



LA GESTION INFORMATISÉE DES CONNAISSANCES DE PUBMED ET PUBMED CENTRAL: EXTRACTION

Turki Houcemeddine

Externe en médecine, Faculté de Médecine de
Sfax, Tunisie

Étudiant en informatique, University of the People,
États-Unis d'Amérique

Consultant en web sémantique, Wikimedia
Medicine, États-Unis d'Amérique

Consultant en sciences ouvertes, SisonkeBiotik,
Afrique du Sud

Ancien chercheur et co-fondateur, Data
Engineering and Semantics Research Unit,
Université de Sfax, Tunisie

PLAN

Introduction

 Informatique médicale

 Evidence-Based Medicine

 Ingénierie des Données

Gestion des connaissances

 PubMed

 PubMed Central

 Modèles de langage

Conclusion

A low-angle, upward-looking photograph of a massive, rusted steel structure, likely a bridge or tower, against a bright sky. The structure is composed of large, riveted steel plates. A prominent red horizontal bar with slanted ends is superimposed over the center of the image, containing the word 'INTRODUCTION' in white capital letters. Thin red diagonal lines are also visible on the white background areas.

INTRODUCTION



INFORMATIQUE MÉDICALE

Un domaine d'expansion rapide

INFORMATIQUE MÉDICALE

Branche de l'Informatique Biomédicale

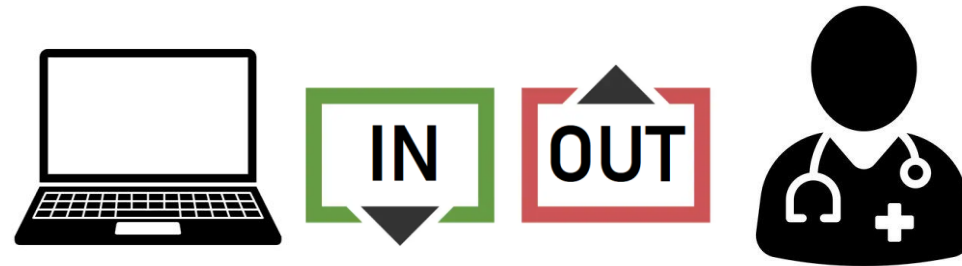
- Évolution exponentielle en rapport avec les dernières avancées de l'Intelligence Artificielle.
- S'intéresse aux données manipulées de façon régulière par les Médecins.
- Essaie d'apporter une aide à la décision clinique.
- Cherche à automatiser quelques tâches biomédicales avec une grande précision.



PLACE DANS L'INFORMATIQUE BIOMÉDICALE

Niveau	Domaine médical	Informatique biomédicale
Population	Santé Publique	Informatique Médicale (Épidémiologie et Données Cliniques)
Individu	Médecine Clinique	
Tissu, Organe	Physiologie et Anatomie	Informatique Médicale (Images Biomédicales)
Cellule	Biologie Cellulaire et Histopathologie	
Molécule	OMICS	Bioinformatique
Atome	Nanomédecine	Chémoinformatique

INFORMATIQUE ET CORPS MÉDICAL



Fausse Perception



Bonne Perception

6D: CE QUE L'AUTOMATISATION VISE À ÉLIMINER

- Dull → Ennuyeux, terne, monotone
 - Dirty → Sale, impur, malpropre
- Dangerous → Dangereux, risqué, périlleux
 - Deep → Profond, intense, abyssal
 - Dear → Cher, précieux, coûteux
- Duration → Durée, laps de temps, période



