**Mukesh Patel School of Technology Management & Engineering**
**School of Technology Management and Engineering**

**Course Policy Document**

**Course Name - (Code): Natural Language Processing (702AI0C009)**

| Program and Semester: B Tech/MBA Tech (Artificial Intelligence) Sem 5 | Pre-requisite Course : Data Structures and Algorithms, Python Programming | | | | | |
|---|---|---|---|---|---|---|
| **Academic Year:** 2024-25 | **Credit Details:** | L | T | P | C | H |
| | | 2 | 0 | 2 | 3 | 4 |
| **Name of Course Faculty:** Ami Munshi **Program:** **Contact Details:** ami.munshi@nmims.edu **Office Hours: 9am-4pm** | **Faculty associated with the course: -** **1.** **2.** | | | | | |

**Pre-Course Activity:**

1. Identify real time applications of Natural Language Processing in various domains (You may use web search and generative AI tools to do this activity)
2. Recognize future scope of Natural Language Processing in emerging domains of Artificial Intelligence

**Course link:**

# 1. Introduction to the Course
## 1.1 Importance of the Course

    **1.1.1** **Domain** Relevance**:** Natural Language Processing (NLP) is essential in Artificial Intelligence (AI), enabling machines to understand, interpret, and respond to human language, thus enhancing human-computer interaction through conversational AI and improving user experiences. It plays a key role in information retrieval, sentiment analysis, and text analytics, making data analysis more efficient. NLP also powers machine translation, automates customer support, and streamlines healthcare processes by analyzing medical records and generating clinical documentation. Additionally, it aids in security by detecting fraud, ensures regulatory compliance, and personalizes content and recommendations, making it a cornerstone of modern AI applications.

    **1.1.2** **Industry** Relevance**:** Natural Language Processing (NLP) is highly relevant across various industries, driving advancements in customer service through chatbots and virtual assistants, enhancing search engines and information retrieval systems, and improving sentiment analysis for market research. In healthcare, NLP aids in analyzing medical records and automating clinical documentation. It also facilitates

machine translation and localization in global business, detects fraud in financial services, ensures compliance in regulatory environments, and personalizes user experiences in e-commerce and content platforms. Overall, NLP's ability to process and understand human language makes it indispensable for modern industry applications.

### 1.2 Objectives of the Course:

**1.2.1** Enable students to understand and apply NLP techniques for real-world applications such as information extraction, language translation, sentiment analysis, and the development of Large Language Models.

**1.2.2** Introduce and explain various NLP algorithms and methods, including text preprocessing, word level analysis, syntax analysis, semantic analysis, pragmatic analysis, and discourse analysis.

**1.2.3** Equip students with practical, hands-on experience in contemporary NLP, highlighting its relevance and application in modern technology and industry.

## 2. Course Outcomes (CO), Mapping with Program Outcomes (PO), and Program Specific Outcomes (PSO)

### 2.1 Course Outcomes

**2.1.1 CO1:** Describe the fundamental concepts and techniques of natural language processing

**2.1.2 CO2:** Apply text preprocessing, word level analysis and syntax level analysis techniques on textual data

**2.1.3 CO3:** Implement approaches for semantic level analysis and pragmatic level analysis

**2.1.4 CO4:** Illustrate various real world applications of NLP

### 2.2 Program Outcomes(PO) the course contributed to:

**2.2.1 PO-1: Engineering knowledge:** Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

**2.2.2 PO-2: Problem analysis:** Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

**2.2.3 PO-3: Design/Development of solutions:** Design solutions for complex engineering problems and design system components or processes that meet

the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

**2.2.4** **PO-4: Conduct Investigations of Complex problems:** Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

**2.2.5** **PO-5: Modern Tool Usage:** Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

**2.2.6** **PO-6: The engineer and society:** Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

**2.2.7** **PO-7: Environment and sustainability:** understand the impact of the professional engineering solutions in societal and environmental contexts, demonstrate the knowledge of, and need for sustainable development.

