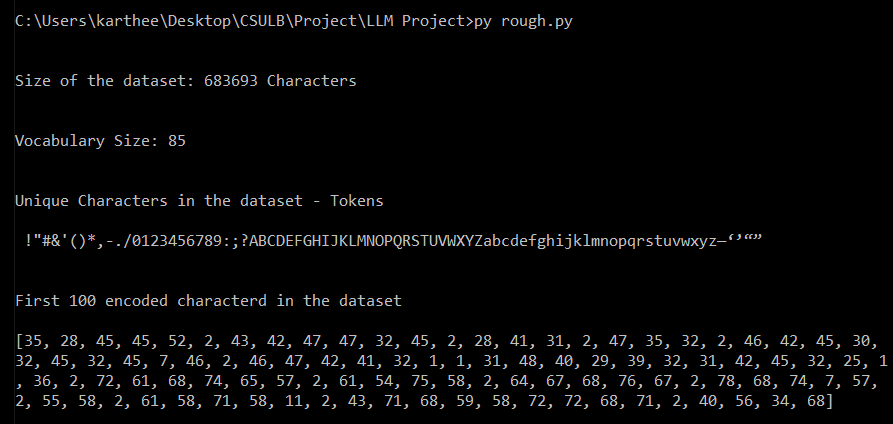
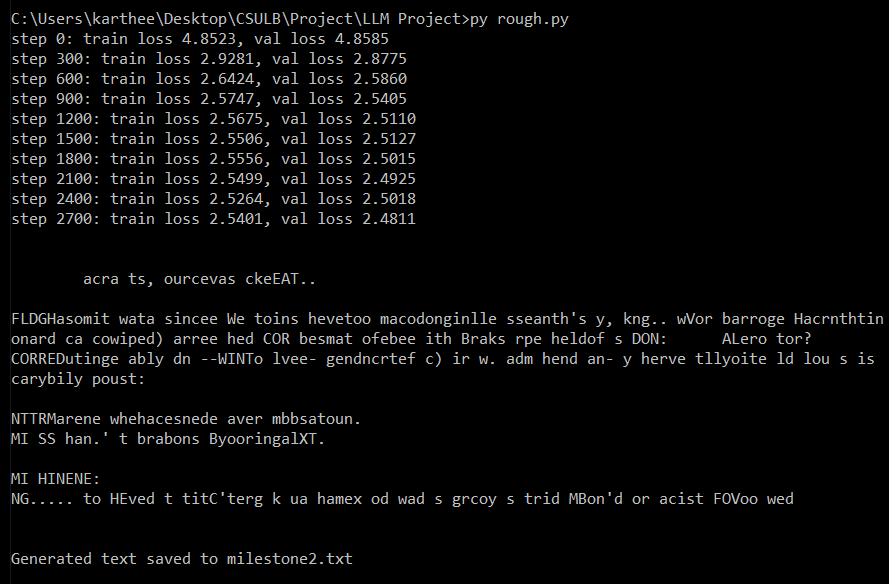
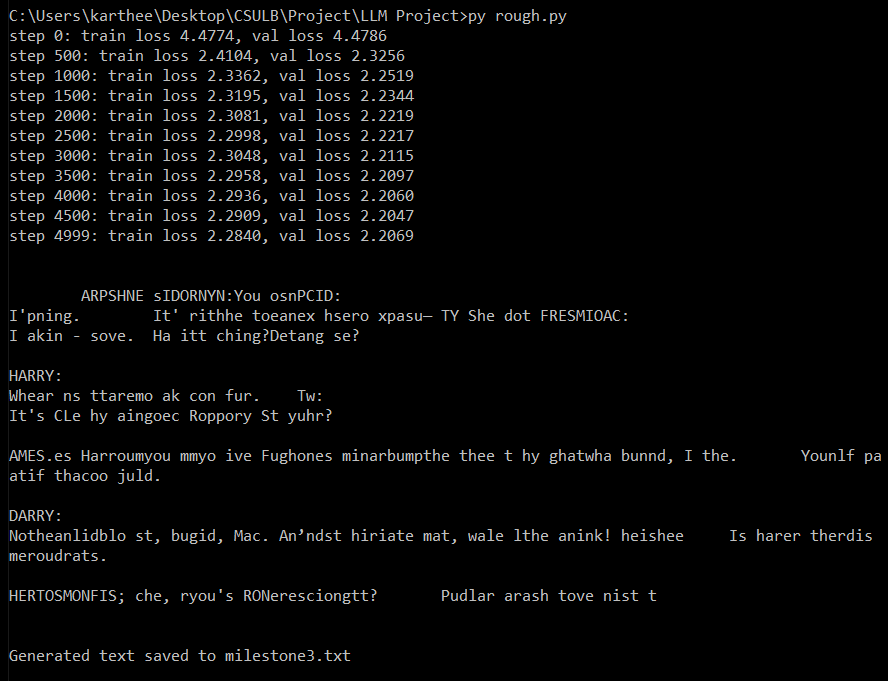
# MILESTONE 1: Dataset Exploration and Preparation



