2eme année Second Cycle Option : IASD
Module: Natural Language Processing

Fiche TP N° 04: Natural Language Processing with Sequence Models

Objectif :

➔ Le but de ce TP est de concevoir et de mettre en œuvre un modèle de reconnaissance d'entités nommées basé sur les réseaux de neurones récurrents (RNN, LSTM, GRU, etc..) à partir d'un jeu de données d'entraînement.

Données : Le jeu de données CoNLL-2003 sera utilisé pour ce TP. Il se compose de phrases annotées pour les entités nommées de type personne, organisation, lieu et divers.

Tâches à réaliser :

1. Prétraitement des données : vous devrez prétraiter les données en effectuant des opérations telles que la tokenization, la normalisation de la casse, etc.
2. Vectorisation des données: vous devez effectuer une vectorisation des données textuelles afin de les transformer en vecteurs numériques.
3. Entraînement du modèle : vous devrez entraîner un modèle de reconnaissance d'entités nommées à l'aide des caractéristiques extraites et du jeu de données d'entraînement CoNLL-2003.
4. Évaluation du modèle : vous devrez évaluer les performances de votre modèle en termes de précision, de rappel et de F1-score sur un jeu de données de test.
5. Amélioration du modèle : vous pourrez améliorer les performances de votre modèle en utilisant différentes techniques, telles que la sélection de caractéristiques, la régularisation, etc.

Livrables : Vous devrez rendre un rapport détaillé décrivant votre approche, vos résultats et vos analyses. Vous devrez également fournir votre code source et vos modèles entraînés.