Rapport d'Avancement

Projet : Implémentation et Analyse de Différents Types de Tokenizers pour les Modèles de Langage

Réalisé par : Elkadiri Anas
Encadrant : LAMGHARI Nidal

Date : 29 Avril 2025

Table des matières

Introduction

La tokenisation constitue une étape clé dans le traitement des données textuelles des modèles de langage (LLMs). Ce projet vise à implémenter et comparer différentes approches de tokenisation : caractère par caractère, Byte-Pair Encoding (BPE) et basée sur les expressions régulières (RegexTokenizer). Ce rapport présente l'état d'avancement des travaux jusqu'à la fin du mois d'avril 2025.

Chapitre 1

Objectifs du Projet

— Implémenter trois types de tokenizers : Caractère, BPE et Regex.
— Concevoir une architecture modulaire en Python.
— Comparer les performances des tokenizers sur des critères définis (taille des séquences, compression, vitesse...).
— Déployer une interface Streamlit pour démonstration.

Chapitre 2

Travaux Réalisés

— Étude bibliographique sur les méthodes de tokenisation utilisées dans les LLM.
— Conception de l'architecture logicielle et diagramme de classes.
— Implémentation des trois tokenizers avec une interface commune (`BaseTokenizer`).
— Mise en place des workflows d'entraînement (BPE) et d'utilisation pour chaque tokenizer.
— Premiers tests fonctionnels sur des jeux de données textuels.

Chapitre 3

Difficultés Rencontrées

— Ajustement des hyperparamètres pour l'entraînement du tokenizer BPE.
— Définition et optimisation du motif Regex pour le `RegexTokenizer`.
— Gestion de l'équilibrage entre performance (vitesse) et qualité de compression.

Chapitre 4

Travaux Restants

— Analyse expérimentale approfondie (temps d'encodage/décodage, taille des séquences).
— Déploiement final de l'application via Docker et intégration complète dans Streamlit.
— Rédaction du rapport final détaillé avec visualisations.

Chapitre 5

Perspectives

L'étape suivante consistera à intégrer ces tokenizers dans un pipeline complet de modèle de langage (LLM) incluant l'embedding, l'architecture Transformer et l'entraînement global du modèle.