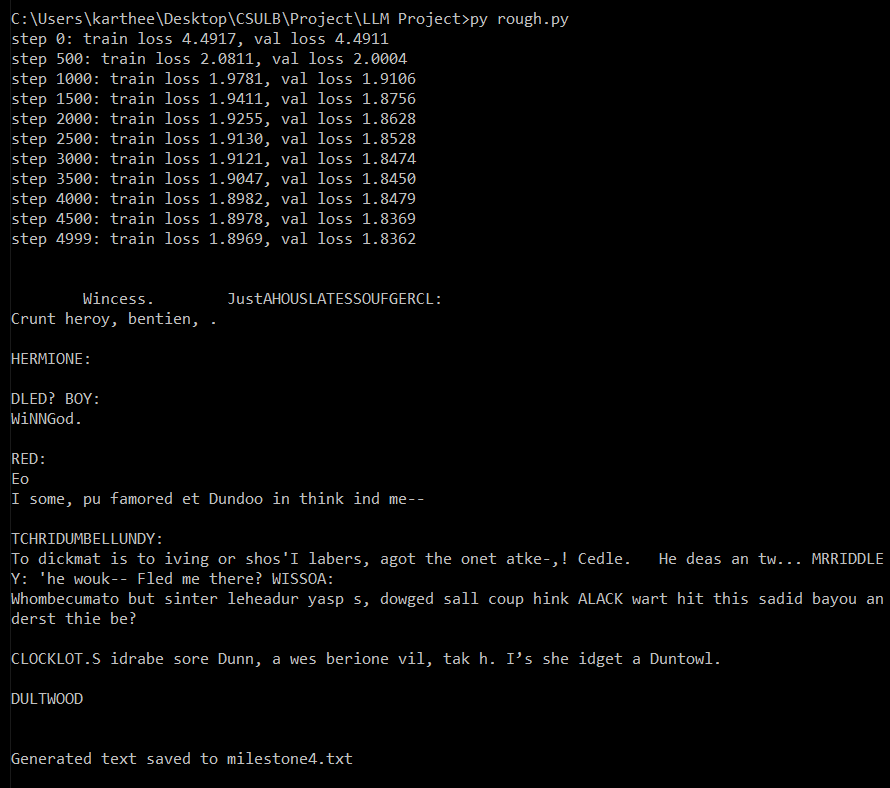
# MILESTONE 2: Bigram Language Model



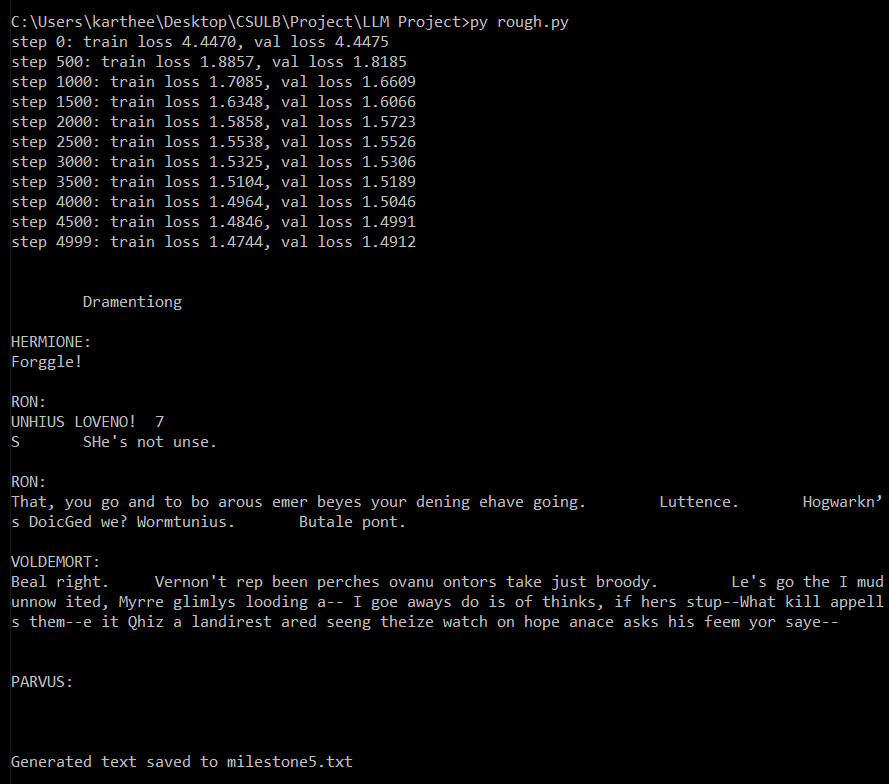
# MILESTONE 3: Self-attention & Softmax Iteration



