Clément Hue

TP Multimédia

Classification de texte

Les textes d'apprentissage ne sont pas rangés dans un ordre particulier

La vectorisation de texte ne nettoie pas les résultats, par conséquent, on retrouve des chiffres, et plusieurs fois les mêmes mots (conjugué, pluriel).

Les résultats prédit grâce à la démanche de bayes naïf est plutôt efficace, il y a un peu près 80% de bonne prédiction
Les thèmes proches admettent plus de confusions, car les mots utilisés sont potentiellement similaire. On peut constater qu'il y a quasiment pas de confusion entre le thème graphique et médecine, à l'inverse entre le hokey et le baseball, il y a 44 mauvaises déductions . Un autre possibilité d'erreur est le manque de données pour obtenir une bonne estimation.

La classification d'un texte, ce fait essentiellement par mot clé, si on entre juste la phrase « goal keeper », la probabilité que ce soit du hokey est de 80 %, si on rajoute d'autre mots dans la phrase la détection est moins évidente Le thème motorcyle et autos, sont également très proche, si une phrase contient le mot 'vehicle', il parvient difficilement à identifier lequel des 2 thèmes est le bon.

Traitement de textes

Pour découper un texte en phrase, l'algorithme doit faire attention aux espace après les points et à la présence de majuscule après chaque point. On remarque que l'absence de majuscule est pris en compte, par contre, l'oublie d'un espace après un point ne l'est pas. On remarque que pour la détection de mot, la ponctuation est bien prise en compte.

Pour identifier une collocation, on peut regarder le nombre de fois où deux mots sont juxtaposés, si ceux-ci le sont souvent, alors ils ont un sens sémantique. Ceci à une importance pour la traduction notamment en anglais, à cause des phrasals verbs.

Le stemming a son importance pour le calcul d'occurrence car un mot peut avoir plusieurs forme (conjugué, pluriel…), sans ça, le calcul du nombre d'occurrence serait biaisé