EXTRACTION BASED TEXT SUMMARIZATION

By - Sayan Das

JULY, 2023

Contents

1. Introduction

In today's information-driven world, the rapid growth of digital content has led to an overwhelming flood of data. Accessing and comprehending the vast volumes of information available poses a significant challenge, requiring efficient methods to distill key insights. Text summarization, a crucial branch of natural language processing (NLP), offers a compelling solution by extracting essential information from extensive textual data.

This project focuses on extractive text summarization, a technique that aims to condense the original text by selecting and assembling the most critical sentences or phrases that convey the core meaning. Unlike abstractive summarization, which generates new sentences based on the understanding of the text, extractive summarization directly leverages existing content.

The primary objective of extractive summarization is to retain the original intent and context of the text while significantly reducing its length. By identifying salient sentences or key phrases, extractive summarization enables users to grasp the main points of a document quickly, making it an invaluable tool for information retrieval, document skimming, and content organization.

This project leverages the power of data-driven techniques to build an extractive summarization system capable of handling a wide range of text sources, including news articles, research papers, online forums, and social media posts. By harnessing the vast amounts of text available online and the advancements in NLP, we aim to develop an accurate, efficient, and scalable solution that delivers concise summaries with minimal loss of information.

The proposed extractive summarization system will rely on Natural Language Processing (NLP). The NLP models will learn to identify sentence importance to extract the most significant content from a given text.

In summary, this project seeks to harness the power of data-driven extraction techniques to develop a extractive text summarization system. Through the exploration of NLP methodologies, we aim to unlock the essential insights hidden within lengthy documents, empowering users with the ability to distill complex information into concise summaries.

2. Types of Text Summarization

Text summarization is a process of extracting important information from a long sequence of text by reducing the length of the sequence but without altering the meaning of the text sequence. Text summarization can be broadly categorized into two main types: extractive summarization and abstractive summarization.

2.1. Extractive Summarization:

Extractive summarization involves selecting and extracting the most important sentences or phrases from the original text to create a summary. The extracted sentences are typically chosen based on various criteria such as relevance, importance, and coherence within the document. This approach relies on the assumption that the most critical information is

already present in the original text and can be effectively conveyed through the selected sentences. Extractive summarization methods often employ statistical or machine learning techniques to determine sentence importance, such as using algorithms that analyze word frequency, sentence position, or semantic similarity.

2.2 Abstractive Summarization:

Abstractive summarization, on the other hand, involves generating a summary that may include new sentences not present in the original text. Instead of selecting sentences from the source material, abstractive summarization aims to understand the content and generate concise, coherent, and human-like summaries. This technique requires a deeper level of natural language understanding and the ability to paraphrase and rephrase information. Abstractive summarization often relies on advanced machine learning models such as recurrent neural networks (RNNs), transformer models, or deep learning architectures that can capture the semantic and contextual nuances of the text.

Both extractive and abstractive summarization have their advantages and limitations. Extractive summarization tends to produce summaries that are more faithful to the original text since they directly rely on existing sentences. However, they may lack fluency and coherence, as the selected sentences are stitched together without the ability to generate new phrasing. Abstractive summarization, while capable of producing more fluent and concise summaries, often faces challenges in maintaining factual accuracy and may introduce unintended biases or errors.

3. Main Principle of Text Extraction in this Project

In this section, we are going to discuss what methodology we have followed in this project to build up a extraction based text summaries of a sequence of text.

We are simply going to find the sentences or phrases in the given text that are of very importance. Once we have found that we take those sentences into our summary according to their importance. There are ------ main steps of these idea of extracting the contents – explained in the following –

1. We are going to find the unique words and their corresponding frequencies from the text passage.
2. In the next step we are going to score each and every sentences by the sum of the frequencies of the word that are present in the sentences. It is clearly understandable that the sentences having more frequent words, will have higher scores. In this way we will easily sort out the sentences according to how much frequent words are present in the sentences.
3. Next, we find out the number of words in the original text say that is N and then we fix a proportion of N (say Np ($0<p<1$)) as the number of words, of the summary text.
4. Next, we will be choosing the sentences with the higher sentence score, calculated previously to be included in the summary of the text, provided that the no. of words in the summary after adding a new sentence in the summary is either less that Np or have just crossed the value of Np.

5. Then we will add up the sentences selected for the summary as a simple text in the order they have appeared in the original text.

4. Pre-processing

Pre-processing is a crucial stage of any data-driven task. The real-life data are present in the form that is not very useful to extract information. For this reason, to enhance the power of information that resides in a data, the data need to be prepared. Here our data is a single sequence of text. We have gone through the following pre-processing stages with our data.

4.1. Converting to Lower Cases

In the initial stage of pre-processing, we have converted all the words that are present in the text sequence, into the lower cases. This technique will help us to reduce duplication in the words. Since according to the methodology proposed in section-3 we will be obtaining the word frequencies, lowercases will help use avoid tallying any word, say for example the word “cat”, for several instances. (For this word the instances can be “cat” and “Cat”).

4.2. Removing the stop words

Stop words are those words in any language that can present in a sentence but provide no meaningful information. Like the words “The”, “in”, “at” etc. These words are distraction to other meaningful words that are present in the sentence. During the pre-processing stages we have removed a large list of stop words.

4.3 Lemmatisation

In language a single word have so many variations over the language. For example the word “go” can be present in the sentence in the form of “go”, “going”, “went”, “gone” etc. All those variation directs us to understand the single word “go”, which is called the root word. Lemmatisation is actually the converting a word to its root/base form. After applying the lemmatisation on the text, all the instances like “go”, “going”, “went”, “gone” will increase the frequency of “go”.

