Sentiment Analysis. Comparison of models.

Julia Domańska

Background and context

The idea of the project is to check if the following neural networks: LSTM, GRU and RNN can distinguish emotion-related text written in American vs British English. The "Humans of New York" Facebook page contains posts about people living the big American city, especially their thoughts, struggles and emotions. Similarly, "Humans of London" includes narratives of people living in London. The post are usually full of emotive expressions, both negative and positive, related to different personal issues. The data should exhibit some differences in emotion-related vocabulary in American and British English.

Objectives

- ✓ Web scraping of "Humans of New York" and "Humans of London" posts, tokenization and conversion into tensors for PyTorch models.
- ✓ Implementation, training and testing of PyTorch's nn.LSTM, nn.GRU, and nn.RNN architectures on the previously collected and prepared data.
- ✓ Comparison of texts in American English vs British English in different models, with the focus on the attention layer.
- ✓ Evaluation whether the models identify if the text was written by an inhabitant of New York or London.

Scope and deliverables

DATASET: Web scraping of the Facebook page "Humans of New York" and "Humans of London".

MODELS: LSTM, GRU, and Vanilla RNN with the corresponding PyTorch layers, and attention layer.

Milestones

1. DATA COLLECTION AND PROCESSING: Extract posts from the “Humans of New York” and “Humans of London” Facebook page (web scraping). Tokenize the collected data using a pre-trained tokenizer (AutoTokenizer), and convert the tokenized text into tensors for training PyTorch models.

2. MODEL IMPLEMENTATION AND TRAINING: Implement three neural network architectures (LSTM, GRU, and Vanilla RNN) using PyTorch layer, and an attention layer. Train for 5 epochs.

3. MODEL EVALUATION AND COMPARISON: Test each trained model on a sample of sentences and whole narratives from the web scraped data and compare the architectures, especially the differences in the attention layer.