**2.2.8** **PO-8: Ethics:** Apply ethical principles and commit to professional ethics, responsibilities, and norms of the engineering practice.

**2.2.9** **PO-9: Individual and teamwork:** Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

**2.2.10** **PO-10: Communication:** Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

**2.2.11** **PO-11: Project management and finance:** Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

**2.2.12** **PO-12: Life-long learning:** Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

## 2.3 Program Specific Outcomes (PSO):

      **2.3.1   PSO-1:**

      **2.3.2   PSO-2:**

**CO-PO Mapping**

|     | PO1 | PO2 | PO3 | PO4 | PO5 | PO6 | PO7 | PO8 | PO9 | PO10 | PO11 | PO12 |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|------|------|
| CO1 | 1   | 1   |     |     |     |     |     |     |     |      |      |      |
| CO2 | 1   | 1   | 1   | 2   | 3   |     |     |     |     |      |      |      |
| CO3 | 1   | 1   | 1   | 1   | 2   | 3   |     |     |     |      |      |      |
| CO4 | 1   | 1   | 1   | 1   | 2   | 3   |     | 3   | 3   | 3    | 3    | 3    |

*Mapping Levels: 1- High, 2-Medium, 3-Low*

    **2.4 Student Outcomes (SO) (For ABET accredited Programs):**

      **2.4.1**    SO-1: an ability to identify, formulate, and solve complex engineering problems by applying principles of engineering, science, and mathematics.

      **2.4.2**    SO2-an ability to apply engineering design to produce solutions that meet specified needs with consideration of public health, safety, and welfare, as well as global, cultural, social, environmental, and economic factors.

      **2.4.3**    SO3-an ability to communicate effectively with a range of audiences.

      **2.4.4**    SO4-an ability to recognize ethical and professional responsibilities in engineering situations and make informed judgments, which must consider the impact of engineering solutions in global, economic, environmental, and societal contexts.

      **2.4.5**    SO5-an ability to function effectively on a team whose members together provide leadership, create a collaborative and inclusive environment, establish goals, plan tasks, and meet objectives.

      **2.4.6**    SO6-an ability to develop and conduct appropriate experimentation, analyze and interpret data, and use engineering judgment to draw conclusions.

      **2.4.7**    SO7-an ability to acquire and apply new knowledge as needed, using appropriate learning strategies.

# 3. Teaching-learning methodology

## 3.1 Instruction Plan

| Lecture No. | Topic | Teaching Method * | Blooms Level | Resources | COs mapped | Assessment and Evaluation |
|-------------|-------|-------------------|--------------|-----------|------------|---------------------------|
| **Unit 1: Introduction to Natural Language Processing** | | | | | | |

| | | | | | |
|---|---|---|---|---|---|
| 1 | Definition and significance of NLP, Real-world, applications of NLP NLP tasks and applications | Lecture based learning Presentation/ Smartboard | BL1,2 | TB1, TB2 | CO1 |
| 2 | Building blocks of language, Stages of linguistic analysis in NLP, Approaches to NLP, Concepts of Natural Language Understanding (NLU) and Natural Language Generation (NLG) Generic NLP pipeline and challenges in NLP | Blended Learning | | TB1 | CO1 |

| | | | | | |
|---|---|---|---|---|---|
| **Unit 2: Text Pre-processing and Text Representation** | | | | | |
| 3 | Tokenization Lower casing | Blended Learning | BL3,4 | TB1, TB2 | CO2 |
| 4 | Stopword and punctuation removal, Stemming and lemmatization | | | | |
| 5 | Label encoding One hot encoding | | | | |
| 6 | Bag of Words (BoW),Term Frequency and Inverse Term Frequency (TF-IDF) | | | | |
| 7 | Lexical similarity measures | | | | |
| **Unit3: Accessing Text Corpora and Word Level Analysis** | | | | | |
| 8 | Text Corpus Structure, Gutenberg Corpus, Web and Chat Text | | BL3,4 | TB2 | CO2 |

