Abdelmalek Essaadi University
Faculty of Science and Technology Tangier
Computer Engineering Department
Master: AIDS
Realized by**:** SABBAHI Mohamed Amine
Framed by: Pr . ELAACHAk LOTFI

Synthesis of lab 2 (NLP)

1. Introduction:

In this lab, we explored various Natural Language Processing (NLP) techniques and word embedding models to analyze text data. Our objective was to gain insights into different methods of text representation and understand how they can be applied in real-world scenarios. Here’s a synthesis of my key learnings:

2. Rule-Based NLP and Regex:

We began by implementing a rule-based NLP approach using regular expressions to extract structured information from unstructured text data. This approach involved defining patterns to capture specific entities and their corresponding attributes, such as product names, quantities, and prices, from a given text. By applying predefined rules and patterns, we were able to generate a bill summarizing the purchased items and their total costs.

3. Word Embedding Techniques:

We then delved into word embedding techniques, which aim to represent words or phrases as dense vectors in a continuous vector space. We explored three fundamental approaches:

- **One-Hot Encoding:** This technique converts words into binary vectors, with each word represented by a vector where only one element is 1 (indicating the presence of the word) and the rest are 0s.
- **Bag of Words (BoW):** BoW represents text data by counting the frequency of each word in a document without considering the order of words.
- **TF-IDF (Term Frequency-Inverse Document Frequency):** TF-IDF assigns weights to words based on their frequency in a document relative to their frequency across all documents in a corpus, aiming to highlight the importance of rare words.

4. Word2Vec (Skip Gram and CBOW), GloVe, and FastText:

Next, we explored advanced word embedding models:

- **Word2Vec:** Word2Vec learns word embeddings by predicting the context (surrounding words) of each word in a given text. We experimented with both Skip Gram and Continuous Bag of Words (CBOW) architectures.
- **GloVe (Global Vectors for Word Representation):** GloVe is another word embedding model that combines global co-occurrence statistics to learn word representations.
- **FastText:** FastText extends the Word2Vec model by considering word substructures (character n-grams), enabling it to capture morphological similarities between words.

5. Conclusion:

In conclusion, this lab provided a comprehensive overview of various NLP techniques and word embedding models, ranging from basic rule-based approaches to advanced neural network-based methods. We learned how to extract structured information from unstructured text data, represent words as dense vectors, and capture semantic relationships between words in a high-dimensional space. These techniques are essential for numerous NLP applications, including sentiment analysis, document classification, machine translation, and information retrieval, among others.