ourdiu.com | Shared by SUPAN ROY

Daffodil International University

Faculty of Science & Information Technology
Department of Computer Science and Engineering
Midterm Examination, Fall-2026
Course Code: CSE432, Course Title: Natural Language Processing
Level: 4 Term: 2 Batch: 65

Time: 1.5 Hours **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)	A news-analytics pipeline processes the sentence *S: "Apple will not ship the new iPhone to US stores."* Explain how processing the sentence through a pipeline of "*Case folding → Stop-word removal → NER*" alters its original meaning. Illustrate at least two critical pieces of information lost during processing, and distinguish why reordering these pipeline stages preserves the sentence's semantic context.	[2]	
	b)	Consider the following context-free grammar (CFG): Production rules: S ⇒ NP VP; NP ⇒ PRP | NNP | DT NN; VP ⇒ VB | VB NP | VP PP; PP ⇒ IN NP Lexicon: DT ⇒ the | a; PRP ⇒ I | she; IN ⇒ in | at; VB ⇒ book | fly; NN ⇒ book | hotel | room		
	i)	Construct a Parse Tree based on LMD for the sentence 'I book the room in the hotel' using the provided grammar.	[2]	CO1
	ii)	Explain whether the provided grammar can generate the imperative sentence *'Book the hotel'* or not.	[1]	
	c)	Consider the corpus of four documents: D_1: rain rain rain stops play D_2: play resumes after rain D_3: crowd cheers the play D_4: rain delays the final		
	i)	A student claims that "rain" is the most important term of D_1 because it occurs the most. Verify the claim by computing the TF-IDF of *rain*, *stops* and *play* in D_1.	[2]	
	ii)	Explain why raw term frequency is insufficient for determining a word's overall importance in a document corpus. Interpret the student's claim by comparing the TF-IDF scores of *'rain'*, *'stops'*, and *'play'*, and illustrate how Inverse Document Frequency (IDF) balances term frequency to reflect true semantic weight.	[3]	
2.		A trigram language model is trained on the following corpus (each sentence is padded with two start tokens and one end token): <s> <s> I love NLP </s> <s> <s> I love deep learning </s> <s> <s> I hate exams </s> <s> <s> students love NLP </s>		CO2

ourdiu.com

ourdiu.com | Shared by SUPAN ROY

	a)	Using MLE, compute P(I love NLP) and P(students love deep learning). Every bigram of the second sentence occurs in the corpus, yet its probability is zero. Explain why, and state which model order would give it a non-zero probability.	[2]	
	b)	Apply Laplace (add-one) smoothing to the given language model to demonstrate whether P(</s> | love NLP) > P(I | <s> <s>) remains true.	[2]	
	c)	For the test sentence <s> <s> I love Python </s>, explain why Laplace smoothing with the V of this corpus still cannot assign P(Python | I love).	[1]	
3.		A multinomial Naïve Bayes spam filter is trained with Laplace smoothing on: *Description* / *Class*: win cash now — Spam; free cash — Spam; team meeting now — Ham; project report today — Ham; lunch meeting today — Ham Test e-mail: "free meeting now urgent"		
	a)	Compute the prior probabilities and the smoothed likelihoods of the test words that occur in the training vocabulary.	[3]	
	b)	Engineer A discards the unseen word "urgent". Determine the predicted class.	[2]	CO2
	c)	Engineer B keeps "urgent" and gives it the smoothed likelihood $\frac{0+1}{N_c+\|V\|}$ in each class. Demonstrate whether the prediction changes, and explain in one sentence why an unseen word is *not* class-neutral here.	[2]	
	d)	On a test set of 1000 e-mails (50 spam), a filter X gives $TP = 30$, $FN = 20$, $FP = 40$, $TN = 910$ (positive class = Spam). The manager prefers a dummy classifier that labels *every* e-mail as Ham "because it is more accurate". (i) Compute the accuracy of both classifiers. (ii) Compute the F_1 score (Spam class) of both classifiers. (iii) If deleting a genuine e-mail is the costliest mistake, should precision or recall be prioritized? Justify your answer.	[3]	