

Neural Networks applied in NLP

Lab 3

N-gram Language Model:

1. (0.20p) Implement a byte-pair encoding language model for the sentences discussed in class.
Useful info: <https://huggingface.co/learn/nlp-course/chapter6/5?fw=pt>
2. (0.40p) Implement a generalized N-gram language model for the Romanian language using a smoothing technique at your choice, build on a corpus you collect from the web (at least 1000 words).
3. (0.20p) Compute the probability of a new Romanian sentence, given at input, using the n-gram model you developed.
4. (0,20p) Use a pre-trained neural language model to predict the next two words after a sequence of 4 words given as input.

Useful info: <https://medium.com/mlearning-ai/an-illustration-of-next-word-prediction-with-state-of-the-art-network-architectures-like-bert-gpt-c0af02921f17>