PROCESSAMENTO DE LINGUAGEM NATURAL

Trabalho Prático 1

Ana Patrícia Costa pg53062
Inês Mendes pg53875
Luís Cunha pg54020

FICHEIROS PROCESSADOS

- **GLOSSÁRIO DE TERMOS MÉDICOS TÉCNICOS E POPULARES**
- **GLOSSÁRIO DO MINISTÉRIO SAÚDE**
- **MINIDICIONÁRIO DE CARDIOLOGISTA DO AUTOR RICARDO SILVEIRA MELLO**

01.

GLOSSÁRIO DE TERMOS MÉDICOS TÉCNICOS E POPULARES

EXPRESSÃO POPULAR (itálico), TERMO (negrito)

OU

TERMO (negrito), EXPRESSÃO POPULAR (itálico)

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**

abdómen , *barriga, ventre* (pop)

abdominal , *ventral* (pop)

aberrante , *anormal* (pop)

PRIMEIRA ABORDAGEM

- **Limpeza** e correção de incoerências no *XML* **manualmente** (termos a negrito consecutivos e expressões divididas pela quebra de página).

- Função ***re.sub()*** para eliminar porções indesejadas

novo_xml.xml

```
<b>antidiabético</b>
<i>substância que reduz a concentração de açúcar no sangue</i>
<b>antidiurético</b>
<i>que não faz urinar</i>
<b>antídoto</b>
<i>contraveneno</i>
<b>antiemético</b>
<i>contra os vómitos</i>
<b>antiepiléptico</b>
<i>medicamento contra a epilepsia</i>
```

- Função ***re.findall()*** para extrair os termos **(...)** e as respetivas expressões populares **(...)**, para as listas “termos” e “pop”.

- Ciclo ***while*** para formar o dicionário **{”termo” : ”expressão popular”}** que foi usado para criar o ficheiro JSON.

glossário.json

```
"antidepressivo": "substância que alivia a depressão",
"antidiabético": "substância que reduz a concentração de açúcar no sangue",
"antidiurético": "que não faz urinar",
"antídoto": "contraveneno",
"antiemético": "contra os vómitos",
"antiepiléptico": "medicamento contra a epilepsia",
"antiestrogénico": "que impede ou contraria o efeito das hormonas estrogénicas",
"antiexudativo": "que impede a exsudação, a transpiração",
"antiflogístico": "que combate a inflamação, anti-inflamatório",
```

SEGUNDA ABORDAGEM

Limpeza manual de conteúdo não importante no início e fim do ficheiro XML.

texto.xml

```
<i>a milionésima parte de um grama</i>
 (pop) ,
<b>micrograma</b>

<i>à volta da boca</i>
 (pop) ,
<b>perioral</b>

<i>à volta da órbita</i>
 (pop) ,
<b>periorbital</b>
```

- O ficheiro XML é separado através do caracter '**\n**' usando o método ***split***.
- Nova limpeza, substituindo entradas indesejadas por '**$**' e todas as entra contendo '(pop)' por 'SEPARADOR'.

- **Exclusão** de todas as entradas da lista iguais a '**$**'.
- Ciclo ***while*** que verifica a existência de um dado **padrão** nas entradas atual e seguintes (até um máximo de 4 entradas futuras) incrementando o índice com base no número de entradas no padrão.
- Criação de uma nova entrada no dicionário "**texto_dict**".
- Criação do dicionário "**texto_dict_organizado**" cujo conteúdo corresponde ao primeiro mas organizado alfabeticamente, que é usado para criar o ficheiro ***JSON***.

texto.json


```
"ACTH": "hormônio adreno-córticotrófico, corticotrofina",
"Gram-negativo": "que não toma o corante de Gram",
"Gram-positivo": "que toma o corante de Gram",
"Petit mal; epilepsia menor": "pequeno mal, epilepsia com ataques pouco intensos",
"abcesso": "abcesso, tumor",
"abdominal": "ventral",
"abdómen": "barriga, ventre",
"aberrante": "anormal",
"aborto": "abortamento, desmancho",
"abrupto": "repentino, brusco",
"absorção": "absorvimento, absorvência",
"abstinência": "jejum",
"acatisia": "incapacidade em permanecer sentado",
"acidental": "por acaso, sem importância",
"acidez": "acidade, azedume",
"acidose": "alteração do equilíbrio ácido básico do sangue e líquidos teciduais",
```

02.

