

École Nationale Supérieure d’Informatique et d’Analyse des Systèmes – RABAT

Mise en place d'un système de recommandation hybride pour les applications de santé en utilisant le NLP

**Date de soutenance**

**Réalisé par :**

EL OUARTI Mouad  
FIALI Mouad

**Encadré par :**

Pr. SABIRI BIHI

Jury

**Année scolaire :** 2022/2023



Remerciements

Nous tenons tout d'abord à exprimer notre profonde gratitude envers notre professeur EL ASRI BOUCHRA pour avoir proposé ce sujet de recherche captivant et pour sa confiance en notre capacité à mener à bien ce projet. Nous sommes également reconnaissants envers le Pr SABIRI BIHI pour son encadrement précieux tout au long de ce travail.

Nos remerciements sincères vont également aux membres du jury qui ont accepté de consacrer leur temps et leur expertise pour évaluer notre travail. Leur présence et leur contribution sont grandement appréciées.

Nous souhaitons également adresser nos remerciements chaleureux à l'ensemble du corps enseignant de l'ENSIAS. Leur dévouement et leurs efforts inlassables dans la transmission du savoir nous ont permis d'acquérir une formation de qualité.



Résumé

Ce rapport constitue une synthèse de notre projet de fin d'année qui vise à approfondir et à améliorer nos connaissances dans le domaine des systèmes de recommandation, en mettant en avant l'utilisation du traitement automatique du langage naturel. Notre projet se concentre sur l'application de ces techniques dans le domaine de la santé. Nous avons développé un système de recommandation hybride en combinant le filtrage collaboratif, l'approche basée sur le contenu et le traitement automatique du langage naturel. En exploitant ces différentes approches, notre système prédit les maladies des utilisateurs en analysant leurs symptômes et leur propose des solutions personnalisées. Cette approche hybride, intégrant le traitement automatique du langage naturel, nous permet d'améliorer la précision et la pertinence des recommandations fournies.

Abstract

This document provides a summary of our end-of-year project aimed at deepening and enhancing our knowledge in the field of recommendation systems. We specifically focused on applying these techniques in the healthcare domain. Our approach relied on a combination of several methods, including collaborative filtering and content-based approaches. By utilizing these approaches, we developed a hybrid recommendation system that predicts users' diseases based on their symptoms and offers personalized solutions. This hybrid approach allows us to leverage the advantages of both methods to provide more accurate and relevant recommendations. Furthermore, our system incorporates natural language processing techniques to further enhance its performance and effectiveness.

TABLE DES MATIÈRES

[TABLE DES FIGURES 7](#_Toc136523086)

[INTRODUCTION 8](#_Toc136523087)

[CONTEXTE GÉNÉRAL DU PROJET 9](#_Toc136523088)

[1. Contexte général 10](#_Toc136523089)

[2. Problématique 10](#_Toc136523090)

[3. Objectifs 10](#_Toc136523091)

[4. Définitions et étude de l’existant 11](#_Toc136523092)

[**4.1.** **Définitions** 11](#_Toc136523093)

[**4.1.1.** **Systèmes de recommandation** 11](#_Toc136523094)

[**4.1.2.** **Apprentissage automatique** 12](#_Toc136523095)

[**4.2.** **Étude de l’existant** 12](#_Toc136523096)

[**4.2.1.** **DeepHealth** 12](#_Toc136523097)

[**4.2.2.** **Clinical Decision Support Systems (CDSS)** 13](#_Toc136523098)

[BIBLIOGRAPHIE 14](#_Toc136523099)

TABLE DES FIGURES

[Figure 1: Illustration - système de recommandation 11](#_Toc136521731)

[Figure 2: Illustration - Machine learning 12](#_Toc136521732)

INTRODUCTION

Au cours des dernières années, les systèmes de recommandation ont suscité un intérêt considérable en raison de leur capacité à fournir des suggestions personnalisées et à améliorer l'expérience des utilisateurs dans différents domaines. Un domaine où ces systèmes sont particulièrement prometteurs est celui des applications de santé. Avec la disponibilité croissante des données de santé numériques et la demande croissante de solutions de santé personnalisées, la mise en place d'un système de recommandation efficace devient essentielle.

