

# Système de Recommandation → de Médicaments

Mossab Arektout  
Abderrahim Mabrouk →



# Agenda



- 01 - Introduction
- 02 - Les données
- 03 - Le Pipeline NLP
- 04 - Le Modèle Final
- 05 - Évaluation du Système
- 06 - Limitations & Défis
- 07 - Conclusion

# ↗ Introduction

## Points clés

- Notre monde génère des quantités astronomiques de texte
- Le NLP permet aux machines de "comprendre" le langage humain
- Problématique : Comment organiser automatiquement ce contenu ?

## Objectif du projet

- Mettre en œuvre un système de recommandation basé sur le contenu utilisant le traitement automatique du langage naturel (TALN)
- Analyser la similarité textuelle à l'aide de la vectorisation TF-IDF
- Évaluer la qualité des recommandations à l'aide de métriques de similarité

# 2- Les données ↗

## Caractéristiques du dataset

- **Source:** Medicine dataset with drug information
- **Size:** ~9,720 medicines
- **Fields:**
  - `Drug\_Name`: Name of the medicine
  - `Reason`: Medical condition/reason for use (50 categories)
  - `Description`: Detailed description (290 unique descriptions)

## Repartition

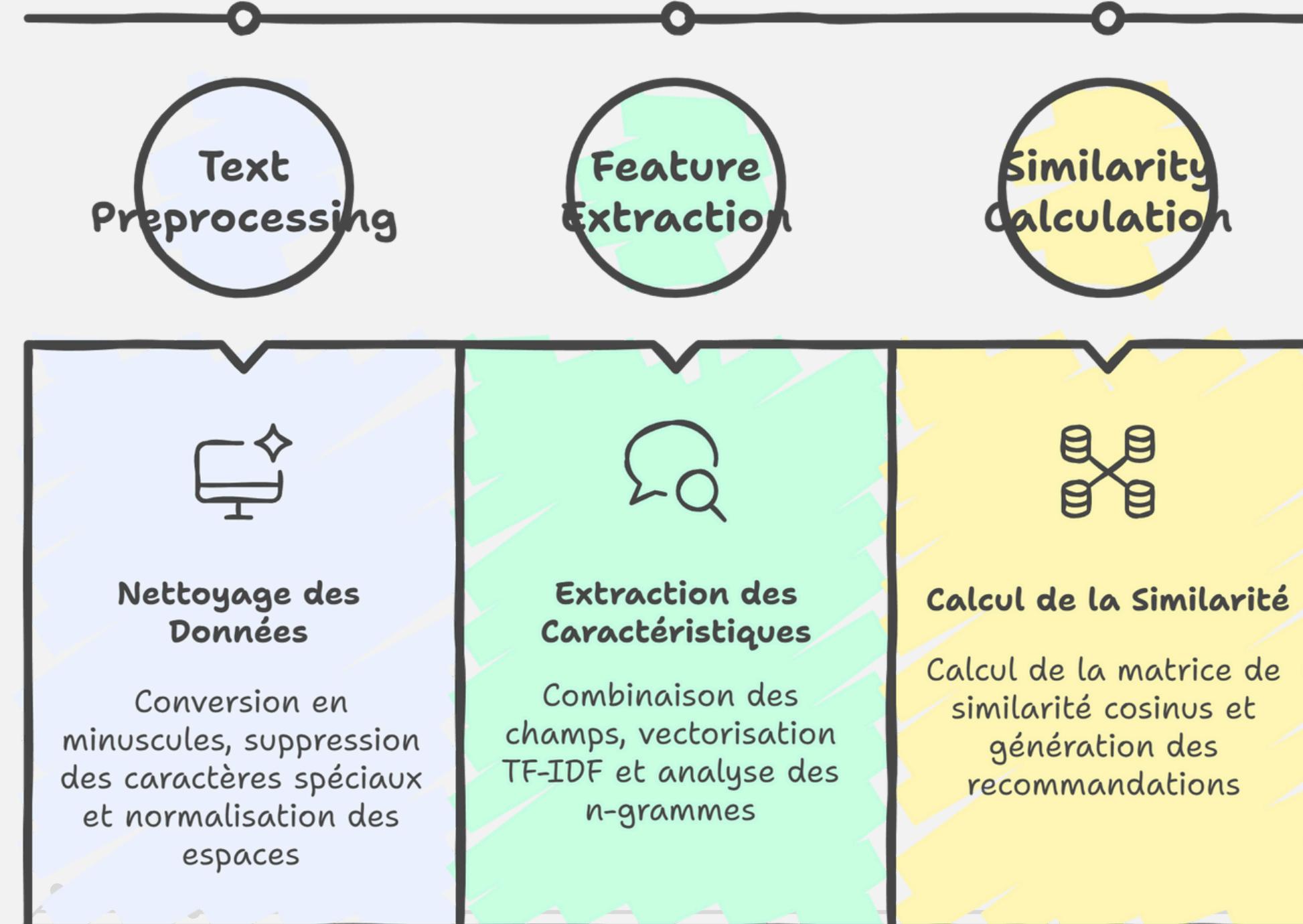
### ===== TOP 10 MEDICAL REASONS (CONDITIONS) =====