GLOSSÁRIO DO MINISTÉRIO SAÚDE

SIGLAS + TERMOS

SIGLA (negrito) - DESCRIÇÃO

+

TERMO (negrito)
CATEGORIA (itálico): categoria
DESCRIÇÃO

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
AIDS – Síndrome da Imunodeficiência Adquirida
AIH – Autorização de Internação Hospitalar
AIS – Ações Integradas de Saúde
ANCED – Associação Nacional de Centros de Defesa
ANS – Agência Nacional de Saúde
ANVISA – Agência Nacional de Vigilância Sanitária

CAPS – Centro de Assistência Psicossocial
CAT – Comunicação de Acidente de Trabalho
CBO – Conselho Brasileiro de Oftalmologia
CCIH – Comissão de Controle de Infecção Hospitalar
CCPDM – Controle de Cadeia Produtiva e de Distribuição de Medicamentos
CCPDS – Controle de Cadeia Produtiva e de Distribuição de Substâncias
CDMS – Comitê de Desburocratização do Ministério da Saúde
CENADI – Centro Nacional de Armazenagem e Distribuição de Imunobiológicos
CENEPI – Centro Nacional de Epidemiologia

Alcoólatra
Categoria: Drogas de Uso Terapêutico e Social
Este termo refere-se tanto aos bebedores-problema quanto aos dependentes do álcool.

Alcoolismo
Categoria: Drogas de Uso Terapêutico e Social
Significa dependência do álcool e/ou problemas relacionados ao consumo de bebidas alcoólicas.

Alimentação equilibrada
Ver Alimentação saudável.

Alimentação saudável
Categoria: Alimentação e Nutrição
É o mesmo que dieta equilibrada ou balanceada e pode ser resumida por três princípios: variedade, moderação e equilíbrio. Variedade significa comer diferentes tipos de alimentos pertencentes aos diversos grupos. Moderação

Alimento dietético
Categoria: Alimentação e Nutrição
São alimentos isentos de algum tipo de nutriente, preparados para atender a restrições dietéticas específicas de várias doenças. Ex.: produtos sem açúcar, para diabéticos; sem sal, para hipertensos; sem colesterol, para portadores de colesterol sangüíneo alto; e assim por diante.

Alimento *in natura*
Categoria: Alimentação e Nutrição
Todo alimento de origem vegetal ou animal, para cujo consumo imediato se exija, apenas, a remoção da parte não comestível e os tratamentos indicados para a sua perfeita higienização e conservação.

Alimento integral
Categoria: Alimentação e Nutrição

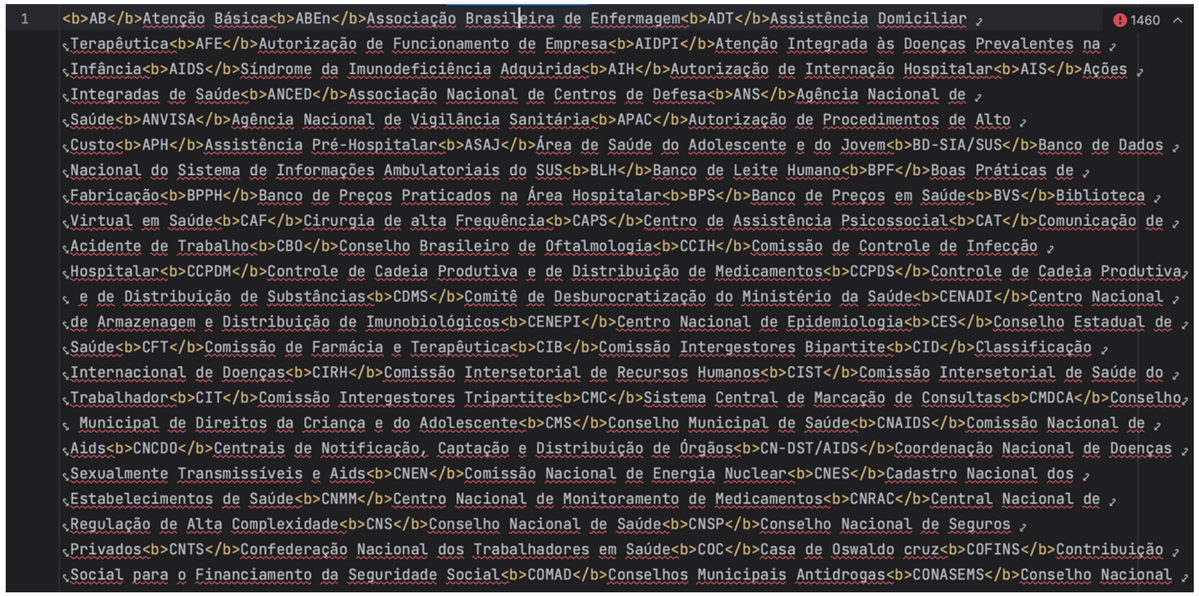
ABORDAGEM

- **Limpeza manual** do início (pré-siglas) e fim (pós-dicionário) do documento XML e substituição da zona de separação entre as duas secções por '**$$$**'.