EVIDENCE-BASED MEDICINE

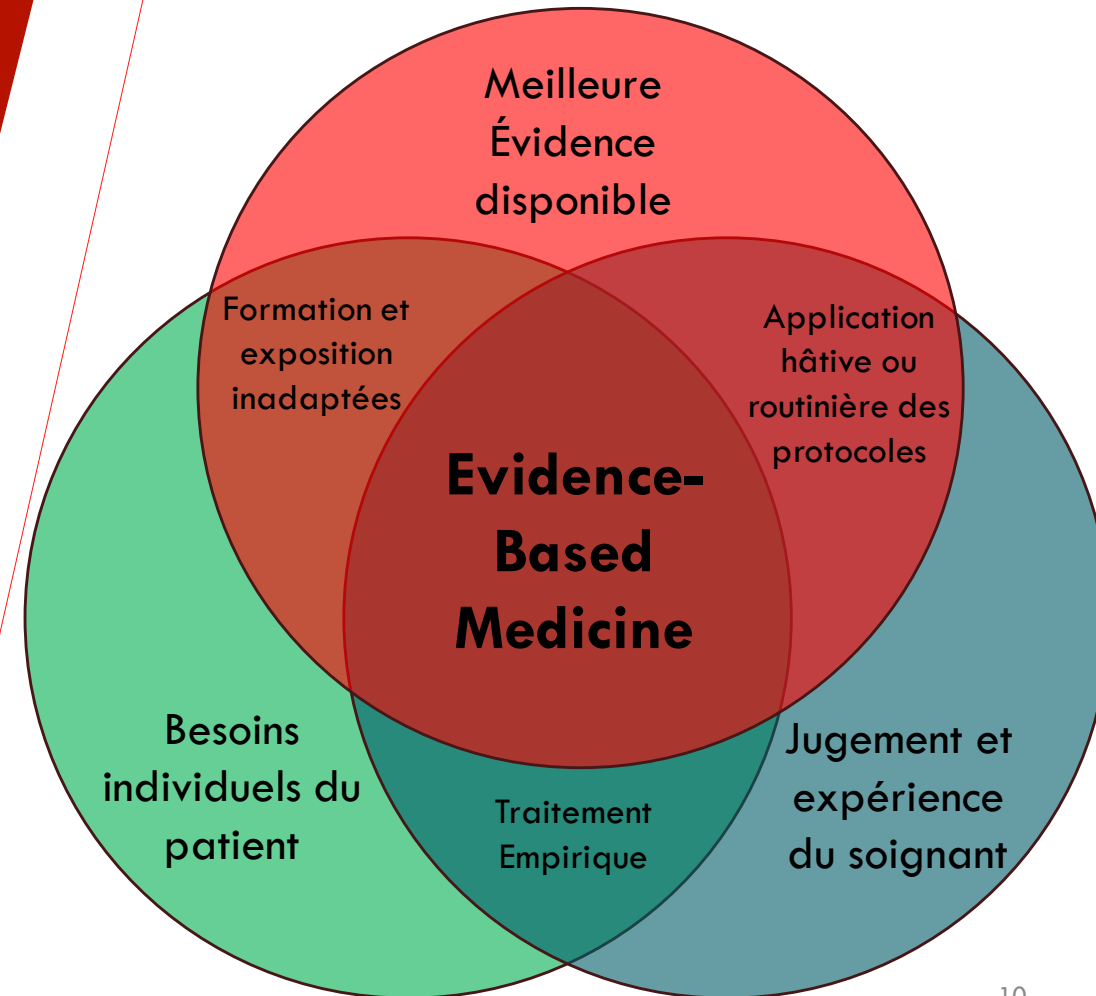
Une pratique basée sur les connaissances



EVIDENCE-BASED MEDICINE

Branche de la pratique médicale personnalisée

- Repose sur l'adaptation de la prise en charge thérapeutique aux pratiques réussies pour les cas similaires.
- Se base sur une lecture exhaustive de la littérature biomédicale.
- Les revues de littérature et les articles de synthèse ont le plus haut niveau d'évidence.
- Les études de cas ont un apport très minime.



