TD : Nettoyage des Données Textuelles en Python pour le NLP

Objectifs du TD :

— Comprendre l'importance du nettoyage des données textuelles pour le NLP.
— Manipuler des bibliothèques Python pour normaliser et nettoyer des textes.
— Préparer les données collectées pour des applications de traitement automatique du langage.

Durée : 1h30

Prérequis :

— Données textuelles collectées via une API ou du web scraping (par exemple, à partir du premier TD).
— Python 3.7+
— Bibliothèques nécessaires : NLTK, Pandas.

Plan du TD :

1. Introduction au nettoyage des données
2. Nettoyage des textes : suppression des caractères inutiles
3. Tokenisation et normalisation
4. Sauvegarde des données nettoyées

1 Introduction au nettoyage des données

Les données brutes collectées contiennent souvent du bruit (caractères spéciaux, liens, mentions inutiles) qui doit être nettoyé avant leur utilisation dans des modèles NLP.

Questions

1. Pourquoi est-il important de nettoyer les données textuelles pour le NLP ?
2. Quels sont les principaux types de bruit dans les textes collectés ?

2 Nettoyage des textes : suppression des caractères inutiles

Étapes à suivre :

— Identifier et supprimer les caractères spéciaux et hyperliens.
— Convertir les textes en minuscules.
— Supprimer les espaces inutiles et normaliser le format.

Analyse des résultats :

— Comparez les textes bruts aux textes nettoyés.
— Identifiez les limites potentielles du nettoyage effectué.

3 Tokenisation et normalisation

Étapes à suivre :

— Divisez les textes en unités lexicales (tokens) à l'aide d'une bibliothèque comme NLTK.
— Effectuez une normalisation des mots (racine ou radical).
— Comparez les résultats obtenus entre le stemming et la lemmatisation.

Questions :

— Quelle est la différence entre le stemming et la lemmatisation ?
— Quels sont les avantages et les inconvénients de chaque méthode ?

4 Sauvegarde des données nettoyées

Étapes à suivre :

— Organisez les textes bruts, nettoyés et tokenisés dans une structure de données (ex. pd.DataFrame).
— Sauvegardez les résultats dans un fichier CSV.

Ressources complémentaires

— Documentation re (expressions régulières)
— Guide NLTK
— Documentation Pandas