



Relatório - Trabalho Prático 1

Processamento de Linguagem Natural

Mestrado em Engenharia Biomédica - Informática Médica

Grupo :

PG57810, Dinis Mesquita

PG52290, Flávio Ribeiro

PG56152, Sofia Correia

Docentes:

Luís Filipe Cunha

José João Almeida



Índice

1	Introdução	2
2	Documentos	2
2.1	Glossário de Neologismos Terminológicos da Saúde Humana	2
2.2	Diccionari multilingüe de la COVID-19	2
2.3	Glossário de Termos Médicos Técnicos e Populares	3
3	Estrutura da informação	3
4	Tratamento de dados	3
4.1	Glossário de Neologismos Terminológicos da Saúde Humana	3
4.2	Diccionari multilingüe de la COVID-19	5
4.3	Glossário de Termos Médicos Técnicos e Populares	7
5	Conclusão	9

1 Introdução

Neste projeto aplicaram-se técnicas de Processamento de Linguagem Natural (PLN) envolvendo a extração e manipulação de informação a partir de documentos em formato PDF. Para tal, são desenvolvidos parsers capazes de identificar e estruturar dados relevantes, recorrendo a expressões regulares e outras estratégias de pré-processamento. A informação extraída é posteriormente organizada num ficheiro em formato JSON. O projeto é implementado em Python e para além de envolver o processamento obrigatório de ficheiros fornecidos, como o Dicionari Multilingue de La Covid 19 e o Glossário de Neologismos Terminológicos da Saúde Humana, foi também processado o documento Glossário de Termos Médicos Técnicos e Populares.

2 Documentos

2.1 Glossário de Neologismos Terminológicos da Saúde Humana

O Glossário de Neologismos Terminológicos da Saúde Humana é um dos dois glossários propostos para este projeto, com categoria obrigatória. Este inicialmente fornecido em formato pdf, apresenta uma estrutura de tese, fator levado em consideração no processamento do mesmo. Consta com 306 conceitos e cada um dos mesmos apresenta as traduções inglesas e espanholas, o significado, significado enciclopédico e em alguns casos um contexto textual. Parâmetros estes que podem ser verificados no exemplo na Figura 8.

adjuvante genético *s.m.*
genetic adjuvant [ing]; adjuvante genético [esp]
Componente viral que potencializa os efeitos da *vacina de DNA* que pode ser utilizada contra a brucelose.
Inf. encicl.: o adjuvante genético também pode ser utilizado em *terapias genéticas* em humanos.
“...composta de um ou mais genes da bactéria causadora da doença, associados a um componente viral - o chamado adjuvante genético, que potencializa os efeitos da vacina....” (213)

Figura 1: Exemplo estrutural do Glossário de Neologismos Terminológicos da Saúde Humana

2.2 Dicionari multilingüe de la COVID-19

Edocumento, escrito em catalão, aborda termos associados à pandemia, incluindo um número identificador, a classificação da palavra, o(s) termo(s) equivalente em diversas outras línguas, a área temática e a definição, bem como os sinónimos em catalão, quando aplicável.

Em termos de estrutura, o documento é bastante complexo, sendo constituído por uma parte inicial, onde são abordadas as motivações e as entidades que permitiram a publicação do dicionário, o prefácio, o prólogo, a introdução, a lista contendo as diferentes áreas temáticas e a lista das abreviações utilizadas. Depois, o dicionário propriamente dito e, por fim, um conjunto de índices, que associam cada palavra (separadas por cada língua e cada área temática) ao número identificador da palavra.

Algumas particularidades do documento são a língua em que está escrito (catalão), e o facto de o dicionário estar maioritariamente organizado em duas colunas por página.

2.3 Glossário de Termos Médicos Técnicos e Populares

O terceiro documento escolhido foi o *Glossário de Termos Médicos Técnicos e Populares.pdf*.

Este glossário é bastante simples em termos de estrutura, apresentando palavras e expressões populares, identificadas por um (*pop*), associadas a conceitos mais técnicos, sendo todas as entradas organizadas alfabeticamente por um índice que se apresenta no início de cada divisão.

