SYNTHESE

Réalisé par : Aymane MAHRI

Encadré par : Pr.Lotfi EL AACHAK

During my work on this lab, I dived into the world of web scraping but with a focus on Arabic sources. Using tools like Beautiful Soup, I learned how to extract data from Arabic websites related to a specific field of interest. This involved navigating through web pages, collecting relevant information, and storing it in a database like MongoDB using tools like pymongo for further analysis.

Once the data was collected, I shifted my focus to natural language processing (NLP) tasks. I cleaned the text, broke it down into words, and removed unnecessary ones such as articles and conjunctions. Then, I explored techniques like stemming and lemmatization to simplify words and prepare them for analysis, for this i used the ISRIStemmer specifically adapted for arabic language and other NLTK libraries.I can conclude that Lemmatization was way more accurate and with a meaning compared to Stemming but took more time.

Additionally, I delved into parts-of-speech tagging using a rule-based approach, as for the machine learning approach, it wasnt possible since spacy isnt really well adapted for Arabic language yet. This allowed me to understand the grammatical structure of Arabic sentences and how words function within them. For this i used tools like Farasa that is mode developped towards arabic language compared to spacy.

Finally, I explored named entity recognition (NER) methods to identify and classify specific entities within the text, such as names of people, places, or organizations. This step was crucial for extracting structured information from unstructured text data.I used Farasa here aswell

Overall, this lab provided me with valuable hands-on experience in web scraping Arabic sources and applying NLP techniques to process and analyze the collected data effectively.