Roll No.												

PRESIDENCY UNIVERSITY

BENGALURU

End - Term Examinations – December 2025

Date: 18 – 12- 2025 **Time:** 01:00pm – 04:00pm

School: SOCSE	**Program:** B.Tech. - CST Big Data	
Course Code : CBD3410	**Course Name:** Natural Language Processing for Big Data	
Semester: VII	**Max Marks**: 100	**Weightage**: 50%

CO - Levels	**CO1**	**CO2**	**CO3**	**CO4**	**CO5**
Marks	**26**	**24**	**26**	**24**	**-**

Instructions:

(i) Read all questions carefully and answer accordingly.

(ii) Do not write anything on the question paper other than roll number.

Part A

Answer ALL the Questions. Each question carries 2marks. **10Q x 2M=20M**

1.	What are the main components of an NLP system?	**2 Marks**	**L1**	**CO1**
2.	Define stop-word removal and state its purpose in NLP.	**2 Marks**	**L1**	**CO1**
3.	What is lemmatization? Provide a suitable example.	**2 Marks**	**L1**	**CO1**
4.	Define smoothing in N-gram models. Why is it needed?	**2 Marks**	**L2**	**CO2**
5.	What is chunking in NLP? Give one real-world application.	**2 Marks**	**L2**	**CO2**
6.	What is text classification? Mention one suitable algorithm.	**2 Marks**	**L2**	**CO3**
7.	What are the main objectives of sentiment analysis in NLP? Give one real-life example	**2 Marks**	**L3**	**CO3**
8.	Define text clustering. Mention one common algorithm used for it.	**2 Marks**	**L3**	**CO3**
9.	Define Spark MLlib. Mention its use in text analytics.	**2 Marks**	**L1**	**CO4**
10.	What is the purpose of cloud-based NLP services like AWS Comprehend?	**2 Marks**	**L1**	**CO4**

Part B

Answer the Questions. **Total Marks 80M**

11.	**a.**	Describe the various stages of text preprocessing in NLP.	**5 Marks**	**L1**	**CO1**
	b.	Explain the importance of normalization and case-folding in text cleaning.	**5 Marks**	**L1**	**CO1**
		Or			
12.	**a.**	Differentiate between stemming and lemmatization with illustrations.	**5 Marks**	**L1**	**CO1**
	b.	Explain the process of tokenization and stop word removal with a suitable example.	**5 Marks**	**L1**	**CO1**

13.	**a.**	Using the given corpus, estimate Trigram probabilities and find the most probable next word for the sequence: <s> Data scientists build ___ ?` Corpus <s> Data scientists build models </s> <s> Data scientists build pipelines </s> <s> Data scientists build models </s> <s> Data scientists evaluate models </s> <s> Teams build models </s> <s> Data engineers build pipelines </s> <s> Data scientists build models </s>	**5 Marks**	**L4**	**CO2**
	b.	Describe the role of grammar rules in building language models.	**5 Marks**	**L2**	**CO2**
		Or			
14.	**a.**	Explain how an N-gram model predicts the next word in a sentence.	**5 Marks**	**L3**	**CO2**
	b.	Discuss how parsing helps in improving machine translation systems.	**5 Marks**	**L2**	**CO2**

15.	**a.**	Explain the working of the Naïve Bayes algorithm for text classification.	**5 Marks**	**L3**	**CO3**
	b.	Descibe the steps involved in performing sentiment analysis.	**5 Marks**	**L2**	**CO3**
		Or			
16.	**a.**	Explain the steps in building a text summarization system.	**5 Marks**	**L3**	**CO3**
	b.	Discuss how document clustering helps in organizing large text data.	**5 Marks**	**L3**	**CO3**

17.	**a.**	Explain the components and architecture of Spark NLP.	**5 Marks**	**L3**	**CO4**

	b.	Describe how MLlib supports NLP-related machine learning tasks.	**5 Marks**	**L2**	**CO4**
		Or			
18.	**a.**	Compare spaCy and NLTK in terms of design, speed, and use cases.	**5 Marks**	**L3**	**CO4**
	b.	Discuss the significance of using Azure Text Analytics in enterprise NLP.	**5 Marks**	**L2**	**CO4**

19.	**a.**	Describe the importance of text preprocessing in Natural Language Processing.Explain each step of the pipeline using the sentence "Data Science is intresting."	**10 Marks**	**L5**	**CO1**
	b.	Discuss how word embeddings improve semantic understanding in NLP	**5 Marks**	**L4**	**CO2**
	c.	Analyze the use of linguistic resources in improving parsing performance.	**5 Marks**	**L3**	**CO2**
		Or			
20.	**a.**	Differentiate between Stemming and Lemmatization. Explain how both techniques reduce words to their root form with examples.	**10 Marks**	**L3**	**CO1**
	b.	Explain the working of N-gram models.	**5 Marks**	**L2**	**CO2**
	c.	Discuss syntax parsing and Word2Vec with an example.	**5 Marks**	**L2**	**CO2**

21.	**a.**		**10 Marks**	**L5**	**CO3**
	b.	Evaluate the use of Spark NLP for real-time social media data analysis.	**5 Marks**	**L3**	**CO4**
	c.	Explain the role of Hugging Face Transformers in modern NLP workflows.	**5 Marks**	**L2**	**CO4**

ID	**Keywords in the Document**	**Class h**
7	Fun Enjoy Game Smile	**?**

Using the given dataset, apply the Naïve Bayes text classification technique to classify the test document (Document ID 7) into the correct class (h or –h) with proper calculation and justification.

ID	**Keywords in the Documents**	**Class h**
1	Play Fun Enjoy Happy	Yes
2	Fun Laugh Enjoy Smile	Yes
3	Enjoy Game Win Fun	Yes
4	Work Stress Pain Hard	No
5	Tired Pain Work Sleep	No
6	Pain Stress No Sleep	No

		Or			
22.	**a.**	Define sentiment analysis and explain the purpose of using it in NLP. Using the given word list and sentence, calculate the overall sentiment score and classify the sentiment. Positive words: good (+1), amazing (+2), happy (+1), love (+1) Negative words: bad (–1), boring (–1), hate (–1), predictable (–1) Sentence: "The movie was good and the performances were amazing, but the ending was boring and predictable."."	**10 Marks**	**L3**	**CO3**
	b.	Compare AWS Comprehend and Azure Text Analytics.	**5 Marks**	**L2**	**CO3**
	c.	Discuss about NLP pipelines in spaCy	**5 Marks**	**L2**	**CO3**