SYNTHESE

Réalisé par : Aymane MAHRI

Encadré par : Pr.Lotfi EL AACHAK

During our work on this lab 3, we immersed ourselves in the world of natural language processing (NLP) with a focus on language modeling and regression/classification tasks. We aimed to gain proficiency in NLP techniques using the Sklearn library and to understand how different encoding methods and machine learning models can be applied to textual data.

We began by establishing a comprehensive preprocessing pipeline that included tokenization, stemming, lemmatization, and stop words removal. This process helped in cleaning and preparing the text data for subsequent analysis. We transformed the textual data into numerical vectors using various encoding methods such as Word2Vec (both CBOW and Skip Gram models), Bag of Words (BoW), and TF-IDF.

Next, we implemented regression models to predict numerical outcomes based on the text data. We trained several models, including Support Vector Regression (SVR), Naive Bayes, Linear Regression, and Decision Tree Regression. These models were evaluated using metrics like Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). The Linear Regression model with Skip Gram Word2Vec embeddings achieved the best performance, highlighting the importance of context-aware embeddings in regression tasks.

We then shifted our focus to classification tasks using a different dataset. We applied the same preprocessing techniques and encoded the data using Word2Vec, BoW, and TF-IDF methods. We trained various classification models, including SVM, Naive Bayes, Logistic Regression, and AdaBoost. These models were evaluated using metrics such as Accuracy, F1 Score, and Log Loss. The SVM model with Skip Gram Word2Vec embeddings outperformed others, demonstrating its effectiveness in capturing contextual information for classification.

Throughout this lab, we gained a comprehensive understanding of NLP techniques and their application in machine learning tasks. We explored various word embedding strategies and learned how to preprocess, encode, and model textual data effectively. This experience allowed us to hone our skills in using popular tools such as Spacy, NLTK, gensim, and Sklearn.

Overall, this lab provided us with valuable insights into the intricacies of NLP and machine learning, equipping us with the knowledge to tackle complex textual data analysis tasks in the future. We successfully applied different models and techniques to real-world datasets, enhancing our proficiency in NLP and gaining practical experience in handling regression and classification problems.