Prétraitement NLP : Guide pédagogique

Par Dr. Asmaa Benguedddach

1. Objectifs pédagogiques

Comprendre et appliquer les étapes de nettoyage, tokenisation et lemmatisation sur des données textuelles médicales, en vue de construire un chatbot ou une base de recherche médicale efficace.

2. Pourquoi préparer les données textuelles ?

Les données brutes contiennent souvent du bruit : majuscules, ponctuation, doublons, formes conjuguées... Ces éléments rendent l?analyse inefficace si non traités. Le prétraitement transforme le texte en une structure exploitable pour l?analyse ou le machine learning.

3. Étapes principales

? Nettoyage : suppression de la ponctuation, mise en minuscules, retrait des caractères spéciaux

? Tokenisation : découpage du texte en unités (tokens) significatives

? Stopwords : retrait des mots fréquents mais peu informatifs (ex: le, de, et...)

? Lemmatisation : réduction des mots à leur racine lexicale (ex: 'running' ? 'run')

4. Exemple de pipeline Python (spaCy)

```
import spacy
import pandas as pd
import re

# Nettoyage basique
def clean_text(text):
    text = text.lower()
    text = re.sub(r'[^a-zA-Z\s]', '', text)
    text = re.sub(r'\s+', ' ', text)
    return text.strip()

# Tokenisation & Lemmatisation
nlp = spacy.load("en_core_web_sm")
```

```
def lemmatize(text):
    doc = nlp(text)
    return [token.lemma_ for token in doc if not token.is_stop and token.is_alpha]
```

5. Résultat attendu

Les colonnes du dataset seront enrichies avec :

- Texte nettoyé
- Liste des tokens
- Liste des lemmes

Cela permet une meilleure correspondance sémantique dans les étapes suivantes : TF-IDF, similarité cosinus, embeddings SBERT, etc.

6. Lien avec le cours

Ce guide est complémentaire au notebook `04_text-preprocessing-nlp-health.ipynb`. Il constitue la base textuelle pour construire un chatbot médical performant.