COMP 7970

Special Topics: Natural Language Processing

Instructor: Shubhra ("Santu") Karmaker

Assignment #3: Implementation of Text Summarization [100 points]

⚠ **Notice:** This assignment is due **Monday, October 1, 2021 at 11:59pm**.

Please submit your solutions via Canvas (https://auburn.instructure.com/). You should submit your assignment as a **typeset PDF**. Please do not include scanned or photographed equations as they are difficult for us to grade.

1. Implementation of Text Summarizer [80 pts]

Implement two different text summarizer model using 1) Google PEGASUS (https://github.com/google-research/pegasus) and 2) Facebook BART (distillbart)
Below is a repository that you can refer for the implementation details of distillbart:

- https://github.com/huggingface/transformers/tree/master/examples/research_projects/seq2seq-distillation
- You do not need to implement the model from scratch. You can use the pretrained model and fine tune on your dataset.

2. Evaluation of Text Summarizer [20 pts]

Report Rouge-1, Rouge-2, Rouge-3 and Rouge-L score based on the summarization. Please refer to below link for more details on Rouge metric:

- https://en.wikipedia.org/wiki/ROUGE_(metric)

Dataset

The CNN / DailyMail Dataset is an English-language dataset containing just over 300k unique news articles as written by journalists at CNN and the Daily Mail. We have sampled a small portion of the big dataset for convenience. Your dataset is updated on the canvas in 3 parts as below:

- cnn_dailymail_train_assignmemt3.csv – contains 2500 articles
- cnn_dailymail_val_assignmemt3.csv – contains 500 articles
- cnn_dailymail_test_assignmemt3.csv – contains 500 articles

Train the model on 2500 articles, validate on 500 articles and test on 500 articles.