PJE : Analyse de comportement avec Twitter

Classification automatique de Tweets par KNN

Objectif : concevoir un algorithme de type KNN pour classifier des tweets comme positif, négatif ou neutre en fonction de la base d'apprentissage.

Prérequis :un fichier CSV étiqueté et nettoyé contenant une liste de tweets français sur les sujets suivant Vlille, Equipe de France, Election présidentielle

KNN

1. Implémenter une fonction de distance entre deux tweets qui prend deux tweets en paramètre.

   On rappelle la fonction de distance considérée :

   D(t1,t2) = (Nombre total de mots – nombre de mots communs) / nombre total de mots

   Vous pouvez utiliser d'autres distance et vous comparerez la qualité de la classification obtenue (voir https://en.wikibooks.org/wiki/Algorithm_Implementation/Strings/Levenshtein_distance, https://github.com/Simmetrics/simmetrics)

2. Réaliser un algorithme de type KNN qui prend en paramètre un tweet à étiqueter, une base d'apprentissage et le nombre de voisin à considérer. La sortie de l'algorithme sera le tweet étiqueté.

```
Données : x le tweet à étiqueter, k le nombre de voisins
pour i allant de 1 à k
    mettre le point i dans proches_voisins
fin pour
pour i allant de k+1 à N
    si la distance entre i et x est inférieure à la distance d'un des
points de proches_voisins à x
        supprimer de proches_voisins le point le plus éloigné de x
        mettre dans proches_voisins le point i
    fin si
fin pour
proches_voisins contient les k plus proches voisins de x
vote(proche_voisins) donnent la classe majoritaire des voisin
```

Evaluation de la qualité de classification de KNN

Dans cette partie nous allons tester la qualité de classification de KNN. Vous partionnerez votre base d'apprentissage en 2/3 et 1/3. Vous vous servirez des 2/3 pour trouver les plus proches voisins des tweets du tiers restant. Le 1/3 étant déjà étiqueté dans la base d'apprentissage, vous pourrez comparer vos résultats entre ce que KNN trouve comme polarité et ce qu'il y avait dans la base d'apprentissage.

Faire varier le nombre de voisin et donnez les matrices de confusion pour les différents tests réalisés. Voici un modèle de tableau :

	Classe Estimée (polarité)		
	Positif	Négatif	Neutre

Classe réelle (polarité)	Positif			
	Négatif			
	Neutre			