SYNTHESE

Réalisé par : Aymane MAHRI

Encadré par : Pr.Lotfi EL AACHAK

During our work on this lab, we immersed ourselves in the world of natural language processing (NLP) with a focus on Rule-based approaches, Regex, and Word Embedding techniques. We began by creating a Python program that uses Regex to extract structured information from a given user-provided text string. This involved identifying product names, quantities, and prices from the input string to generate a detailed bill. This exercise improved our proficiency in manipulating text and leveraging pattern matching techniques.

Once we successfully generated the bill using Regex, we shifted our attention to NLP Word Embedding methods. This involved transforming textual data into numerical representations for subsequent machine learning applications. We implemented a variety of encoding techniques including one-hot encoding, bag of words, and TF-IDF to create feature vectors from the data collected during the lab.

Further, we explored advanced word embedding approaches such as Word2Vec, both Skip Gram and CBOW models, on our dataset to understand different methods of embedding words based on context. Additionally, we applied GloVe and FastText approaches to the dataset, which gave us deeper insights into the nuanced relationships between words and their contexts.

To visualize the encoded and vectorized data, we employed the t-SNE algorithm, which allowed us to effectively compare and evaluate the various approaches. This visualization helped us observe the performance and efficiency of different techniques in transforming text data into meaningful numerical vectors.

Overall, this lab offered us a comprehensive understanding of how NLP techniques, including Rule-based methods and various word embedding strategies, can be utilized effectively to process, analyze, and gain insights from textual data. We honed our skills in applying these techniques and gained valuable experience in handling NLP tasks using popular tools such as Spacy, NLTK, matplotlib, gensim, mongopy, regex, pandas, numpy, and Sklearn.