#### Reason

Hypertension	2505
Infection	1109
Pain	1072
Fungal	509
Anaemia	252
General	243
Angina	219
Acne	204
Allergies	203
Glaucoma	197

Name: count, dtype: int64

# 3- Le Pipeline NLP ↗



# 3- Le Pipeline NLP

## 3-1- Prétraitement du Texte

**Avant**: "WorldCom ex-boss launches defence! Lawyers defended..."

**Étapes de transformation :**

1. **Gérer les valeurs manquantes (NaN)** → les remplacer par une chaîne vide
2. **Convertir le texte en minuscules** → pour normaliser la casse.
3. **Supprimer les caractères spéciaux** → ne conserver que les lettres, chiffres et espaces.
4. **Normaliser les espaces** → supprimer les espaces, tabulations et sauts de ligne en trop.
5. **Combiner les champs de texte nettoyés** → fusionner plusieurs colonnes en une seule pour faciliter l'analyse NLP.

**Après**: "worldcom exboss launch defence lawyer defend..."

# 3- Le Pipeline NLP

## 3-2- Feature Extraction

**But:** De Mots à Nombres : La Magie du TF-IDF

**Concept :**

- TF-IDF Vectorization :
  - TF (Term Frequency) : Fréquence du mot dans le document
  - IDF (Inverse Document Frequency) : Rareté du mot dans le corpus
- Analyse N-gram :
  - unigrams et bigrams

