Project Title : Text Summarization

**Project Overview:**

The implementation of a text summarization tool using Spacy, a popular Natural Language Processing (NLP) library in Python. The objective of the project is to automatically generate a summary from a given text by identifying and extracting the most important sentences. The approach involves tokenizing the text, removing Stop Words and punctuation, calculating word frequencies, and scoring sentences based on these frequencies.

**Text Input and Pre processing:**

* **Text Input:** The input text is provided as a string and stored in a variable. The text can be any document that needs to be summarized.
* **NLP Model Loading:** The SpaCy library is used to load a pre-trained English language model (en\_core\_web\_sm), which processes the text to tokenize it into words and sentences.

**Stopwords and Punctual Removal:**

* **Stopwords:** Common words that do not carry significant meaning (e.g., "and," "the," "is") are removed using SpaCy's built-in list of stopwords.
* **Punctuation:** All punctuation marks are removed to ensure that only meaningful words are considered in the summarization process.

**Word Frequency Calculation:**

* **Word Frequencies:** The script calculates the frequency of each word in the text, excluding stopwords and punctuation. Words that appear more frequently are considered more important.
* **Normalization:** The word frequencies are normalized by dividing each word's frequency by the maximum frequency found in the text.

**Sentence Scoring:**

* **Sentence Tokens:** The text is split into sentences, which are then scored based on the frequency of the words they contain.
* **Scoring Mechanism:** Each sentence is assigned a score based on the sum of the frequencies of the words it contains. Higher scores indicate more important sentences.

**Selection of Top Sentences:**

* **Top Sentences:** The script selects the top 30% of sentences with the highest scores. This percentage is chosen to create a concise yet informative summary.
* **Sentence Extraction:** The selected sentences are extracted and combined to form the final summary.

**Summary Generation:**

* **Final Summary:** The extracted sentences are concatenated to create a summary of the input text.
* **Length Comparison:** The length of the original text and the summary are compared to assess the effectiveness of the summarization.

**References:**

**DataCamp courses:**

1. Introduction to natural language preprocessing in python
2. Natural language preprocessing with spaCy