



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 vômitos</i>  
<b>antiepiléptico</b>  
<i>medicamento contra a epilepsia</i>
```

- ➡ Função *re.findall()* para extrair os termos (**...**) e as respectivas expressões populares (**<i>...</i>**), 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 vômitos",  
"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

```
2  <i>a milionésima parte de um grama</i>
3  (pop)
4  <b>micrograma</b>
5
6  <i>à volta da boca</i>
7  (pop)
8  <b>perioral</b>
9
10 <i>à volta da órbita</i>
11 (pop)
12 <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 entradas 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": "acidez, 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

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

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

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 sanguíneo 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>**' constituindo a lista de termos.

```
1 <b>AB</b>Atenção Básica<b>ABEn</b>Associação Brasileira de Enfermagem<b>ADT</b>Assistência Domiciliar
<b>Terapêutica<b>AFE</b>Autorização de Funcionamento de Empresa<b>AIDPI</b>Atenção Integrada às Doenças Prevalentes na
<b>Infância<b>AIDS</b>Síndrome da Imunodeficiência Adquirida<b>AIH</b>Autorização de Internação Hospitalar<b>AIS</b>Ações
<b>Integradas de Saúde<b>ANCED</b>Associação Nacional de Centros de Defesa<b>ANS</b>Agência Nacional de
<b>Saúde<b>ANVISA</b>Agência Nacional de Vigilância Sanitária<b>APAC</b>Autorização de Procedimentos de Alto
<b>Custo<b>APH</b>Assistência Pré-Hospitalar<b>ASAJ</b>Área de Saúde do Adolescente e do Jovem<b>BD-SIA/SUS</b>Banco de Dados
<b>Nacional do Sistema de Informações Ambulatoriais do SUS<b>BLH</b>Banco de Leite Humano<b>BPF</b>Boas Práticas de
<b>Fabricação<b>BPPH</b>Banco de Preços Praticados na Área Hospitalar<b>BPS</b>Banco de Preços em Saúde<b>BVS</b>Biblioteca
<b>Virtual em Saúde<b>CAF</b>Cirurgia de alta Frequência<b>CAPS</b>Centro de Assistência Psicossocial<b>CAT</b>Comunicação de
<b>Acidente de Trabalho<b>CBO</b>Conselho Brasileiro de Oftalmologia<b>CCIH</b>Comissão de Controle de Infecção
<b>Hospitalar<b>CCPDM</b>Controle de Cadeia Produtiva e de Distribuição de Medicamentos<b>CCPDS</b>Controle de Cadeia Produtiva
<b>e de Distribuição de Substâncias<b>CDMS</b>Comitê de Desburocratização do Ministério da Saúde<b>CENADI</b>Centro Nacional
<b>de Armazenagem e Distribuição de Imunobiológicos<b>CENEPI</b>Centro Nacional de Epidemiologia<b>CES</b>Conselho Estadual de
<b>Saúde<b>CFT</b>Comissão de Farmácia e Terapêutica<b>CIB</b>Comissão Intergestores Bipartite<b>CID</b>Classificação
<b>Internacional de Doenças<b>CIRH</b>Comissão Intersetorial de Recursos Humanos<b>CIST</b>Comissão Intersetorial de Saúde do
<b>Trabalhador<b>CIT</b>Comissão Intergestores Tripartite<b>CMC</b>Sistema Central de Marcação de Consultas<b>CMDCA</b>Conselho
<b>Municipal de Direitos da Criança e do Adolescente<b>CMS</b>Conselho Municipal de Saúde<b>CNAIDS</b>Comissão Nacional de
<b>Aids<b>CNCDO</b>Centrais de Notificação, Captação e Distribuição de Órgãos<b>CN-DST/AIDS</b>Coordenação Nacional de Doenças
<b>Sexualmente Transmissíveis e Aids<b>CNEN</b>Comissão Nacional de Energia Nuclear<b>CNES</b>Cadastro Nacional dos
<b>Estabelecimentos de Saúde<b>CNMM</b>Centro Nacional de Monitoramento de Medicamentos<b>CNRAC</b>Central Nacional de
<b>Regulação de Alta Complexidade<b>CNS</b>Conselho Nacional de Saúde<b>CNSP</b>Conselho Nacional de Seguros
<b>Privados<b>CNTS</b>Confederação Nacional dos Trabalhadores em Saúde<b>COC</b>Casa de Oswaldo Cruz<b>COFINS</b>Contribuição
<b>Social para o Financiamento da Seguridade Social<b>COMAD</b>Conselhos Municipais Antidrogas<b>CONASEMS</b>Conselho Nacional
```



Método **re.findall()** aplicado a expressões entre '**<\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

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



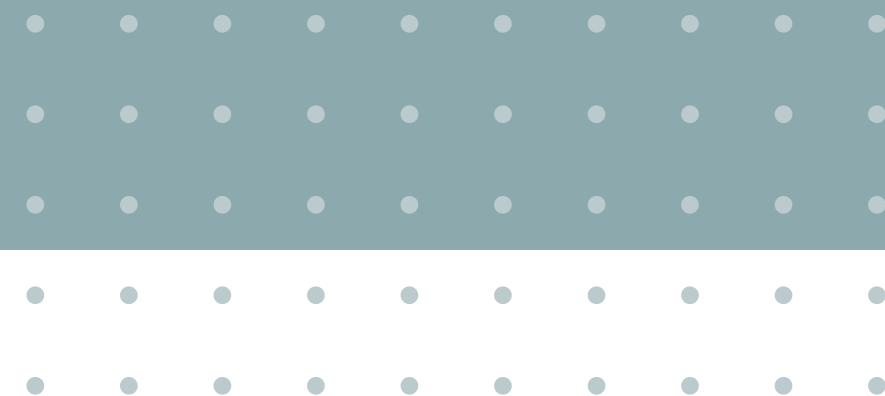
Formação de um dicionário final através das listas criadas, usado para criar o documento **JSON**.

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 outro  
},  
"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 im-
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-
ditadamente** – OVERTLY

SUPPORT – Apoio / Colaboração

SURGICAL SUITE – Centro cirúrgico

SURROGATE ENDPOINTS RESULT – Resul-
tados secundários

SURROGATE POINTS – Desfechos subs-
titutivos

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.**