**Résultat :** Une matrice numérique où chaque ligne = un article, chaque colonne = un mot pondéré

Analyse N-gram

# 3- Le Pipeline NLP

## 3-3- Calcule de Similarité

**But**: mesurer la similarité entre tous les médicaments pour recommandation ou recherche

**Méthode :**

- **Cosine Similarity sur TF-IDF :**

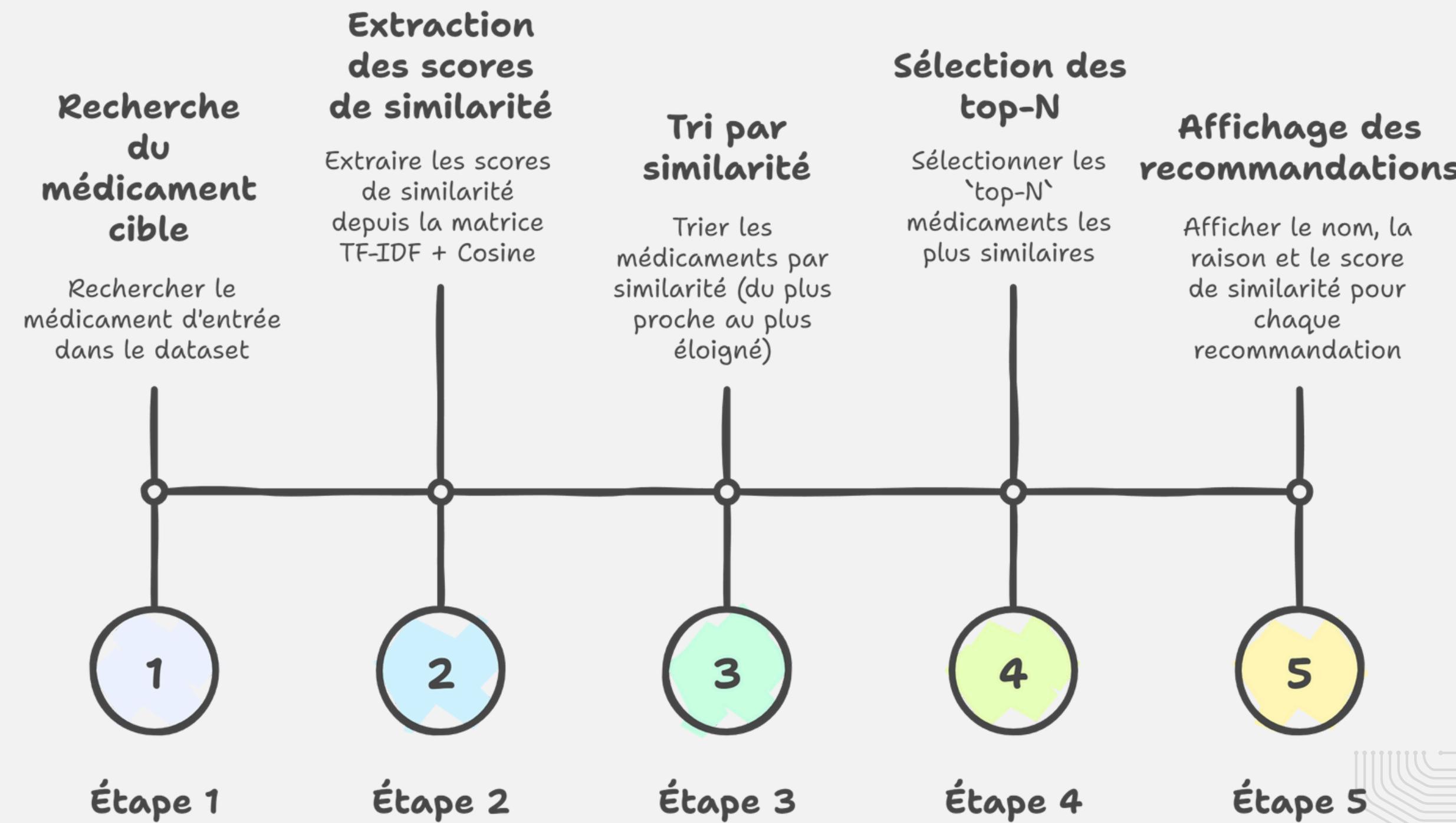
- 0 : Les vecteurs sont orthogonaux (complètement différents)
- 1 : Les vecteurs sont colinéaires (identiques)

$$\text{cosine similarity} = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \times \sqrt{\sum_{i=1}^n B_i^2}}$$

