Dataset generator & Chatbot assistant for Instruction Manuals

## Objective:

In our day to day lives, we spend a considerable amount of time on reading instruction manuals for various purposes. The main **objective** of this model is to reduce man hours invested on reading those instructions for instruments, appliances, automobiles etc.

## Abstract:

A chatbot is a software application that’s used to engage in human conversation in a natural way. Chatbots are commonly used across many different industries for many different purposes like web searches, clearing queries etc. The project proceeds through three major stages of text processing.

1. Dataset generation from corpus.
2. Generating text summaries.
3. Building the chatbot based on data from stage 1 and 2.

The manual is fed into the dataset block for preparing a dataset for both the text summarizer and chatbot. The dataset is to be developed based on the template provided by huggingface.co [1]. Then the native summarizer built using LexRank algorithm [2], gives text summaries as input to the chatbot. The chatbot works on a simple semantic matching technique [3] and feeds on the data from the dataset generator and the text summaries to give out instructions in forms of either single sentence or text summaries of instructions.

## Key Features:

1. Inbuilt dataset generator exempts the need for pre-built datasets.
2. Text summaries help to give out detailed answers for queries.
3. Provides only required instructions in an interactive environment.

## Timeline:

The timespan proposed for the project is 8 weeks during which we will be:

# Developing dataset generator (2 weeks):

We will be developing the foresaid dataset generator and will be testing and training it with multiple corpus for evaluation.

# Working on text summarizer (2 weeks):

After finishing up with the dataset generator we will be focusing on the development of text summarizer and testing it with the datasets from the dataset generator. The text summarizer will also be tested with other documents for evaluating its accuracy.

# Building chatbot (3 weeks):

Based on the said semantic matching model a chatbot is to be developed and tested to give out single line answers to text summaries based on the user queries. The chatbot is tested and evaluated based on multiple criteria.

# Tweaking the whole model (1 weeks):

After the successful development of the model, it might still need to be perfected based on training data and test results. Multiple corpus is used to evaluate the model and necessary changes are made if found any.

## Evaluation criteria:

The model is primarily evaluated on the following criteria:

1. Chatbot response volume.
2. Conversation length.
3. Comprehension capabilities.

## Conclusion:

The motive of the project is to develop a user-friendly tool to retrieve the required information from a vast data source or corpus. The proposed model does this seamlessly and reduces the manual work involved.

## Citations:

[1] [Writing a dataset loading script — nlp 0.4.0 documentation (huggingface.co)](https://huggingface.co/docs/datasets/v0.3.0/add_dataset.html)

[2] [LexRank method for Text Summarization (opengenus.org)](https://iq.opengenus.org/lexrank-text-summarization/)

[3] [A simple chatbot using Python and NLTK | by Iris Jestin | Analytics Vidhya | Medium](https://medium.com/analytics-vidhya/a-simple-chatbot-using-python-and-nltk-c413b40e9441)