- Função ***split*** através da sequência '**$$$**'.

- Função ***re.sub()*** para substituição de porções indesejadas.

siglas.xml



- Método **re.findall()** aplicado a expressões entre '**<b>...<\b>**' constituindo a lista de termos.

- Método **re.findall()** aplicado a expressões entre '**<\b>...<b>**' ou '**<\b>...[fim_do_doc]**' constituindo a lista de descrições.

- Criação de um ficheiro *JSON* (**siglas.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",
"AIDS": "Síndrome da Imunodeficiência Adquirida",
"AIH": "Autorização de Internação Hospitalar",
"AIS": "Ações Integradas de Saúde",
"ANCED": "Associação Nacional de Centros de Defesa",
"ANS": "Agência Nacional de Saúde",
"ANVISA": "Agência Nacional de Vigilância Sanitária",
"APAC": "Autorização de Procedimentos de Alto Custo",
"APH": "Assistência Pré-Hospitalar",
"ASAJ": "Área de Saúde do Adolescente e do Jovem",
```

dicionario.xml

```
<text top="426" left="176" width="270" height="14" font="21"><b>Abordagem médica tradicional do adulto </b></t
<text top="444" left="176" width="88" height="14" font="21"><b>hospitalizado</b></text>
<text top="461" left="176" width="67" height="16" font="16"><i>Categoria: </i></text>
<text top="461" left="244" width="105" height="16" font="14">Atenção à Saúde</text>
<text top="479" left="176" width="290" height="16" font="14">Focada em uma queixa principal e o hábito </text>
<text top="497" left="176" width="290" height="16" font="14">médico de tentar explicar todas as queixas </text>
<text top="515" left="176" width="290" height="16" font="14">e os sinais por um único diagnóstico, que é </text>
<text top="533" left="176" width="289" height="16" font="14">adequada no adulto jovem - não se aplica em </text>
<text top="551" left="176" width="105" height="16" font="14">relação ao idoso.</text>
<text top="570" left="176" width="181" height="14" font="21"><b>Abuso financeiro dos idosos</b></text>
<text top="587" left="176" width="67" height="16" font="16"><i>Categoria: </i></text>
<text top="587" left="244" width="136" height="16" font="14">Acidentes e Violência</text>
<text top="605" left="176" width="289" height="16" font="14">Exploração imprópria ou ilegal e/ou uso não </text>
<text top="623" left="176" width="285" height="16" font="14">consentido de recursos financeiros dos idosos.</text>
<text top="642" left="176" width="115" height="14" font="21"><b>Abuso incestuoso</b></text>
<text top="659" left="176" width="67" height="16" font="16"><i>Categoria: </i></text>
<text top="659" left="244" width="136" height="16" font="14">Acidentes e Violência</text>
<text top="677" left="176" width="289" height="16" font="14">Consiste no abuso sexual envolvendo pais ou </text>
<text top="695" left="176" width="285" height="16" font="14">outro parente próximo, os quais se encon</text>
<text top="713" left="176" width="285" height="16" font="14">tram em uma posição de maior poder em re</text>
<text top="731" left="176" width="91" height="16" font="14">lação à vítima.</text>
<text top="750" left="176" width="192" height="14" font="21"><b>Abuso sexual na adolescência</b></text>
<text top="767" left="176" width="186" height="16" font="14">Ver Abuso sexual na infância.</text>
```

- Método ***split*** através do caracter '**\n**'.
- **Substituição** das entradas indesejadas por '**$**' e subsequente **exclusão** destas.

Ciclo ***for*** que procura encontrar, a cada iteração um **padrão** nas entradas atual e anterior ou atual e posterior, adicionando **termos**, **categorias** e **descrições** às suas respetivas listas.

Formação de um dicionário final através das listas criadas, usado para criar o documento **JSON**.

dicionario.json


