Report about Data Science Internship Test

Maksym Bratsiun

Task 1. Natural Language Processing. Named entity recognition

In this task, we need to train a named entity recognition (NER) model for the identification of mountain names inside the texts.

Link to solution

During this task, I learned a lot about of BERT training pipeline, tokenizer and NLP preprocessing and model inference. Own RNN or LSTM required a lot of text and time to train. That is why I have chosen BERT. I had some trouble with train dataset (duplication [CLS] mark when call outside training what broke score calculating) but I have fixed that. It was an interesting task.

Improvements:

- Increase the amount of training data.
- Change the model to multilingual.
- Add marks to other geographical terms: ridges, valleys, tracts.
- Uploading text as a file.
- Inference the model as an app like Streamlit.

Task 2. Computer vision. Sentinel-2 image matching

In this task, we have to work on the algorithm (or model) for matching satellite images. For the dataset creation, I download Sentinel-2 images from the official source here or use our dataset from Kaggle. My algorithm should work with images from different seasons.

Link to solution

During this task, I became familiar with the capabilities of the OpenCV library for figure matching. This was my first experience using this part of OpenCV's capabilities. I liked the brevity of the library's documentation, the large number of materials (medium.com, ai.stanford.edu), the ability to download data from the site and explore them. I was running out of time.

Improvements:

- Maybe MaxPool needed to be converted to a lower resolution or other scaling.
- Optimize the TCI channel when collapsing to one channel and refine the contrast.
- Exclude parts of the images where there are a lot of clouds.
- Calculate the geolocation of the points on images and compare for validation.

I do not want to make publication on GIT links to googledrive resources so it is here:
S2A_MSIL1C_20180611T083601_N0206_R064_T37UCR_20180611T104008.SAFE
S2A_MSIL1C_20180611T083601_N0206_R064_T36UYA_20180611T104008.SAFE