Introduction:

The project focuses on creating an automatic text summarization tool using Python. Text summarization is the process of condensing a longer piece of text into a shorter version while retaining its key information. The project employs techniques such as text cleaning, sentence tokenization, word tokenization, and word-frequency analysis to generate a concise summary of the input text.

Project Steps:

1. Text Cleaning:

The project starts by cleaning the input text to remove unnecessary characters and elements. It utilizes the spaCy library to process the text and prepare it for further analysis. Stop words and punctuation are removed to isolate meaningful content.

2. Word Tokenization:

The text is then tokenized into individual words to create a list of tokens. This step is crucial for analyzing word frequencies and building the word-frequency table.

3. Word-Frequency Analysis:

A word-frequency table is constructed to capture the frequency of each unique word in the text. The project calculates the frequency of each word while excluding stop words and punctuation. This table forms the basis for evaluating the significance of words in the text.

4. Sentence Tokenization:

The input text is split into individual sentences using sentence tokenization. This step is essential to analyze the importance of each sentence in the context of the entire text.

5. Scoring Sentences:

Sentences are scored based on the frequency of significant words they contain. Each sentence's score is calculated by summing the frequencies of the words it contains. The higher the score, the more relevant the sentence is to the overall content.

6. Summarization:

To generate the final summary, the project uses a technique called "extractive summarization." It involves selecting the most important sentences based on their scores. A fraction of the total sentences (e.g., 30%) is chosen to create the summary.

Conclusion:

The automatic text summarization project demonstrates how to leverage Python and the spaCy library to create an efficient summarization tool. By performing text cleaning, word and sentence tokenization, word-frequency analysis, and scoring sentences, the project successfully generates a concise summary that captures the essence of the input text. This summarization approach can be valuable for quickly extracting key information from lengthy documents or articles.