

Classification de commentaires

Présentation du projet

Pour une entreprise proposant à ses utilisateurs d'écrire des commentaires, il est intéressant de savoir analyser les traces que laissent les utilisateurs sur leur plateforme.

C'est d'autant plus important car vous pourriez être responsable autant que ceux qui l'ont écrit des propos injurieux et toxiques de vos utilisateurs.

L'application est assez simple, il vous suffit d'entraîner un modèle capable d'analyser automatiquement du texte et de le classer dans plusieurs catégories.
Pour chaque commentaire, à vous de déterminer la classe la plus plausible.

Consignes

- Créer un répertoire [Github](#) et sauvegarder votre travail régulièrement.
- Utiliser ce jeu de [données](#) pour l'entraînement.
- Commencer par entraîner un modèle très simple pour déterminer la difficulté du problème.
- Progresser de manière itérative vers un modèle de deep learning utilisant de l'embedding et des RNNs.
- Créer un pipeline qui prend en entrée une phrase brute, qui effectue tous les traitements et effectue une classification

Conseils

- Dans un premier temps, prenez seulement une petite partie des données pour réduire le temps de calcul.
- Réutilisez au maximum les exemples vus dans les TPs précédents.
- Inspirez-vous du [code](#) des autres personnes ayant déjà résolu ce problème sur Kaggle.
- N'hésitez pas à chercher de l'aide sur internet.

Rendus

Un notebook :

- disponible sur votre Github,
- un sommaire en markdown doit être mis en place au début du notebook.
- chaque étape doit être justifiée en utilisant le markdown,
- Le code doit être clair et commenté

Un rapport sera rédigé reprenant :

- le contexte, le sujet du projet,
- les données et l'approche utilisées,
- les détails de votre solution,
- les améliorations possibles de cette version.

Le rapport ne devra pas faire plus de 5 pages.

Le rapport et le notebook seront à rendre par email à l'adresse suivante (gautherotmorgan0@gmail.com) au plus tard le 15 février à minuit.