Este documento apresenta várias inconsistências, sendo das mais notáveis a repetição de entradas no glossário e, por vezes, apresenta só uma palavra ou expressão seguida de uma vírgula e de um (*pop*), tornando-se ambíguo do que se trata. Estas últimas ocorrências são raras e, por isso, durante a fase de tratamento de dados, foram retiradas manualmente.

Existe também palavras em negrito que se supõe serem os termos técnicos usados, mas por falta de fontes e desatualização das referências apresentadas no documento, considerou-se somente duas diferentes classificações de palavras em cada entrada do glossário: termos ou expressões populares e conceitos técnicos.

3 Estrutura da informação

Depois de consultar as diferentes informações e estruturas dos variados documentos, pretendia-se organizar toda a informação relevante num só ficheiro *json*. Este ficheiro *json* deverá ser composto por uma lista de dicionários, em que cada dicionário seguirá a seguinte forma:

"conceito": "neologismo",

Língua :

"traducao_na_respeitva_Lingua_1", "traducao_na_respeitva_Lingua_2", ...

"significado": "definição clara e contextualizada do conceito",

"contexto": "exemplo textual de uso do conceito",

"outras associacoes a 'termo referente'":

"termo_relacionado_1", "termo_relacionado_2", ...

4 Tratamento de dados

4.1 Glossário de Neologismos Terminológicos da Saúde Humana

Para este glossário procurou-se fazer o tratamento em formato html, o que requereu transformar este documento pdf em html com recurso ao comando `pdftohtml` na linha de comandos. Após o documento se encontrar num formato viável, Figura 2, o passo inicial para a limpeza do documento revolveu-se em retirar toda a informação irrelevante ao nosso objetivo. Neste caso é tanto a informação que se encontra antes e depois do intervalo dos conceitos e os respetivos parâmetros como os elementos XML, Figura 3.

```

5 <page number="1" position="absolute" top="0" left="0" height="1263" width="892">
6   <fontspec id="0" size="18" family="GAAIBX+Times" color="#000000"/>
7   <fontspec id="1" size="18" family="YSWOZB+Times" color="#000000"/>
8   <text top="128" left="310" width="320" height="21" font="0"><b>AURI CLAUDIONEI MATOS FRÜBEL </b></text>
9   <text top="148" left="468" width="4" height="21" font="0"><b></b></text>
10  <text top="169" left="468" width="4" height="21" font="0"><b></b></text>
11  <text top="190" left="468" width="4" height="21" font="0"><b></b></text>
12  <text top="210" left="468" width="4" height="21" font="0"><b></b></text>

```

Figura 2: Excerto do ficheiro XML

```

@abeta
» abeta [ing]; abeta [esp]
*Proteína que pode ser encontrada em todos os tipos de células do
organismo humano. Ao acumular-se excessivamente no córtex cerebral
do ser humano pode contribuir para o aceleração do
mal de alzheimer .
£"... Pesquisadores alemães da Universidade de Bonn ajudaram a entender
como a proteína abeta se acumula no córtex cerebral de portadores do
mal de Alzheimer..."

@ação vasoconstritora
» vasoconstriction [ing]; acción vasoconstritora [esp]
*Redução do diâmetro das veias artérias do organismo humano, o que
implica na elevação da pressão sanguínea.
£"...descobriram como atuam diferentes versões dos genes que controlam
a produção de duas enzimas essenciais para a sobrevivência por fazer a
pressão arterial subir ou cair: a enzima conversora da angiotensina
(ECA), que reduz o diâmetro das artérias (ação vasoconstritora) e eleva
a pressão..."

