# Internship Report: Text Summarization Tool

### Introduction

This project focuses on the development of a text summarization tool designed to generate concise summaries by extracting and combining the most relevant sentences from longer documents. The tool utilizes advanced natural language processing (NLP) techniques, including TF-IDF and logistic regression, to analyze, rank, and summarize text effectively. A user-friendly interface was implemented using Streamlit to enable users to interact with the summarization system.

### Background

Text summarization plays a critical role in information processing and retrieval by distilling large amounts of text into easily digestible summaries. This project combines traditional NLP methods, such as TF-IDF vectorization, with machine learning algorithms to predict sentence importance. By integrating these techniques, the tool provides dynamic summarization tailored to varying levels of granularity (Low, Medium, High).

### Learning Objectives

1. Develop a robust text summarization tool capable of handling diverse document inputs.
2. Gain proficiency in TF-IDF vectorization and logistic regression for NLP applications.
3. Understand and implement dynamic text summarization with user-configurable levels.
4. Create an interactive graphical user interface using Streamlit.

### Activities and Tasks

1. Preprocessing Pipeline: Developed a preprocessing function that removes punctuation, converts text to lowercase, and eliminates stopwords to prepare text for vectorization.
   - Libraries used: NLTK for tokenization and stopword removal.

2. TF-IDF Vectorization: Employed TF-IDF to convert sentences into numerical vectors, enabling machine learning models to evaluate their importance.

- TF-IDF captures the significance of words relative to the entire document.

3. Logistic Regression Model:

- Trained a logistic regression model to classify sentences based on importance.
- Implemented `predict_proba` to rank sentences dynamically.

4. Dynamic Summarization Levels:

- Implemented three summarization levels (Low, Medium, High) to adjust the granularity of generated summaries.
- Defined thresholds (10%, 30%, 50% of sentences) based on user selection.

5. Streamlit Interface:

- Built an intuitive interface allowing users to input text, select summarization levels, and view generated summaries.
- Included error handling for empty inputs and clear instructions for use.

### Challenges and Solutions

1. Challenge: Ensuring accurate preprocessing of diverse text inputs.

- Solution: Utilized NLTK for robust tokenization and stopword removal.

2. Challenge: Balancing summary conciseness and informativeness.

- Solution: Provided dynamic summarization levels and used logistic regression to rank sentence importance effectively.

3. Challenge: Efficient integration of the ML pipeline with the user interface.

- Solution: Streamlits lightweight framework allowed seamless integration of the model and preprocessing pipeline.

### Outcomes and Impact

1. Delivered a functional text summarization tool with an interactive GUI.
2. Achieved effective summarization through sentence ranking and dynamic levels of granularity.
3. Enhanced proficiency in NLP techniques and machine learning applications for text analysis.
4. Demonstrated the potential of combining traditional NLP with ML for extractive summarization.

### Conclusion

This project showcases the effective application of TF-IDF and logistic regression in text summarization. The integration of a user-friendly interface with a robust summarization backend highlights the importance of combining technical accuracy with usability. This work lays the foundation for further enhancements, including support for multilingual summarization and integration with external data sources.

### Skills and Competencies

1. Proficiency in natural language processing and text preprocessing.
2. Expertise in implementing TF-IDF and machine learning for text analysis.
3. Advanced Python programming for data-driven applications.
4. Development of interactive applications using Streamlit.
5. Application of logistic regression to ranking and classification tasks.