Ce projet vise à relever ce défi en développant un système de recommandation hybride qui combine différentes approches pour fournir des recommandations personnalisées dans le domaine de la santé. Notre approche repose sur l'utilisation du traitement automatique du langage naturel (NLP) pour analyser les symptômes des utilisateurs et prédire les maladies potentielles. De plus, nous intégrons des techniques de filtrage collaboratif et de recommandation basée sur le contenu pour améliorer la pertinence des suggestions fournies.

L'objectif principal de ce projet est de proposer un système de recommandation performant qui aide les utilisateurs à prendre des décisions éclairées en matière de santé. En fournissant des recommandations personnalisées basées sur les symptômes des utilisateurs et en proposant des solutions adaptées, nous visons à améliorer l'efficacité des soins de santé et à optimiser l'expérience des utilisateurs dans le domaine de la santé.

CONTEXTE GÉNÉRAL DU PROJET

*Introduction du chapitre :*

Ce chapitre situe notre projet dans son contexte en identifiant la problématique à résoudre et en définissant notre objectif spécifique. Nous examinons les travaux et les systèmes de recommandation existants dans le domaine de la santé. Nous détaillons également la planification de notre projet, en décrivant les étapes et les ressources nécessaires.

# **Contexte général**

Le projet se situe dans le domaine des applications de santé et vise à développer un système de recommandation hybride utilisant le traitement automatique du langage naturel (NLP). Ce système sera conçu pour prédire les maladies des utilisateurs en se basant sur leurs symptômes et leur fournir des recommandations personnalisées de solutions de santé.

# **Problématique**

La disponibilité croissante de données de santé numériques et la demande croissante de solutions de santé personnalisées nécessitent la mise en place d'un système de recommandation efficace. Cependant, les systèmes de recommandation traditionnels se basant uniquement sur le filtrage collaboratif ou l'approche basée sur le contenu peuvent présenter des limitations en termes de pertinence et de précision. Il est donc nécessaire de développer un système de recommandation hybride qui combine ces approches pour améliorer la pertinence des recommandations dans le domaine de la santé.

# **Objectifs**

On pourra résumer les objectifs de notre projet dans ce qui suit :

* Développer un système de recommandation hybride : L'objectif principal est de concevoir un système de recommandation qui combine le filtrage collaboratif et l'approche basée sur le contenu pour fournir des recommandations personnalisées de solutions de santé aux utilisateurs.
* Utiliser le NLP pour l'analyse des symptômes : Le projet vise à intégrer des techniques de traitement automatique du langage naturel pour analyser les symptômes des utilisateurs et prédire les maladies potentielles.
* Améliorer l'efficacité des soins de santé : En fournissant des recommandations personnalisées et adaptées aux utilisateurs, le projet cherche à améliorer l'efficacité des soins de santé et à faciliter la prise de décision éclairée des utilisateurs.
* Optimiser l'expérience des utilisateurs : L'objectif est de créer un système convivial et intuitif qui offre une expérience utilisateur fluide et satisfaisante lors de l'obtention de recommandations de solutions de santé.

# **Définitions et étude de l’état de l’art**

## **Définitions**

### **Systèmes de recommandation**

Un moteur de recommandation, également appelé système de recommandation, est un outil ou un algorithme qui analyse les données sur les préférences, les intérêts ou les comportements d'un utilisateur pour lui fournir des recommandations personnalisées. Son objectif est d'aider les utilisateurs à découvrir de nouveaux produits, services, contenus ou informations qui correspondent à leurs goûts et à leurs besoins.

Le moteur de recommandation utilise des techniques d'apprentissage automatique, telles que le filtrage collaboratif, l'approche basée sur le contenu, le traitement automatique du langage naturel (NLP) ou d'autres méthodes analytiques, pour analyser les données disponibles. Il compare les informations de l'utilisateur avec celles d'autres utilisateurs similaires ou avec des caractéristiques similaires, afin de générer des recommandations personnalisées et pertinentes.