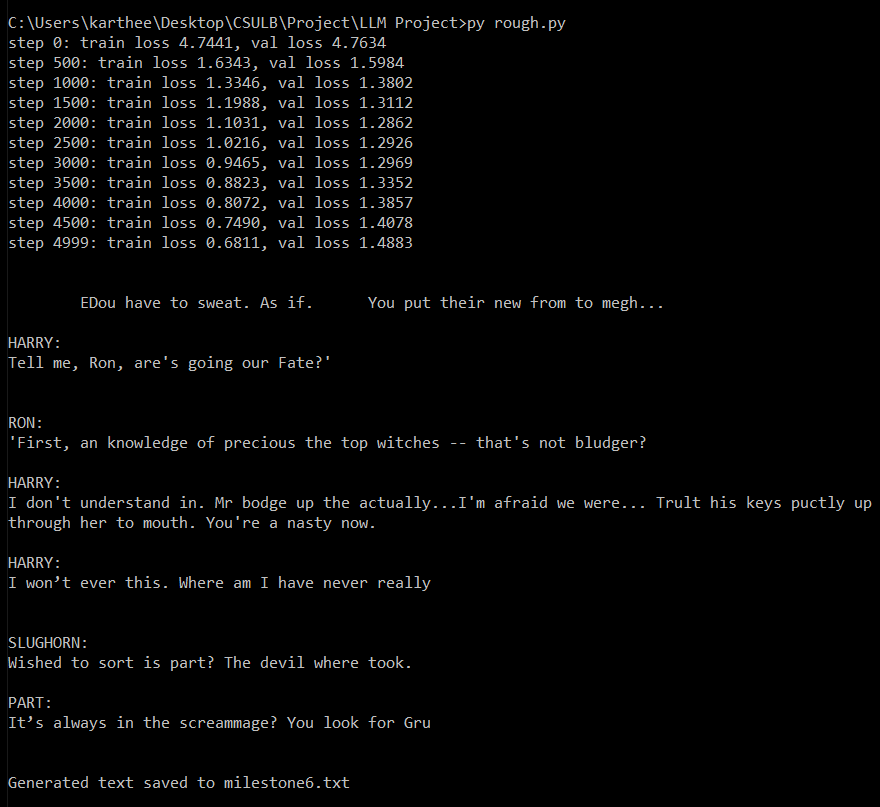
# MILESTONE 4: Multi-head Attention



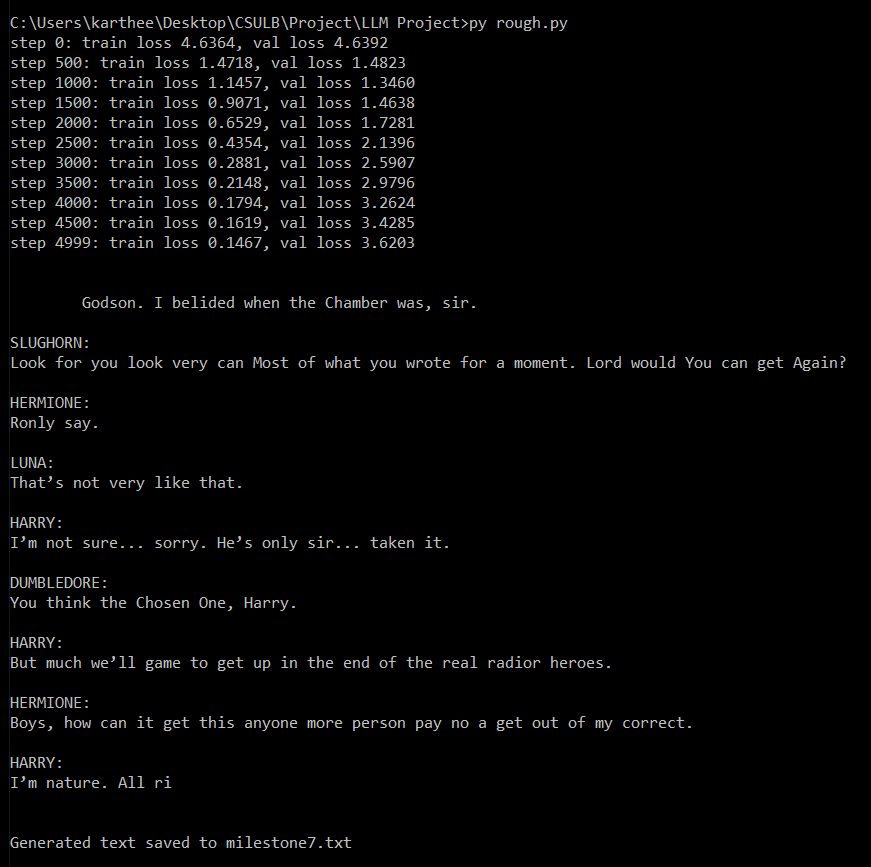
# MILESTONE 5: Feed Forward Layers



# MILESTONE 6: Residual Connections

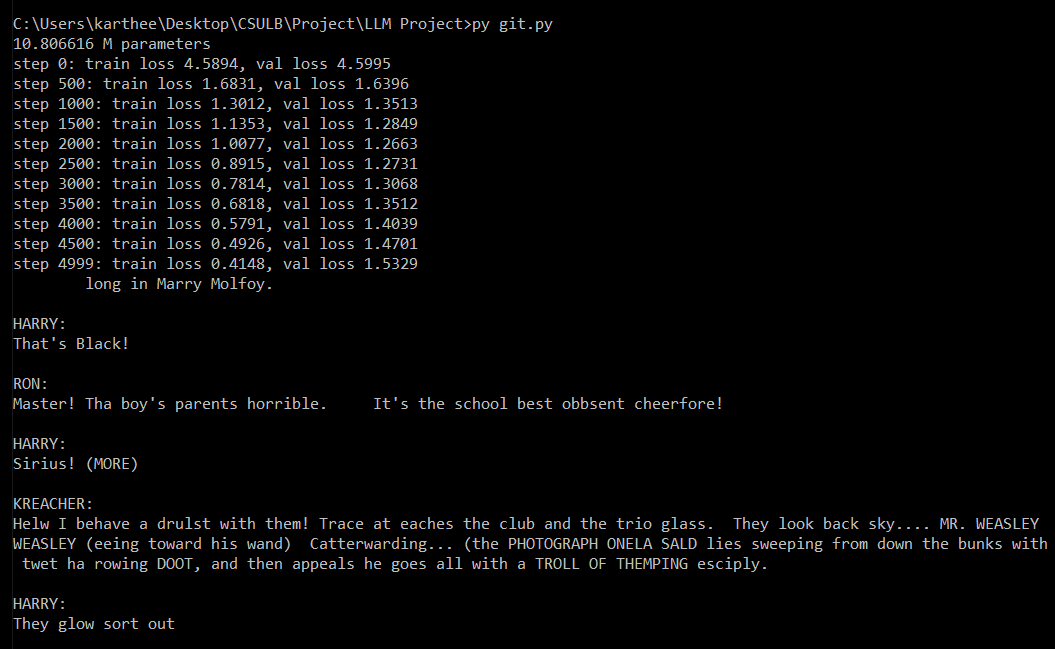


