Automatic summarization

Automatic summarization: Automatic 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.

Types of text summarization: There are 2 types of text summarization,

1. Extractive Summarization: content is extracted from the original data, but the extracted content is not modified in any way.
2. Abstractive summarization: Abstractive methods build an internal semantic representation of the original content, and then use this representation to create a summary that is closer to what a human might express.

Extractive Summarization: In my sample project, I focused in extractive summarization. Where the summary is a set of important sentences from original document, which might represent the whole document briefly.

Solving Approach: The approach I follow has 3 steps,

Step 1, First we tokenize the sentences from given document. Then we need the numerical representation of the sentences. For that I use BERT from transfer learning. Basically I used hugging's pre-trained bert-base-uncased model [1]. It will give us encoded representation of the sentences.

Step 2, when I got the encoded representation of the document. I can feed my data to any machine learning model. Here, I used K-Means Clustering to cluster similar sentences. But here is a problem, for K-Means Clustering I need the value of K.

How could I find the appropriate value for K?

For that, I use elbow and silhouette method. It will give me optimize K value for K-Means Clustering.

Step 3, when I got the cluster the next procedure is to extract the most important sentences for the summary.

How did I decide that which sentences to extract?

Here I extract one sentence from the cluster as they all represent the similar context.

But, which one?

Sentence, which is closest to the centroid of the clusters.

To improve readability, I sorted the sentences and marge them. Finally we got our expected summary. Below, there is a diagram to represent the whole approach.

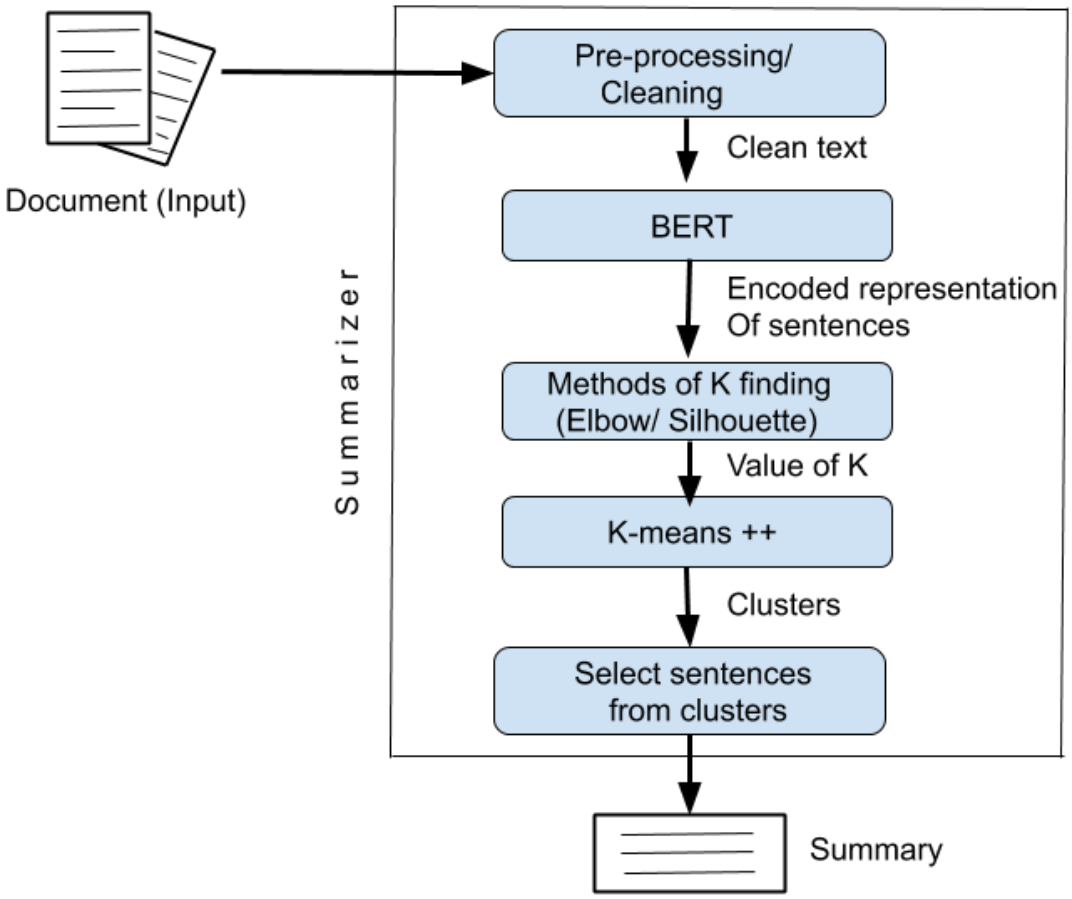

Fig: The flowchart of Solving Approach.

Conclusion: This a simple implementation of extractive text summarization. There is a lot of ways to improve it. To evaluate generated summary we can use ROUGE matrix. Which can tell us how good or bad the generated summary is.

Project Link: https://github.com/nowshad7/SampleProject_Infolytx

References:

1. https://huggingface.co/bert-base-uncased