Assignment 3:
Large Language Models

- This assignment is due on **July 3rd, 2024 (23:59, CET)**
- You can discuss the problems with your classmates or browse the Internet to get help. However, copy and paste is cheating.
- You need at least 50% of the points of every assignment to take part in the exam.
- There are 3 assignments in total.
- Submit at `https://elearn.informatik.uni-kiel.de/course/view.php?id=274`
  - only pdf and jupyter files and only one zip file per assignment.
  - put your names and matriculation numbers on *each* page in the pdf file

Task 1: Name Entity Recognition (NER)

Implement an Named Entity Recognition (NER) system for GMB (Groningen Meaning Bank) corpus using the BERT model . You can download this data from this link: NER Data. Please use ner.csv file for this task.

a) Dataset Preparation **2 P**

- Load the dataset and explore its structure and contents.
- Display the first few data samples and provide basic statistics about the data. (like number of samples and labels., etc)
- Use the BERT tokenizer to tokenize the text data.
- Ensure the tokens align with the original words and the corresponding labels.

b) Features **3 P**

- Explore the data to identify additional features beyond text and target that might improve the model, or consider creating a new feature.

c) Data Splitting **1 P**

- Split the dataset into training and testing sets to evaluate the model later. Data split ratio can be 80 to 20 percent.

d) BERT (NER) Model Training **5 P**

- Choose and load the pre-trained BERT model and configure it for token classification.
- Train the model and display the training process and evaluate the model on the validation set periodically.

e) Model Evaluation **2 P**

- Save the trained BERT model to disk for future use.
- After training, evaluate the model's performance on the test set.
- Calculate metrics such Accuracy, and F1-score for each entity class.

f) Bonus (Optional) **5 P**

- Compare the BERT-based NER model with other baseline models (e.g., LSTM-CRF) and report the findings.
- Deploy the trained NER model as an API or a web application.

Task 2: NER Theory

a) What are the three common tagging schemes used for NER? **3 P**

b) Should stemming or lemmatization be applied in NER? If so, why? If not, why not? **2 P**

c) Is converting text to lowercase beneficial in NER? Please explain in your own words. **2 P**

Task 3: Student Project Showcase

In this task, you need to work on your own project. The goal is to inspire creativity, independent thinking, and problem-solving. So you can explore what interests you and use what you have learned in the course. **20 P**

Here are some **ideas** you might like: using a transformer-based model for multi-label text classification, creating summaries and question-answering systems, or training models on both characters and tokens for embedding.

The marks distribution will be similar to this template

- Data pre-processing **5 P**
- Exploratory data analysis (EDA) **5 P**
- Data Splitting **1 P**
- Model Training **5 P**
- Model evaluation **4 P**

Note: For this task, along with the code file (Jupyter notebook), please provide a PDF file. The PDF should include a project description explaining what the project is about, what you have done, the type of dataset used (including a link to the dataset), the model used, the evaluation metrics, and a brief conclusion.