Processamento de Linguagem Natural em Engenharia Biomédica

Trabalho Prático 2

Ana Andrade PG56107 Filipa Costa PG56123 Leonor Pereira PG57813 Professores: Luís Filipe Cunha José João Almeida

Introdução

- Descrição do Dataset Inicial
- Enriquecimento do Dataset
 - Junção dos Datasets
 - Web Scraping
 - Integração dos Dados
 - Validação e Qualidade dos Dados
- Análise de Conceitos e Relações
 - Similaridade semântica
 - Agrupamento por domínios
- Desenvolvimento da Ferramenta Web
 - Tecnologias utilizadas
 - Funcionalidades
- Testes e Validação
- Conclusão

Dataset Inicial

```
"letra": {
   "num_conceito":{
       "designação": "designação",
       "categoria lexical da designação":
       "n (nom) / n pl (nom plural) / n m (nom masculí) / n m pl (nom masculí plural)
       / n f (nom femení) / n f pl (nom femení plural) / n m, f (nom masculí i femení)
       / n m/f (nom masculí o femení) / adj (adjectiu) / v tr (verb transitiu)
       / v tr/intr (verb transitiu o intransitiu) / v intr (verb intransitiu)".
       "complementos designação":{
           "siglas": ["sigla"],
           "sinónimos absolutos": "sin.",
           "sinónimos complementares":["sin. compl."],
           "denominação comercial": "den. com.",
           "consultar outra entrada": "veg."
       "traduções":{
           "occità": "oc",
           "basc": "eu",
           "gallec": "gl",
           "castellà": "es",
           "anglès": "en",
            "francès": "fr",
            "portuguès": "pt",
           "[PT]portuguès de Portugal": "pt [PT]",
           "[BR]portuguès del Brasil": "pt [BR]",
           "neerlandè": "nl",
           "àrab": "ar"
        "códigos alternativos":{
           "símbolo": "sbl",
           "nome científico": "nc",
           "número CAS": "CAS"
        "áreas temáticas": {
           "área": "Conceptes generals/Epidemiologia/Etiopatogènia/Diagnòstic/
           Clínica/Prevenció/Tractament/Principis actius/Entorn social",
           "descrição": "descrição"
        "notas":[]}
```

Processo de junção do Dataset Inicial

Regras de conversão:

diccionari-multilinguee-de-la-covid-19 permaneceu com a mesma estrutura!

glossario.json:

- tipo + número → categoria lexical da designação
 - \circ fem + vazio \rightarrow n f
 - \circ masc + vazio \rightarrow n m
 - \circ fem + plural \rightarrow n f pl
 - \circ masc + plural \rightarrow n m pl
- Outros mapeamentos:
 - o descrição → descrição
 - \circ notas \rightarrow informações_notas
 - espanhol → castellà
 - \circ inglês \rightarrow anglès
 - sinônimo → sinónimos absolutos
 - remissiva → remissiva
 - expandida → expandida

glossario_completo.json:

- referencia_gramatical → categoria lexical da designação
 - \circ s.f. \rightarrow n f
 - \circ s.m. \rightarrow n m
- Outros mapeamentos:
 - espanhol → castellà
 - \circ inglês \rightarrow anglè
 - \circ Sigla \rightarrow siglas
 - Definicao → descrição
 - Sinonimos → sinónimos absolutos
 - Informacao_enciclopedica, Abonação,
 Numero_identificacao e Marcas_Tipograficas foram colocados em "notas"

Web Scraping

Fonte externa: Vocabulaire de la Médecine (OQLF, Canadá) Etapas do scraping:

- Extração de links por ordem alfabética
- Colheita do conteúdo terminológico, incluindo:
 - Termo, sinónimos, traduções (FR/EN), definições, área médica, notas

• Ferramentas utilizadas:

o requests, BeautifulSoup, re, deep_translator

• Pós-processamento:

- Tradução automática;
- Padronização para o mesmo formato JSON;
- Nova fusão com dicionario_unificado.json → Ficheiro final: dicionario_unificado_final.json



Figura: Captura de Ecrã de uma parte do website com a informação retirada.

