Assignment: Convolutional Neural Networks for Text Mining

Course: Text Mining (Summer Term 2025)
Instructor: [Your Name]
Submission Deadline: [Insert Date]

Part A: Conceptual Questions (Short Answer)

1. Explain the difference between **word embeddings** and **character-level representations**. In which cases might one outperform the other?

2. Why are CNNs more parameter-efficient than fully connected networks? Illustrate with an example.

3. Describe the role of **stride** and **padding** in convolution operations. How do they affect the output feature map dimensions?

4. Discuss the intuition behind **max pooling** in CNNs. Why is it especially useful for text classification tasks?

5. Compare and contrast the following embedding strategies:

   ○ cnn-rand

   ○ cnn-static

   ○ cnn-nonstatic

   ○ cnn-multichannel

Part B: Applied Exercises (Programming)

Use **Python** with **PyTorch** or **TensorFlow**. Submit your code along with explanations.

1. **Word-level CNN for Sentiment Classification**

   ○ Train a CNN-based classifier on the IMDB sentiment dataset.

   ○ Experiment with at least two embedding setups:

      ■ Pre-trained Word2Vec or GloVe embeddings (fixed).

      ■ Trainable embeddings (initialized randomly).

   ○ Compare results and discuss your findings.

2. **Effect of Convolution Parameters**

   ○ Implement a toy CNN where you vary **filter size** (e.g., 2, 3, 5 words) and **stride** values.

   ○ Visualize how these changes affect the extracted n-gram features.

3. **Character-level CNN**

   ○ Implement a simple character-level CNN for text classification (you may use a subset of data if training is heavy).

   ○ Reflect on performance compared to the word-level CNN.

Part C: Critical Reflection (Essay, 800–1000 words)

● Evaluate the impact of CNNs in **NLP** compared to earlier models (bag-of-words, n-grams, RNNs).

● Discuss the limitations of CNNs for text classification. Where might RNNs or Transformers still be preferable?

● Suggest one novel research direction or modification to CNNs that could improve their performance on modern NLP tasks.