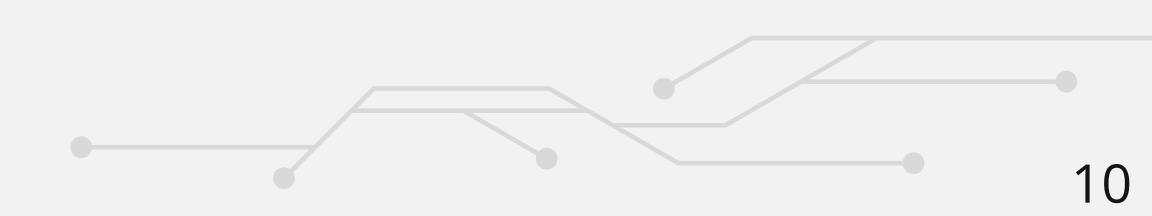
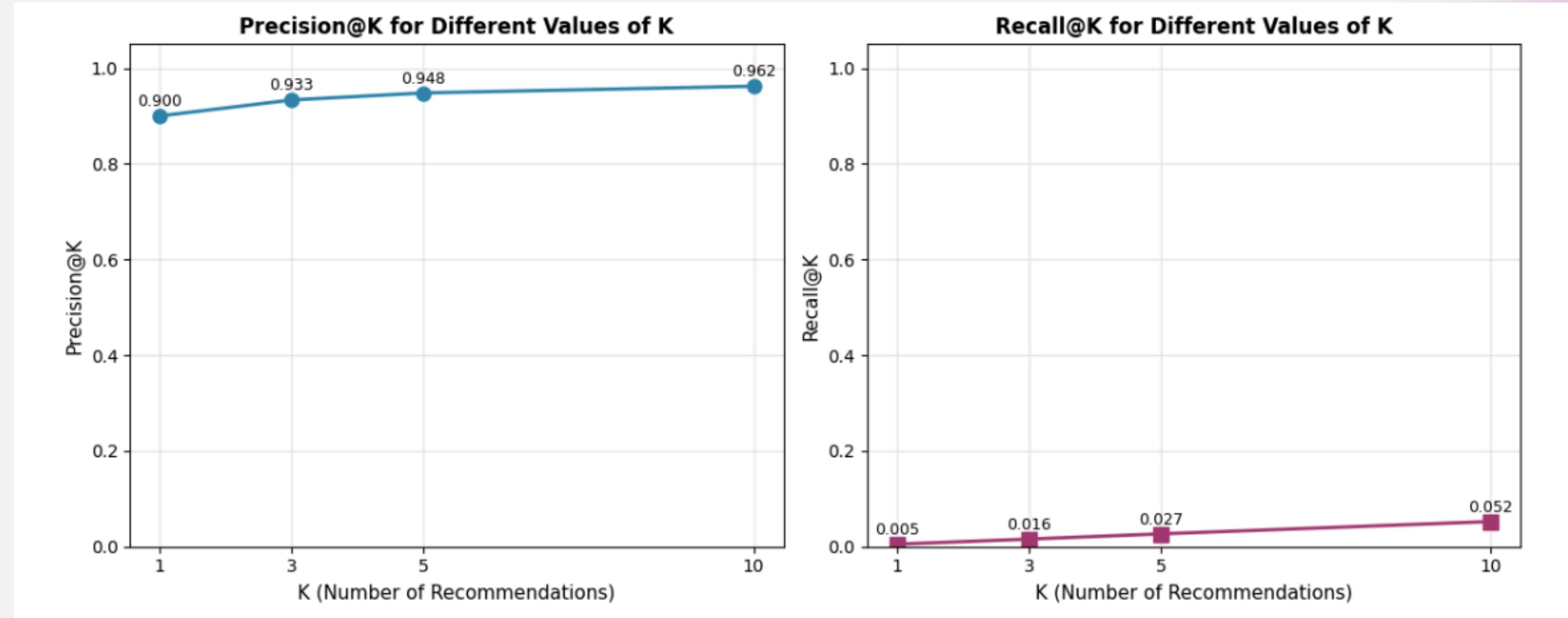
**Après**: matrice symétrique + statistiques (moyenne, médiane, percentiles).

