MpToVideo Documentation

Necessary Resources

1. **Python3. Any version of Python 3.8 or above will work**
2. **Pandas. Necessary for data parsing**
3. **Sci-kit-learn. Required for creating TF-IDF vectorizer model.**

Documentation Overview:

Libraries: The script uses pandas for data manipulation, sklearn's TfidfVectorizer for text vectorization, and cosine_similarity for measuring the similarity between the query and lecture transcripts.

preprocess_text Function: This function is responsible for cleaning and preparing the text data. It currently converts text to lowercase. Additional preprocessing steps like removing punctuation or stopwords can be added here.

Data Loading and Preprocessing: The CSV file containing lecture transcripts is loaded into a DataFrame. The transcripts are then cleaned using the preprocess_text function.

TF-IDF Vectorization: The TfidfVectorizer is used to convert the text data into a numerical format (TF-IDF vectors) that can be used for similarity comparisons.

find_relevant_lectures Function: This function takes a query and returns the top N relevant lectures based on the query's cosine similarity with the lecture transcripts. It incorporates a relevance threshold to filter out less relevant results.

Example Usage: Demonstrates how to use the find_relevant_lectures function with a sample query.

***** For detailed usage steps, please refer the source code as documentation headers are provided for each function *****