**TASK1**

**NATURAL LANGUAGE PROCESSING AND ITS APPLICATION**

Natural Language Processing (NLP) is a branch of artificial intelligence (AI) that deals with the interaction between computers and humans through natural language. It enables computers to understand, interpret, and generate human language in a way that is both meaningful and useful. NLP encompasses a variety of tasks such as text classification, sentiment analysis, machine translation, named entity recognition, and question answering.

NLP has a wide range of applications across various industries. In customer service, it can be used for chatbots and virtual assistants to provide instant support and improve user experience. In healthcare, NLP helps extract valuable insights from medical records, aiding in diagnosis and treatment recommendations. In finance, it facilitates sentiment analysis of news articles and social media data for making investment decisions. Additionally, NLP powers voice recognition systems like Siri and Google Assistant, enabling users to interact with devices through spoken language.

In the field of machine learning, NLP is crucial for tasks that involve textual data. With the exponential growth of unstructured data on the internet, NLP algorithms play a vital role in extracting valuable information from text documents. By leveraging techniques such as deep learning and natural language understanding, machine learning models can accurately process and analyze large volumes of text, leading to advancements in automated language translation, text summarization, and information retrieval.

Overall, NLP is important in machine learning because it bridges the gap between human language and machine understanding, enabling computers to comprehend and generate text data in a way that enhances human-computer interaction and facilitates decision-making processes in various domains.

**TASK2**

**What is tokenization?**

Tokenization is the process of replacing sensitive data with unique identification symbols that retain all the essential information about the data without compromising its security. Tokenization, which seeks to minimize the amount of sensitive data a business needs to keep on hand, has become a popular way for small and midsize businesses to bolster the security of credit card  transactions while minimizing the cost and complexity of with industry standards and government regulations.

**What is NLTK?**

NLTK refers to Natural language ToolKit, an open-source library for python, written by Steven Bird, Edward Loper, and Ewan Klein to use in development and education. Many tasks can be performed by NLTK, like tokenizing, parse tree, visualization, etc.

It has tons of algorithms from which we can choose any to perform a task. NLTK supports stemming, Part-Of-Speech (POS), entity recognition, etc.

**What is spaCy?**

spaCy is implemented in Cython and developed by Matt Honnibal. It is an open-source library for advanced Natural Processing Library(NLP) in python.

If you are working with lots of text and need to know what is the text about, which words belong to who, or what is the context in the text, then spaCy is quite helpful. It can also find similar text and search for any specific word in the text. Many more tasks can be performed by spaCy, like sentence detection, tokenization, lemmatization, etc.

**TASK3**

**What are Stop words?**

A stop words  is a commonly used word (such as “the”, “a”, “an”, or “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query.

The necessity of removing stopwords in NLP is contingent upon the specific task at hand. For [text classification](https://www.geeksforgeeks.org/rnn-for-text-classifications-in-nlp/) tasks, where the objective is to categorize text into distinct groups, excluding stopwords is common practice. This is done to channel more attention towards words that truly convey the essence of the text. As illustrated earlier, certain words like “there,” “book,” and “table” contribute significantly to the text’s meaning, unlike less informative words such as “is” and “on.”

Conversely, for tasks like machine translation and text summarization, the removal of stopwords is not recommended. In these scenarios, every word plays a pivotal role in preserving the original meaning of the content.

**TASK4**

**Stemming versus lemmatization**

Stemming and lemmatization function as one stage in text mining pipelines that convert raw text data into a structured format for machine processing. Both stemming and lemmatization strip affixes from inflected word forms, leaving only a root form.[4](https://www.ibm.com/topics/stemming-lemmatization#f04) These processes amount to removing characters from the beginning and end of word tokens. The resulting roots, or base words, are then passed along for further processing. Beyond this basic similarity, stemming and lemmatization have key differences in how they reduce different forms of a word to one common base form.

The practical distinction between stemming and lemmatization is that, where stemming merely removes common suffixes from the end of word tokens, lemmatization ensures the output word is an existing normalized form of the word (for example, lemma) that can be found in the dictionary.[7](https://www.ibm.com/topics/stemming-lemmatization#f07)