

Série de TP 2 : Prétraitement du texte

ENNONCE :

Vous êtes chargé de traiter un jeu de données textuelles provenant de différents commentaires sur des produits en ligne. Ces commentaires sont bruts, avec des erreurs de typographie, des fautes d'orthographe, des mots inutiles et un usage incohérent des majuscules et des ponctuations. Votre objectif est de nettoyer, normaliser et préparer ces textes pour les utiliser dans un modèle d'analyse de sentiment.

Vous devez :

1. **Prétraitement des données textuelles** : Vous devez effectuer plusieurs étapes de préparation pour rendre les textes utilisables pour une analyse automatique.
2. **Normalisation du texte** : Vous devez standardiser la forme des textes, ce qui inclut la mise en minuscule, la correction des fautes d'orthographe, et la suppression des symboles inutiles.
3. **Nettoyage des données textuelles** : Vous devrez éliminer les éléments qui peuvent nuire à l'analyse du texte, tels que les mots vides, la ponctuation, les chiffres et les caractères spéciaux.

Exemple

Commentaire Original	Commentaire Prétraité et Nettoyé
"J'ai adoré ce produit ! C'était vraiment super 123 !!!"	"adoré produit vraiment super"
"Le service était OK mais le produit était défectueux."	"service ok produit défectueux"
"Bien, mais un peu cher !!!"	"bien cher"