Daffodil International University

Faculty of Science & Information Technology
Department of Computer Science & Engineering
Midterm Examination, Fall 2025
Course Code: CSE445, Course Title: Natural Language Processing
Level: 3 Term: 2 Batch: 63
Time: 01:30 Hrs **Marks: 25**

Answer ALL Questions

[The figures in the right margin indicate the full marks and corresponding course outcomes. All portions of each question must be answered sequentially.]

1.	*a)*	An IT company is developing a new **AI agent** for text generation. The team is going to follow proper guidelines for this task. As an AI Architect (**NLP Engineer**), define what should be the proper Text-Processing-Steps to follow?	2.5	CO1
	b)	Another research team is developing a document classification model for academic articles. The words "**studies**" and "**studying**" occur frequently. After applying stemming, the model's accuracy decreases. Now they would like to modify the text-processing steps. Which text normalization approach would be more appropriate for this corpus, and why?	2.5	
2.		**D1: AI improves healthcare service** **D2: AI improves education system** **D3: Education and healthcare need AI**		CO2
	a)	i. List all **unique terms** (vocabulary).	1	
		ii. Compute **DF** (document frequency) and **IDF** for each term.	1.5	
		iii. Identify which term is most **common** and most **rare**.	1	
		iv. Calculate **TF–IDF** values for each term in all documents.	1.5	
	b)	For the query "**AI healthcare system**", sum the TF–IDF scores of the matching terms in each document, and decide which document is most **relevant** to the query and why?	5	
3.		**S → NP VP** **NP → Det N | Det Adj N | NP PP** **VP → V NP | VP PP** **PP → P NP** **Det → the** **Adj → smart** **N → students | papers | library** **V → read** **P → in**		CO3
	a)	Drive a top-down parse tree '**the smart students read papers in the library**' using given Context Free Grammar.	7.5	
	b)	A Multinomial Naive Bayes email classifier is trained to detect Spam and Ham messages. When tested on **8,650 unseen emails**: The model predicted **60% of the emails as Spam and 40% as Ham**. Among the actual Spam emails (**50% of total**), the model correctly identified **80% as Spam**. Among the actual Ham emails, the model correctly classified **60% as Ham**. Is the dataset **balanced or unbalanced**? If it were the opposite case, analyze how the overall accuracy might change and why accuracy could become misleading for this model.	2.5	