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 1 (NLP)

In this lab, I embarked on an insightful journey through the complexities of natural language processing (NLP), focusing on the Arabic language. The lab guided me through several foundational NLP tasks, including text scraping, data storage, preprocessing techniques, and advanced NLP applications such as part-of-speech (POS) tagging, named entity recognition (NER), and the nuances of working with language-specific features. Here’s a synthesis of my key learnings:

1. Web Scraping and Data Storage:

The lab commenced with the task of efficiently scraping web content, for which I utilized Python's **requests** and **BeautifulSoup** libraries. This exercise underscored the importance of ethical web scraping practices, notably adhering to the guidelines specified in robots.txt. My scraping efforts were concentrated on extracting Arabic text from **AlJazeera** articles that focused on the Russia-Ukraine conflict. This content was then meticulously stored in a **MongoDB** database, offering me hands-on experience with managing unstructured web data in a structured repository.

2. Text Preprocessing:

A significant portion of the lab was dedicated to familiarizing myself with vital text preprocessing steps, especially tailored for the Arabic language. Through the application of **text cleaning, tokenization, stop word removal, normalization, stemming, and lemmatization**, I prepared the raw scraped text for deeper NLP analysis. This process illuminated the critical need to comprehend Arabic's morphological and syntactical complexities to ensure the effective preparation of text data for subsequent processing.

3. POS Tagging and NER:

In exploring **POS tagging** and **NER**, I engaged machine learning methodologies, employing the **Stanza** library for practical application. This phase of the lab was particularly enlightening, revealing the challenges posed by Arabic's richly inflected nature. It emphasized the significance of context and precise word form in achieving accurate linguistic annotations.

4. Language-Specific Tools and Libraries:

A key takeaway from the lab was the importance of choosing suitable tools and libraries for language-specific NLP tasks. While **NLTK** offers a broad range of NLP capabilities, the lab highlighted the superior performance of language-specific libraries like **Stanza** for Arabic text processing, particularly for **lemmatization** and **stemming**.

Overall, this lab was a comprehensive exploration of Arabic NLP, offering valuable hands-on experience and insights into the challenges and considerations specific to working with Arabic text. It has equipped me with a solid foundation in NLP that I can build upon in my future projects and studies.