

Plateforme d'analyse des documents en arabe

Bochra NAHALI

Résumé :

Ce mémoire présente la conception et le développement d'une plateforme avancée d'analyse des documents académiques en arabe, combinant les techniques de traitement du langage naturel (NLP) et de stylométrie. La plateforme propose une solution automatisée pour la classification des documents académiques et l'analyse des variations stylistiques. Elle repose sur trois modèles principaux : le premier utilise des algorithmes d'apprentissage supervisé et des modèles pré-entraînés, tels qu'AraBERT, pour classer les documents en différentes catégories académiques en exploitant les caractéristiques lexicales et sémantiques des textes ; le deuxième modèle se spécialise dans la détection des changements de style d'écriture au fil du temps, en identifiant les évolutions stylistiques ou les indices de ghostwriting ; enfin, le troisième modèle applique des méthodes de clustering pour regrouper des segments textuels selon leurs caractéristiques stylistiques, détectant ainsi les variations dues à l'intervention de plusieurs auteurs ou à la coexistence de styles différents dans un même document. La plateforme est construite sur une architecture 3-tiers robuste, comprenant un front-end dynamique en React, un back-end performant avec FastAPI, et une base de données MongoDB, permettant ainsi une gestion fluide et évolutive des données. Cette solution complète et performante représente un outil précieux pour l'analyse, l'organisation et l'évaluation des documents académiques en arabe.