Text Summarization

Automatic text summarization is the process of shortening a set of data computationally, to create a subset (a summary) that represents the most important or relevant information within the original content. Automatic summarization as a field is not limited to text. In fact, we can 'summarize' images and videos as well as text. We have focused on the most common type of automatic summarization: automatic *text* summarization. We will be making an NLP application which will summarise the text. We have two main for automatic text summarization:

- Extractive summarization
- Abstractive summarization

Extractive summarization

Extractive summarization algorithms perform a seemingly very simple task: they take in the original text document and extract parts of it that they deem important. This means that they do not create new data (new sentences). Instead, these models simply select parts of the original data which are most important and combine them to form a summary.
This is in contrast to how most humans summarize text. Instead of simply copying over the most important sentences, a well written summary authored by a human will contain new sentences which include just the most important points in the original text.
For achieving extractive summarization we have used the Text rank algorithm.

Text Rank Algorithm : Text Rank chooses important sentences by "ranking" all sentences in the text. After sentences are ranked, the top n ranked sentences are used to create a summary.

The process of ranking sentences is fairly straightforward. We start by creating a graph in which every node represents a sentence from the original text we want to summarise. We then link each sentence in this graph to other similar sentences. These links are the edges of the resulting graph. In this graph, each sentence will point to other sentences that hold similar information. The resulting edges of the graph are weighted. We then run a complex graph-based ranking formula over this weighted graph to determine the most important sentences in the original text and create the final summary.

Abstractive summarization

Instead of just rewriting parts of the original text document, abstractive summarization methods mimic humans by creating completely new sentences to describe key concepts from the original text document. These new sentences can often use new words, not present in the original text, and aim to contain just the core information, with everything unimportant removed.

Abstractive summarization techniques have a more human-like approach to text summarization, they primarily rely on deep learning models. While initially these models used RNN (recurrent neural networks) based architectures, as of recent, the models that have taken over the world of natural language processing are called Transformers.

Transformers can, given some input text, generate completely new text. In the case of abstractive summarization, transformers take the original text as the input and generate the summary text as the output.

In our application we have used Hugging Face offers models based on Transformers for PyTorch and TensorFlow 2.0. Here, there are thousands of pre-trained models to perform tasks such as text classification, extraction, question answering, and more.