Name: --- (Student’s Name Here) ---

CSCI S-89B Introduction to Natural Language Processing

Assignment 2

**Problem 1 (25 points)**

In this problem, we will use a recurrent neural network for predictions. Please consider observations of temperature (in Celsius) measured every 10 minutes in Jena, Germany, between January 2009 and January 2017. The dataset is available at https://www.bgc-jena.mpg.de/wetter/.

Please load the time series:

df = pd.read\_csv('jena\_climate\_2009\_2016.csv', parse\_dates=True, index\_col='Date Time')

xt = df['T (degC)']

xt = xt.reset\_index(drop=True)

Reserve the last 1,440 observations for testing. Build a recurrent neural network of your choice for one-step ahead prediction. For this, you will need to minimize the validation Mean Squared Error (MSE). Please note that the validation set should not contain the 1,440 test observations.

1. Plot the 1,440 reserved observations of temperature (10 days) along with the corresponding one-step ahead predictions.
2. Compute the test MSE using these 1,440 observations and compare it with the test MSE obtained by a persistent forecast, in which the next observation is predicted to be the same.

To receive full credit, please create a network that beats the persistent forecast according to the test MSE.

SOLUTION:

--- (Student’s Solution Here) ---

**Problem 2 (15 points)**

Please consider the example of movie review classification found in

` 3.5-classifying-movie-reviews.ipynb`.

Change the number of most frequently used words (the `num\_words` parameter) from the suggested 10000 to 200 and run the classification.

1. Plot the results for training and validation accuracy versus the number of epochs.
2. Report the test accuracy of the model when trained with the optimal number of epochs.
3. Explain what `x\_train[0]` represents.

SOLUTION:

--- (Student’s Solution Here) ---

**Problem 3 (25 points)**

Modify the movie review pre-processing (“Preparing the data”) part of Problem 2 so that each input to the network is a sequence of vectors. The dimensions of `x\_train` must be (25000, 500, 200).

Next, change the architecture of the network to a recurrent network of your choice.

1. Plot the results for training and validation accuracy versus the number of epochs.
2. Report the test accuracy of the model when trained with the optimal number of epochs.
3. Explain what `x\_train[0]` and `x\_train[0,0]` represent now.

To receive full credit, please create a recurrent network (Simple RNN, LST, GRU, or mixture) that produces test accuracy of at least 0.75.

SOLUTION:

--- (Student’s Solution Here) ---