Lab assessment, Quiz, CT1 (for rows 1–2)

Lab assessment, Quiz, CT1 (for rows 3–7)

Lab assessment, Quiz, CT1 (for row 8)

| | | | | | | |
|---|---|---|---|---|---|---|
| 9 | Annotated Text Corpora Pronouncing Dictionary, Wordlist Corpora, Regular Expressions | Lecture based learning Presentation/ Smartboard | | TB2 | | |
| 10 | N-gram language models | | | TB2 | | |
| 11 | Smoothing models | | | TB2 | | |
| **Class Test 1 will be conducted in the 6th Week** | | | | | | |
| **Unit4: Syntax Level Analysis** | | | | | | |
| 12 | Introduction to POS tagging, Rule-based and stochastic tagging | Blended Learning | BL2,3,4 | TB2 | CO2 | Lab assessment, Quiz, CT2 |
| 13 | Machine learning approaches to POS tagging | | | | | |
| 14 | Concept of chunking and chinking | | | SCAN ME | | |
| 15 | Syntax parsing techniques | | | SCAN ME | | |
| 16 | Introduction and approaches to NER | | | TB1 | | |
| **Unit 5: Semantic Analysis and Vector Space Models** | | | | | | |
| 17 | Relations among lexemes: Homonymy, Polysemy, Synonymy, Hyponymy, | Blended Learning | BL2 | | CO3 | Lab assessment, Quiz, CT2 |
| 18 | Introduction to WordNet | | BL2 | | CO3 | |
| 19 | WordNet Hierarchy | | BL2 | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 20 | Word Sense Disambiguation (WSD), ☐ Approaches to WSD | | BL3,4,5 | | CO3, CO4 | |
| 21 | Distributional Semantics, ☐ Lexical Semantics | | | | CO3 | |
| 22 | Vector Semantics | | | | | |
| 23 | Word embeddings: word2vec | | | | CO3, CO4 | |
| **Class Test 2 will be conducted in the 13ᵗʰ week** | | | | | | |

**Unit 6: Pragmatic and Discourse Analysis**

| | | | | | | |
|---|---|---|---|---|---|---|
| 25 | Reference resolution | Blended Learning | BL3,4 | TB2 | CO3 | Lab assessment, Quiz |
| 26 | Co-reference resolution | | | | | |
| 27 | Pronoun resolution | | | | | |
| 28 | Approaches to pronoun resolution | | | | | |

**Unit 7: Real word applications of NLP**

| | | | | | | |
|---|---|---|---|---|---|---|
| 28 | Text Classification, Sentiment analysis, | Flipped Classroom | BL3,6 | | CO4 | Lab assessment, Quiz |
| 30 | Machine translation, Question answering systems, Information extraction and retrieval | | | | | |

BL1- Remembering , BL2-Understanding, BL3- Applying, BL4- Analyzing, BL5- Evaluating, BL6- Creating

# Resources

| Text Books |
| --- |
| 1.  Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, Harshit Surana, *Practical Natural Language Processing*, O'Reilly Publications, 2020. |
| 2.  Daniel Jurafsky and James H. Martin, *Speech and Language Processing*, 2nd Edition, Pearson Publications, 2014 (Classic). |
| **Reference Books** |
| 1.  Hobson Lane, Cole Howard, Hannes Hapke, *Natural language processing in action*, 1st Edition, Manning Publications, 2019. |
| 2.  Benjamin Bengfort, *Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning*, 1st Edition, O'Reilly Media, 2018. |
| 3.  Antic Z., *Python Natural Language Processing Cookbook*, Packt Publishing, 2021. |

# 4. Assessment and Evaluation Scheme

| | Internal Continuous Assessment (ICA) (50 Marks) | | | | |
| --- | --- | --- | --- | --- | --- |
| | **Class Tests (20 Marks)** | | **Term-work (30 Marks)** | | |
| **Assessment Component** | Class Test-1 | Class Test-2 | Mini project (case study/research paper) | Lab performance and practicum submission | Quiz /Assignment |
| **Marks** | 10 | 10 | 10 | 10 | 10 |