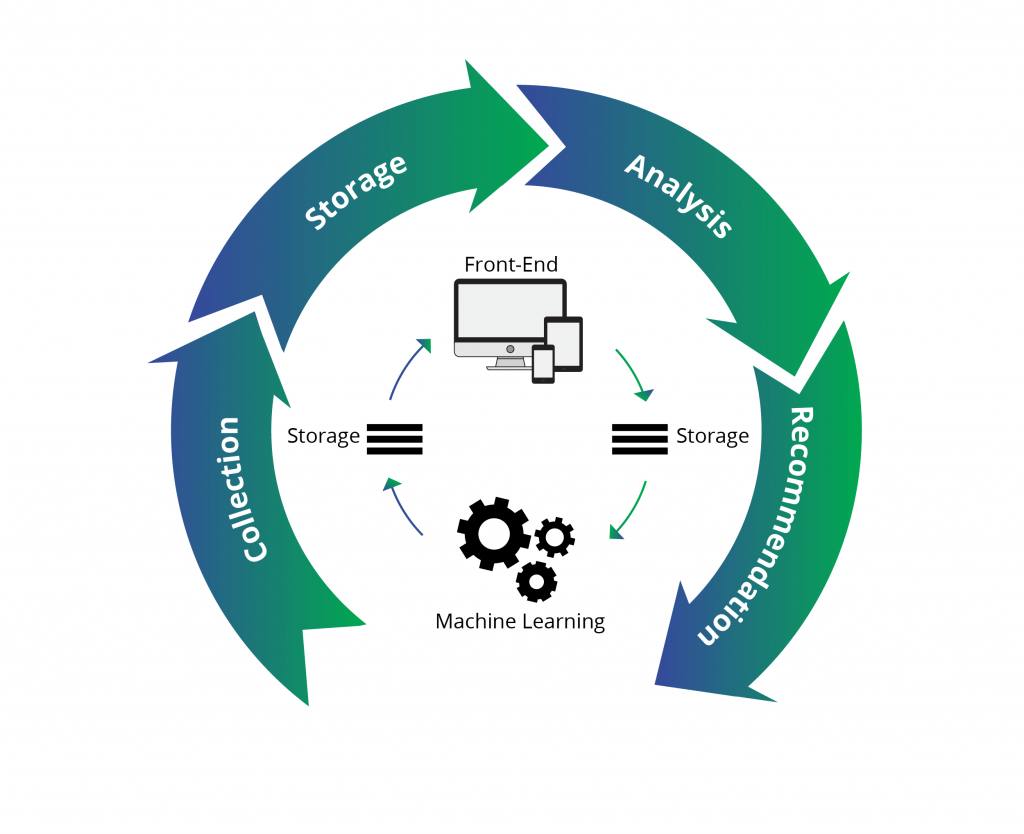


Figure 1: Illustration - système de recommandation

### **Apprentissage automatique**

L'apprentissage automatique, également connu sous le nom de machine learning, est une discipline de l'intelligence artificielle (IA) et de l'informatique qui se focalise sur l'utilisation de données et d'algorithmes pour reproduire de manière progressive le processus d'apprentissage humain, en améliorant continuellement la précision des résultats obtenus.

De plus, l'apprentissage automatique joue un rôle prépondérant dans le domaine en plein essor de la science des données. En utilisant des méthodes statistiques, les algorithmes sont entraînés à réaliser des classifications ou des prédictions, permettant ainsi de mettre en évidence des informations cruciales dans le cadre de projets d'exploration de données. Ces informations servent ensuite de guide pour la prise de décisions au sein des applications et des entreprises.