JSON 3 DATASETS + WEB SCRAPING

```
"A": {
  "Acidente isquêmico de transição": {
   "fonte": "https://vitrinelinguistique.oqlf.gouv.qc.ca/fiche-gdt/fiche/8369586/accident-ischemique-transitoire",
   "dominio": [
      "medicamento > Semiologia e Patologia",
      "medicamento > neurologia"
   "autor": "Escritório de Quebec da língua francesa",
   "data_ultima_atualizacao": "2021",
   "Definição:": "Isquemia temporária em uma parte do cérebro, que não causa infarto cerebral e resulta de uma obstr
   "Notas:": "Os sinais e sintomas de um acidente isquêmico transitório aparecem repentinamente e desaparecem em men
   "termos_privilegiados": [
     "accident ischémique transitoire n. m.",
      "AIT n. m.",
      "ischémie cérébrale transitoire n. f.",
      "ICT n. f."
    "termos_em_contexto": [
     "mini-accident vasculaire cérébral n. m.",
      "mini-AVC n. m."
   "ingles": [
     "transient ischemic attack",
      "TIA"
   "ingles_associado": [
      "mini-stroke"
```

```
Acidente isquêmico de transição": {
"dicionario": "vocabulaire-de-la-medecine-oqlf",
"categoria lexical da designação": "",
"complementos designação": {
  "siglas": [],
  "sinónimos absolutos": "",
  "sinónimos complementares": [],
  "denominação comercial": "",
  "consultar outra entrada":
"traduções": {
 "castellà": "",
"anglès": "transient ischemic attack, TIA",
  "anglès_associado": "mini-stroke",
  "francès": "accident ischémique transitoire n. m., AIT n. m., ischémie cérébrale transitoire n. f., ICT n. f.",
  "francès_contexte": "mini-accident vasculaire cérébral n. m., mini-AVC n. m."
"códigos alternativos": {
  "símbolo": "",
  "nome científico": "",
  "número CAS": ""
"áreas temáticas": {
  "área": "medicamento > Semiologia e Patologia, medicamento > neurologia",
  "descrição": "Isquemia temporária em uma parte do cérebro, que não causa infarto cerebral e resulta de uma obstrução arterial."
"notas": {
    "Os sinais e sintomas de um acidente isquêmico transitório aparecem repentinamente e desaparecem em menos de 24 horas. Pode, por exemplo
  "Informacao_enciclopedica": "",
  "Abonacao": "",
  "Numero identificacao": "",
  "Marcas_Tipograficas": [],
  "remissiva": "",
  "expandida": ""
```

Figura: Dataset Web Scraping

Figura: Dataset Unificado

Estrutura Dataset Final

```
"letra":{
   "termo":{
       "dicionario": "...",
       "categoria lexical da designação": "n/ n pl/n m/n m pl/n f/n f pl/n m, f/n m/f/adj/v tr/v tr/intr/v intr",
       "complementos designação":{
           "siglas": ["sigla"],
           "sinónimos absolutos": "sin.",
           "sinónimos complementares":["sin. compl."],
           "denominação comercial": "den. com.",
           "consultar outra entrada": "veg."
       "traduções":{
           "occità": "oc",
           "basc": "eu",
           "gallec": "gl",
           "castellà": "es",
           "anglès": "en",
           "francès": "fr",
           "portuguès": "pt", //igual ao nome do termo nos outros dois
           "[PT]portuguès de Portugal": "pt [PT]",
           "[BR]portuguès del Brasil": "pt [BR]",
           "neerlandè": "nl",
           "àrab": "ar"
       "códigos alternativos":{
           "símbolo": "sbl",
           "nome científico": "nc",
           "número CAS": "CAS"
       "áreas temáticas": {
           "área": "...",
           "descrição": "descrição"
       },
       "notas":{
           "informacoes_notas":[...],
           "Informacao_enciclopedica": "...",
           "Abonacao": "...",
           "Numero_identificacao": "...",
           "Marcas_Tipograficas": [ ... ],
           "remissiva":"...",
           "expandida":"..."
```

