WeHelp

Assignment - Week 5

Task Submission:

Submit your final crawler, cleaner, tokenizer, training code for embedding model and classification model, sample documents, and answer the following questions based on what you have actually done.

A. Total Number of Source Titles: 1108440 Total Number of Tokenized Titles: 850059

B. If A and B are different, what have you done for that?

I removed articles that had no valid tokens.
I filtered out reply posts (titles starting with "re:", "fw:", "r: [").

C. Parameters of Doc2Vec Embedding Model.

   a. Total Number of Training Documents: 850059
   b. Output Vector Size: 300 Min Count: 1 Epochs: 100 Workers: 8
   c. First Self Similarity: 78.5% Second Self Similarity: 82.6%

D. Parameters of Multi-Class Classification Model.

   a. Arrangement of Linear Layers: 300x128x64x9
   b. Activation Function for Hidden Layers: ReLU
   c. Activation Function for Output Layers: Softmax
   d. Loss Function: Categorical Cross Entropy
   e. Algorithms for Back-Propagation: Adam
   f. Total Number of Training Documents: 680047
   g. Total Number of Testing Documents: 169974
   h. Epochs: 30 Learning Rate: 0.001
   i. Accuracy on Testing Documents: 79.90%
   j. Any other parameters you think are important. Dropout Rate: 0.2 / Batch Size: 128 / Random Seed: 42

E. Share your experience of optimization, including at least 2 change/result pairs.

1. Change: Added custom dictionary with domain-specific terms (77 stock-related words) to improve tokenization accuracy.
   Result: Accuracy decreased from 79.98% to 79.90%.
2. Change: Increased training epochs from 30 to 50.
   Result: Accuracy improved from 79.90% to 80.00%.
3. Change: Added learning rate scheduler that reduces learning rate by 50% every 10 epochs (StepLR with step_size=10, gamma=0.5).
   Result: Accuracy decreased from 80.00% to 79.86%.