ourdiu.com | Shared by Md. Shefat Al Mahmud

Daffodil International University

Faculty of Science & Information Technology
Department of Computer Science & Engineering
Mid Examination, Spring 2025
Course Code: CSE445, Course Title: Natural Language Processing
Level:4 Term: 2 Batch: 60

Time: 01:30 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)***	What are the main challenges in Natural Language Processing (NLP), and why is it difficult for machines to understand and process human language effectively? Discuss key issues such as ambiguity, context understanding, syntactic and semantic variability, and computational complexity.	**3**	**CO1**
	b)	A financial news platform wants to develop an automated sentiment analysis system to classify stock market news articles as "Bullish" (positive), "Bearish" (negative), or "Neutral." However, the raw text data contains informal writing, redundant words, and different grammatical variations of the same term (e.g., invest, investing, invested). To improve model performance, the NLP team decides to apply text preprocessing techniques. As an NLP specialist, answer the following: a) "The role of each preprocessing step (tokenization, stop-word removal, stemming, and lemmatization) in transforming raw text data." b) Why is stop-word removal important for improving the model's focus on meaningful words? c) The trade-offs between stemming vs. lemmatization and when one might be preferred over the other in real-world applications.	**3**	
2.	***a)***	Compare the role of Part-of-Speech (POS) tagging, Named Entity Recognition (NER), and Parsing in text classification. How do these techniques help improve text in Clustering?	**3**	**CO2**
	b)	A research team is building a news article categorization system that automatically classifies articles into topics like Politics, Sports, and Technology. Initially, they use the Bag-of-Words (BoW) model to represent text, but they notice that the model struggles with understanding contextual meaning and handling large vocabulary sizes. As an NLP expert, answer the following: a) Explain the ways of the Bag-of-Words (BoW) model representation text and evaluate its strengths and limitations in document classification. b) How does the Vector Space Model (VSM) improve document similarity measurement compared to BoW? c)Analyze the role of N-grams (bigrams/trigrams) in enhancing text classification. Provide a practical example where N-grams improve model accuracy compared to BoW.	**2+2+1**	
3.	***a)***	A company is building a document search engine that retrieves relevant documents based on user queries. They use the TF-IDF (Term Frequency-	**5**	**CO3**

ourdiu.com | Shared by Md. Shefat Al Mahmud

Inverse Document Frequency) method to compute word importance in documents.
Given the following corpus of three documents:
Doc1: "Deep learning improves NLP models."
Doc2: "TF-IDF helps NLP document search."
Doc3: "Deep learning and TF-IDF are used in text analysis."
A user searches for: "Deep learning NLP". Now perform the relevant operation.

b) A company wants to automatically classify customer reviews as Positive (P) or Negative (N) using a Naïve Bayes classifier. Given a dataset of labeled reviews, you need to predict the sentiment of a new review: "service is bad" — **6**

Review	Sentiment (Level)
"food is great"	Positive (P)
"service is good"	Positive (P)
"bad food"	Negative (N)
"food is bad"	Negative (N)
"bad service"	Negative (N)