Requirements:

1. Overall knowledge of NLP and deep learning

   - It might be necessary to read more in the next steps

   - Going forward to more advanced concepts

2. Understanding of Large Language Models (LLMs), and have a good understanding of Transformers in NLP

   - Knowing the structure, input/output

   - How to use and train them in a real problem

   - Different types of them

   - …

3. A Python deep learning framework such as Pytorch to implement neural networks

   - We strongly intend to use this framework, but we may utilize other framework in case of requirement

   - Each framework has its own documentation, so you can learn a lot from them

4. Working with Google Colab platform

   - All codes could be implemented on it (as an Ipython notebook)

   - On Google Colab notebook you could create and connect to a session in order to run your code

   - After connecting to a session, a Linux machine is accessible in the backend, so it is possible to execute Linux commands in order to connect to Google Drive, download and upload file, etc. Also, it is possible to use Python programming language to do so.

   -

5. Working with Huggingface platform

   - A comprehensive hub for models, datasets and more in the AI field

   - You can easily load a LLM and use it

6. Search for similar implementations, codes and available material on Internet!

   - Websites: A very good example is Machine Learning Mastery which contains lots of tutorials. Furthermore, there are lots of other sources as you explore deeply!

   - YouTube: There are lots of videos! Just look for what you want. For example, consider the following videos:

     - https://www.youtube.com/watch?v=c36lUUr864M

- https://www.youtube.com/watch?v=SZorAJ4I-sA
- GitHub: My suggestion is to explore a lot on this platform. For example, these links seem suitable:

    - https://github.com/bentrevett/pytorch-sentiment-analysis

    - https://github.com/Shivanandroy/T5-Finetuning-PyTorch

- E-books

Tasks definition:

- We want to solve a typical supervised learning problem

- Given text at the input, we want to generate labels as output

- Data comes from a subset of SST-2 dataset, which is a sentiment classification dataset

- Dataset is divided into train and test sets. You can access and download them via the links

- You should access the files and analyze samples in this dataset, using Google Colab platform

- Check data distribution, labels, etc.

- Particularly, you have to do a prediction task for sentiment

- Probably do some data preparation methods on data for input and output! And other procedures you have learned from courses or other sources about NLP and deep learning

- Use a LLM as the core of your model in order to solve the task, for instance you can look at BERT, RoBERTa, BART, T5, Flan-T5, etc.

- Design of the model(s) is up to you. You can change model structure, tune hyperparameters and use any other method (based on your previous knowledge of deep learning and NLP) to improve the results

- Training and inference are essential parts of this assignment

- You should adopt one or more suitable evaluation metrics to measure the quality of predictions

- You should simply train your model on the train data, and then use the test data to measure the quality of predictions on unseen samples

- Analysis of predictions is necessary, such as using confusion matrix

- Clean code is always appreciated

- After you have done the tasks, you should be able to present your work

- For necessary questions, you can contact me at my Telegram or Skype account