# 4- Le Modèle Final

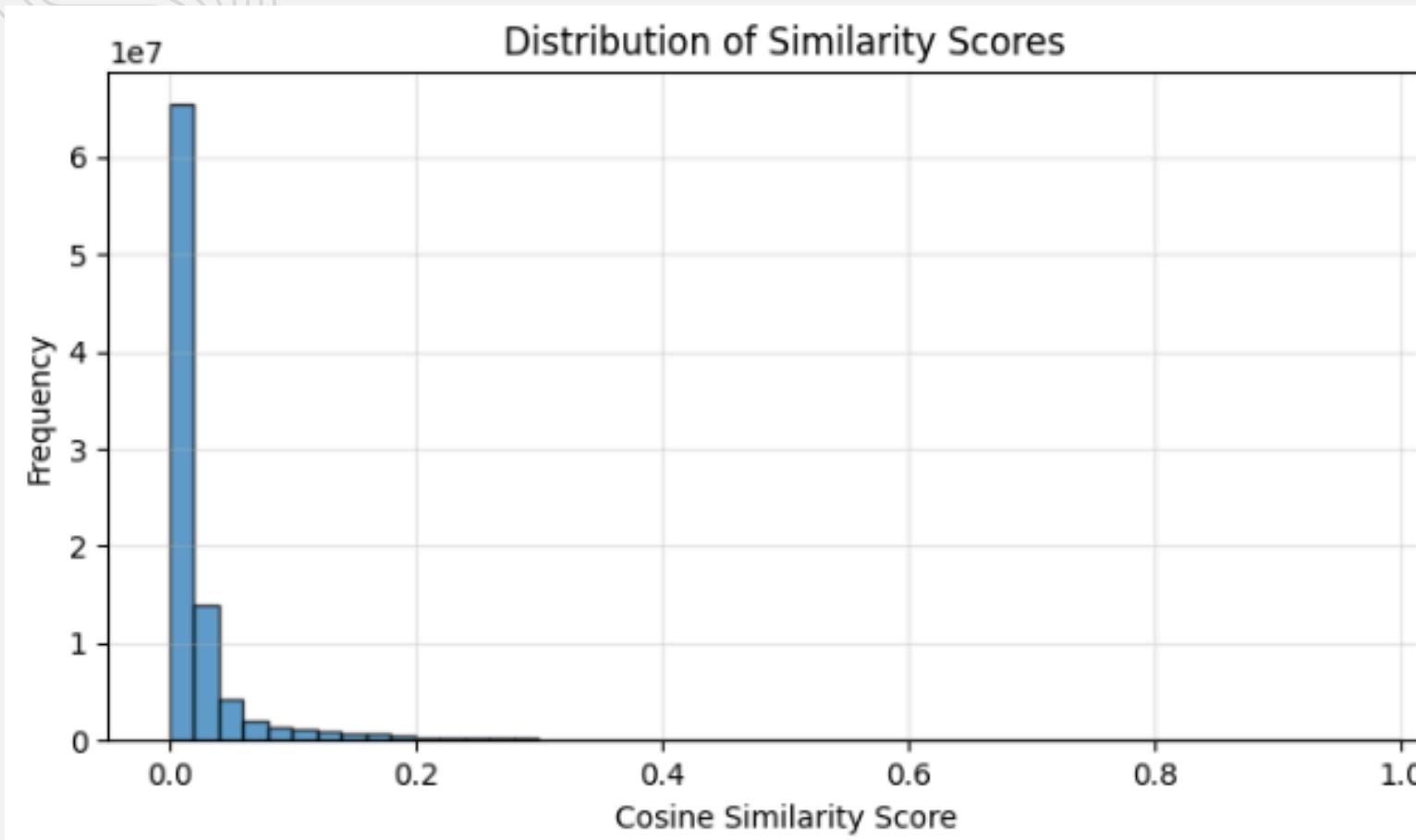
## Étapes du système de recommandation de médicaments



