Higher school of computer and digital sciences and technologies
Module: Natural Language Processing (NLP)

2024/2025
3rd year specialty class

TP n° 2

Exercise 01: *Tokenization*

Objective: Understand the process of tokenization using spaCy and analyze token properties.

Before starting the exercise, make sure you have spaCy installed and the English language model downloaded.

Using This sentence: **Google is planning to purchase an U.S. software company for $120 million**.

Q1) What are the steps to tokenize a given text using spaCy and how can you access various properties of each token?

Q2) List and explain at least five different properties of tokens that can be accessed using spaCy.

Q3) How does spaCy handle special cases in tokenization, such as punctuation, numbers, and abbreviations?

Q4) How does spaCy's tokenization differ from simple string splitting? Provide an example to illustrate the difference.

Q5) Do the tokenization this time with word_tokenize from NLTK, what are the differences?

Exercise 02: *Sentence Segmentation*

Objective: Understand the process of sentence segmentation using various NLP libraries and analyze different approaches.

Before starting the exercise, make sure you have spaCy, NLTK, and TextBlob installed.

Using this text:

"Mr. Smith bought cheapsite.com for 1.5 million dollars, i.e. he paid a lot for it. Did he mind? Adam Jones Jr. thinks he didn't. In any case, this isn't true... Well, with a probability of .9 it isn't."

Q1) What are the steps to perform sentence segmentation using spaCy, NLTK, and TextBlob? How do you access the segmented sentences in each case?

Q2) Compare the results of sentence segmentation from spaCy, NLTK, and TextBlob. Are there any differences in how they handle abbreviations, ellipsis, or other special cases?

Q3) How do these libraries handle sentence boundaries in the presence of quotation marks, parentheses, or other punctuation marks?

Q4) What challenges arise in sentence segmentation when dealing with informal text, such as social media posts or chat messages? How might you address these challenges?

Q5) Implement a simple rule-based sentence segmentation function. How does its performance compare to the results from spaCy, NLTK, and TextBlob? What are the limitations of a rule-based approach?

Exercise 03: *Part-of-Speech*

Objective: Understand the process of POS tagging using spaCy and analyze the POS tags assigned to words.

Sentence to Use: **The NLP system accurately classified 95% of the customer feedback as positive.**

Q1) What are the steps to perform POS tagging using spaCy, and how can you access various POS tags for each token?

Q2) What are the different POS tags in spaCy, and what do they represent? List and explain at least five POS tags from the sentence

Q3) How does spaCy handle multi-word expressions and abbreviations in POS tagging, such as "NLP" or "95%"?

Q4) Perform POS tagging using pos_tag from NLTK. What are the differences?

Exercise 04: *[Stemming, Lemmatization, Name Entity Recognition, Stop words]*

Q1) Do the same to explore Stemming, lemmatization, NER and Stop words

By: Mr. MAMMASSE Amine