TD : Collecte des Données en Python pour le Traitement du Langage Naturel (NLP)

Objectifs du TD

— Comprendre les méthodes de collecte de données pour le NLP.
— Manipuler des outils pour extraire des textes à partir de sources variées.
— Explorer les sources de données en ligne et apprendre à extraire des informations.
Durée : 1h30

Prérequis

— Python 3.7+
— Bibliothèques nécessaires : Requests, BeautifulSoup, Pandas.

Plan du TD

1. Introduction et mise en place
2. Extraction de données à partir d'une API
3. Web scraping de textes

1 Pour commencer

L'objectif de ce TD est de collecter des données textuelles à partir de différentes sources et de les préparer pour des applications NLP.

Questions

— Quelles sont les principales sources de données textuelles pour le NLP ?
— Pourquoi la collecte de données est-elle essentielle pour le NLP ?

2 Extraction de données à partir d'une API

Pour ce TD, nous utiliserons une API comme **NewsAPI**, une API gratuite pour récupérer des articles d'actualité.

Étapes à suivre :

— Créer un compte API et récupérer une clé API.
— Effectuer une requête pour collecter des articles sur un thème donné.
— Explorer les résultats obtenus, en identifiant les titres et les descriptions des articles.
— Sauvegarder les résultats dans un fichier CSV pour une utilisation ultérieure.

3 Web scraping de textes

Dans cette partie, nous utiliserons un outil comme **BeautifulSoup** pour extraire des données textuelles à partir d'un site web.

Étapes à suivre :

— Choisir une source web adaptée (par exemple, un blog ou un site d'actualités).
— Identifier les balises HTML contenant les informations à extraire (par exemple, les titres des articles).
— Extraire les données textuelles et les afficher.
— Enregistrer les données extraites dans un fichier CSV.

Pour aller plus loin

— Récapitulez les étapes principales de la collecte des données.
— Collectez des données sur un thème de votre choix (à partir d'une API ou d'un site web).
— Sauvegardez les résultats dans un fichier CSV.

Ressources complémentaires

— Documentation Requests
— Documentation BeautifulSoup
— API NewsAPI