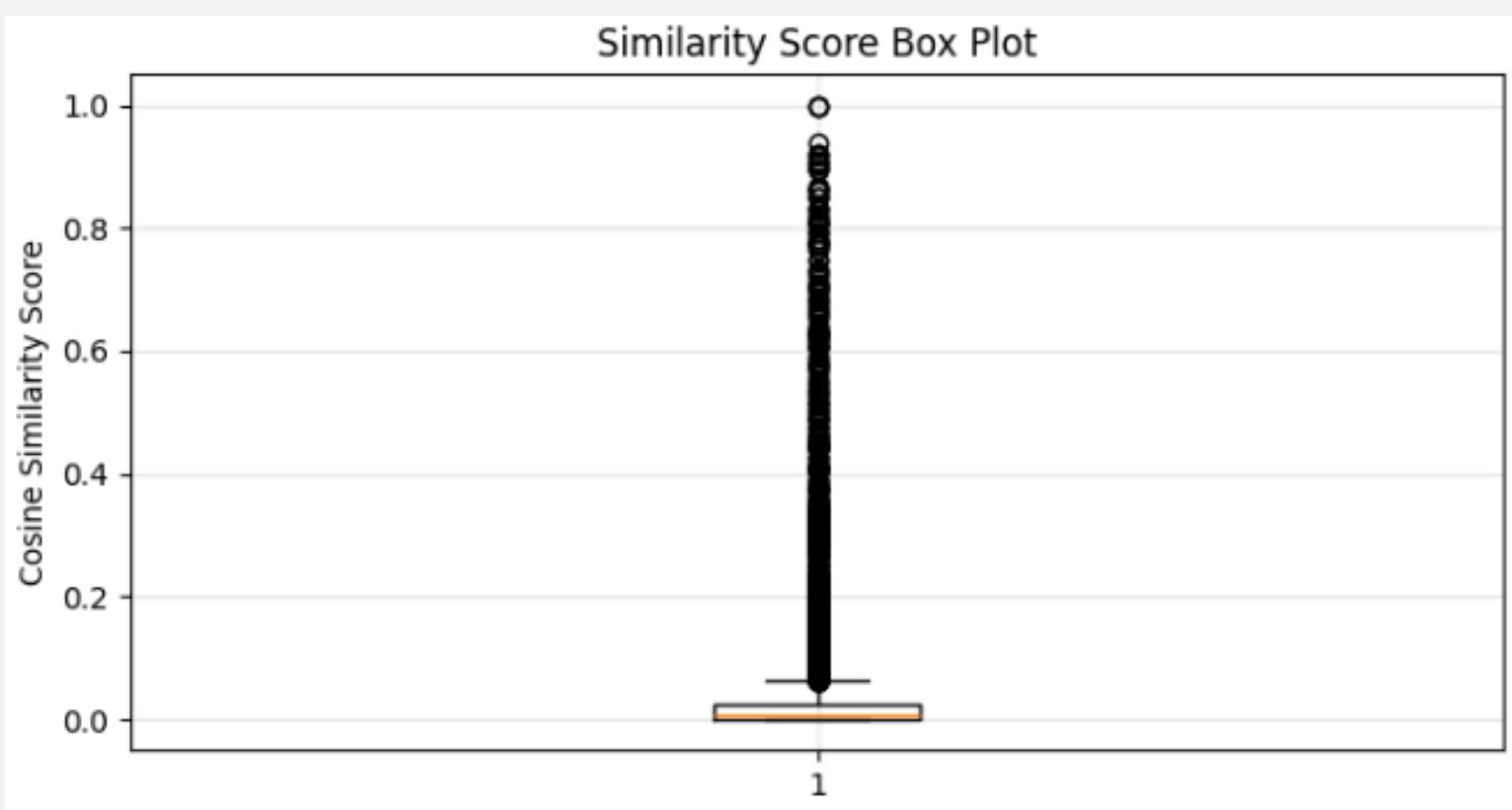
# 5- Évaluation du Système



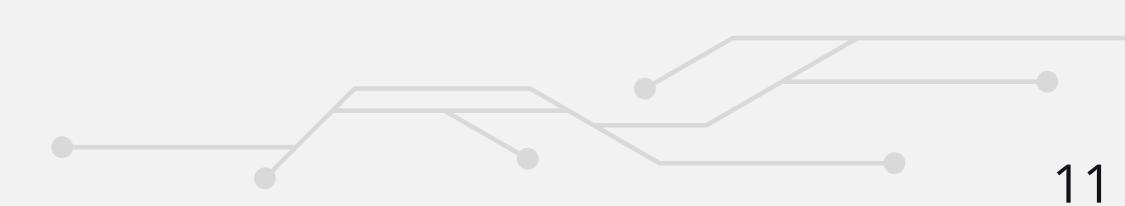
# 5- Évaluation du Système



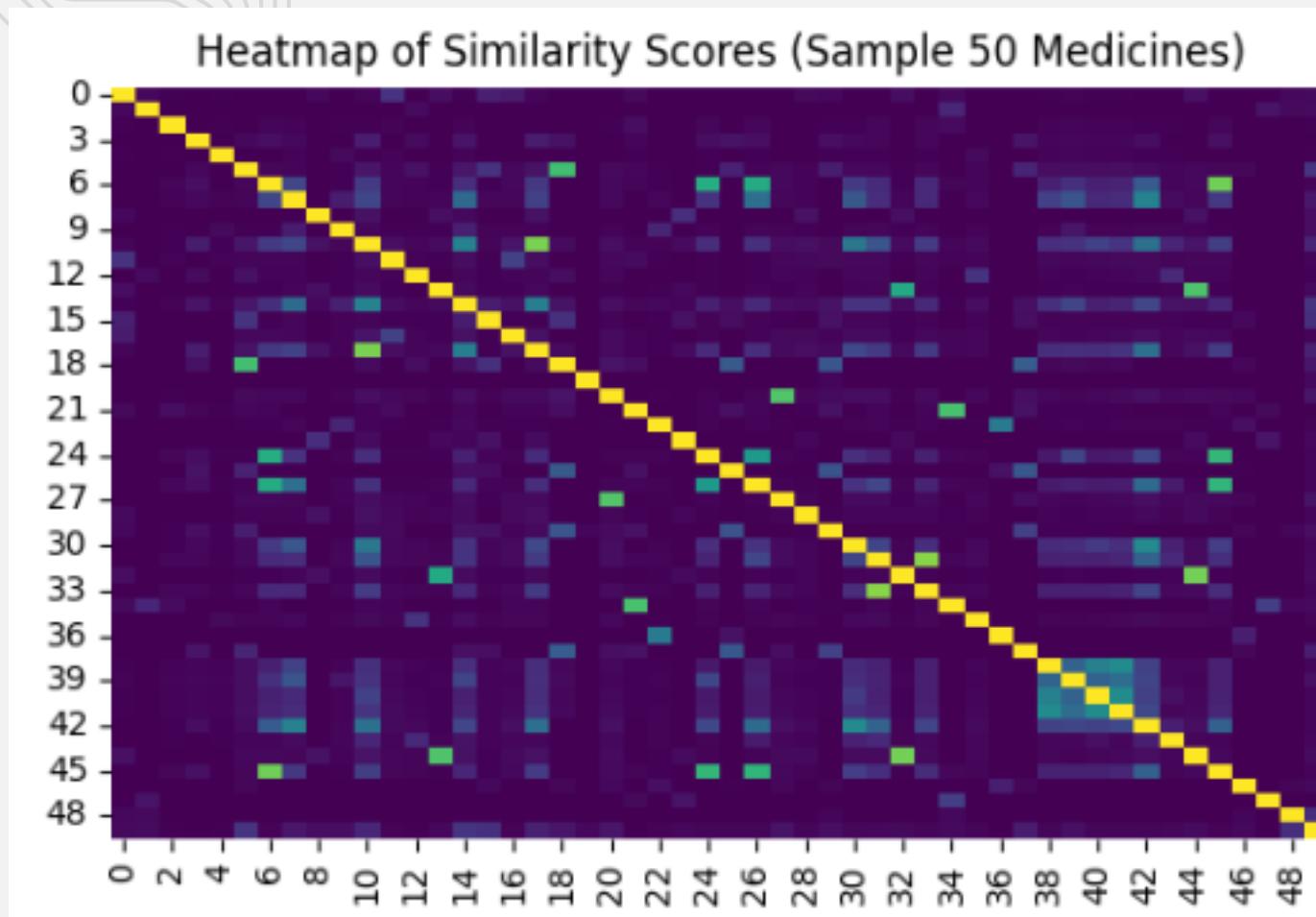
- Concentration massive de scores proches de 0 (~67M paires)
- La majorité des médicaments sont très différents entre eux
- Cela est attendu : un médicament pour le diabète  $\neq$  un antifongique
- Distribution asymétrique : peu de médicaments vraiment similaires