## 4.1 Internal Continuous Assessment (ICA) – 50 marks

### 4.1.1 Mini Project implementation and presentation (15 marks)

- A group of 2-3 students will identify an application relevant to the topics of Natural Language Processing After topic approval, students will start working towards implementation

- Evaluation of mini project
  - o Midterm evaluation (5 Marks)
  - o Final evaluation -Implementation and presentation (5 Marks)

### 4.1.2 ClassTest (10 marks+10marks)

- Class Test 1 will be conducted in the 6th week and Class 2 in the 13th week for 10 marks, and the duration would be 1 hour.

- **Lab performance and submission (10 marks)**

  - Continuous assessment for laboratory experiments will be conducted. There are 9 lab exercises, each carrying weightage of 10 marks. Practicum submission

also carries 10 marks. At the end of the course, average of total marks will be taken to obtain marks out of 10.

- Discussion of your work with your peers is allowed. However, each student is expected to submit his/her original work. Submissions which are very similar will be marked zero. Assessment of the lab work will be carried out based on parameters like timely completion of lab work file, understanding of the experiment performed, originality in the work, involvement of the student, regularity, discipline etc. during the session. There is a 40% penalty on late submission.

### 4.1.3 Lab performance and submission (10 marks)

- Continuous assessment for laboratory experiments will be conducted. There are 9 lab exercises, each carrying weightage of 10 marks. Practicum submission also carries 10 marks. At the end of the course, average of total marks will be taken to obtain marks out of 10.

- Discussion of your work with your peers is allowed. However, each student is expected to submit his/her original work. Submissions which are very similar will be marked zero. Assessment of the lab work will be carried out based on parameters like timely completion of lab work file, understanding of the experiment performed, originality in the work, involvement of the student, regularity, discipline etc. during the session. There is a 40% penalty on late submission.

### 4.1.4 Quiz/Assignment (10 marks)

- In total 3 quizzes will be conducted in the $5^{th}$ week, $11^{th}$ week and $14^{th}$ week to cover the entire syllabus. At the end of the course, average of total marks will be taken to obtain marks out of 5.

- One or two assignments would be given based on the syllabus

## 4.2 Term End Examination (TEE) – (100 marks scaled down to 50)

TEE conducted at the end of the semester will be for 100 marks (3-hour duration). Marks obtained scaled down to 50. There will be 7 questions (each question will be 20 marks). Q1 will be compulsory and any 4 from the remaining to be solved.

## 4.3 Course Passing Criteria

    **4.3.1** ICA (50 marks) – No minimum marks

    **4.3.2** TEE (100 marks scaled to 50) – 40% required for passing

**4.3.3** (ICA + TEE) (100 marks) -

### 4.4 Assessments and Mapping to Course Outcomes

| | Internal Continuous Assessment (ICA) | | | | | Term End Examination (TEE) |
|---|---|---|---|---|---|---|
| Course Outcomes | CT-1 | CT-2 | Mini Project | Lab Performance and Submission | Quiz | TEE |
| CO-1 | Y | | Y | Y | Y | Y |
| CO-2 | Y | | Y | Y | Y | Y |
| CO-3 | | Y | Y | Y | Y | Y |
| CO-4 | | Y | Y | Y | Y | Y |

## 5. Laboratory details

The following 10 programming exercises will form the submission for laboratory coursework. Each programming exercise will contain 3 to 5 programs.

