

**BRAC UNIVERSITY**  
**Department of Computer Science and Engineering**

Examination: Midterm  
Duration: 75 minutes

Semester: Summer 2023  
Full Marks: 30

**CSE 440: Natural Language Processing II**

Figures in the right margin indicate marks.

Name:	ID:	Section:
-------	-----	----------

Answer all 3

1. [CO1]      A. **Explain** (with examples) one major challenge for each of the following [4]  
NLP pipeline components:
  - a. Named entity recognition
  - b. PoS tagging
  - c. Word tokenization
  - d. Stemming
  
2. [CO1]      B. You have two classifiers ( $y = \mathbf{w}^T \mathbf{x} + b$ ), classifier A with weight  $\mathbf{w}_A = [1 \ 0 \ 2]^T$  and bias  $b_A = 1$ ; and classifier B with weight  $\mathbf{w}_B = [1 \ 2 \ 0]^T$  with bias  $b_B = -1$ . For one example  $\mathbf{X} = [1 \ 1 \ 0]^T$ , predict  $y_A$  and  $y_B$ . Which classifier incurs lower cross entropy loss if  $\mathbf{X}$ 's original label is 1? **Calculate. Show your work.** [6]
  
2. [CO1]      A. **Explain** the three refinement techniques of bag-of-words features: [3]
  - a. Binarization
  - b. N-grams
  - c. Lexicon features

Write at least one advantage and one disadvantage for each of these refinement techniques.
  
2. [CO1]      B. **Derive** Bayes' rule using chain rule. [2]
  
2. [CO1]      C. **Explain** overfitting and underfitting. [3]
  
2. [CO1]      D. Explain why we use harmonic mean to calculate F1 score, not regular mean. [2]
  
3. [CO2]      A. **Show, with examples**, why we prefer to work with word-level [2]  
representations instead of sentence-level representations.
  
3. [CO2]      B. Let's say we are working with Shakespeare's plays, and we have two [4]  
plays in our hand: Anthony and Cleopetra and Julius Caesar, with  
three key characters: Anthony, Brutus and Caesar. These characters  
appear in the plays as many times given in table 1. Consider this as

your bag-of-words. Now build a term-term co-occurrence matrix. **Explain** what each value in the term-term co-occurrence matrix means.

- C. You are tasked to design an ML app that will try to identify cyberbully in social media. Your app's specific target is to identify name-calling, that is, identify derogatory words directed towards a specific person. After building your model, you could see that your model is associating traditionally feminine words like gendered pronouns (she, her) with name-calling compared to traditionally masculine gendered pronouns (he, his). Do you think your model has a problem? Why do you think your model was facing this problem? What solution can you propose to solve this issue? [4]

Table 1

	Anthony	Brutus	Caesar
Anthony and Cleopetra	16	13	7
Julius Caesar	12	9	21