DL4NLP - AI text detection

We have one month in total for the whole project.
Workflow: We have human and AI written texts: first step is to find a dataset with both these texts (check on google scholar) - these datasets are usually obtained by asking AI to re-write human-written text; second step we need to define our methodology: we can for example implement a simple NN as a classification task; we can also use the model internal metrics (perplexity and loss) with no training (DetectGPT paper): we do perturbations on both AI and human generated texts and then we see how the loss and perplexity of the model compares between the perturbed and non-perturbed texts: for AI generated text loss and perplexity between the two will be high and for human generated text will be low; we can also train a detector with a contrastive learning loss; we can fine-tune an LLM for this task. We need to choose one of these methods to implement. In the first week we have to do the proposal: choose dataset, set goal and submit a mini-project proposal. Then we need to implement, get results and get a 4 page report.

Methodology

Distinguishing the text using the model's internal metrics will be our project because it allows us to to more exploratory work. Some possible research questions are:

- Can we use LLM's internal metrics to distinguish between AI and human generated texts according to their change before and after perturbations?
- Will different kinds of perturbations make the internal metric's differences sharper and the detection better?
- How does the AI text detection performance task change with AI text generated text from different models? Given a fixed model used for detection, is AI generated text from some models easier to detect than from others?
- Are some models better than others at detecting AI generated texts? Are smaller models better at this kind of agentic task? Given a fixed AI text that we want to detect, are some models better at detecting AI text than others?

The original paper repo is in https://github.com/eric-mitchell/detect-gpt.