Analise de Conceitos e Relações

1. Similaridade Semântica

Conversão de texto em vetores

- Utilizámos o modelo 'pt-mteb/average_fasttext_cc.pt.300' da biblioteca 'sentence-transformers'
- Cada conceito é transformado num vetor de embedding que representa o seu significado semântico

Componentes do embedding

- Vetor gerado a partir de múltiplos atributos:
 - Termo principal
 - Categoria lexical
 - Sinónimos e denominações comerciais
 - Traduções e descrição temática
- Cada componente tem um peso, reforçando os elementos mais relevantes

Análise de Conceitos e Relações

1. Pesquisa por Similaridade

Mecanismo de busca inteligente

- O termo pesquisado é convertido em vetor
- Calcula-se a similaridade de cosseno com todos os conceitos existentes
- Resultados ordenados por grau de similaridade

Ajustes de relevância (boosting)

• Reforço aplicado se houver correspondência direta no nome (+0.2) ou descrição (+0.1)

Vantagens

- Identificação de relações semânticas subtis
- Supera a simples correspondência textual

2. Agrupamento por Domínios

Filtragem avançada de conceitos

- Conceitos contêm metadados:
 - Dicionário de origem
 - Categoria lexical
 - Área temática

0

Funcionalidade de agrupamento

- Utilizador pode filtrar por domínio (ex: só conceitos da cardiologia)
- Filtros atuam isoladamente ou em conjunto com a pesquisa semântica

1. Tecnologias utilizadas

Arquitetura leve e modular

- Backend: Flask (Python)
 - Criação de rotas e páginas dinâmicas
 - Processamento das consultas e cálculos de similaridade
- Frontend: HTML + Bootstrap 5
 - Design responsivo
 - Interface acessível em desktop e mobile

Integração eficaz

- Comunicação entre frontend e backend via HTTP:
 - GET para pesquisa
 - POST para inserção de dados

Interface e Layout

Componentes interativos

- Filtros com caixas de seleção e botões
- Áreas com scroll para agrupamento e navegação
- Layout responsivo com navegação intuitiva

Objetivo

 Tornar a interação fluída para utilizadores de qualquer dispositivo

1. Funcionalidades da Ferramenta

Divisão funcional em 4 áreas principais:

- 1. Consulta
- 2. Navegação
- 3. Gestão de conteúdos
- 4. Visualização

Base tecnológica:

 $Flask + Bootstrap \rightarrow$ Interface intuitiva para utilizadores

Consulta de Conceitos

3 modos de busca:

- Similaridade semântica
- Filtragem por atributos (dicionário, categoria, tema)
- Modo híbrido: filtros + semântica combinados

Resultados ordenados por relevância

- Exibem nome, descrição e score de similaridade
- Boosting aplicado se o termo surgir no nome/descrição

Navegação e Exploração

Exploração sem consulta textual

- Listagem alfabética completa
- Paginação com 20 conceitos por página

Visualização detalhada por conceito:

- Dados básicos: nome, categoria, dicionário
- Dados extra: sinónimos, traduções, códigos, notas

