

27/12/2024

Préparé par :

* BELLEBNA Kheira
* GUIROUAOU Nour

**Chatbot médical**

**Covid\_chatbot**

Plan de travail :

1. Introduction

2. Public visé de ce chatbot

3. Méthodologie

3.1 Préparation des données

3.2 Prétraitement du texte

3.3 Vectorisation TF-IDF

3.4 Similarité cosinus

3.5 Seuil pour les questions hors du jeu de données

3.6 Intégration Streamlit

4. Améliorations et orientations futures

5. le résultat final de covid\_chatbot

6. Conclusion

1. Introduction :

Ce rapport fournit une analyse complète du chatbot médical développé pour répondre aux questions des utilisateurs concernant la COVID-19. Le chatbot exploite un ensemble de questions et réponses fréquemment posées (FAQ) sur la COVID-19 afin de fournir des informations pertinentes aux utilisateurs. Des techniques de traitement automatique du langage naturel (TALN) telles que le nettoyage de texte, la tokenisation, la lemmatisation, le TF-IDF et la similarité cosinus sont employées pour comprendre les requêtes des utilisateurs et récupérer les réponses les plus appropriées.

1. Public visé :

D'après le rapport que vous avez fourni, le chatbot médical a été développé pour répondre aux questions des utilisateurs sur la COVID-19.

En se basant sur cette information, on peut déduire que le public cible de ce chatbot est le grand public, c'est-à-dire toute personne cherchant des informations sur la COVID-19.

Cependant, le rapport mentionne également que le chatbot pourrait évoluer pour devenir un outil précieux pour les professionnels de la santé.

Donc, le public cible peut être divisé en deux catégories :

1. Le grand public :
   * Personnes souhaitant obtenir des informations générales sur la COVID-19 (symptômes, transmission, prévention, vaccins).
   * Personnes ayant des questions spécifiques sur la COVID-19 et cherchant des réponses rapides et fiables.
2. Les professionnels de la santé :
   * Médecins, infirmiers et autres professionnels de la santé qui ont besoin d'accéder rapidement à des informations médicales sur la COVID-19.
   * Chercheurs et étudiants en médecine qui souhaitent explorer les données et les informations liées à la COVID-19.

Il est important de noter que le chatbot est encore en développement et que son public cible pourrait évoluer à mesure que ses fonctionnalités s'étoffent. Par exemple, si le chatbot est étendu pour couvrir d'autres sujets médicaux, son public cible pourrait s'élargir pour inclure des personnes atteintes de maladies chroniques, des personnes âgées, etc.

1. Méthodologie :

3.1 Préparation des données :

La base de connaissances du chatbot est extraite du fichier "covid\_faq.csv" contenant des questions et des réponses relatives à la COVID-19.

J’ai importé dataset sur colab pas google drive

* Gestion des valeurs manquantes : La fonction df.info() est utilisée pour identifier et gérer les valeurs manquantes dans le jeu de données, garantissant l'exhaustivité des données.
* Nettoyage des données : Le jeu de données est soumis à un processus de nettoyage rigoureux pour éliminer les caractères non pertinents, convertir le texte en minuscules et corriger les incohérences. Cela garantit la qualité et la cohérence des données.

3.2 Prétraitement du texte :

* Tokenisation : Le texte est divisé en mots ou jetons individuels, permettant au chatbot d'analyser les composants du texte.
* Lemmatisation : Les mots sont réduits à leur forme de base (lemmes) en utilisant la librairie spaCy. Cela normalise les mots et améliore l'efficacité du modèle.
* Suppression des stop words : Les mots courants comme "le", "la", "et" sont supprimés à l'aide de la liste de stop words de spaCy, ce qui permet au modèle de se concentrer davantage sur les mots significatifs.

3.3 Vectorisation TF-IDF :

Le TF-IDF (Term Frequency-Inverse Document Frequency) est appliqué pour convertir le texte en vecteurs numériques, représentant l'importance des mots dans les questions et les réponses.

