Experiment No.1
Study various applications of NLP and Formulate the Problem Statement for Mini Project based on chosen real world NLP applications: Machine Translation, Text Categorization, Text summarization, Chat Bot, Plagiarism, Spelling & Grammar Checkers, Sentiment / Opinion analysis, Question answering, Personal Assistant, Tutoring Systems, etc.
Date of Performance:
Date of Submission:

Aim: Study various applications of NLP and Formulate the Problem Statement for Mini Project based on chosen real world NLP applications: Machine Translation, Text Categorization, Text summarization, Chat Bot, Plagiarism, Spelling & Grammar Checkers, Sentiment / Opinion analysis, Question answering, Personal Assistant, Tutoring Systems, etc.

Objective: Understand the different applications of NLP and their techniques by reading and critiquing IEEE/ACM/Springer papers.

Theory:

1. **Machine Translation**

Machine translation is a process of converting the text from one language to the other automatically without or minimal human intervention.

2. **Text Summarization**

Condensing a lengthy text into a manageable length while maintaining the essential informational components and the meaning of the content is known as summarization. Since manually summarising material requires a lot of time and is generally difficult, automating the process is becoming more and more popular, which is a major driving force behind academic research.

Text summarization has significant uses in a variety of NLP-related activities, including text classification, question answering, summarising legal texts, summarising news, and creating headlines. Additionally, these systems can incorporate the creation of summaries as a middle step, which aids in shortening the text.

The quantity of text data from many sources has multiplied in the big data era. This substantial body of writing is a priceless repository of data and expertise that must be skillfully condensed in order to be of any use. A thorough investigation of NLP for automatic text summarization has been necessitated by the increase in the availability of documents. Automatic text summarising is the process of creating a succinct, fluid summary without the assistance of a human while maintaining the original text's meaning.

3. Sentiment Analysis

Sentiment analysis, often known as opinion mining, is a technique used in natural language processing (NLP) to determine the emotional undertone of a document. This is a common method used by organisations to identify and group ideas regarding a certain good, service, or concept. Text is mined for sentiment and subjective information using data mining, machine learning, and artificial intelligence (AI).

Opinion mining can extract the subject, opinion holder, and polarity (or the degree of positivity and negative) from text in addition to identifying sentiment. Additionally, other scopes, including document, paragraph, sentence, and sub-sentence levels, can be used for sentiment analysis.

Businesses must comprehend people's emotions since consumers can now communicate their views and feelings more freely than ever before. Brands are able to listen carefully to their customers and customise their products and services to match their demands by automatically evaluating customer input, from survey replies to social media chats.

4. Information Retrieval

A software programme that deals with the organisation, storage, retrieval, and evaluation of information from document repositories, particularly textual information, is known as information retrieval (IR). The system helps users locate the data they need, but it does not clearly return the questions' answers. It provides information about the presence and placement of papers that may contain the necessary data. Relevant documents are those that meet the needs of the user. Only relevant documents will be pulled up by the ideal IR system.

5. Question Answering System (QAS)

Building systems that automatically respond to questions presented by humans in natural language is the focus of the computer science topic of question answering (QA), which falls under the umbrella of information retrieval and natural language processing (NLP).

Conclusion:

Comment on the Pros and Cons of each paper and also summarize your findings from the review of literature.

1. **Paper 1: Leveraging NLP Approaches to Define and Implement Text Relevance Hierarchy Framework for Business News Classification**
    - **Summary:** This paper proposes a five-level text relevance hierarchy framework to help companies prioritize and manage business news data. It uses NLP approaches like entity recognition, topic modeling, and similarity analysis to assess relevance based on various criteria, including topics, organizations, people, and locations. The framework helps companies focus on critical texts and identify new areas of interest.
    - **Pros:**
        - Provides a structured and systematic approach to classifying business news, making it easier to identify relevant texts.
        - Incorporates various NLP techniques, such as keyword matching and machine learning, to improve classification accuracy.
        - The hierarchical framework allows for easy updates and adaptations to changing business needs.
        - Offers practical applications for companies to better handle business news and align with market trends.
    - **Cons:**
        - The proposed model may require a comprehensive initial setup and continuous updating to maintain its relevance.
        - The implementation may be computationally intensive, particularly with large datasets.
        - Framework relies heavily on a predefined taxonomy, which could limit its adaptability to new or unforeseen business topics.
2. **Paper 2: Open versus Closed: A Comparative Empirical Assessment of Automated News Article Tagging Strategies**
    - Summary: This paper compares different machine learning approaches for automated news article tagging using both closed-ontology and open-ontology

models. The evaluation focuses on how well these models perform in accurately categorizing and tagging news articles, and it highlights the trade-offs between closed and open methods.

- Pros:
  - Provides a comprehensive comparison of closed and open approaches for news tagging, offering insights into their respective strengths and weaknesses.
  - Highlights how closed models ensure consistent tagging, while open models offer flexibility and adaptability to new content.
  - Uses state-of-the-art models like BERT, which demonstrate high performance in classification tasks.
- Cons:
  - Closed-ontology models might limit flexibility and fail to capture new or evolving topics.
  - Open-ontology models can generate more variability in tags, potentially leading to inconsistent tagging.
  - The evaluation focuses on specific datasets and might not generalize well to other types of news content.

3. Paper 3: Category Classification and Topic Discovery of Japanese and English News Articles
   - Summary: This paper presents algorithms for topic analysis, including category classification and topic discovery. The proposed approach is adaptable for different languages and can handle online training for both category and topic classification, meeting special requirements for news, such as dynamic classification and sparse training data.
   - Pros:
     - The proposed algorithms support online training, making it suitable for real-time news streams.
     - Can handle sparse training data, which is advantageous for scenarios where limited labeled data is available.

  - Adaptable to multiple languages, enhancing its versatility for international news classification.

- Cons:
  - The model may face challenges in scalability when applied to larger datasets.
  - Online training approaches may need frequent retraining, which could increase computational costs.
  - Limited focus on integrating sentiment analysis or other contextual factors that might enhance classification.