LABORATORY SESSION 3

Transformers - BERT

Objective:

- Check embedding related to context
- Classification using BERT
- Train a tokenizer / Word2Vec
- Train a transformer

Upload your results (code & accuracy) and a small explanation of your work to “PoliformaT / Espacio”

Your have to do the 3 parts in two lab sessions.

In the case of part 3, b) section is optional.

1. Distances of sentences

There is a Jupiter notebook here:

https://github.com/bhattbhavesh91/word2vec-vs-bert

- Copy the different lines to Colab and try to understand the purpose of everyone.
- Explain what the code does
- Try with other pair of sentences in the same sense

2. Spam Classification using BERT

Carry out this tutorial

https://github.com/codebasics/deep-learning-keras-tf-tutorial/blob/master/47_BERT_text_classification/BERT_email_classification-handle-imbalance.ipynb

After obtaining the results, try to improve it using a different architecture of neurons or changing the hyperparameters.

3. Train tokenizers and transformers

a) Try this example of tokenizer training

https://medium.com/@ankiit/word2vec-vs-bert-d04ab3ade4c9

b) Train a transformer (optional)

https://huggingface.co/blog/how-to-train