My motivation is to create a more efficient software that allows the user to:
- Identify the key-points of a text
- Avoid paraphrasing information
- Capture the author’s voice

Text data can be summarized using various algorithms, but no standard software exists to highlight key information. A Neural Network, a type of machine learning algorithm, can be trained to identify important parts of text for highlighting.

The TREC (Text Retrieval Conference) collection is a widely used, publicly accessible collection of text documents for information retrieval research. I will use Disks 4 and 5 to train the Neural Network and generate the Chat GPT summaries.

My project uses a variety of algorithms that start by defining four unique features. Each feature requires separate training on a set of flat documents. The process starts by defining the individual features, pre-processing the text using Gensim, and mapping words with Word2vec to create the final model.

After developing the feature models, a Neural Network is constructed that takes in four inputs, the previously trained models. The Network has five hidden layers and uses a Sigmoid function to normalize and determine the most important sentences. This non-linear approach enables the Neural Network to learn complex patterns in the data. The resulting output is analyzed.

The results demonstrate the efficacy of our approach in discerning the salient aspects of a wide range of texts, empowering readers to grasp the core message with ease.

- Improving the feature extraction process
- Improving the summarization process
- Evaluating the performance
- Integrating the software into a larger system
- Adding more features as an input