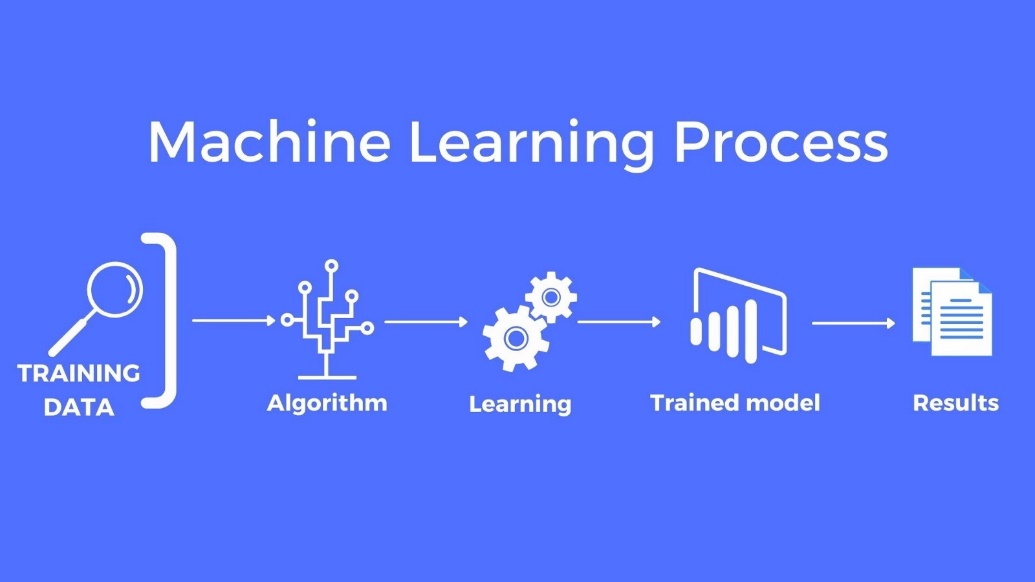


Figure 2: Illustration - Machine learning

## **État de l’art**

### **DeepHealth**

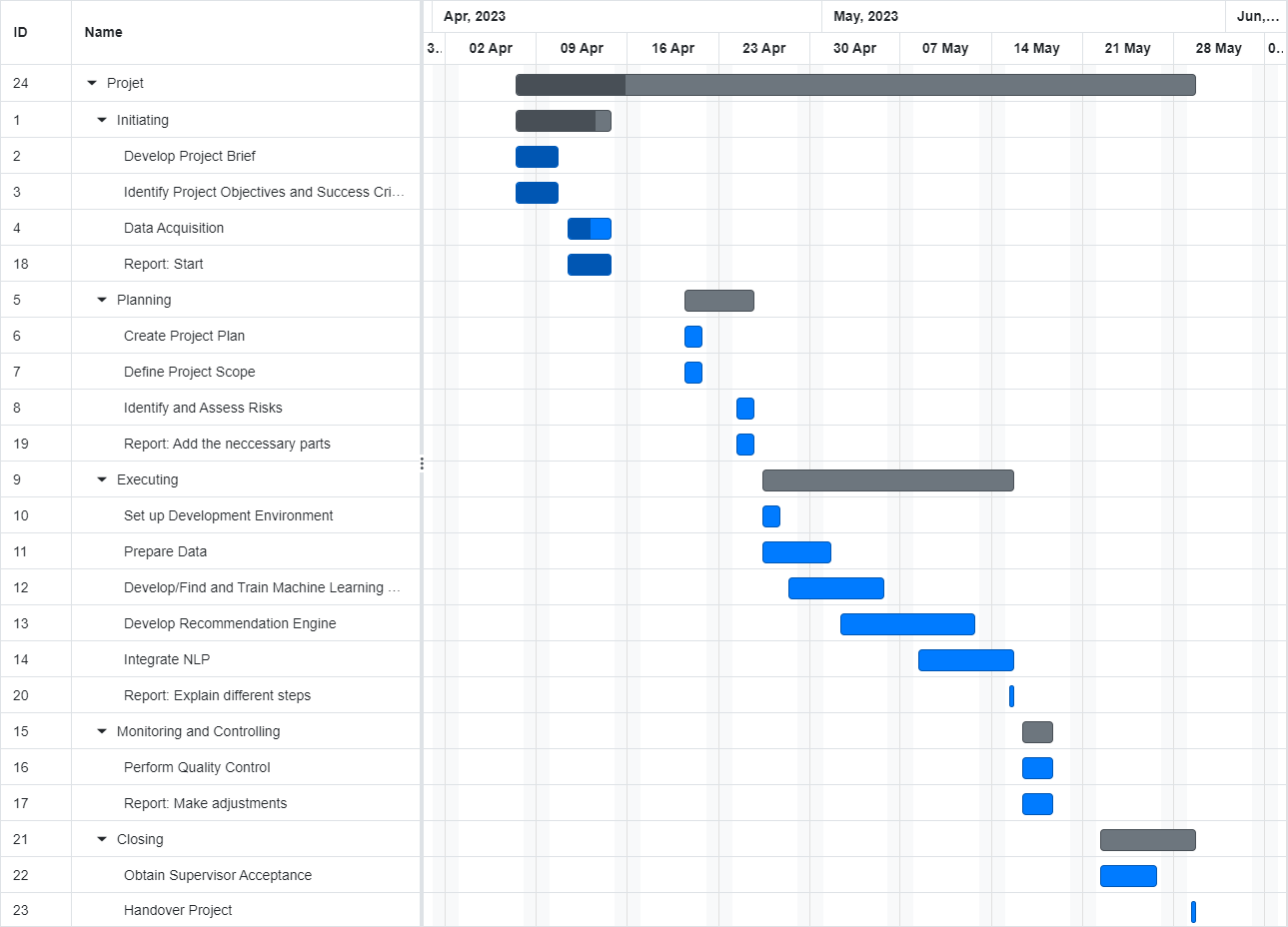
C'est une initiative de recherche qui se concentre sur l'utilisation de l'apprentissage en profondeur (Deep Learning) pour améliorer les systèmes de recommandation dans le domaine de la santé. Il explore des techniques avancées pour la prédiction de maladies, la recommandation de traitements et l'optimisation des soins de santé personnalisés. [1]

