19.03.2024

CENG 464 – Text Mining

Term Project

For the term project, you are required to showcase your ability to train datasets with multiple models. You should be able to perform training with both older models (RNN, LSTM, etc.) and newer models (Transformer, BERT, RoBERTa, ERNIE, GPT, T5, etc.).

Firstly, you should select an easily accessible dataset from the Internet for your work. You can select more than one dataset for your training as well. We suggest using multiple datasets on similar topics so that you can showcase learning on these datasets better. Keep in mind that these datasets should be for Text Mining and/or NLP tasks. We will not accept datasets for tasks like image recognition. Also, remember that your selected dataset(s) may need to be preprocessed before they can be trained on your models.

Then, you should select an older model like RNN, LSTM, their variations, etc. You need to prepare and train your dataset(s) on this model and acquire the metrics for your success. Generally, it is a good idea to select your metric according to the needs of your selected dataset(s). For example, perplexity is a good metric for language models (specifically, task of mask filling).

Afterwards, you should perform the previous step with a newer model like BERT, RoBERTa, ERNIE, GPT, T5, etc. You should use Transformers library and directly finetune on the pretrained models from this library. Make sure the model of your choice can be downloaded from the Transformers library (of HuggingFace). You should use the same metric from before to show the success of your dataset(s).

After you complete all these experiments, you can write a report on all the steps you performed for these tasks and discuss the success of your models using your metrics. We expect a report close to scientific papers. For example, you should at least have the following part in your report: Introduction, Methodology, Experimental Results & Discussion and Conclusion.

During the week of the finals, you are expected to present your results using a PowerPoint presentation or something similar. We will reveal more details about this in a future date.

<u>Helpful Extra Resources:</u>

- You can acquire some example datasets from the following link (from HuggingFace). Other sources are also acceptable.
  https://huggingface.co/datasets
- Some example codes for RNN training can be found on these links. Other types of older models are also acceptable.
  https://www.kaggle.com/code/kcsener/8-recurrent-neural-network-rnn-tutorial

https://www.datacamp.com/tutorial/tutorial-for-recurrent-neural-network

- You can check this simple example tutorial for a training with Transformers (newer models) library.
  https://huggingface.co/docs/transformers/training

Notes:

- We will be holding a lab session on these topics sometime in May. You can also start working on finding your dataset(s) and models beforehand as well.
- **You will be working on this project in Teams of two students. Please determine your groups before 5th of April (at 23:59). You can send your group information to Research Assistant Güliz Akkaya through Teams chat. Those of you who do not have a group will be matched randomly.**
- You should submit your report before 23:59, 2nd of June. We will be opening a Teams assignment page for submissions sometime before this date.