POINT FAIBLE

Les connaissances médicales évoluent rapidement

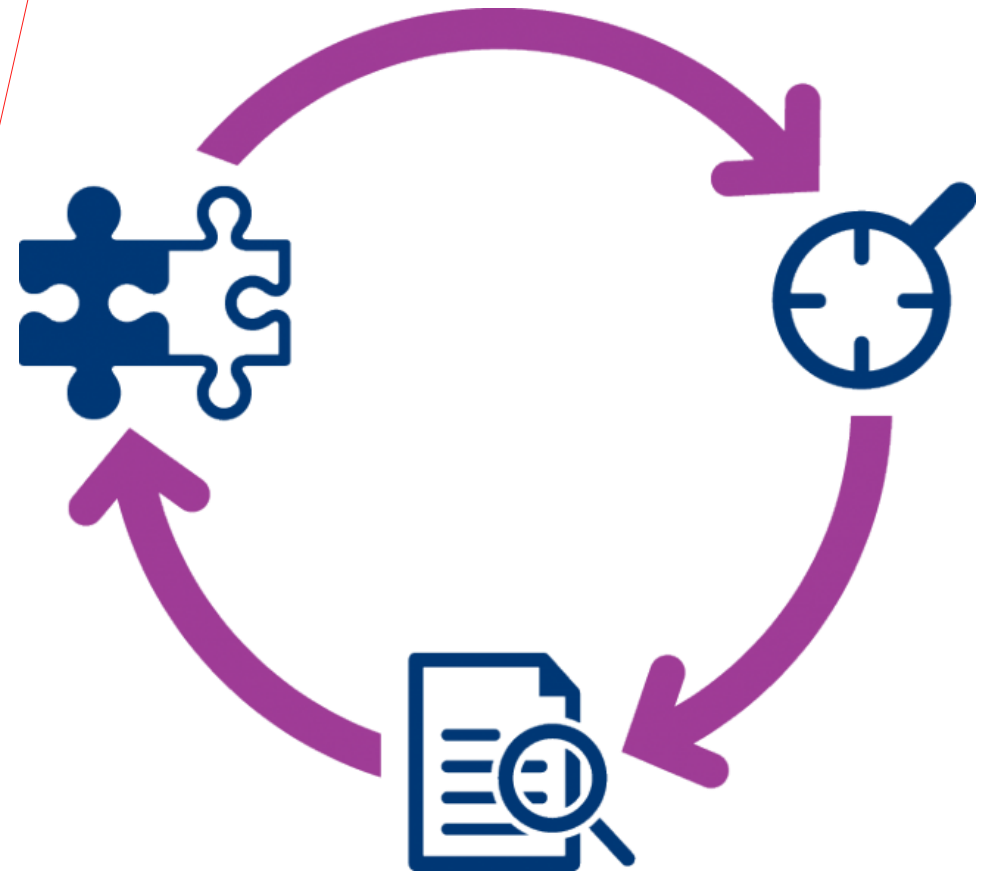
- Ce qui est valable aujourd'hui peut devenir obsolète.
- La production scientifique biomédicale ne peut pas être analysée par un seul chercheur.
- La terminologie médicale change.
- Les exigences de l'intégrité scientifique deviennent plus importantes.



REVUES SYSTÉMATIQUES VIVANTES

Une revue systématique qui se met à jour automatiquement d'une façon très régulière.

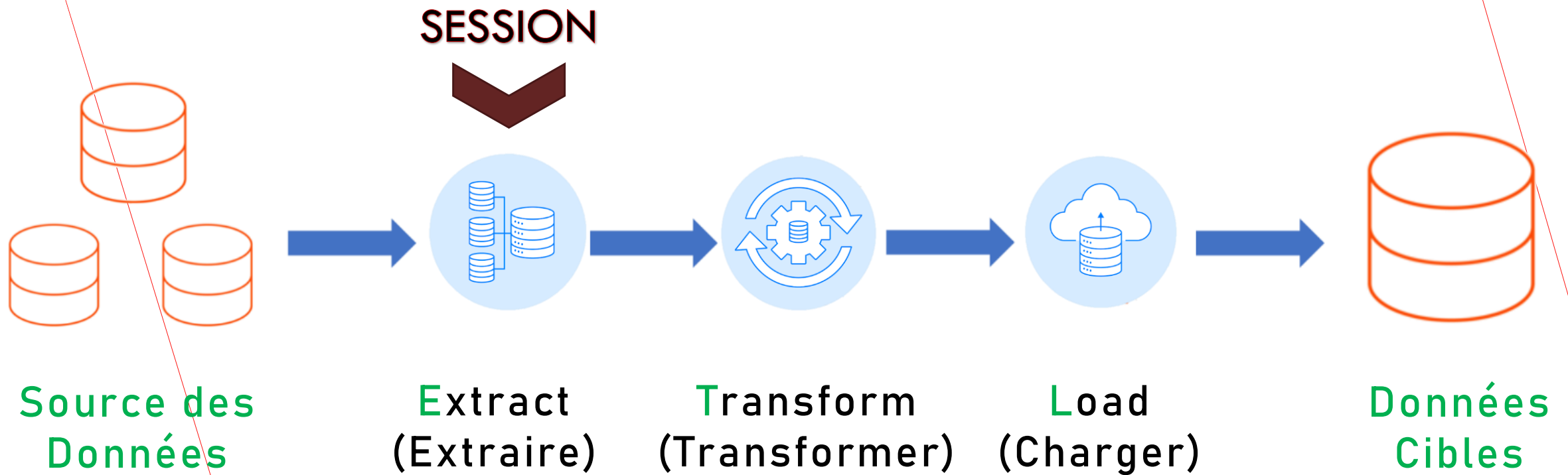
- Est constituée grâce à un cycle de développement: Recherche, Crible, Aspiration, Extraction, Rapport.
- Repose sur les techniques de l'Intelligence Artificielle et des algorithmes de l'extraction des connaissances.
- Utilise quelques pratiques des sciences de l'information comme le Snowballing.
- Concept soutenu par Cochrane.



