

Text summarization is a very useful and important part of Natural Language Processing (NLP). First let us talk about what text summarization is. Suppose we have too many lines of text data in any form, such as from articles or magazines or on social media. We have time scarcity so we want only a nutshell report of that text. We can summarize our text in a few lines by removing unimportant text and converting the same text into smaller semantic text form.

Now let us see how we can implement NLP in our programming. We will take a look at all the approaches later, but here we will classify approaches of NLP.

TEXT SUMMARIZATION

In this approach we build algorithms or programs which will reduce the text size and create a summary of our text data. This is called automatic text summarization in machine learning.

Text summarization is the process of creating shorter text without removing the semantic structure of text.

There are two approaches to text summarization.

1. Extractive approaches
2. Abstractive approaches

EXTRACTIVE APPROACHES:

Using an extractive approach we summarize our text on the basis of simple and traditional algorithms. For example, when we want to summarize our text on the basis of the frequency method, we store all the important words and frequency of all those words in the dictionary. On the basis of high frequency words, we store the sentences containing that word in our final summary. This means the words which are in our summary confirm that they are part of the given text.

ABSTRACTIVE APPROACHES:

An abstractive approach is more advanced. On the basis of time requirements we exchange some sentences for smaller sentences with the same semantic approaches of our text data.

EXTRACTIVE APPROACHES:

We will take a look at a few machine learning models below.

TEXT SUMMARIZATION USING THE FREQUENCY METHOD

In this method we find the frequency of all the words in our text data and store the text data and its frequency in a dictionary. After that, we tokenize our text data. The sentences which contain more high frequency words will be kept in our final summary data.