# Rapport Tp1

## Les algorithmes les plus adéquats

Nous avons supprimé les caractères et les mots répétitifs ainsi que les homoglyphes afin de faciliter le traitement du texte et d'améliorer les performances. De même, nous avons transformé le texte en forme canonique, mais on fait ca uniquement pour les balises HTML, afin de les supprimer. En revanche, les emails et les liens ont été conservés, car ils sont essentiels pour préserver le sens du texte.

Pour normaliser le texte, nous l'avons converti en minuscules et traduit dans une seule langue.

Concernant la segmentation, j'ai constaté que la segmentation en sous-mots est plus performante, car elle est compatible avec les modèles modernes en NLP et gère mieux les mots rares que les autres méthodes.

Enfin, pour la réduction des formes, la méthode de lemmatisation s'est avérée la plus adéquate.