Third assignment

Motivation: Software engineers ask and answer questions in Stack Overflow. Some of these discuss design decisions. However, software engineers do not explicitly tag posts about design decisions, which make it hard for software engineers to find this knowledge.

Goal: Develop a machine learning classifier that can identify and classify architectural posts.

Dataset: In a previous work, we manually classified architectural posts, and have a dataset of classified posts. Each post is either architectural or programming post. Also, architectural posts are classified based on their purpose and type of solution. The list of **architectural posts** and their manual categories are attached in an Excel sheet. Also, a list of the IDs of **programming posts** are attached in a text document.

Given the time constraints, each group will focus on one deep learning model.

Group	Model	Focus
Group 1	CNN	Solution
Group 2	RNN	Purpose
Group 6	BERT	Purpose
Group 7	CNN	Solution
Group 8	BERT	Purpose
Group 10	RNN	Solution

To achieve the goal of the assignment, I propose these guiding steps in the following weeks:

1st week: **Prepare dataset of posts**
In this week, you will perform the following:
1. *Download the data of all posts in the dataset*. Use the query interface of Stack Overflow to download post title, question and the answer with the highest vote.
2. *Perform NLP pre-processing*. Concatenate post title, question and the answer with the highest vote in one string. Then perform text filtering. This includes removing HTML tags and stop words, lemmatization and stemming. Please make sure to remove or replace source code in the post.
3. *Determine the number of terms per post*. This will be required to decide on the size of the input to the deep learning model.
4. *Write the source code that classify posts into training, validation and test set*. This code will be required to train deep learning models in the next week. You may re-use and take this code as an example. Make sure to follow a stratified random sampling in splitting the dataset. The code should divide the dataset for k-cross validation.

2nd week: **Develop deep learning models**
In this week, you will train and evaluate your specified deep learning model.
1. *Design the deep learning model*: This can be different from one model to another. Check the lecture slides on the design of your model and check this repository which can help you write the code for your model. For the assignment, it is enough to experiment with an initial “reasonable” model (e.g., like the one in the

repo). Bonus points will be given, if you performed optimization for your parameters.
2. *Train and evaluate deep learning model*: This is also different from one model to another. You can use or re-use the code in the repository.

You need to specify the following:

a. Hyper parameters such as number of layers and their size, batch size, epochs, training parameters such as early stopping, optimizers and loss functions.
b. Create word embedding matrix from the pre-trained Stack Overflow word embedding (for models other than BERT). See lecture slides.
c. Perform the training.
d. Run the trained model on the test set and calculate metrics.

Perform the training using 10 cross validations.

Note: You might need to run the training on the HPC and use GPUs. If you run your scripts on the HPC, then it is enough to use A40 on Noctua 1. Remember, if you requested lots of resources, your job will be queued for a long period. Please consider that you need to load the required libraries such as Tensorflow before running your script.

3rd week, **Predict architectural posts, and answer research questions**.
In this week, you will answer research questions. The first question is obligatory, and the second and third are optional to get the bonus for this assignment.

1. How accurate is your model to classify architectural posts?

Here, you will present the precision, recall and F1-score. Try to present the full numbers for the 10 iterations. Use charts or tables. We require the metrics to classify architectural posts, and to classify each category of the posts.
Also present a confusion matrix that shows how the model classifies different types of posts compared to the test set.

2. How many architectural posts exist in Stack Overflow? And what are the most common types of these posts?

Here, you will use the trained model to predict the posts in the full dataset of Stack Overflow posts and determine the number of architectural posts for each type. Provide statistics on these numbers.

3. What are the characteristics of the architectural posts in Stack Overflow?

Here, you will analyze the characteristics of the architectural posts, such as the number of answers, votes, and tags. Provide these characteristics per type and conduct significant tests.

To answer RQs 2 and 3, you will analyze the full XML dump of Stack Overflow. It has been copied on the scratch folder on Noctua 1. The file is called posts.xml.
As a results of RQs 2 and 3, please provide an excel table or csv table with the IDs of all posts in the XML, their predictions based on the classifier, and their characteristics.