**UNIVERSITE DE KINSHASA**

**FACULTE DES SCIENCES ET TECHNOLOGIE**

**Département de Mathématiques, Statistiques et Informatique**

Sujet 5 :

Traitement du langage naturel (NLP) : Création d'un programme utilisant NLTK ou SpaCy pour l'analyse de texte ou la création de Chatbots.

**Rapport du travail pratique de Python**

**Par MWAMBA KANDE Franklin**

**Apprenant en DEA**

**Prof.. MASAKUNA Jordan**

**Année académique**

**2022 - 2023**

# 1. Problématique

Le Traitement du Langage Naturel (NLP) représente un domaine de recherche et de développement essentiel dans le domaine de l'intelligence artificielle. Toutefois, il est confronté à une série de défis complexes qui nécessitent une compréhension approfondie et une ingénierie minutieuse pour être surmontés. Ces défis découlent de la nature même du langage humain et des exigences rigoureuses pour son traitement automatique.

Tout d'abord, le langage naturel est intrinsèquement ambigu et variable. Les mots et les phrases peuvent revêtir des significations multiples en fonction du contexte, ce qui rend la tâche de compréhension et d'interprétation difficile pour les machines. Cette ambiguïté est exacerbée par la diversité linguistique, qui englobe des variations syntaxiques, grammaticales et lexicales considérables dans différentes langues et cultures.

De plus, le langage humain est riche en nuances, avec des expressions figurées, des métaphores et des figures de style qui nécessitent une compréhension approfondie de la sémantique et de la pragmatique. L'ironie, le sarcasme et d'autres formes de non-littéralité ajoutent une couche supplémentaire de complexité à la détermination du sens réel des énoncés.

Un autre défi majeur réside dans la gestion des entités nommées et des références, qui sont omniprésentes dans le langage naturel. Identifier et comprendre correctement les noms de personnes, d'organisations, de lieux, etc., ainsi que leurs relations dans le texte, représente une tâche non triviale, en particulier dans des contextes spécialisés ou multilingues.

Par ailleurs, la diversité des expressions linguistiques et des styles de communication soulève des défis pour la détection des intentions et des émotions exprimées dans le langage. Comprendre non seulement ce qui est dit, mais aussi pourquoi cela est dit, est crucial pour une interprétation précise et contextuellement appropriée.

Enfin, sur le plan technique, le traitement efficace du langage naturel implique des algorithmes et des modèles capables de gérer des volumes massifs de données textuelles tout en maintenant des performances rapides et précises. L'évolution rapide des technologies et des besoins en matière d'application ajoute une dimension supplémentaire à cette problématique, nécessitant une adaptation continue et une innovation constante dans le domaine.

Ainsi, la problématique du Traitement du Langage Naturel est profondément enracinée dans la complexité inhérente du langage humain et les exigences rigoureuses pour son traitement automatique. Son exploration et sa résolution exigent une approche multidisciplinaire, combinant des méthodes et des techniques provenant de l'informatique, de la linguistique, de la psychologie cognitive et d'autres domaines connexes pour surmonter ces défis et avancer vers des systèmes NLP plus efficaces et sophistiqués.

# 2. Revu de littérature

## 2.1. Introduction

Le traitement du langage naturel (NLP) a émergé comme un domaine interdisciplinaire de recherche qui vise à permettre aux machines de comprendre, interpréter et générer un langage humain de manière naturelle. Au fil des années, le NLP a connu des avancées significatives grâce à l'application de techniques d'apprentissage automatique et de modèles de traitement du langage.

## 2.2. Modèles Traditionnels de NLP

Les premières approches du NLP étaient basées sur des règles et des algorithmes heuristiques. Des systèmes tels que les arbres de dépendances et les grammaires formelles ont été utilisés pour analyser la structure syntaxique du langage. Cependant, ces modèles étaient limités dans leur capacité à gérer la variabilité et la complexité du langage naturel.

## 2.3. Évolution vers l'Apprentissage Automatique

L'avènement de l'apprentissage automatique a considérablement transformé le paysage du NLP. L'utilisation de techniques telles que les réseaux de neurones et les modèles de langue statistiques a permis d'améliorer la capacité des systèmes à comprendre le contexte et la sémantique du langage. L'émergence des embeddings de mots et des représentations vectorielles a également ouvert la voie à des modèles plus efficaces.