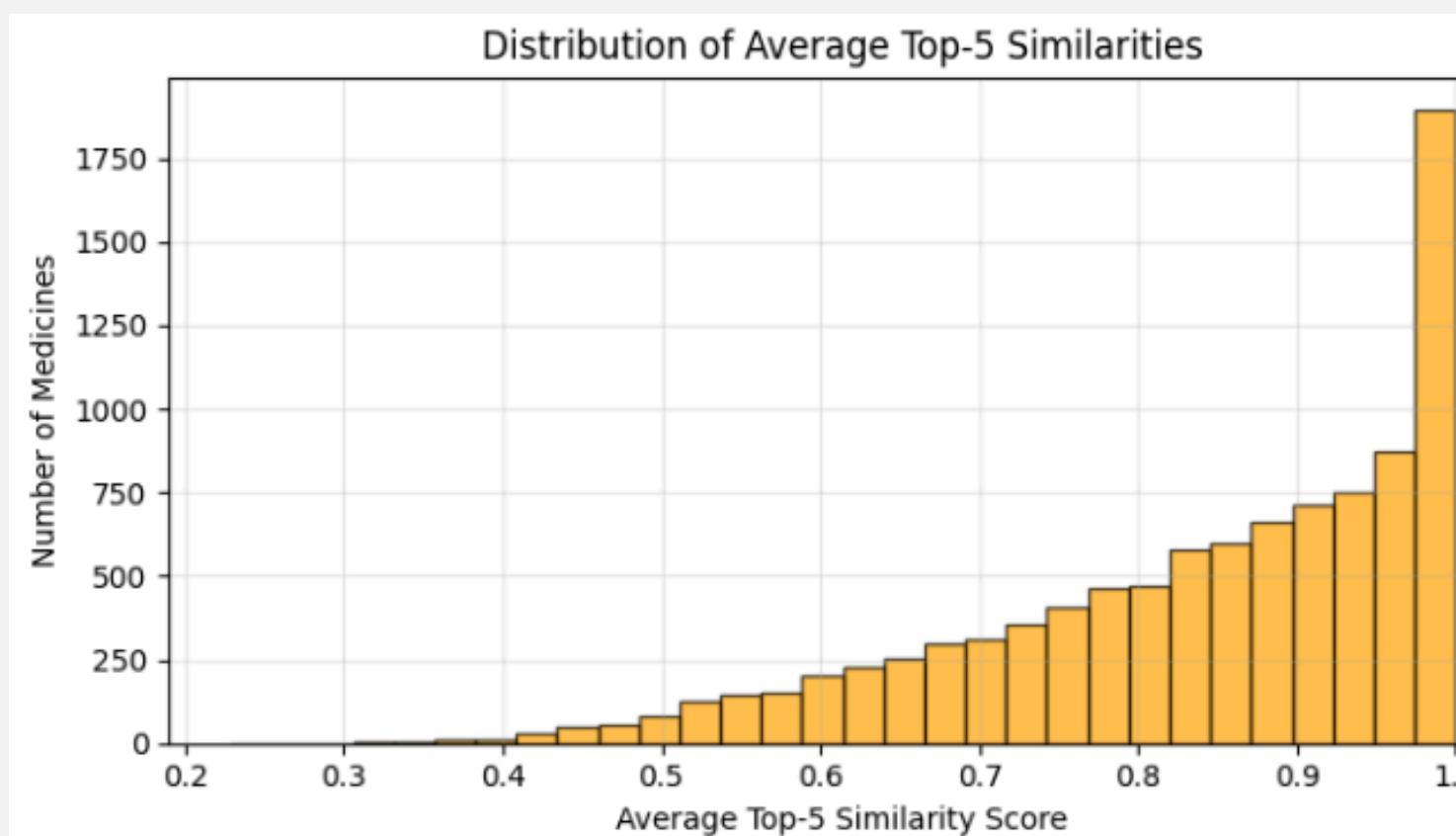
- Médiane  $\approx 0.02$  : la plupart des paires ont une similarité très faible
- - Outliers à 1.0 : médicaments avec descriptions identiques
- - Range interquartile très étroit  $\rightarrow$  forte hétérogénéité du dataset



# 5- Évaluation du Système



- Diagonale jaune = similarité parfaite avec soi-même (score = 1.0)
- - Blocs de couleur : clusters de médicaments similaires (même raison médicale)
- - Majorité violet foncé : confirmation que la plupart des paires sont dissimilaires



- Pic à 1.0 (~1800 médicaments) : beaucoup ont au moins 1 "jumeau" parfait
- - Distribution décroissante 0.9-1.0 : les recommandations sont de haute qualité
- - Très peu de médicaments avec top-5 moyen < 0.5
- - **\*\*Interprétation clé\*\*** : Pour chaque médicament, on trouve facilement 5 alternatives très similaires

# 6- Limitations et Défis

## Terminologie Médicale

- Les noms de médicaments peuvent avoir plusieurs variantes (noms commerciaux vs génériques)
- **Exemple** : "Paracétamol" vs "Acétaminophène" vs "Doliprane"
- La vectorisation TF-IDF ne capture pas cette synonymie médicale

## Vocabulaire Technique Limité

- Les descriptions courtes réduisent la richesse sémantique
- 290 descriptions uniques pour ~9,720 médicaments → beaucoup de duplications
- Risque de sur-recommandation de médicaments avec descriptions identiques

## Complexité Computationnelle

- Matrice de similarité :  $9,720 \times 9,720 = \sim 94M$  comparaisons
- Temps de calcul croît quadratiquement avec le nombre de médicaments
- **Solution** : approximation avec KNN ou indexation (Annoy, FAISS)

# Merci pour votre attention ↗

