetros com múltiplas entradas representados como listas (de strings ou dicionários, consoante a necessidade de distinguir a classe gramatical);
- A definição do conceito (em catalão), incluída como string quando disponível;
- A secção de notas, representada como uma lista de strings.

2.3 Glossário do Ministério da Saúde

Para preparar o documento para processamento, foi realizada a sua conversão para formato de xml. Este procedimento foi efetuado através da execução do comando "pdftohtml –xml" na linha do terminal.

2.3.1 Limpeza do documento

Para a extração das Siglas, começou por se fazer a limpeza e marcação do texto, removendo hífen ('-') e espaços brancos antes e depois deles. Foi extraído do documento apenas a porção que contem as siglas, sendo este segmento delimitado pela primeira sigla, AB e por </page> após a última sigla.

2.3.2 Extração de siglas

Após definir o segmento que contém a informação desejada, as siglas e as suas definições são extraídas. A extração desta informação revelou-se direta, estando as siglas em bold, delimitadas por e , sendo as suas definições igualmente fáceis de encontrar por se encontrarem entre a sigla a que pertencem e a sigla seguinte.

2.3.3 Extração de termos

A extração de termos não utiliza a limpeza do documento realizada no início, sendo a informação desejada num segmento diferente do documento. Os termos são definidos por três componentes: nome do termo, a sua categoria e a sua descrição.

Como o nome e a descrição podem estar contidos em uma ou mais linhas, foi utilizado *ElementTree* para facilmente detetar informação como elementos em bold.

A estrutura dos termos contêm algumas exceções. O nome encontra-se em bold, seguido de "Categoria:" em itálico, com a categoria do termo na linha seguinte e por fim, a descrição, sendo esta, o restante texto até ao próximo texto bold (nome do termo seguinte), encontrando-se sempre duas linhas após "Categoria:" devido à categoria nunca ocupar mais de uma linha. Utilizando esta regra, o texto foi extraído diretamente.

Porém, existem algumas exceções entre os termos que seguem esta estrutura, existem alguns casos em que não apresentam categoria e a sua descrição referencia um termo diferente: Ver "termo" (ex. "Ver Alta Complexidade"). Sendo que os termos serão ordenados por categoria, foi necessário desenvolver uma solução para estes casos.

Para tal, foi adicionada uma exceção no código: caso não encontre categoria, atribua a categoria "Sem Categoria" a esse termo, sendo o texto encontrado de seguida até ao nome do seguinte termo definido.

2.4 Glossário de Termos Técnicos e Populares

O script `extract_glossário_de_termos_técnicos_e_populares.py` extrai, com o auxílio de expressões regulares, termos médicos técnicos e as suas correspondentes definições "populares" de um ficheiro de texto de um glossário português, convertendo esta informação num formato JSON estruturado.

2.4.1 Limpeza do documento e marcação de secções

A primeira ação era normalizar os espaços em branco, substituindo um ou mais caracteres de nova linha consecutivos por uma única nova linha. Também removendo os caracteres de alimentação de formulário.

O segundo objetivo tem como alvo detetar letras maiúsculas isoladas que estão rodeadas por caracteres de nova linha. Inserindo uma string "###" imediatamente antes e depois dessas letras maiúsculas isoladas. Assim conseguimos criar blocos de texto facilmente isoláveis.

Por último, foi necessário processar cada linha dentro de cada bloco de texto para tratar de casos mais raros antes de começar a extração de termos. É preciso tomar decisões sobre como tratar cada linha com base no fato de conter ou não "(pop)" e com base no seu comprimento.

- Para linhas sem "(pop)" que excedam 35 caracteres, a função move-as para o início da linha seguinte.
- Para linhas mais curtas (menos de 35 caracteres) que não têm "(pop)", anexa-as à linha anterior.

Este processo é fundamental para tratar de casos mais raros.

2.4.2 Extração de blocos e termos

A Extração de blocos identifica blocos de conteúdo entre os marcadores, anteriormente mencionados, e analisa pares termo-definição utilizando uma correspondência de padrões, ao mesmo tempo que lida com múltiplas variações de formatação presentes no documento de origem.

As expressões regulares utilizadas nesta tarefa, foram concebidas para analisar as linhas dentro do bloco que correspondem a uma de duas estruturas específicas que o seguintes padrões:

- Estrutura 1: uma descrição do termo (pop) , termo que pode ter espaços.
- Estrutura 2: termo que pode ter espaços , uma descrição do termo (pop).

O resultado armazenado das entradas será uma lista de tuplos. Cada tuplo terá 4 elementos, correspondentes aos 4 grupos de captura: (Grupo 1, Grupo 2, Grupo 3, Grupo 4).

- Se uma linha corresponder à estrutura 1: O tuplo terá o seguinte aspeto ('uma descrição do termo', 'termo que pode ter espaços', '', ''). Os grupos 3 e 4 não participaram, pelo que são cadeias de caracteres vazias.

- Se uma linha corresponder à Estrutura 2: O tuplo terá o seguinte aspeto (‘’, ‘’, ‘termo que pode ter espaços’, ‘uma descrição do termo’). Os grupos 1 e 2 não participaram.

3 Concatenação do Conteúdo

A concatenação do conteúdo foi concebida para consolidar 4 glossários e dicionários médicos guardados em ficheiros JSON diferentes que contêm terminologia médica.

O processo de fusão começa por utilizar o glossário de termos médicos técnicos e populares como base e depois incorpora a secção das siglas do glossário do ministério da saúde. Para os termos do glossário do ministério da saúde, acrescenta os que ainda não existem no dicionário de base.

Os termos do glossário neologismos de saúde são então integrados, preservando as suas estruturas e metadados específicos.

A integração do dicionário multilingue da covid-19 é a parte mais complexa do processo. O código extrai cuidadosamente termos portugueses deste dicionário (tratando variações como o português do Brasil e o português europeu) e adiciona-os como novas entradas, se não existirem, ou enriquece as entradas existentes com informações adicionais. Para cada termo, o código preserva as classificações gramaticais, definições, traduções em várias línguas (catalão, occitano, basco, galego, espanhol, inglês, francês, holandês e árabe), notas e números de registo CAS, quando disponíveis.

4 Conclusão

Os scripts analisados representam uma abordagem para transformar glossários médicos pouco estruturados em formatos legíveis por computador. Como parte de uma iniciativa mais ampla de normalização da terminologia, este trabalho contribui para melhorar a comunicação e a acessibilidade da informação sobre cuidados de saúde. As técnicas de expressão regular empregues demonstram capacidades sofisticadas de correspondência de padrões, embora o tratamento adicional de erros possa aumentar a robustez para ambientes de produção.