### **Clinical Decision Support Systems (CDSS)**

Les systèmes de support à la décision clinique sont des outils informatiques conçus pour aider les professionnels de la santé dans leurs prises de décision en fournissant des recommandations basées sur des données cliniques, des protocoles médicaux et des directives de pratique. [2]

# **Plannification du projet**

La planification du projet revêt une importance capitale dans la phase préliminaire. Elle englobe la prévision du déroulement du projet tout au long des différentes étapes qui composent le cycle de développement, dans notre cas, un cycle en cascade. À cet effet, nous avons élaboré un diagramme de Gantt afin de structurer les différentes phases du projet et d'établir leur échéancier respectif.



Matériel & Méthodologie

*Introduction du chapitre :*

Ce chapitre sert à décrire la population étudiée et son ampleur. Notre objectif est d ‘évaluer des aspects spécifiques, déterminés par des critères de jugement précis. Nous détaillons le déroulement de l’étude étape par étape, tout en mettant en lumière notre approche méthodologique.

# **Population étudiée**

## **Modèle d’extraction de symptômes (NLP)**

Le modèle choisi pour l’extraction des symptômes est un modèle **NER** (**N**ame **E**ntity **R**ecognition) utilisant la bibliothèque de traitement automatique des langues **spaCy**.  
Le modèle a été entrainé sur une base de données de patients présentants différentes conditions médicales ainsi que le code médical auquel le symptome est associé.

# **Objectifs**

## **Modèle NER d’extraction des symptômes**

L’objectif de cette étape était de développer un modèle **NLP**  basé sur **spaCy NER** pour extraire avec précision les symptômes des textes médicaux des patients.  
Les symptômes extraits devraient être utilisées comme entrée pour notre modèle (input hna modèle dyalk), afin d’identifier la maladie correspondante et fournir d’autres informations cliniques et recommandations au patient.

# **Critères de jugement**

## **Modèle NER d’extraction des symptômes**

Les critères de jugement pour l'évaluation de l'extraction des symptômes étaient les suivants :

* **Précision:** La capacité du modèle à extraire les symptômes de manière précise.
* **Rappel:** La capacité du modèle à identifier tous les symptômes présents dans le texte.

# **Déroulement**

## **Collection de données**

### **Description de l'évidence**

Chaque évidence dans le fichier "release\_evidences.json" est décrite en utilisant les entrées suivantes :

**name:** nom de l'évidence.

**code\_question:** un code permettant d'identifier les évidences liées. Les évidences ayant le même code\_question forment un groupe de symptômes connexes. La valeur du code\_question fait référence à l'évidence qui doit être simulée/activée pour que les autres membres du groupe soient éventuellement simulés.