INGÉNIERIE DES DONNÉES

Une pratique basée sur les connaissances

ETL: PIPELINE DE DONNÉES



CHAMPS D'APPLICATION

RECHERCHE SCIENTIFIQUE

LES DONNÉES BIBLIOGRAPHIQUES PEUVENT ÊTRE UTILISÉES POUR ÉVALUER LA **PRODUCTIVITÉ SCIENTIFIQUE** DANS UN DOMAINE PARTICULIER.

LES TEXTES INTÉGRAUX, LES ABSTRACTS, LES TITRES ET LES MOTS-CLÉS PEUVENT ÊTRE UTILISÉS DANS LE CADRE DE L'AUTOMATISATION DES **REVUES SYSTÉMATIQUES** ET LA CRÉATION DES **GRAPHES DES CONNAISSANCES** BIOMÉDICALES.

PRATIQUE CLINIQUE

LES ARTICLES CONTIENNENT DES DONNÉES CLINIQUES DÉJÀ ANONYMES.

ON PEUT LES UTILISER DANS DES **SYSTÈMES D'AIDE À LA DÉCISION CLINIQUE**.

A low-angle, upward-looking photograph of a massive, rusted steel structure, likely a bridge or industrial building. The structure is composed of large, riveted steel beams and plates, showing significant weathering and rust. The perspective creates a sense of height and scale. A prominent red banner with the word 'RESSOURCES' in white capital letters is superimposed over the center of the image. The background is a clear, light blue sky. Thin red lines are visible in the corners, possibly part of a design template.

RESSOURCES

BASES DES DONNÉES BIBLIOGRAPHIQUES



PUBMED

<https://pubmed.ncbi.nlm.nih.gov/>

Bases des données des métadonnées des publications scientifiques.

Automatisation des recherches en utilisant BioPython.

Extraction des données bibliographiques en utilisant BioPython.



PUBMED CENTRAL

<https://pmc.ncbi.nlm.nih.gov/>

Bases des données des métadonnées et des textes intégraux des publications scientifiques en accès ouvert.

Automatisation des recherches en utilisant BioPython.

Extraction des textes intégraux, des PDFs et des figures en utilisant le dump de PMC.

BIOPYTHON

Bibliothèque Python

- Permet de naviguer toutes les bases des données fournies par National Center for Biotechnology Information (NCBI).
- Ces bases des données incluent plusieurs bases des données en rapport avec la bioinformatique ainsi que PubMed et PubMed Central.
- Ne requiert pas une clé API. Authentification possible grâce à votre adresse de courrier électronique.
- Utilisation relativement intuitive pour les débutants.