Gestão de Conteúdos

Adicionar novos conceitos

- Formulário com validação dos campos
- Geração automática dos embeddings

Editar conceitos existentes

 Atualização de campos e processamento automático de listas

Persistência

Todas as alterações guardadas num ficheiro JSON central

Usabilidade e Acessibilidade

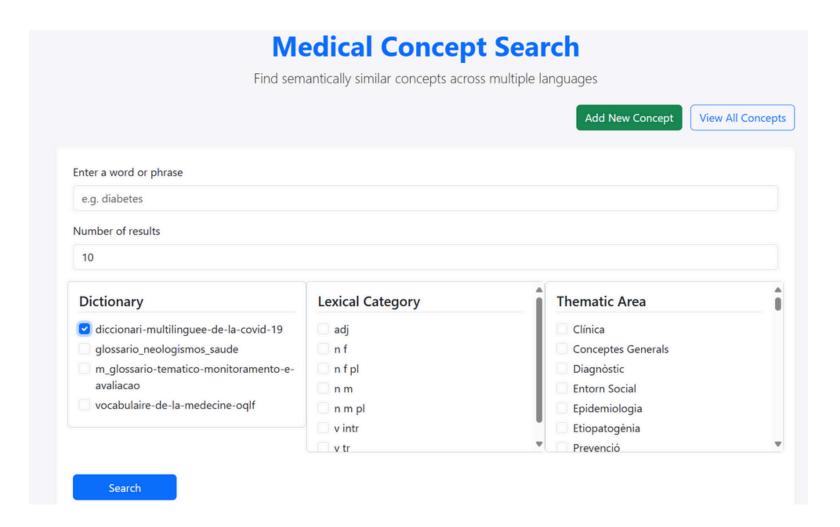
Design centrado no utilizador

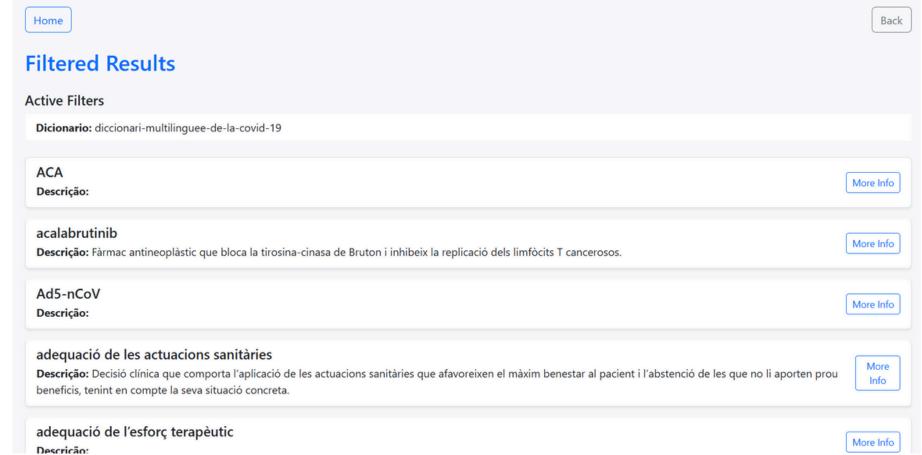
- Botões de navegação e *feedback* visual imediato
- Organização clara com uso de tabelas
- Mensagens para casos sem resultados

Consistência dos dados

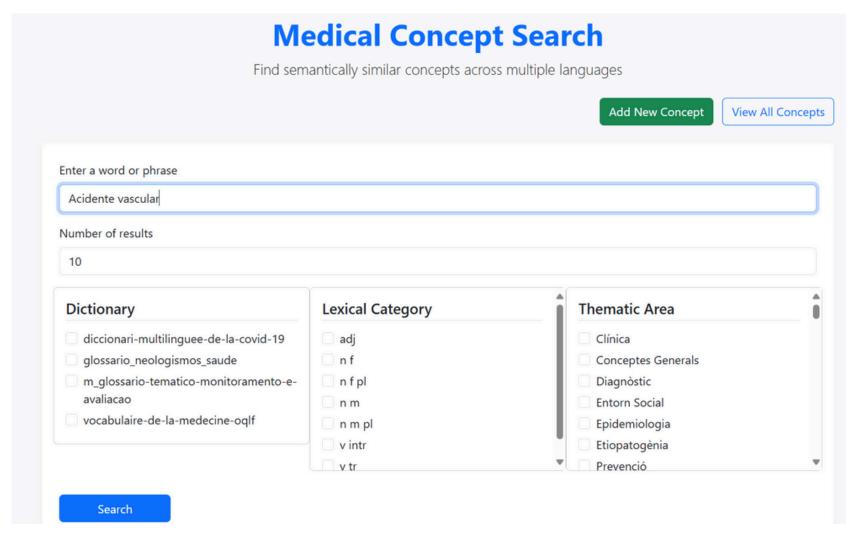
 Todas as ações têm persistência automática no backend