```

Figura 3: Ficheiro após a limpeza e com tags

Num segundo momento, foi essencial incrementar tags únicas no início de cada parâmetro relevante para a nossa extração de informação. Os parâmetros e as respetivas tags deste glossário estenderam-se entre: conceitos (@), traduções (»), significado (*), significado enciclopédico (£), e contexto (£). Por fim, foi criado um script, para armazenar as informações relevantes deste glossário em formato JSON, Figura 4. Realizou-se o split a partir da tag @, de forma a separar cada conceito e o respetivo conteúdo. A partir daí extraiu-se cada parâmetro com base na sua tag.

```

[
  {
    "conceito": "abeta",
    "en": "abeta",
    "es": "abeta",
    "significado": "Proteína que pode ser encontrada em todos os tipos de células do organismo humano. Ao acumular-se",
    "contexto": "... Pesquisadores alemães da Universidade de Bonn ajudaram a entender como a proteína abeta se acumula no"
  },
  {
    "conceito": "ação vasoconstritora",
    "en": "vasoconstriction",
    "es": "acción vasoconstritora",
    "significado": "Redução do diâmetro das veias artérias do organismo humano, o que implica na elevação da pressão sanguínea.",
    "contexto": "...descobriram como atuam diferentes versões dos genes que controlam a produção de duas enzimas essenciais para"
  },
  {
    "conceito": "acidente vascular cerebral isquémico",
    "en": "ischemic cerebrovascular accident",
    "es": "accidente vascular cerebral isquémico",
    "significado": "Sigla: AVCI Lesão que causa a morte de parte do cérebro humano, em virtude da falta de oxigénio e nutriente:",
    "significado_enciclopédico": "Uma vez privados do sangue, os neurónios morrem e liberam glutamato, uma substância química que",
    "contexto": "...Conhecido como acidente vascular cerebral isquémico (AVCI) ou isquemia cerebral, esse problema pode levar à in"
  }
],

```

Figura 4: Documento JSON

No processo de extrair os dados e verificar que toda a informação foi mantida, verificou-se a existência de falhas, relativamente às traduções espanholas. Havendo dois momentos onde

estas traduções não estavam presentes, Figura 5. Este acontecimento deveu-se à incoerência do documento original, onde nestas exceções ou não estava presente a tag total [esp] ou a ausência de certos caracteres da mesma, Figura 6 e Figura 7.

```

10 # Verificar os campos sem "tradução_esp"
11 conceitos_sem_traducao_esp = [
12     conceito["conceito"] for conceito in conceitos
13     if "tradução_esp" not in conceito or not conceito["tradução_esp"].strip()
14 ]
15
16 # Resultado
17 if conceitos_sem_traducao_esp:
18     print("Conceitos sem tradução_esp:")
19     for c in conceitos_sem_traducao_esp:
20         print(c)

```




Figura 5: Conceitos sem tradução espanhola

câncer gástrico *s.m.*
gastric cancer [ing]; **câncer gástrico** [es]
Ver este termo *câncer de estômago*.
“...Finalmente, deve ser comentado que há muito tempo que a metaplasia intestinal é universalmente considerada uma condição com risco aumentado para o **câncer gástrico**...”. (51, 195, 71)

Figura 6: Conceito com tag incompleta

encefalopatia espongiiforme *s.f.*
spongiform encephalopathy [ing]; **encefalopatía espongiiforme**
Ver este termo *doença da vaca louca*.

Figura 7: Conceito sem tag

4.2 Dicionari multilingüe de la COVID-19

Para extrair a informação relevante deste documento, começou-se por excluir o que não tem interesse. Então, converteu-se para texto apenas as páginas que dizem respeito ao dicionário propriamente dito.

Executou-se, portanto, o comando `pdftotext` com as opções `-f 30 -l 182`.

O documento resultante foi à primeira vista satisfatório, no entanto, com uma análise um pouco mais profunda, percebeu-se que haviam muitas linhas trocadas, as colunas não

estavam a ser bem processadas, e era impossível distinguir alguns números de página dos identificadores.

Optou-se, então, por utilizar as opções `-f 30 -l 182 -raw`.

A opção `-raw` foca-se no fluxo de leitura, descartando completamente o layout do pdf original.

Seria a opção ideal, uma vez que a formatação de cada entrada é constante (muito útil para utilizar regex) . No entanto, mostrou-se com falhas graves, uma vez que, por cada página, transcrevia primeiro a coluna da direita, e só depois a da esquerda, comprometendo, por vezes, a completude de cada bloco de informação. Não havendo uma marca constante de fim ou início de coluna, teve de utilizar-se uma nova solução.

