BRAC UNIVERSITY
Department of Computer Science and Engineering

Examination: Midterm | Semester: Spring 2025
Duration: 75 minutes | Full Marks: 30

CSE 440: Natural Language Processing II

Figures in the right margin indicate marks.

Answer all 3

1. A. How can overfitting be detected in a machine learning model? What are the general techniques you can use to reduce overfitting? [4]

   B. You have built a spam detector model, and that model was tested on 100 emails. Out of these 100 emails, 80 were spams and 20 were not spams. Your model identified 90 of them as spam and 10 of them not spams. All the not spams the model identified are correct. Now, build a confusion matrix and calculate precision, recall and F-score for the spam class. [6]

2. A. Suppose you have a collection of 3 documents and a term "AI" appears in them as follows: Document 1 (D1): it has 100 words and "AI" appears 3 times, Document 2 (D2): it has 200 words and "AI" appears 5 times, Document 3 (D3): it has 150 words and "AI" appears 0 times in a document with 150 words. Now calculate TF-IDF for “AI” in all three of these documents. [4]

   B. With appropriate examples, write short notes on (what it is, what are the challenges, how to solve them etc.): [6]
      a. Named Entity Recognition
      b. Tokenization
      c. Stemming and lemmatization

3. A. The Book presents three semantic properties of word embeddings: Different types of Similarity or Association, Analogy/Relational Similarity, and Historical Semantics. Write short notes on each of them with appropriate examples. [6]

   B. You have 2 models, model A: $y_A = \sigma(2x_1 + 4x_2 - 3)$ and model B: $y_B = \sigma(3x_1 + 4x_2)$, and one data point x: $[x_1, x_2] = [0, 1]$ that has a label $y = 1$. Which model incurs higher binary cross entropy loss for x? [4]