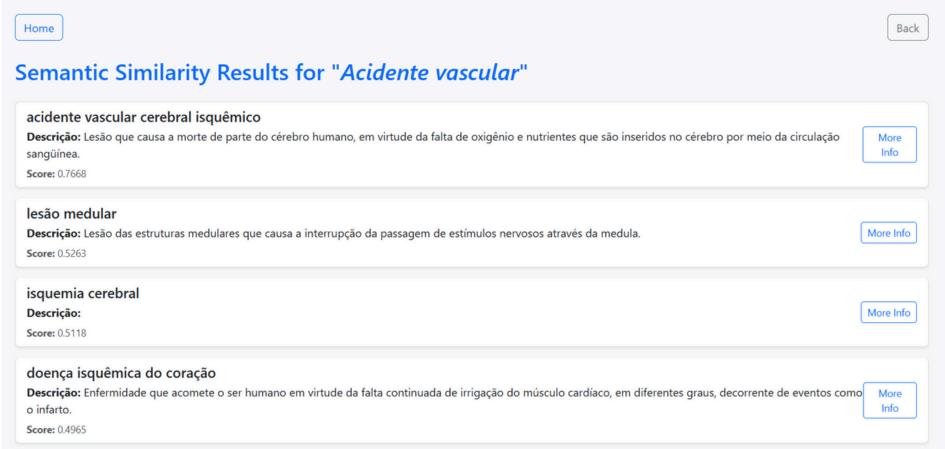
Pesquisa por filtro



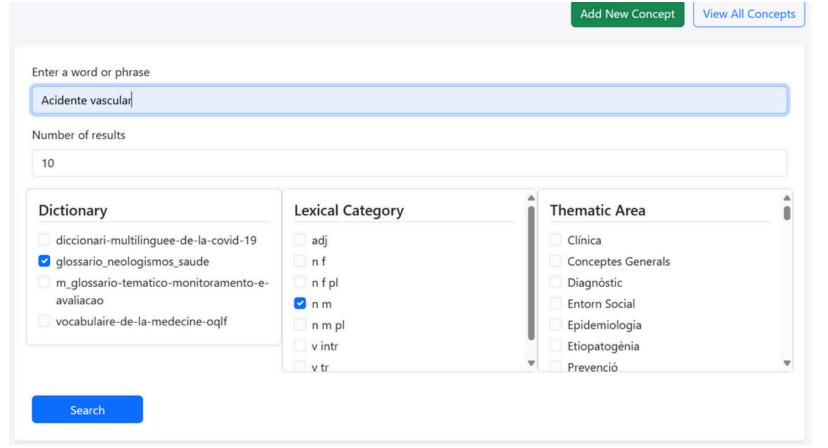


• Pesquisa por similaridade





• Pesquisa por similaridade e aplicação de filtros





• Verificação de detalhes

Concept Details: acidente vascular cerebral isquêmico

Dictionary

glossario_neologismos_saude

Lexical Category

n m

Complementary Info

• Acronyms: AVCI

Translations

- Spanish (Castilian): accidente vascular cerebral isquêmico
- English: ischemic cerebrovascular accident