**question\_fr:** la requête associée à l'évidence en français.

**question\_en:** la requête associée à l'évidence en anglais.

**is\_antecedent:** un indicateur indiquant si l'évidence est un antécédent ou un symptôme.

**data\_type:** le type d'évidence. Nous utilisons B pour binaire, C pour catégorique et M pour les évidences à choix multiple.

**default\_value:** la valeur par défaut de l'évidence. Si cette valeur est utilisée pour caractériser l'évidence, c'est comme si l'évidence n'était pas synthétisée.

**possible-values:** les valeurs possibles pour les évidences. Valide uniquement pour les évidences catégoriques et à choix multiple.

**value\_meaning:** La signification, en français et en anglais, de chaque code qui fait partie du champ possible-values. Valide uniquement pour les évidences catégoriques et à choix multiple.

**Description de la pathologie**

Le fichier "release\_conditions.json" contient des informations sur les pathologies dont les patients des ensembles de données peuvent souffrir. Chaque pathologie possède les attributs suivants :

**condition\_name:** nom de la pathologie.

**cond-name-fr:** nom de la pathologie en français.

**cond-name-eng:** nom de la pathologie en anglais.

**icd10-id:** code ICD-10 de la pathologie.

**severity:** la gravité associée à la pathologie. Plus la valeur est basse, plus la pathologie est grave.

**symptoms:** structure de données décrivant l'ensemble des symptômes caractérisant la pathologie. Chaque symptôme est représenté par son nom correspondant dans le fichier "release\_evidences.json".

**antecedents:** structure de données décrivant l'ensemble des antécédents caractérisant la pathologie. Chaque antécédent est représenté par son nom correspondant dans le fichier "release\_evidences.json".

**Description du patient :**

Chaque patient dans chacun des 3 ensembles de données possède les attributs suivants :

**AGE:** l'âge du patient synthétisé.

**SEX:** le sexe du patient synthétisé.

**PATHOLOGY:** nom de la pathologie réelle (propriété "condition\_name" dans le fichier "release\_conditions.json") dont souffre le patient synthétisé.

**EVIDENCES:** liste des évidences vécues par le patient. Une évidence peut être binaire, catégorique ou à choix multiple. Une évidence catégorique ou à choix multiple est représentée sous la forme [nom-de-l'évidence]\_@\_[valeur-de-l'évidence] où [nom-de-l'évidence] est le nom de l'évidence (entrée "name" dans le fichier "release\_evidences.json") et [valeur-de-l'évidence] est une valeur de l'entrée "possible-values". Notez que pour une évidence à choix multiple, il est possible d'avoir plusieurs éléments [nom-de-l'évidence]\_@\_[valeur-de-l'évidence] dans la liste des évidences, chaque élément étant associé à une valeur d'évidence différente. Une évidence binaire est représentée sous la forme [nom-de-l'évidence].

**INITIAL\_EVIDENCE:** l'évidence fournie par le patient pour démarrer une interaction avec un système ASD/AD (Trouble du Spectre Autistique / Déficience Intellectuelle). Cela est utile lors de l'évaluation d'un modèle pour une comparaison équitable des systèmes ASD/AD, car ils commenceront tous une interaction avec un même patient à partir du même point de départ. L'évidence initiale est sélectionnée de manière aléatoire parmi les évidences binaires trouvées dans la liste des évidences mentionnée ci-dessus (c'est-à-dire EVIDENCES) et fait partie de cette liste.

