NATURAL LANGUAGE PROCESSING
LAB 1

Master's in Artificial Intelligence & Data Science

Done by:
MOHAMED AMINE FAKHRE-EDDINE,
Supervised by:
Lotfi EL AACHAK

08/04/2024

1 What have we learned?

1.1 Web Scraping and data storing

Using Beautiful Soup (BS4), we've managed to scrape a number of articles about different topics from the "Al Jazzera" news website to power the demonstrations in our lab.

With that we've learned that web scraping is one of the most if not the most important method of gathering data, especially for NLP. Since most textual data can be hard to find in a usable format outside the web, extracting content from it would be the best solution we have.

Even if it didn't have a real use case in our lab, saving data is also important in data related tasks, that's why mongodb, with the flexibility and scalability for handling large volumes of unstructured data it provides, is one of the best tools for storing data in our context.

1.2 NLP Pipeline

Cleaning data is ONE of many important steps in any data related task, including NLP.

Arabic is also special since it has more complex elements than other languages, like "Harakat", "Tachkeel", Special characters, complex grammar and vocabulary which makes it beautiful but hard to work with. That's why an extensive form of cleaning is required for Arabic, that includes **Removing Punctuation**, **Discretization**, **Tokenization**, **Normalization** and **Removing Stopwords**.

While this may sacrifice some context, it streamlines the process.

1.3 Stemming and Lemmatization

Stemming and Lemmatization present both advantages and drawbacks, contingent upon the desired outcomes. In this experiment, I've implemented a light Stemmer and explored various established processing libraries for Arabic, assessing their precision and efficiency.

Stemming being reducing to their base or root form by removing prefixes or suffixes, will transform a word into it's simplest form which is a verb in most words, that changes the type which is not a good idea since we are planning on POS tagging and recognizing Named Entities.

On the other hands, Lemmatization converts words to their base or dictionary form, known as the lemma. Unlike stemming, which chops off prefixes and suffixes without considering the context, Lemmatization uses a vocabulary and morphological analysis to ensure valid words and take context into account. This results in more accurate root words, which can be useful for our use case.

1.4 POS tagging and NER

Part-of-Speech (POS) tagging involves identifying the grammatical category of each word in a text, and various methods can be employed to achieve this, including rule-based and machine learning approaches.

In the context of Arabic, a rule-based approach poses significant challenges and is inherently incomplete due to several factors. Arabic's complexity, as mentioned earlier, is

compounded by the presence of "Awzan" – patterns that dictate the structure and sound of words, also determining their grammatical type. However, Arabic encompasses a vast array of "Awzan", with patterns that can sound different yet share the same structure (different "Tachkeel"), leading to ambiguity. Consequently, discretizing an input text inevitably oversimplifies its linguistic depth.

To illustrate this, I developed a basic rule-based POS tagger, showcasing its simplicity. However, fully integrating all the necessary rules would be a time-consuming endeavor.

On the other hand, the machine learning approach offers a solution to these challenges, abut with its own set of constraints. It necessitates a substantial amount of training data to achieve satisfactory results, effectively addressing the complexities inherent in Arabic POS tagging.

References

[1] T. Zerrouki, "pyarabic, an arabic language library for python." [Online]. Available: https://pypi.python.org/pypi/pyarabic,year={2010}

[2] ——, "Tashaphyne, arabic light stemmer," 2012. [Online]. Available: https://pypi.python.org/pypi/Tashaphyne/

[3] ——, "qalsadi, arabic mophological analyzer library for python." 2012. [Online]. Available: https://pypi.python.org/pypi/qalsadi

[4] ——, "Towards an open platform for arabic language processing," 2020.

[5] A. L. T. G. at Qatar Computing Research Institute (QCRI), "farasa, the state-of-the-art full-stack package to deal with arabic language processing." 2020. [Online]. Available: https://farasa.qcri.org/

[6] M. H. Btoush, A. Alarabeyyat, and I. Olab, "Rule based approach for arabic part of speech tagging and name entity recognition," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 6, 2016. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2016.070642