| Exp. No. | Week No.# | Programming Topic | Mapped CO |
|---|---|---|---|
| 0 | | **Revisiting Python Concepts –File handling and Regular expressions**<br>**File handling**<br>a. Read all the content present inside the file<br>   i. Read only the first 5 characters of the file.<br>   ii. Read the content of the file on a line by line basis.<br>   iii. Read all the lines present inside the text file including the newline characters.<br>   iv. ReadaspecificlinefromaFile<br>b. Write content in a file<br>   i. Write the String 'Hello World' into the 'test.txt' file<br>   ii. Write "Hello World" in first line and "Hello Python" in second line<br>   iii. Write list of fruits in .txt file<br>c. Write a program to read resume and its details<br>**Regular expression**<br>a. Search for a sequence that starts with "he", followed by two (any) characters, and an "o": Txt="hello world"<br>b. For txt = "The rain in Spain" , search the string to see if it | |

| | | | |
|---|---|---|---|
| 2 | | i.     starts with "The"<br>ii.     ends with "Spain"<br>iii.     starts with "The"  and ends with "Spain"<br>c.   Find all lower/Upper case characters alphabetically between "a" and "m": txt = "The rain in Spain"<br>d.   Search for the first white-space character in the string: txt = "The rain in Spain"<br>e.   Replace every white-space character with the number 9 in any given string<br>f.   Check if an email ID is valid or not | |
| 1 | | **Text preprocessing level-1**<br><br>i.     To implement operations such as change of case, sentence tokenization, word tokenization, stop word removal, punctuation mark removal, stemming, lemmatization, Parts of Speech (PoS) tagging using NLTK (Natural Language Tool Kit) platform<br>ii.     To implement tokenization without using built in function of nltk<br>iii.     To comprehend the difference between stemming and lemmatization<br>iv.     To count frequency of each word in the given document | |
| 2 | | **Text pre-processing level-2**<br><br>i.     To implement label encoding and one hot encoding on textual data<br>ii.     To implement Bag of Words (BoW) feature engineering technique on textual data<br>**iii.**     To implement TF-IDF feature engineering technique To analyze and comprehend the effect of various approaches to convert text into vectors | |
| 3 | | **Analyzing Gutenberg and Brown corpus with python**<br><br>i.     Working with corpus file lists<br>**ii.**     Working with file contents<br>**iii.**     Visualization | |
| 4 | | **Language Modelling- N gram model** | |
| 5 | | **Information extraction -Part of Speech(POS) tagging**<br><br>**i.**     Identify the Part of speech like noun, verb, adjective, adverb and tag it<br>**ii.**     Identify Named Entity in text data | |

| 6 | | **Named Entity Recognition** | |
|---|---|---|---|
| 7 | | **Word Sense Disambiguation** | |
| 8 | | **Word Embedding** <br><br> **i.** Implementation of Word2Vec word embedding technique to observe similarity between two words/sentences <br> **ii.** Implementation of GloVe word embedding technique to measure semantic similarity | |
| 9 | | **Sentiment Analysis** <br><br> **i.** Implement sentiment analysis on the given dataset in Natural Language Processing <br> **ii.** Analyze and comprehend the results obtained | |
| 10 | | **Practicum** | |

# 6. Tutorial Plan

# 7. Course Material

References and Lab Manuals would be uploaded on LMS by faculty every week.

# 8. GenAI Usage
## 8.1 Pre-class Activity:

The student can use the AI tool to gather background information on the topic of the laboratory assignment as follows

- **Topic Overview**: Ask the AI to provide a summary or overview of the topic.
- **Key Concepts**: Request explanations of key concepts or theories related to the laboratory assignment.

## 8.2 In-Class Activity:

GenAI should not be used in "In-Class Activity".

# 9. Academic Integrity Statement

Original work expected from students for all of the **assigned assessment work**. Copying in any form not acceptable and will invite strict disciplinary action. Evaluation of corresponding component will be affected proportionately in such cases. Plagiarism detection software will be used to check plagiarism wherever applicable. Academic integrity is expected from students in all components of course assessment**.**

**\* - Only Teaching Method in the Instruction Plan for the course may vary for different faculty teaching the course, rest of the Course Policy Document will not change.**