This report contains information about problems with models from test task and what can be improved.

## Task 1. Natural Language Processing. Named entity recognition

Our model has next problems:

1. Dataset for this task was synthetical. We got mountain names from Wikipedia and then created sentences by making templates. Although there were many templates, which described the mountains from different sides, our model probably learned more about templates structure and not mountain names in sentences. It works great on this dataset, but if some strange sentence, that is very different from our templates, will occur model can make mistakes.
2. Other problems relate to computational power. DistilBERT-base-cased model was chosen as pretrained model for this task. But BERT-base-cased is heavier and computational harder, but better. So, if we had better GPU, we could use BERT-base-cased model and get better results.
3. To get better results, we could also generate more sentences for our dataset. We also must make much more templates, so our model will learn not templates, but finding mountain names. We could also make more sentences without mountains and take more locations (like lakes, some buildings), so our model can better differ mountains from other locations.
4. Also, we can slightly increase number of epochs. We don't want too many because of overfitting, but 5 epochs could give better results.
5. Lastly, we took 25000 mountain names from Wikipedia. We could take more mountain names because our model works better if it already saw mountain name in her training. This could highly increase recall of our model.

## Task 2. Computer vision. Sentinel-2 image matching

Our model has next problems:

1. First, dataset had about 5600 rows at the start, but after cleanup only 1483 remained. From 50 folders of images only 46 are connected to geojson file. So even though our dataset is 36GB, actual data is much lower than it may seems.
2. Some images contained parts, where it was fully black (all zeros in matrix). That was the main reason why from 2582 potential positive pairs only 2300 pairs were created. This is also the main reason why from 4800 potential positive pairs only 3613 pairs were created. So, we had much less pairs than we expected.

3. Sentinel-2 images are huge, they are about 10000x10000 pixels or (100x100 km), but our polygons were much lower. After cleanup all polygons could fit completely in 32x32 pixel patch. For balanced dataset it would be better if we could have much bigger polygons, so we could increase our patch size to 64x64.
4. Model trained quickly, but setting a greater number of epochs could lead to overfitting. What could really improve results of our model is more complex architecture of model. We could add more convolutional layers, or more fully connected layers. Experimenting with neural network model could improve results but take some time.
5. Lastly, we could try to use some pretrained model and fine-tune it like we did in first task. But with such small dataset it could only slightly improve results.