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.	*c)*	**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
	h)	**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

		$S \rightarrow AC \mid BD$ $A \rightarrow AA \mid a$ $B \rightarrow CC \mid b$ $C \rightarrow AB \mid c$ $D \rightarrow DC \mid d$		
	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.>			**CO2**
	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]	
	b)	**Determine** the similarity between Doc2 and Doc3 using cosine similarity and Euclidean distance.	[3]	

Daffodil International University

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Midterm Examination, Fall-2023

Course Code: CSE445, Course Title: Natural LanguageProcessing

Level: 4 Term: 1, 2 Batch: 56, 57

Time: 1 Hour and 30 Minutes **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.]

<table>
<tr><td rowspan="2">1.</td><td>a)</td><td>Define Natural Language Processing. Write down the major areas of research and development in NLP.</td><td>3</td><td rowspan="2">CO1</td></tr>
<tr><td>b)</td><td>Write down the regular expression based on the following scenario:
“First match the word Column, then followed by a number and optional spaces, the whole pattern repeated any number of times”</td><td>2</td></tr>
<tr><td rowspan="2">2.</td><td>a)</td><td>Describe the difference between lemmatization and stemming with proper examples.</td><td>3</td><td rowspan="2">CO2</td></tr>
<tr><td>b)</td><td>Analyze the best suitable algorithm for calculating minimum edit distance to convert “EDITING” to “DISTANCE” (using insertion cost = 1, deletion cost = 1 and substitution cost = 2) .</td><td>5</td></tr>
<tr><td>3.</td><td>a)</td><td>Analyze Naive Bayes Theorem to predict the category of this testing data: Bangladesh cricket team get excellent victory
<table>
<tr><th>words</th><th>class</th></tr>
<tr><td>Bangladesh cricket team victory</td><td>Pos</td></tr>
<tr><td>Cricket matches Bangladesh</td><td>Pos</td></tr>
<tr><td>Cricket players skillful</td><td>Pos</td></tr>
<tr><td>Proud of Bangladesh team</td><td>Pos</td></tr>
<tr><td>Bangladesh cricket match defeated</td><td>Neg</td></tr>
<tr><td>Bangladesh cricket team get struggled</td><td>Neg</td></tr>
<tr><td>Excellent batting Bangladesh team</td><td>Pos</td></tr>
<tr><td>Injuries in Bangladesh cricket team</td><td>Neg</td></tr>
</table></td><td>7</td><td>CO2</td></tr>
<tr><td>4.</td><td>a)</td><td>Explain shortly about “discounting”. Mention the problems regarding HMM and give proper solutions.</td><td>2+3</td><td>CO2</td></tr>
</table>