Name-Krish Shah
Branch-Computer Engineering

Synapse Task 3.2

Language Translators use a process of translating basic text paragraphs or verbal text notes to a simplified understandable language by maintaining the structure and meaning of the sentence.
The steps involved throughout the processing the involves:-

Data Processing /Cleaning :- This involves converting the sentences to words and cleaning the set where there is a frequent use of articles('the','a').

Natural Language Toolkit :- It is use widely used to analyze the structure of the words. NLTK has incorporated most of the tasks like tokenization, stemming, Lemmatization, Punctuation, Character Count, and Word count.

```
from nltk.corpus import stopwords
import nltk
```

Nltk stop words are widely used words (such as "the," "a," "an," or "in") that a search engine has been configured to disregard while indexing and retrieving entries.

BeautifulSoup:- It is a python package which allows us to pull data out of HTML and XML documents.

Open Vocabulary Text Mining:- Open-vocabulary approaches reveal more specific and concrete patterns across a broad range of content domains, better address ambiguous word senses, and are less prone to misinterpretation, suggesting that they are well-suited for capturing the nuances of everyday psychological processes.

Text summarization: NLP techniques are used to summarize long text documents into shorter versions, which is useful for tasks such as news summarization and document indexing.

Techniques for Word Translation:

Word Embeddings: Word Embeddings in NLP is a technique where individual words are represented as real-valued vectors in a lower-dimensional space and captures inter-word semantics. Each word is represented by a real-valued vector with tens or hundreds of dimensions.

Subword Tokenization: To handle out-of-vocabulary words, translation models often use subword tokenization. Words are broken down into smaller units (subwords) that the model has seen during training.