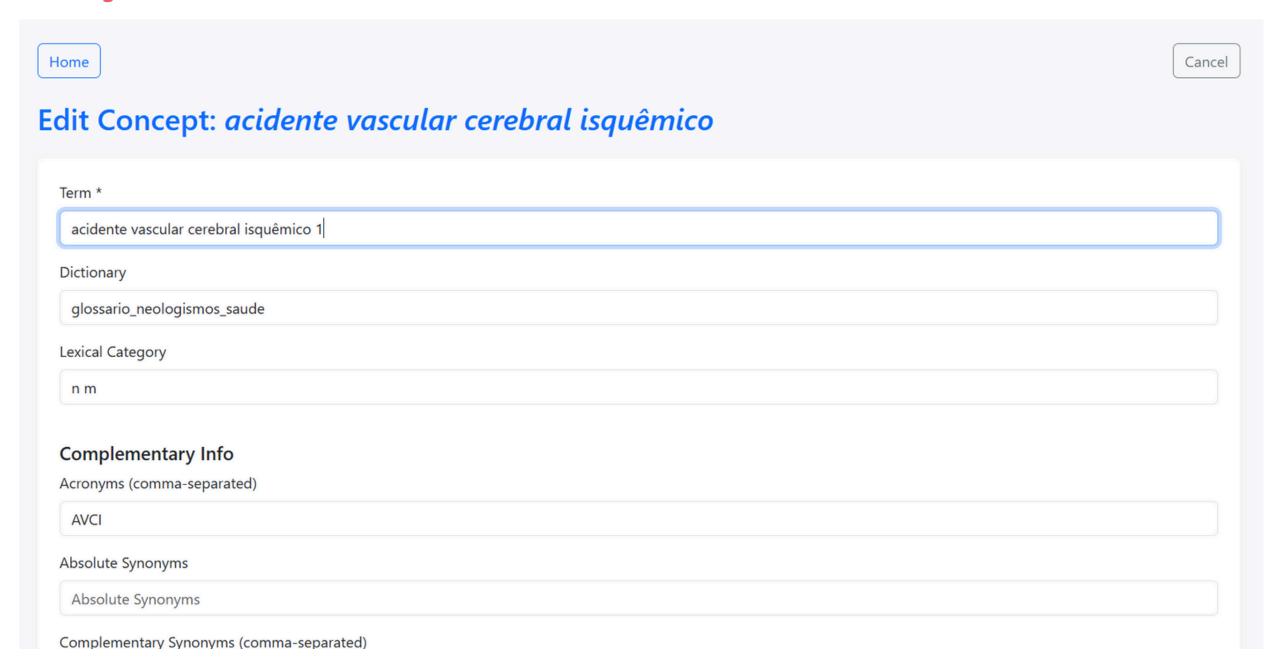
Thematic Area

Description: Lesão que causa a morte de parte do cérebro humano, em virtude da falta de oxigênio e nutrientes que são inseridos no cérebro por meio da circulação sangüínea.

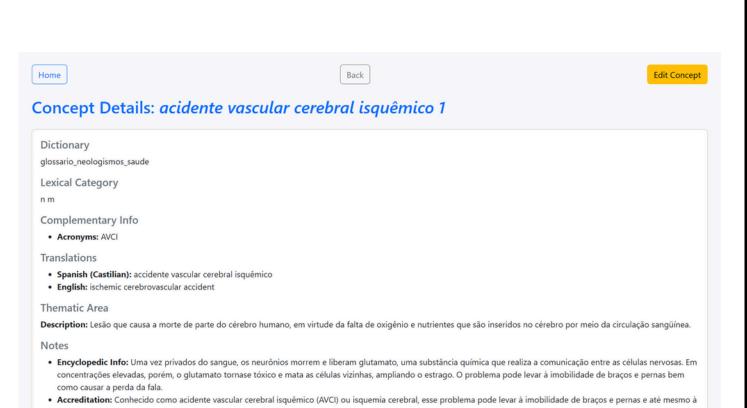
Notes

- Encyclopedic Info: Uma vez privados do sangue, os neurônios morrem e liberam glutamato, uma substância química que realiza a comunicação entre as células nervosas. Em concentrações elevadas, porém, o glutamato tornase tóxico e mata as células vizinhas, ampliando o estrago. O problema pode levar à imobilidade de braços e pernas bem como causar a perda da fala.
- Accreditation: Conhecido como acidente vascular cerebral isquêmico (AVCI) ou isquemia cerebral, esse problema pode levar à imobilidade de braços e pernas e até mesmo à perda da fala....
- ID Number: 37