Experimentou-se o comando `pdftohtml` com as opções `-xml -f 30 -l 182`.

Contudo, o problema da ordem das colunas manteve-se.

Posto isto, converteu-se o pdf usando as opções `-f 30 -l 182 -layout` do `pdftotext`. Esta opção, que tem em conta a parte visual do documento original, fez um bom trabalho a manter o conteúdo organizado como no pdf. Porém, a posição da coluna da direita varia de página em página, o que dificulta o seu processamento.

Avaliando as opções, decidiu-se que o ficheiro txt com layout seria utilizado apenas para extrair as definições (depois de algum pré-processamento), e as informações adicionais (traduções e área médica) seria extraídas dos índices. Por fim os termos em catalão seriam extraídos do documento com opção `-raw`.

Começou por tratar-se o documento com layout.

Desenhou-se uma função que, em cada página, separa as colunas, na posição 60, e re-escreve o texto com uma coluna só. Depois, existem algumas etapas para limpar o texto (remover números de página, cabeçalhos e caracteres invisíveis), e marcá-lo em posições de interesse, nomeadamente, depois da área temática e antes do número identificador. Estas marcações delimitam precisamente a definição do conceito, que é o que se pretende extrair. Alguns casos especiais que foram ou não abrangidos pelas expressões regex utilizadas de forma errada foram também corrigidos.

Por fim, dividiu-se o texto em blocos, em que cada bloco corresponde ao conjunto de informação completa referente a um determinado conceito. Deste processo nasceu o `dic_definicoes.json`, um dicionário em json que faz corresponder a cada identificador a definição da palavra.

O dicionário de conceitos em catalão, por sua vez, surgiu a partir do dicionário em txt com `-raw`. No documento `dicionário_to_JSON.py`, o texto foi limpo e marcado. Mais uma vez, dividiu-se em blocos. Para cada bloco, utilizou-se uma expressão regex para isolar o conceito em catalão, obtendo, assim um dicionário em JSON que tem como chave o identificador e como valor a palavra catalã.

A partir do índice português, obteve-se o dicionário `JSON index_PT.json`, que faz corresponder ao identificador a lista de conceitos em português que são equivalentes à palavra castelã. No documento `index.py`, definiram-se as funções responsáveis pela limpeza do índice e a sua passagem para JSON, que foram reaproveitadas para o processamento dos restantes índices.

No documento `indexprocessamento.py`, criam-se em massa os dicionários de cada área temática e língua restante, com algumas particularidades dependendo da categoria em que se inserem. Destaca-se a criação dos dicionários por área, uma vez que foi necessário um processamento adicional, visto o documento txt de índice de áreas precisou de ser separado

por cada área específica (Conceptes generals, Epidemiologia, Etiopatogènia, etc.).

Para ajudar na visualização, estes documentos, foram, mais uma vez, transformados em documentos JSON.

Por fim, ainda em `index_processamento.py`, criou-se o dicionário final, com toda a informação extraída por cada número identificador. Como não é relevante para o documento final, o identificador foi descartado. O resultado final é uma lista de dicionários, seguindo a estrutura que é mostrada na figura seguinte:

4.3 Glossário de Termos Médicos Técnicos e Populares

Para efetuar o tratamento da informação apresentada neste documento, utilizou-se o auxílio do comando `pdftotxt` com o objetivo de obter um ficheiro texto mais fácil de manipular e retirar o conteúdo.

De seguida, utilizou-se algumas expressões regulares para retirar as quebras de página (caracter `ff`), retirar alguns *new lines* para ter todas as entradas seguidas em cada nova linha, por último, normalizar a disposição de como diferentes expressões e palavras são separadas (Figura 8).

```
texto=re.sub( pattern: r"\f", repl: "", texto) #tira as quebras de pagina
texto=re.sub( pattern: r"\n\n", repl: "\n", texto) #poe todo o texto corrido
texto = re.sub( pattern: r"\s*[\.:]\s*", repl: ", ", texto) # ficar tudo normalizado
```

Figura 8: Expressões regulares utilizadas para normalizar o conteúdo do documento texto.

