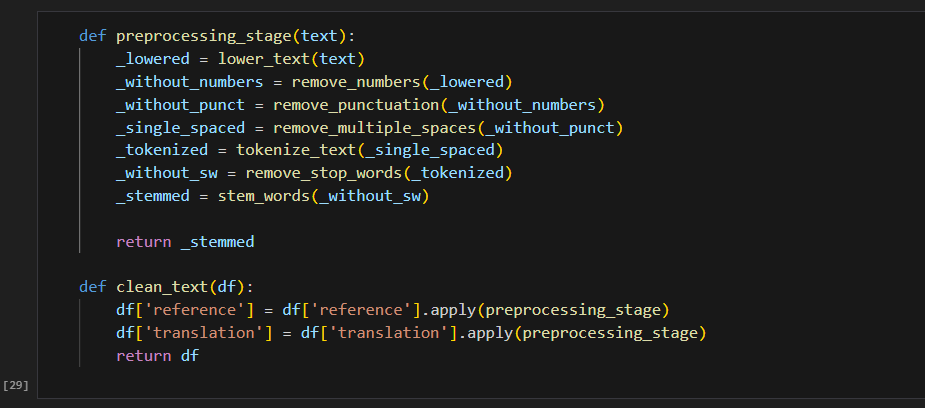
First things first, when I started the assignment I read thepaperprovided in the task description.

I have observed several things:

1. I can use the model that Dale et. al. have trained and uploaded onto their GitHub repository.
2. It is better to use the J-score which is described in the paper for the final solution because BLEU score does not evaluate the style and detoxification.

I have created notebook “1.0-data-observation” and tried to preprocess data here. What I did is I transferred functions from one of the labs, and ran them on the dataset. However, I did not take into consideration that I shouldn’t preprocess translation text because it is what the model will generate as output, that is why it is better to leave it “as it is”.



After the initial preprocessing, I have started to incorporate the ParaGeDi model, and I’ve run out of computational power because my laptop doesn’t have a dedicated GPU.

Later I’ve moved to the Kaggle environment, and have started working there, but I couldn’t make the solution work. Hence, I have decided to use pretrained models from library “Hugging Face” and fine-tune them, on the example of CondBERT in the paper.

But for that, I would have need Hugging Face datasets and tokenizers that is why I have scratched my preprocessing and started again.