• Atualização de termos

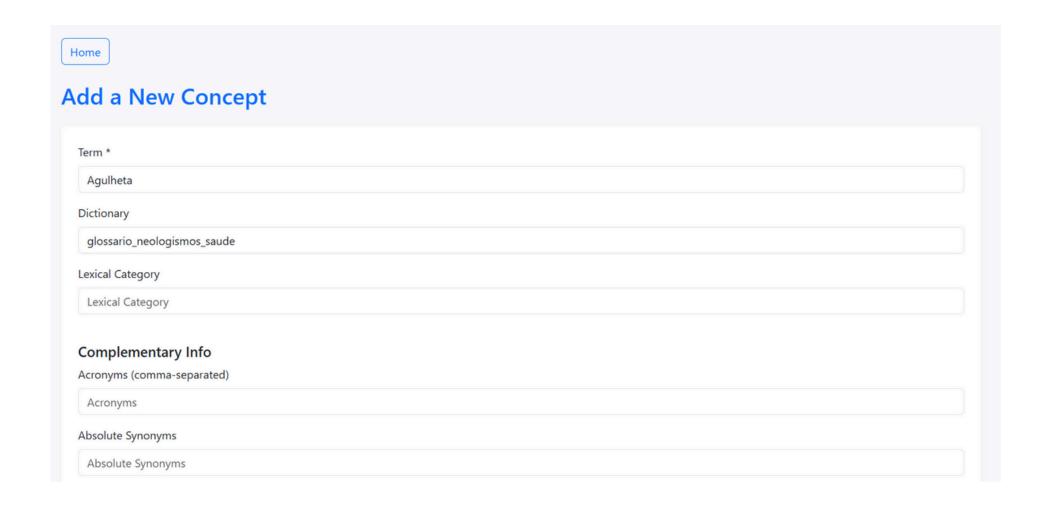


• Atualização de termos



```
acidente vascular cerebral isquêmico 1": {
"dicionario": "glossario_neologismos_saude",
"categoria lexical da designação": "n m",
"complementos designação": {
 "siglas": [
   "AVCI"
  "sinónimos absolutos": "",
  "sinónimos complementares": [],
  "denominação comercial": "",
 "consultar outra entrada": ""
"traduções": {
 "occità": "",
 "basc": "",
  "gallec": "",
  "castellà": "accidente vascular cerebral isquêmico",
  "anglès": "ischemic cerebrovascular accident",
  "francès": "",
  "portuguès": "",
  "[PT]portuguès de Portugal": "",
  "[BR]portuguès del Brasil": "",
 "neerlandè": "",
  "àrab": ""
"códigos alternativos": {
 "sbl": "",
 "nc": "",
  "CAS": ""
"áreas temáticas": {
 "área": "",
  "descrição": "Lesão que causa a morte de parte do cérebro humano, em virtude da falta de oxigênio e nutrientes que são inseridos no cérebro por meio da circulação sangüínea."
"notas": {
 "informacoes_notas": [],
 "Informacao_enciclopedica": "Uma vez privados do sangue, os neurônios morrem e liberam glutamato, uma substância química que realiza a comunicação entre as células nervosas. Em c
  "Abonacao": "Conhecido como acidente vascular cerebral isquêmico (AVCI) ou isquemia cerebral, esse problema pode levar à imobilidade de braços e pernas e até mesmo à perda da fal
  "Numero_identificacao": "37",
  "Marcas_Tipograficas": [],
  "remissiva": "",
  "expandida": ""
```

Adição de termos



```
Agulheta": {
"dicionario": "glossario_neologismos_saude",
"categoria lexical da designação": "",
"complementos designação": {
 "siglas": [],
 "sinónimos absolutos": "",
  "sinónimos complementares": [],
 "denominação comercial": "",
 "consultar outra entrada": ""
},
"traduções": {
  "occità": "",
  "basc": "",
  "gallec": "",
  "castellà": "",
  "anglès": "",
  "francès": "",
  "portuguès": "",
  "[PT]portuguès de Portugal": "",
  "[BR]portuguès del Brasil": "",
 "neerlandè": "",
  "àrab": ""
"códigos alternativos": {
 "sbl": "",
 "nc": "",
  "CAS": ""
"áreas temáticas": {
 "área": "",
 "descrição": "Pedaço de plástico ou metal na ponta de um atacador."
},
"notas": {
 "informacoes_notas": [],
 "Informacao_enciclopedica": "",
  "Abonacao": "",
 "Numero_identificacao": "",
  "Marcas_Tipograficas": [],
 "remissiva": "",
  "expandida": ""
```

Obrigada!