Para além dos problemas apresentados inicialmente, durante a conversão para texto, as entradas que ocupavam mais do que uma linha não apresentavam nenhuma característica textual que as diferenciavam de uma entrada completamente distinta.

Para resolver este problema, foi utilizado o índice alfabético para adicionar a marca `@` às entradas que de facto se inicializavam pela letra do índice em que se situavam.

```
M
@Má digestão (pop),dispepsia
@maceração,empolamento por contacto com líquidos,extracção de drogas por
humedecimento,extracção a frio (pop)
@maciço,grande,amplo (pop)
```

Figura 9: Solução encontrada para diferenciar as entradas no glossário.

Esta solução não é ideal, porém, é eficaz o suficiente para deixar somente algumas anomalias em que, a interrupção da entrada e o começo na linha seguinte, correspondem ao índice atual. Para além disso, existem algumas situações em que algumas entradas, que estavam formatadas de forma diferente no ficheiro pdf das restantes, ao serem convertidas para o ficheiro txt, alteraram a ordem das palavras, separando-as também por *new lines*, deixando de ser perceptível.

Em ambas ocasiões, era possível alterar ou remover, respetivamente, estas entradas do glossário de modo manual, exigindo pouco esforço pela sua raridade.

De seguida, utilizaram-se expressões regulares para retirar as marcas (*pop*) e fazer alguma limpeza adicional, e posteriormente, foi criado um ficheiro *json* que guardava um dicionário

com as várias entradas do glossário, em que as diferentes chaves correspondem aos termos populares de cada entrada, e o valor é uma lista dos distintos termos mais técnicos (Figura 10).

```
"abdominal": [
  "relativo ao ventre",
  "ventral"
],
"ablação": [
  "extracção"
],
"abocamento": [
  "anastomose"
],
"absorvência": [
  "absorção",
  "absorvimento"
```

Figura 10: Ficheiro *json* criado com o conteúdo tratado.

A partir deste ficheiro *json*, criado para organizar o conteúdo temporariamente, foi possível a sua utilização para adicionar a informação relevante ao ficheiro *json* proveniente do *scrapping* dos documentos anteriores.

Para o fazer, utilizaram-se expressões regulares para encontrar situações em que alguma palavra é referenciada (ignorando palavras que não acrescentam conhecimento, por exemplo advérbios), adicionando aos dicionários do ficheiro *json* final, um conjunto de vocabulário relacionado (Figura 11).

```
{
  "conceito": "hantavirose",
  "en": "hantavírus",
  "es": "hantavírus",
  "significado": "Infecção causada no ser humano pelo hantavírus. « a hantavirose é capaz de matar num prazo",
  "significado_enciclopédico": "a hantavirose é capaz de matar num prazo dois a três dias grande parte da",
  "contexto": "...Minas Gerais, Mato Grosso e Goiás - onde houve casos de hantavirose em seres humanos",
  "outras associacoes a 'infecção': [
    "infecção"
  ],
  "outras associacoes a 'parte': [
    "fracção",
    "porção",
    "número fraccionário (número) quebrado"
  ]
},
```

Figura 11: Excerto do ficheiro *json* final criado com o todo o conteúdo tratado



5 Conclusão

O desenvolvimento deste trabalho permitiu uma aplicação prática de técnicas de Processamento de Linguagem Natural em documentos com estruturas e conteúdos bastante distintos, inseridos no contexto da saúde. Através da utilização de ferramentas como expressões regulares e diferentes formatos de conversão (HTML, TXT, XML), foi possível extrair, normalizar e estruturar a informação proveniente de três glossários distintos, culminando na criação de um ficheiro JSON unificado e funcional. Apesar dos desafios associados à inconsistência dos documentos originais, a abordagem adotada demonstrou ser eficaz na maioria dos casos, com soluções adequadas mesmo para exceções. Este projeto evidenciou a importância do pré-processamento cuidadoso e da adaptação às especificidades de cada fonte.