MÉTHODES DE TRAITEMENT

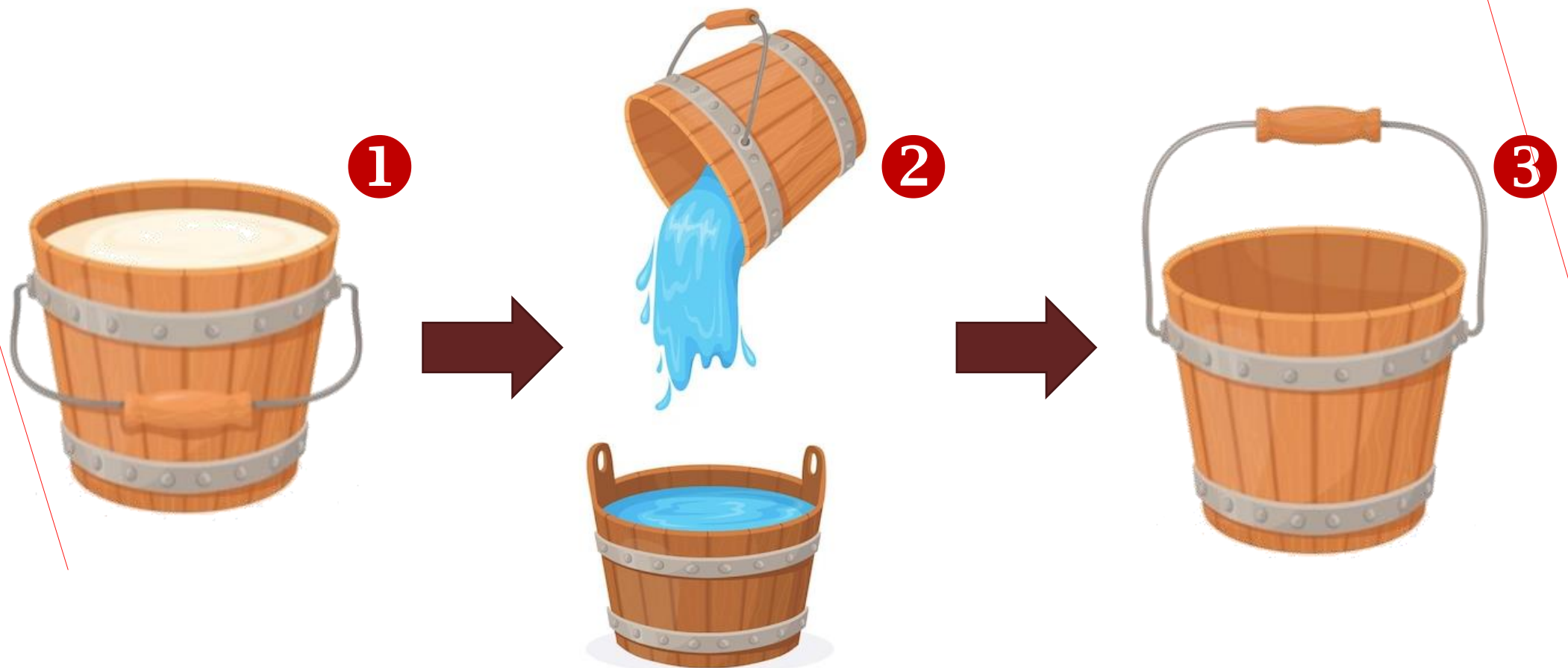
Recherche



Aspiration



PRINCIPE



TRAITEMENT DES TEXTES INTÉGRAUX



Accès au dump PMC



Téléchargement de l'Archive



Décompression de l'Archive



Traitement du fichier XML

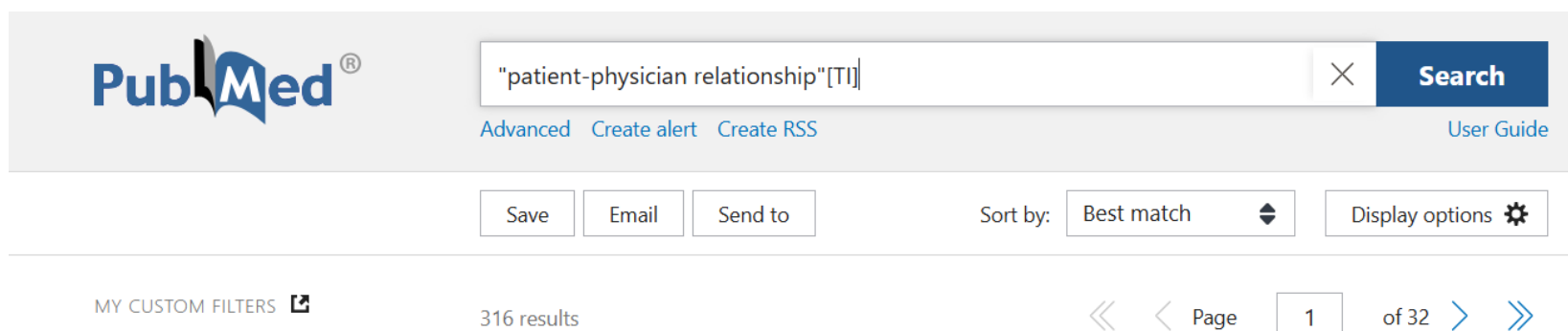
ÉTIQUETTES DU FORMAT PUBMED

Vous pouvez retrouver toutes les étiquettes sur <https://pubmed.ncbi.nlm.nih.gov/help/>.

Étiquette	Nom	Description
AB	Résumé	Résumé en anglais tiré directement de l'article publié
AU	Auteur	Auteurs
CIN	Commentaire dans	Référence contenant un commentaire sur l'article
DP	Date de publication	Date de publication de l'article
JT	Titre complet du journal	Titre complet du journal selon le catalogue NLM
PMID	Identifiant unique PubMed	Numéro unique attribué à chaque citation PubMed
RIN	Rétractation dans	Rétractation de l'article
TI	Titre	Titre de l'article

UTILISATION DES ÉTIQUETTES

Application	Format	Exemple
Requête de recherche	"Valeur"[étiquette]	"patient-physician relationship"[TI]
Extraction des données de BioPython	<code>next(Medline.parse(handle))["TI"]</code>	<code>metadata["TI"]</code>



The screenshot shows the PubMed search interface. At the top left is the PubMed logo. To its right is a search bar containing the query "patient-physician relationship"[TI]. Below the search bar are links for "Advanced", "Create alert", and "Create RSS". To the right of the search bar is a "Search" button. Below the search bar are buttons for "Save", "Email", and "Send to". To the right of these buttons is a "Sort by:" dropdown menu set to "Best