Name: --- (Student’s Name Here) ---

CSCI S-89B Introduction to Natural Language Processing

Assignment 3

**Problem 1 (25 points)**

In this problem, we will explore basic text processing techniques using the NLTK and SpaCy libraries in Python.

Choose a text of your choice (e.g., part of a news article, a movie review, a post, etc.) and perform the following steps using both NLTK and SpaCy.

1. Load and preprocess the text data using NLTK and SpaCy for tokenization, stemming, and lemmatization. Display samples of the processed text.
2. Compare the results of tokenization, stemming, and lemmatization from both libraries and discuss any differences observed.
3. Discuss the importance of tokenization, stemming, and lemmatization in NLP applications.

SOLUTION:

--- (Student’s Solution Here) ---

**Problem 2 (25 points)**

In this problem, we will explore word tokenization using NLTK and implement a Bag of Words (BoW) model using 1-grams, incorporating lemmatization in the preprocessing.

Choose a large text dataset (e.g., a two- or three-page news article) to create the vocabulary and then select a smaller text with some new words for the Bag of Words (BoW) representation. Perform the following steps using NLTK.

1. Apply word tokenization and then lemmatization using NLTK on both the large text dataset and the smaller text with new words. Display samples of the processed smaller text.
2. Using 1-grams, create a vocabulary from the large text dataset and implement a BoW model on the smaller text with new words. Display the resulting dictionary with frequencies that represent the new text.
3. Can you see from this dictionary how many new words the new text has? What is the key that corresponds to any new word?

SOLUTION:

--- (Student’s Solution Here) ---

**Problem 3 (15 points)**

Please consider the following example of a Convolutional Neural Network (CNN) for handwritten digit classification found in ` 5.1-introduction-to-convnets.ipynb`:

model = models.Sequential()

model.add(layers.Conv2D(32, (3, 3), activation='relu',

input\_shape=(28, 28, 1)))

model.add(layers.MaxPooling2D((2, 2)))

model.add(layers.Conv2D(64, (3, 3), activation='relu'))

model.add(layers.MaxPooling2D((2, 2)))

model.add(layers.Conv2D(64, (3, 3), activation='relu'))

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. Explain what test accuracy means in this context. If there are 10,000 images of handwritten digits in the test dataset, please explain how they are classified by your model.
3. Explain why the `Conv2D(32, (3, 3),…)` layer has a 26 × 26 × 32 dimensional output.
4. Explain why there are 320 parameters in this layer.

SOLUTION:

--- (Student’s Solution Here) ---