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.

Problématique

Description

Le NLP pour Natural Language Processing est une discipline qui se concentre sur la compréhension, la manipulation et la génération du langage naturel par des machines. Ainsi, la PNL se situe véritablement à l’interface entre l’informatique et la linguistique . Il s’agit de la capacité de la machine à interagir directement avec les humains.

La PNL est un terme assez générique qui couvre un très large spectre d’applications. Voici les applications les plus populaires :

Traduction automatique

Le développement d’algorithmes de traduction automatique a véritablement révolutionné la manière dont les textes sont traduits aujourd’hui. Des applications, telles que Google Translate , sont capables de traduire des textes entiers sans aucune intervention humaine.

Le langage naturel étant par nature ambigu et variable, ces applications ne reposent pas sur le remplacement mot à mot, mais nécessitent une véritable analyse et modélisation de texte , connue sous le nom de traduction automatique statistique.

Analyse des sentiments

Également connue sous le nom d'« Opinion Mining », l'analyse des sentiments consiste à identifier des informations subjectives dans un texte pour en extraire l'opinion de l'auteur.

Par exemple, lorsqu’une marque lance un nouveau produit, elle peut utiliser les commentaires recueillis sur les réseaux sociaux pour identifier le sentiment global positif ou négatif partagé par les clients.

En général, l'analyse des sentiments est un moyen de mesurer le niveau de satisfaction des clients à l'égard des produits ou services fournis par une entreprise ou une organisation. Cela peut même être beaucoup plus efficace que les méthodes traditionnelles telles que les enquêtes.

En effet, si l’on hésite souvent à consacrer du temps à répondre à de longs questionnaires, une part croissante de consommateurs partage aujourd’hui fréquemment son avis sur les réseaux sociaux. Ainsi, la recherche de textes négatifs et l'identification des principales plaintes permettent d' améliorer les produits , d'adapter la publicité et de réduire le niveau d'insatisfaction des clients.

Commercialisation

Les spécialistes du marketing utilisent également la PNL pour trouver des personnes susceptibles d'effectuer un achat.

Ils s'appuient sur le comportement des internautes sur les sites Internet , les réseaux sociaux et les requêtes des moteurs de recherche. Ce type d’analyse permet à Google de générer un profit important en proposant la bonne publicité aux bonnes personnes. Chaque fois qu'un visiteur clique sur une annonce, l'annonceur paie jusqu'à 50 dollars !

Plus généralement, les méthodes PNL peuvent être utilisées pour construire une image riche et complète du marché existant, des clients, des problèmes, de la concurrence et du potentiel de croissance de nouveaux produits et services d'une entreprise.

Les sources de données brutes pour cette analyse incluent les journaux de ventes, les enquêtes et les médias sociaux…

Revu de littérature

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.

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.

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.

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.

Traitement des Séquences et Réseaux de Neurones Récurents

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.

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.

Conclusion

Cette revue de littérature a exploré l'évolution du traitement du langage naturel, passant des approches traditionnelles aux modèles de pointe basés sur l'apprentissage automatique. Les progrès continus dans ce domaine ouvrent la voie à des applications NLP plus sophistiquées et prometteuses, impactant divers secteurs tels que la recherche d'information, la traduction automatique et l'interaction homme-machine.

Questions de recherche

Discussions