TP n°1 : Prétraitement des données textuelles

Rappel :

L'objectif de ce TP est de comprendre les différentes étapes de prétraitement du texte en s'entrainant avec différents types de datasets.

Le pré-traitement du texte comprend les étapes suivantes :

1. Récupération du corpus
2. Tokenisation
3. Suppression des stopwords (mots vides)
4. Stemming
5. Lemmatisation
6. Convertir le texte en minuscules ou en majuscules.
7. Normalisation

Le prétraitement du texte est une première étape importante et il faut vraiment observer le contenu de votre corpus après transformation pour être sûr que les données correspondent à ce que vous désirez, en vue des traitements ultérieurs.

Exercice 1

1. Le dataset que vous allez utiliser : https://cs.nyu.edu/~kcho/DMQA/
   Les données brutes représentent un corpus d'articles CNN récupérés par des chercheurs pour leurs expérimentations (télécharger le dossier cnn_stories).
2. Vous devrez effectuer les opérations de prétraitement suivantes sur le texte :
   - Créer des paires de document (article, highlights)
   - Suppression de la ponctuation
   - Séparation en token en minuscules
   - Suppression des stopwords pour les articles

Le texte final est tokenisé sans ponctuation et en minuscules

Exercice 2

1. Le dataset que vous allez utiliser : est SPAM_Data.csv
2. Vous devrez effectuer les opérations de prétraitement suivantes sur le texte :
   - Supprimer les ponctuations
   - Transformer le texte en tokens
   - Supprimer les tokens de longueur inférieure ou égale à 3
   - Supprimer les stopwords
   - Appliquer les méthodes du stemming et lemmatization

Le résultat final doit être comme suit :

Out[19]:

	Category	Message	removed_punc	tokens	larger_tokens	clean_tokens	stem_words	lemma_words	clean_text
0	ham	Go until jurong point, crazy.. Available only ...	Go until jurong point crazy Available only in ...	[go, until, jurong, point, crazy, available, o...	[until, jurong, point, crazy, available, only,...	[jurong, point, crazy, available, bugis, great...	[jurong, point, crazi, avail, bugi, great, wor...	[jurong, point, crazy, available, bugis, great...	jurong point crazy available bugis great world...
1	ham	Ok lar... Joking wif u oni...	Ok lar Joking wif u oni	[ok, lar, joking, wif, u, oni]	[joking]	[joking]	[joke]	[joking]	joking
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	Free entry in 2 a wkly comp to win FA Cup fina...	[free, entry, in, 2, a, wkly, comp, to, win, f...	[free, entry, wkly, comp, final, tkts, 21st, 2...	[free, entry, wkly, comp, final, tkts, 21st, 2...	[free, entri, wkli, comp, final, tkt, 21st, 20...	[free, entry, wkly, comp, final, tkts, 21st, 2...	free entry wkly comp final tkts 21st 2005 text...
3	ham	U dun say so early hor... U c already then say...	U dun say so early hor U c already then say	[u, dun, say, so, early, hor, u, c, already, t...	[early, already, then]	[early, already]	[earli, alreadi]	[early, already]	early already
4	ham	Nah I don't think he goes to usf, he lives aro...	Nah I dont think he goes to usf he lives aroun...	[nah, i, dont, think, he, goes, to, usf, he, l...	[dont, think, goes, lives, around, here, though]	[dont, think, goes, lives, around, though]	[dont, think, goe, live, around, though]	[dont, think, go, life, around, though]	dont think go life around though

Exercice 3

Appliquer les opérations de prétraitement nécessaires et convenable aux différents datasets présentés ci-dessous :

- **DrugProt corpus** : est un corpus utilisé dans le domaine biomédical, qui permet l'évaluation de systèmes capables de détecter automatiquement les relations entre composés chimiques/médicament et gènes/protéines. Vous allez utiliser le fichier : drugprot_training_abstracs.tsv
- Corpus des Tweets dédié au support client sur Twitter. Utilisez le fichier twcs.csv