Daffodil International University

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Midterm Examination, Spring-2024
Course Code: CSE445, Course Title: Natural Language Processing
Level: 4 Term: 1 Batch: 58

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.	Assume the given corpus: Walking is a good exercise. Mr. M. A. Sattar walks every morning. His email is sattar_143@diu.edu.bd.			
	a)	**Apply** the following preprocessing techniques on the corpus and determine the final outcome. i) Remove punctuation and special character ii) Tokenization iii) Remove stopwords (determinant, auxiliary verbs) iv) Lemmatization	[3]	CO1
	b)	**Construct** a Regular Expression that can identify any email (email consists of any small letter, number and special character like "_", "." and "@".	[2]	
2.	***a)***	**Apply** the Hidden Markov model to find out the Emission and Transition Probabilities considering the corpus below: 1. Sandy <Noun> Maya <Noun> enjoy <Verb> university <Noun>. 2. Can <Modal> Sandy <Noun> find <Verb> university <Noun>? 3. Will <Modal> Maya <Noun> find <Verb> university <Noun>? 4. Sandy <Noun> can <Modal> observe <Verb> Will <Noun>. 5. Maya <Noun> will <Modal> cherish <Verb> MIT <Noun>. 6. Will <Noun> is <Verb> student <Noun>.	[4]	CO1
	b)	**Determine** the appropriate sequence of tags for the sentence, "Will can find campus." from the above probabilities.	[3]	
3.	***a)***	**Explain** whether the following Context free grammar G is suitable for the CYK algorithm mentioning appropriate reason:	[2]	CO2

		S → AC | BD A → AA | a B → CC | b C → AB | c D → DC | d		
	b)	If possible, **apply** the CYK algorithm step by step using the formula to determine if "cadbba" is in L(G).	[6]	
4.	Consider the following documents: Doc1: <Bangladesh is an independent country.> Doc2: <Bangladesh became independent with a liberation war.> Doc3: <Myanmar become independent this year.>			
	a)	**Build the vocabulary** of the corpus (alphabetic order) and **vectors** for the documents (based on the ordered vocabulary) after applying the following preprocessing: i) Remove punctuation and special character ii) Convert the corpus into small letter iii) Tokenization iv) Remove stopwords (determinants and prepositions)	[2]	CO2
	b)	**Determine** the similarity between Doc2 and Doc3 using cosine similarity and Euclidean distance.	[3]	