NeuralMind is offering 2 positions for the summer internship 2022-2023 with the following subjects:

Program 1: Deployment of a Neural Search Engine
Program 2: Artificial Intelligence Assistant

These projects will be developed in a home-office environment with a mentor to supervise his/her work.

The minimum period of the internship is three months, expected to start December 1st, but can be flexible depending on the academic schedule.

The stipend for working 30 hours per week is R$ 2.200,00 per month.

The application can be done by sending an email to talentos@neuralmind.ai with the subject "Summer Internship 2023"

Please enclose your CV and Academic Records.

Thank you

Roberto Lotufo
CTO, NeuralMind

Summer Internship 2022-2023 at NeuralMind

Title: Deployment of a Neural Search Engine
Period: December 1st 2022 - February 28th 2023
Objective: Optimize the deployment of a deep learning-based search engine
Modality: home-office
Areas of Interest: Natural Language Processing, Information Retrieval, Deployment of Deep Neural Models and Optimization

Summary: Pretrained language models such as BERT are widely used by the top performing academic and industrial search engines (including Google Search and MS Bing). This project aims at deploying and optimizing a state-of-the-art search engine for various types of document search. The project also comprises an exploration of modern compression techniques such as distillation to reduce query latency, which will eventually allow us to run the system on CPUs rather than GPUs. To learn more about our system and the expected activities, check out our paper recently published: "NeuralSearchX: Serving a Multi-billion-parameter Reranker for Multilingual Metasearch at a Low Cost", presented at DesIRes 2022 Conference in last August:
https://desires.dei.unipd.it/2022/papers/paper-05.pdf

Schedule: 12 weeks (excluding Christmas and New Year)

List of Activities:

1. Learn about the inner workings of modern search engines based on deep learning.
2. Create document indexing scripts.
3. Deploy an open source search engine based on high-performance pretrained models.
4. Learn about modern quantization and model compression techniques.
5. Experiments with model optimization methods for low latency inference.
6. Evaluation and report writing

activ/week	1	2	3	4	5	6	7	8	9	10	11	12
1	X											
2		X	X	X								
3				X	X	X						
4						X	X	X				
5								X	X	X	X	
6											X	X

Resources:

Text Ranking with Transformers:

- Document Ranking with a Pretrained Sequence-to-Sequence Model: https://arxiv.org/pdf/2003.06713.pdf

Model Compression:

- Distillation: https://arxiv.org/pdf/1503.02531.pdf

Quantization:

- Q8BERT: Quantized 8Bit BERT

Serverless architectures:

- AWS: https://aws.amazon.com/serverless/
- Google Cloud https://cloud.google.com/serverless
- Serverless BERT: https://towardsdatascience.com/serverless-bert-with-huggingface-and-aws-lambda-625193c6cc04

Summer Internship 2022-2023 at NeuralMind

Title: Artificial Intelligence Assistant
Period: December 1st 2022 - February 28th 2023
Objective: Build an AI Assistant using Contextual Learning
Modality: home-office
Areas of Interest: Foundation Models, Natural Language Processing,

Summary: This project will develop an artificial intelligence (AI) system that analyzes and interprets documents and generates summary reports.

With the introduction of Foundation Models, such as OpenAI GPT-3[1], a new way to generate "Query-based Summarization" from one or more documents has become commercially viable and a number of new applications are now possible.

Neuralmind has new projects on AI Assistant using contextual learning with the use of single language models such as GPT-3. This internship is to help the development of these new projects. The proposed system is based on two technologies developed by NeuralMind, NeuralSearchX[2] and GPTimbau. NeuralSearchX is a DL-based search engine and GPTimbau is a Portuguese GPT-like model that has been pre-trained by NeuralMind.

Schedule: 12 weeks (excluding Christmas and New Year)

List of Activities:

1. Study of Foundation Models such as Generative Pre-trained Transformers (GPT)
2. Study of MultiDoc Question-Answer Systems
3. Learning NeuralMind's NeuralSearchX architecture
4. Deployment of API
5. Monitoring, testing and improvements
6. Final evaluation and report writing

activ/week	1	2	3	4	5	6	7	8	9	10	11	12
1 GPT	X	X	X									
2 MultiDocQA		X	X	X								
3 NeuralSearchX				X	X	X						
4 Deployment						X	X	X	X			
5 Monitoring								X	X	X	X	
6 Evaluation										X	X	X

References:

[1] Tom Brown, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901.

[2] T. Almeida et al. "NeuralSearchX: Serving a Multi-billion-parameter Reranker for Multilingual Metasearch at a Low Cost", DesIRes 2022 Conference, 2022:
https://desires.dei.unipd.it/2022/papers/paper-05.pdf