* Paramètres : Le TfidfVectorizer est configuré avec des paramètres tels que stop\_words, min\_df, max\_df et max\_features pour optimiser le processus de vectorisation. Ces paramètres contrôlent la suppression des stop words, l'inclusion des mots en fonction de la fréquence des documents et le nombre maximal de caractéristiques à prendre en compte (voir dans le document).

3.4 Similarité cosinus :

La similarité cosinus est utilisée pour mesurer la similitude entre la requête de l'utilisateur et les questions du jeu de données.

Ce score de similarité aide à identifier la réponse la plus pertinente à la question de l'utilisateur en comparant les représentations vectorielles de la requête et des questions du jeu de données.

3.5 Seuil pour les questions hors du jeu de données :

Un seuil de similarité est mis en œuvre pour déterminer si la question de l'utilisateur relève du champ d'application du jeu de données.

Si le score de similarité est inférieur au seuil, le chatbot fournit une réponse par défaut ou indique son incapacité à répondre à la question. Cela permet de gérer les attentes des utilisateurs et garantit que le chatbot ne fournit pas d'informations non pertinentes ou incorrectes.

3.6 Intégration Streamlit :

Le chatbot est intégré dans une interface utilisateur en utilisant Streamlit, une librairie Python pour la création d'applications web.

Cela permet aux utilisateurs d'interagir facilement avec le chatbot via un navigateur web.

* Interface utilisateur : L'application Streamlit fournit un champ de saisie de texte permettant aux utilisateurs d'entrer leurs questions et affiche la réponse du chatbot ainsi que le score de similarité. Cela offre une interface conviviale pour interagir avec le chatbot.
* Déploiement : ngrok est utilisé pour créer une URL publique pour l'application Streamlit, la rendant accessible aux utilisateurs sur Internet. Cela permet aux utilisateurs d'interagir avec le chatbot depuis n'importe quel appareil disposant d'un accès Internet.

1. Améliorations et orientations futures :

* Élargir la base de connaissances : Intégrer un plus large éventail de sujets médicaux en incluant davantage de questions et de réponses. Cela peut être réalisé en ajoutant plus de données au fichier "covid\_faq.csv" ou en intégrant des bases de connaissances médicales externes.
* Mettre en œuvre la compréhension contextuelle : Améliorer la capacité du chatbot à comprendre le contexte de la conversation en intégrant des mécanismes de suivi des interactions précédentes de l'utilisateur. Cela permettra au chatbot de fournir des réponses plus personnalisées et pertinentes.
* Intégrer des ressources externes : Connecter le chatbot à des bases de données et des ressources médicales externes afin de donner accès à des informations médicales plus complètes et à jour. Cela pourrait impliquer l'intégration avec des API provenant de sites web médicaux, de bases de données de recherche ou de systèmes d'aide à la décision clinique.

1. Le résultat final de covid\_chatbot :
   1. Conclusion :

Le chatbot médical démontre le potentiel important des techniques de TALN pour fournir aux utilisateurs des réponses rapides et précises aux questions médicales. En élargissant continuellement sa base de connaissances, en affinant ses modèles de TALN et en intégrant des fonctionnalités plus avancées telles que la compréhension contextuelle et l'intégration avec des ressources externes, le chatbot peut évoluer pour devenir un outil précieux pour les professionnels de la santé et le public afin d'accéder à des informations médicales et de prendre des décisions éclairées en matière de soins de santé.

J'espère que ce rapport détaillé offre un aperçu complet du chatbot médical et de son processus de développement. N'hésitez pas à poser d'autres questions.

J'ai fourni une explication plus détaillée des étapes impliquées dans la construction du chatbot médical, en soulignant les techniques et les bibliothèques spécifiques utilisées. Ce rapport enrichi devrait vous donner une compréhension plus approfondie des fonctionnalités du chatbot et de son potentiel d'amélioration. Je suis là pour vous aider si vous avez besoin de plus amples éclaircissements ou si vous avez des aspects spécifiques sur lesquels vous aimeriez que je m'étende.