match". To the right of the dropdown menu is a "Display options" button with a gear icon. At the bottom left is a link for "MY CUSTOM FILTERS". In the center bottom is the text "316 results". At the bottom right is a pagination bar showing "Page 1 of 32" with navigation arrows.

PublMed®

"patient-physician relationship"[TI]

Advanced Create alert Create RSS

User Guide

Save Email Send to

Sort by: Best match

Display options

MY CUSTOM FILTERS

316 results

Page 1 of 32

MODÈLES DE LANGAGE

Une nouvelle vague qui monte en masse dès l'apparition de ChatGPT en 2023.

C'est un modèle probabiliste conversationnel qui complète une discussion en se basant sur des données d'entraînement.

L'utilisation informatisée des grands modèles de langage comme ChatGPT, Claude, Gemini et d'autres est plafonnée à un certain nombre de requêtes.

Par contre, il existe des petits modèles de langage qui peuvent être stockés dans moins que 16 Go. Ces modèles peuvent être utilisés par une machine locale sans recours à un intermédiaire.

L'utilisation de ces modèles requiert l'utilisation d'une carte graphique performante (GPU) et une bibliothèque Python nommée Llama-cpp-python.



PASSONS À LA PRATIQUE

UN TUTORIEL

Atelier 1 : Extraire des métadonnées bibliographiques de PubMed en utilisant BioPython.

Atelier 2 : Traiter des textes intégraux de PubMed Central en utilisant le dump de PubMed Central.

Atelier 3 : Traiter des données en utilisant des modèles de langage de petite taille et Llama-cpp-python.



MERCI POUR VOTRE
ATTENTION



Université de Sfax
TUNISIE