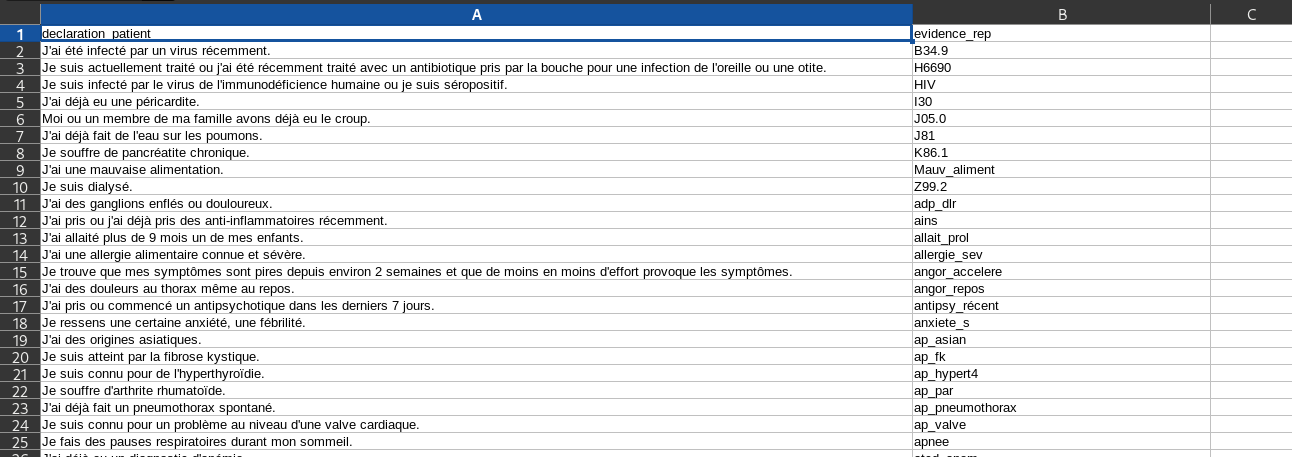
**DIFFERENTIAL\_DIAGNOSIS:** Le diagnostic différentiel réel pour le patient. Il est représenté sous la forme d'une liste de paires de la forme [[patho\_1, proba\_1], [patho\_2, proba\_2], ...] où patho\_i est le nom de la pathologie (entrée "condition\_name" dans le fichier "release\_conditions.json") et proba\_i est sa probabilité associée.

## **Modèle dyalk**

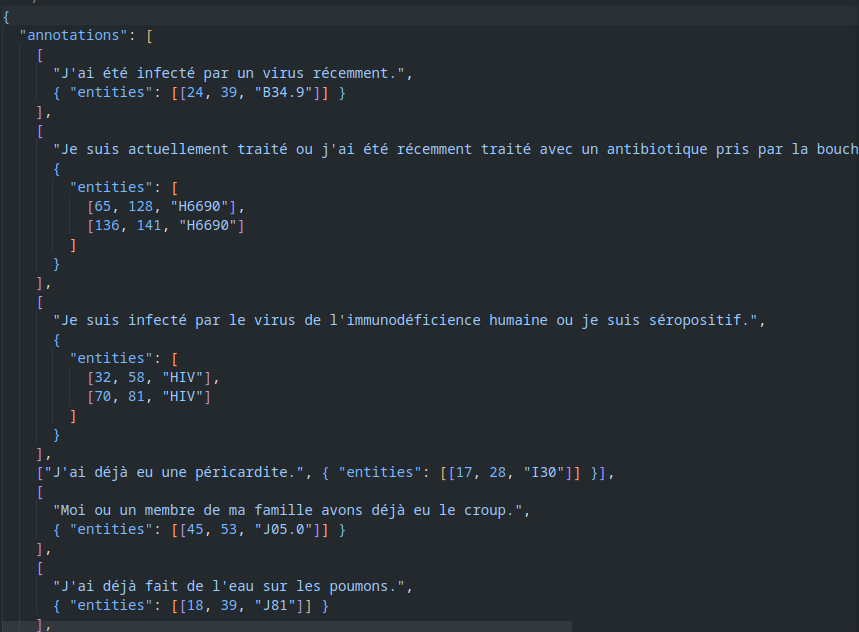
**………………………………………………………………………………………..**

## **Modèle NER**

1. **Adaptation des données :**  Les textes médicaux présents dans la principale base de données ont été personnalisés pour un entrainement du modèle d’extraction des symptômes.

