INTRODUCTION

En partenariat avec Data4good, nous avons utilisé pour identifier les maladies d'origine alimentaire et les épidémies dans les restaurants.

L'algorithme d'apprentissage automatique suit les maladies d'origine alimentaire. Sans intelligence artificielle, ces avertissements potentiels pourraient ne pas être détectés.

Cela permettra aux consommateurs d'avoir accès à des informations importantes au moment où ils prennent une décision.

ETAT DE L’ART

Il a été développé un système pour la découverte des maladies d'origine alimentaire mentionnées dans les critiques de restaurants Yelp en ligne en utilisant la classification textuelle. Le système est utilisé par le Département de la santé et de l'hygiène mentale de la ville de New York pour surveiller le réseau social en ce qui concerne les plaintes de maladie d'origine alimentaire.

Il a été créé des classificateurs pour deux tâches : (1) déterminer si une note mentionnait une personne souffrant d'une maladie d'origine alimentaire et (2) déterminer si c’était plusieurs personnes qui avaient souffert d'une maladie d'origine alimentaire. Il a d'abord été développé un prototype de classificateur en 2012 pour les deux tâches en utilisant un petit ensemble de données labellisés. Au fil des années de déploiement du système, les épidémiologistes du DOHMH ont identifié 13 526 commentaires sélectionnés par ce classificateur. Les chercheurs ont utilisé ces données biaisées et un échantillon de commentaires complémentaires dans un modèle optimisé afin de développer des classificateurs nettement améliorés. Enfin, il a été effectué une analyse d'erreur des meilleurs classificateurs résultants.

Il a été constaté que la régression logistique entrainée sur les données augmentées de biais était la meilleure pour les deux tâches de classification, avec des scores F1 de 87% et 66% respectivement pour les tâches 1 et 2.

L’analyse des erreurs a révélé que l’incapacité des modèles à tenir compte de longues phrases était ce qui entraînait le plus d’erreurs.

FAIT

Filtre de commentaires par mots clés :

- On crée une fonction pour tester la présence de mots clés dans un commentaire

- On récupère les commentaires concernés

Machine learning :

* Word2Vec
* TF-IDF

EN COURS :

Active learning

DIFFICULTES :

* Le code d'apprentissage automatique formé par Word2vec est trop lourd pour les référentiels github classiques, un système de fichiers volumineux est disponible à l'adresse <https://github.com/neohack22/pretrain-word2vec>.
* Quantité de lectures
* Travail d’équipe (communication, méthodes, coordination, revues de code, organisation…)

A FAIRE :

- comparer les résultats avec le stem

- comparer des résultats en modifiant le critère de sélection des restaurants (mots clés, note d'hygiène)

- améliorer la sélection des mots des commentaires notés pour améliorer la pondération des mots et donc la notation des commentaires

- choisir la meilleure méthode de sélection et récupérer l'adresse des restaurants concernés pour les représenter sur une carte