NLP class 2025 Practical Exercises 1

- State of the art Natural Language Processing : Examples of applications
- Basic tools : NLTK, Spacy

Part 1

Question 1 :

text : This movie is beautiful. I would like to watch this movie again.

Predict the sentiment of the given text as positive or negative or neutral, using:

a. two different HuggingFace pipelines
b. one conversational model (e.g chatGPT)

Are the answers the same?

Question 2 :

You receive a collection of text messages, some of which are spam, while others are genuine (not spam). Your task is to analyze these messages and classify them into two categories: Spam or Not Spam.

1. *"Congratulations! You've won a free trip to the Bahamas. Click the link to claim your prize!"*
2. *"Hey, are we still on for dinner at 7?"*
3. *"URGENT: Your bank account has been compromised. Verify your details immediately."*
4. *"Limited-time offer! Buy one, get one free on all electronics. Don't miss out!"*
5. *"Can you send me the notes from today's meeting?"*
6. *"You have been selected for a cash prize! Call now to claim."*
7. *"Let's catch up this weekend. It's been a while!"*
8. *"Earn $500 daily from home! No experience needed. Sign up now!"*

a. Read the following messages carefully. Sort by hand the messages into two groups: Spam and Not Spam. Explain quickly your reasoning.
b. Sort again the sentences using LLMs. Use at least two different models.
c. Compare your sorting with that of the models. Which sentences do you agree on? Do the sentences you disagree with carry a particular pattern?

Part 2

Question 1 :

Given the sentence:

> Union Bank Inc. was originally incorporated as “Union Savings and Mortgage Bank”. Unlike other Austrian universal banks, however, it didn't develop a branch network. It wasn't until 1866 that the bank opened for banking business in London.

a. How do you perform a word level tokenization? Implement it using at least three different methods (hint : check nltk.tokenize package).

b. Compare the results.

c. What modifications would you suggest to improve the tokenization, possibly as a rule or post-processing step?

Question 2 :

a. What rules would you apply to tokenize sentences?

b. Apply them to the text of Question 1. What do you think of the results?

c. Repeat tokenization using nltk. Have the problems been solved?

Question 3 :

Consider the following text:

> text : Adrian Brody was excellent , as were many of the actors. Cinematography was superb, and was the music score! One of the best movies I have ever seen!!! A24 has done it again they deserve all of the Academy awards they have been nominated for!!!

a. Perform Part of Speech (POS) tagging using nltk.

b. What are the POS tags, which might be used for sentiment analysis?