## 2.4. Modèles de Langage et Représentations Distribuées

Les modèles de langage, tels que Word2Vec, GloVe et BERT, ont révolutionné la manière dont les mots et les phrases sont représentés. Ces modèles exploitent des techniques de représentation distribuée pour capturer les relations sémantiques et contextuelles entre les mots. BERT, en particulier, a établi de nouveaux records en exploitant des transformer-based architectures.

## 2.5. Traitement des Séquences et Réseaux de Neurones Récurrents

Les réseaux de neurones récurrents (RNN) ont été largement utilisés dans le traitement des séquences, permettant aux systèmes NLP de gérer des contextes temporels et séquentiels. Cependant, les limitations des RNN en termes de mémoire à long terme ont conduit au développement de modèles récents, tels que les réseaux de neurones à mémoire à court et long terme (LSTM) et les réseaux de transformer.

## 2.6. Défis et Perspectives

Malgré les avancées significatives, le NLP est confronté à divers défis, tels que la compréhension des nuances linguistiques, la résolution d'anaphores et la gestion des biais linguistiques. Les futurs développements pourraient impliquer l'intégration de techniques d'explicabilité, la compréhension des émotions et l'amélioration de la généralisation des modèles.

# 3. Questions de recherche sur le NLP

Quelques questions de recherche pertinentes dans le domaine du NLP pourraient inclure :

Comment améliorer la compréhension des entités nommées dans des domaines spécialisés ou des langues moins courantes ?

Quelles sont les meilleures stratégies pour résoudre le problème de la désambiguïsation lexicale dans les systèmes de dialogue ?

Comment construire des modèles de génération de langage qui produisent des réponses cohérentes et pertinentes dans des contextes complexes ?

Quels sont les impacts éthiques et sociaux de l'utilisation généralisée des systèmes NLP, et comment les atténuer ?

# 4. Fonctionnement du chatbot mis en place

Dans le cadre de ce travail, nous avons mis en place un chatbot capable de répondre aux questions liées à la conjonctivite. Il s’agit d’une affection caractérisée par une inflammation de la conjonctive, la muqueuse qui recouvre la partie blanche de l'œil et la face interne des paupières.

Le chatbot développé utilise un modèle de réseau de neurones artificiels pour classer les questions relatives à la conjonctivite. Il est basé sur un jeu de données d'intentions (patterns et tags) contenu dans le fichier JSON "intents\_conjonctive.json". Le modèle est entraîné à l'aide de la bibliothèque PyTorch et NLTK est utilisé pour la tokenization et le stemming des mots.

Le processus de fonctionnement du chatbot est le suivant :

* L'utilisateur saisit une question sur la conjonctivite.
* Le message est tokenisé et converti en un vecteur de mots.
* Ce vecteur est passé à travers le modèle neuronal pour prédire une intention (tag) associée à la question.
* Sur la base de la prédiction, une réponse appropriée est sélectionnée à partir des réponses disponibles dans le fichier JSON d'intentions.



Figure 1 : Interface du chatbot

# 5. Discussion et commentaire

Le chatbot semble fonctionner de manière satisfaisante pour répondre aux questions simples sur la conjonctivite. Cependant, il existe des limitations, notamment :

* La capacité limitée à comprendre les nuances du langage naturel. Le modèle ne peut pas répondre efficacement aux questions complexes ou ambiguës.
* La nécessité de maintenir et d'enrichir régulièrement le jeu de données d'intentions pour améliorer la qualité des réponses.
* L'absence de gestion poussée des erreurs pour des questions hors du domaine de la conjonctivite.

En résumé, bien que le chatbot fournisse une démonstration de base de l'application des techniques de NLP avec NLTK pour la création d'un système de dialogue, des améliorations futures sont nécessaires pour le rendre plus robuste et capable de gérer une plus grande variété de questions et de contextes.

Contenu

[1. Problématique 2](#_Toc163415795)

[2. Revu de littérature 3](#_Toc163415796)

[2.1. Introduction 3](#_Toc163415797)

[2.2. Modèles Traditionnels de NLP 3](#_Toc163415798)

[2.3. Évolution vers l'Apprentissage Automatique 3](#_Toc163415799)

[2.4. Modèles de Langage et Représentations Distribuées 3](#_Toc163415800)

[2.5. Traitement des Séquences et Réseaux de Neurones Récurrents 3](#_Toc163415801)

[2.6. Défis et Perspectives 3](#_Toc163415802)

[3. Questions de recherche sur le NLP 4](#_Toc163415803)

[4. Fonctionnement du chatbot mis en place 4](#_Toc163415804)

[5. Discussion et commentaire 4](#_Toc163415805)