# MILESTONE 7: Layer Normalization



# MILESTONE 8: Dropout

ITERATION 1



# IMPROVEMENTS

ITERATION 1

learning Rate = 1e-4

droupout = 0.3

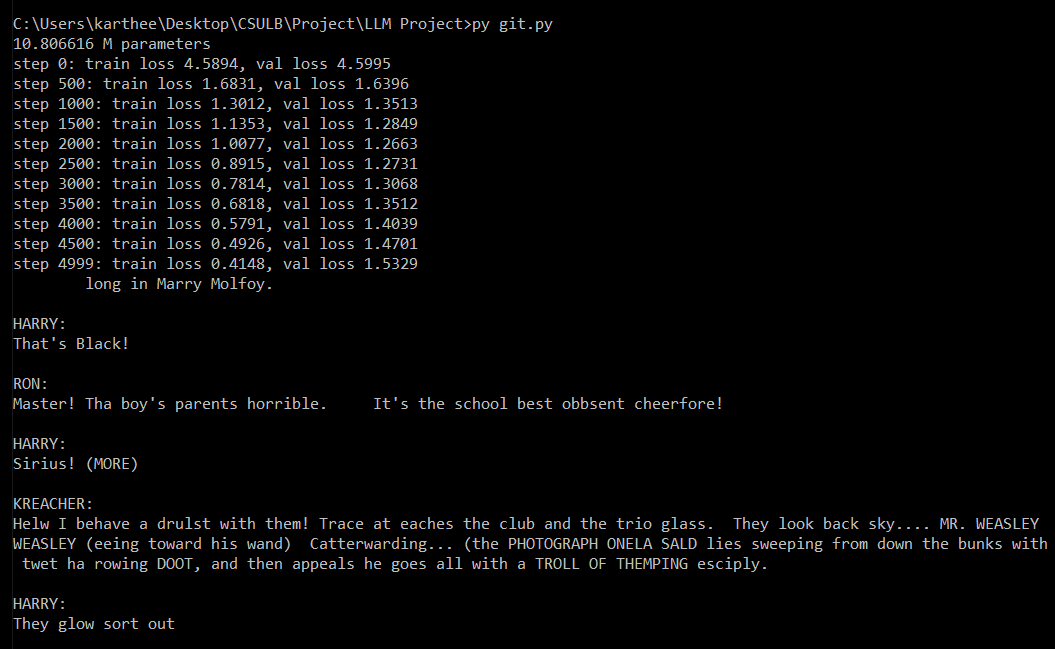
optimizer = torch.optim.AdamW(model.parameters(), lr=