****

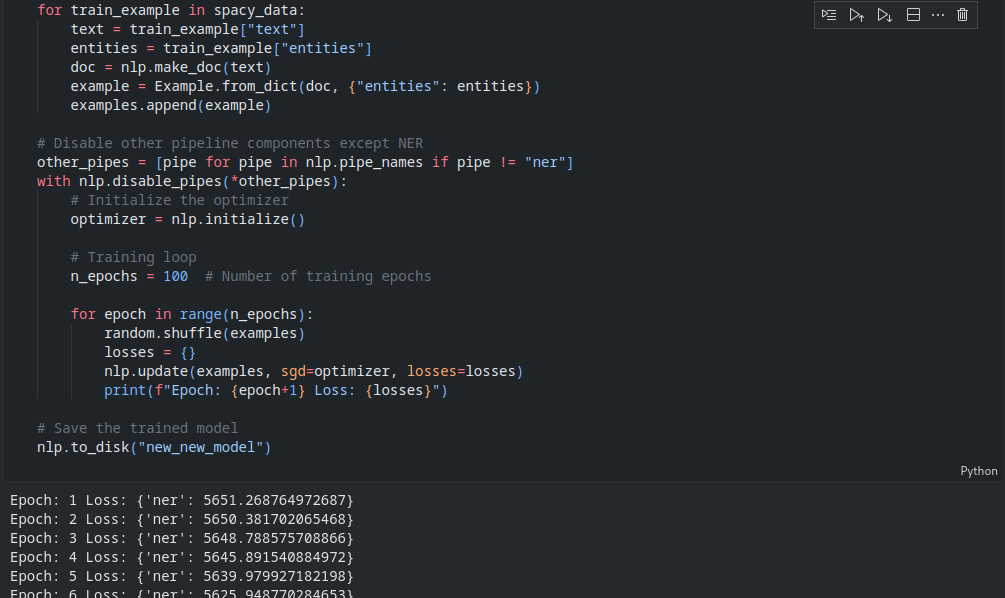
1. **Prétraitement des données :** Les textes sont néttoyés pour supprimer les informations sensibles et les éléments non pertinents.
2. **Annotation des données :** Les symptômes ont été annotés manuellement dans les textes en utilisant des balises spécifiques.

****

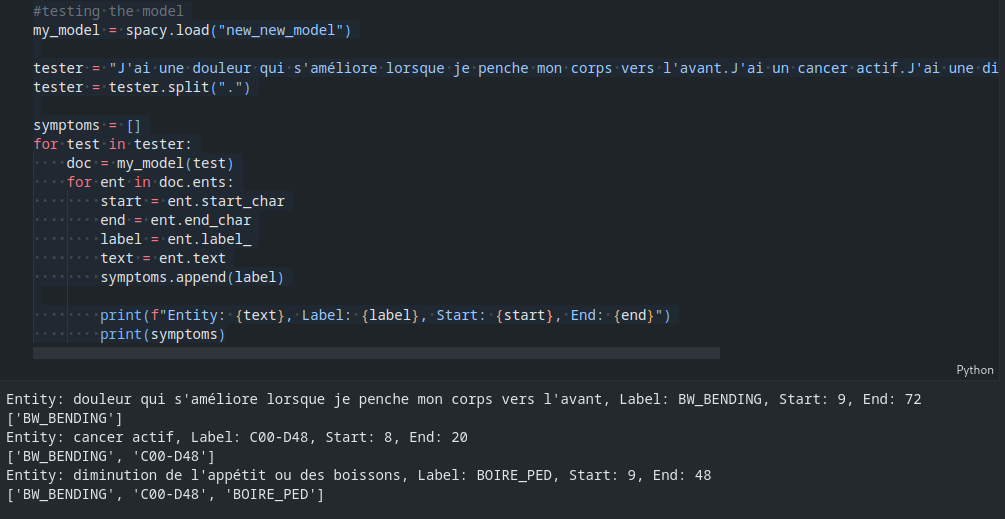
Pour faire, un outil d’annotation en ligne a été utilisé qui est le [NER Text Annotator](https://tecoholic.github.io/ner-annotator/).



1. **Entrainement du modèle :** Le modèle **spaCy NER** a été entraîné sur les données annotées en utilisant un algorithme d’apprentissage automatique.

****

1. **Evaluation du modèle :** Le modèle entraîné a été évalué sur un ensemble de données de test distinct pour mesuré sa performance en termes de précision et de rappel.

****

1. **Analyse de résultats :** Les performances du modèle ont été analysées et interprétées pour évaluer son efficacité dans l’extraction des symptômes.

Résultats

*Introduction du chapitre :*

BIBLIOGRAPHIE

[1] <https://deephealth.com/>

[2] <https://www.techtarget.com/searchhealthit/definition/clinical-decision-support-system-CDSS>