```
    "Abordagem médica tradicional do adulto hospitalizado": {
        "categoria": "Atenção à Saúde",
        "descricao": "Focada em uma queixa principal e o hábito médico
    },
    "Abuso financeiro dos idosos": {
        "categoria": "Acidentes e Violência",
        "descricao": "Exploração imprópria ou ilegal e/ou uso não conse
    },
    "Abuso incestuoso": {
        "categoria": "Acidentes e Violência",
        "descricao": "Consiste no abuso sexual envolvendo pais ou outr
    },
    "Abuso sexual na adolescência": {
        "categoria": "N/A",
        "descricao": "Ver Abuso sexual na infância."
    },
```

03.

MINIDICIONÁRIO DE CARDIOLOGISTA

TRADUÇÕES EN-PT
+
TRADUÇÕES PT-EN

TERMO_EN (negrito) - TRADUÇÃO_PT

TERMO_PT (negrito) - TRADUÇÃO_EN

STRESS – Estresse – tensão, pressão / Acentuar / Dar ênfase

STRETCH (TO) – Alongar-se

STRETCHER – Maca

STRIKING – Notável / Extraordinário / Formidável

STRING – Fio / Cordão

STRIPE – Listra

STROKE – Derrame cerebral

STROKE VOLUME – Volume sistólico

STRUGGLE (TO) – Lutar / Combater

COUNTERCLOCKWISE

A faculdade 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

SUPPORT – Apoio / Colaboraçao

SURGICAL SUITE – Centro cirúrgico

SURROGATE ENDPOINTS RESULT – Resultados secundários

SURROGATE POINTS – Desfechos substitutivos

SURVEILLANCE – Vigilância / Fiscalização

SURVEY – Pesquisa de opinião

SWALLOW (TO) – Engolir

SWEATING – Sudorese

SWITCH (TO) – Trocar / Inverter

Afrouxadamente – LOOSELY

Agulha – NEEDLE

Agulha de buraco largo – LARGE BORE NEEDLE

Ala – WING

Alargar / Largura / Largo – WIDEN (TO) / WIDTH / WIDE

Aleijado – CRIPPLED

Além do mais / Além de que / Aliás – MOREOVER

ABORDAGEM

- **Limpeza** da parte inicial do XML manualmente e substituição da zona de **separação entre as duas secções** do Minidicionário por um ‘@’.
- Função ***re.sub()*** para eliminar porções indesejadas
- Função ***re.split()*** para fazer a divisão do XML pelo ‘@’.
- Marcação dos termos: **@termo@\ntradução\n\n**

EN-PT.xml

```
@DEVICE@
Dispositivo / Instrumento / Algo construído com um “design”

@DIASTOLIC GRUNT@
Estalido / Grunhido (emitir som gutural) /Som inarticulado (que expressa indiferença)

@DIASTOLIC THRILL@
Frêmito diastólico

@DISABILITY@
Incompetência / Falta de habilidade física ou mental

@DISARRAY@
Desordem, desarranjo / Fora de disposição / Fora de conjunto
```

PT-EN.xml

```
@Realização / Conquistas / Feitos@
ACHIEVEMENT

@Receita médica para adquirir medicamentos na farmácia@
MEDICAL PRESCRIPTION

@Receita para farmácia@
PRESCRIPTION

@Receptor do material doado (por exemplo: o sangue)@
DONOR ACCEPTOR

@Recorrente / que se repete@
RECURRENT
```

Função ***re.findall()*** com a expressão ‘**@(.+)@\n(.*)**’ para fazer a captura da informação

EN-PT.json


```
"DEVICE": "Dispositivo / Instrumento / Algo construído com um “design”",
"DIASTOLIC GRUNT": "Estalido / Grunhido (emitir som gutural) /Som inarticulado (que expressa indiferença)",
"DIASTOLIC THRILL": "Frêmito diastólico",
"DISABILITY": "Incompetência / Falta de habilidade física ou mental ",
"DISARRAY": "Desordem, desarranjo / Fora de disposição / Fora de conjunto",
```


PT-EN.json


```
"Realização / Conquistas / Feitos": "ACHIEVEMENT ",
"Receita médica para adquirir medicamentos na farmácia": "MEDICAL PRESCRIPTION",
"Receita para farmácia": "PRESCRIPTION",
"Receptor do material doado (por exemplo: o sangue)": "DONOR ACCEPTOR",
"Recorrente / que se repete": "RECURRENT",
```

CONCLUSÕES

- **Identificação de padrões e características individuais em cada documento.**
- **Adaptação das estratégias de processamento para lidar com desafios específicos.**
- **Sucesso na extração e organização de informações essenciais.**