learning\_rate, weight\_decay=1e-2)

ITERATION 2

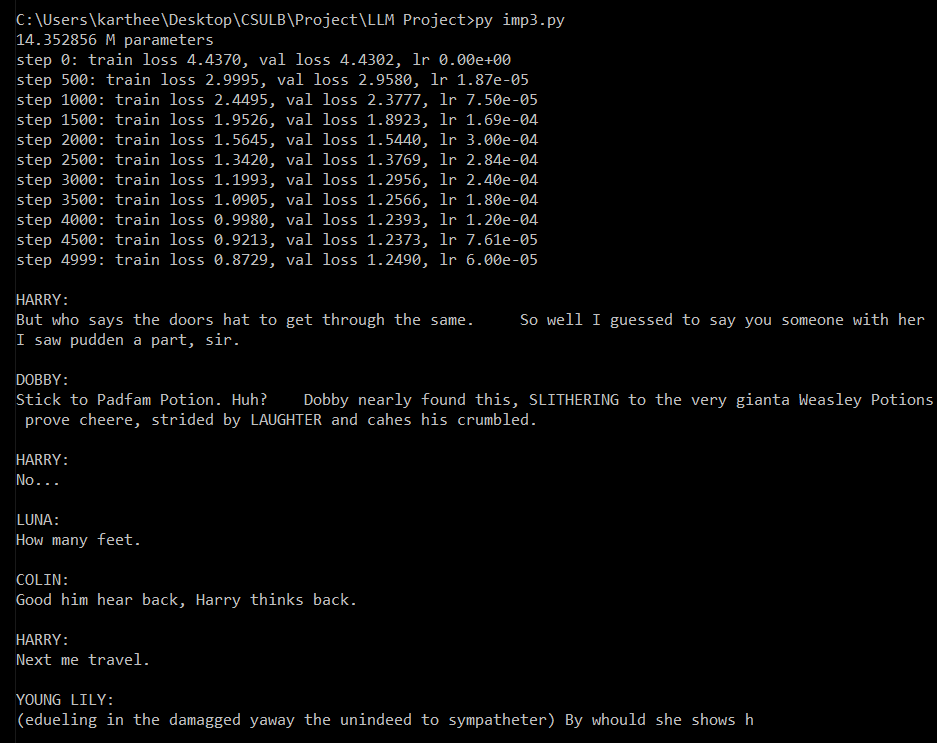
scheduler = torch.optim.lr\_scheduler.ReduceLROnPlateau(optimizer, 'min', factor=0.5, patience=10)

ITERATION 3:

Changes in imp2



Imp3



Imp4