5. Finding the sentence scores

In the next stage of pre-processing we have takes steps of finding the score for each sentence in the given text, which score will represent the importance of the sentence in the text sequence. To do this we have followed two steps:

1. The first step is to create a word frequency dictionary. We considered all the unique words that are present in a text after pre-processing (that means only the words of its basic form, and no stop words). By traversing through these words and counting their number of occurrences in the text we have obtained the frequencies of the words. At the end of these we have a list of words (say W) along with their frequencies.
2. Next, we have traversed through each sentences and for each sentences we Have obtained a score by adding the frequencies of all the word that are present in the sentence. Say, there are n words W_1, W_2, … , W_n with corresponding frequencies f_1, f_2, … , f_n in a sentence then the score corresponding to the sentence will be

$$score = \sum_{i=1}^{n} f_i$$

Note: Here for computational advantages we have take the normalised frequency instead of actual frequency ie. Here,

$$f_i = \frac{frequency\ of\ the\ word}{frequency\ of\ the\ most\ frequenct\ word}$$

Here f_i will be lying between 0 to 1. At the end of this stage we will have the list of sentences along with their scores say S_i (score for the i^{th} sentence)

6. Deciding which sentence to include in the summary

In the next step, we have to decide 1. which sentences we need to include into the summary and 2. how much do we have to include in the summary.

To answer the first question, we would say that we have the scores S's that decides the importance of the sentence. We will only include the sentences with the higher score values. Then comes the question how much do we includes in the summary. For these we have to follow some thresh-hold.

We have counted the total number of words in the original text. Say, the text has N many word. We have decided to keep about a 30% of words in the summary. Which means the number of words in the summary should be around 0.3xN. So according to this thresh-hold we would keep those sentences that has the top scores and the number of words in the whole summary text due to inclusion of those sentences (Here when it comes to the inclusion of sentence in the summary we have to take the original form of the sentence that is given in the text not the pre-processed sentences, pre-processed sentences are only formed calculate the sentence score) have just crossed the value 0.3xN.

7. Forming the summary out of these selected sentences

In the next step we are going to join the sentences side-by-side to form the summary. But while doing this we have keep certain things in mind.

1. Sentences that are joined in the summary must be in their original form, in which they are present in the text and not the form that is created after the pre-processing. Otherwise there will be no meaning of those sentences. Those pre-processed sentences are good for the calculation score but not good for human to understand it.
2. Those sentences must in the order they appeared in the original text. I may happen that some sentence that are in the later part of the text have more importance than another sentence in the initial part of the text. Now, when we choose sentence according to importance then we rank the sentences according to their importance score. In this ranking some sentence appearing in the later part of the text may come first than a sentence from the first part of the text. If both the sentences are selected

for the summary then they must follow their original order of appearance in the text. Otherwise, the meaning of the sentence could alter.

8. Some Examples

Here we have demonstrated some examples of text on which we have applied the text-summarization algorithm, Here are the examples –

8.1 Example 1:

Original Text: (Word Count: 218)

The evolution of the horse is a classic example of natural selection in action. Over millions of years, horses evolved from small, four-toed creatures to the large, powerful animals we know today. This evolution was driven by the need for horses to adapt to their changing environment.

Early horses lived in forests, where they were preyed upon by large predators. To survive, these horses needed to be able to run fast. They also needed to be able to see well in low light conditions. Over time, horses evolved longer legs and necks, which helped them to run faster and see better.

As the climate changed and forests gave way to grasslands, horses needed to adapt again. Grasslands are more open than forests, so horses needed to be able to see predators from a distance. They also needed to be able to run long distances to find food and water. Over time, horses evolved larger bodies and faster running speeds.

The evolution of the horse is a testament to the power of natural selection. Through a process of trial and error, horses evolved into the perfect animals for their environment.

Summary: (Word Count: 71)

This evolution was driven by the need for horses to adapt to their changing environment. Over time, horses evolved longer legs and necks, which helped them to run faster and see better. Grasslands are more open than forests, so horses needed to be able to see predators from a distance. Over time, horses evolved larger bodies and faster running speeds8.2

8.2 Example 2:

Original Text: (Word Count: 193)

The Great Wall of China is one of the most impressive feats of engineering in human history. It stretches for over 13,000 miles, making it the longest man-made structure in the world. The wall was built over centuries by different dynasties, and it served as a defense against invaders from the north.

The Great Wall is not a single, continuous structure. It is made up of a series of walls and fortifications that were built over time. The walls vary in height and width, and they are made from different materials, such as stone, brick, and earth. The most well-known section of the Great Wall is the Badaling section, which is located near Beijing.

The Great Wall is not just a physical barrier. It is also a cultural icon that is synonymous with China. The wall has been featured in many films and television shows, and it is a popular tourist destination. The Great Wall is a reminder of China's rich history and its unique culture.

Summary: (Word Count: 79)

The Great Wall of China is one of the most impressive feats of engineering in human history. The wall was built over centuries by different dynasties, and it served as a defense against invaders from the north. The most well-known section of the Great Wall is the Badaling section, which is located near Beijing. The Great Wall is a reminder of China's rich history and its unique culture.

8.3 Example 3:

Original Text: (Word Count: 166)

The human brain is the most complex organ in the human body. It is responsible for everything we do, from thinking and feeling to moving and breathing. The brain is made up of billions of neurons, which are interconnected by trillions of synapses. These neurons communicate with each other using electrical and chemical signals.

The brain is divided into two halves, called the left and right hemispheres. The left hemisphere is responsible for language, logic, and analysis. The right hemisphere is responsible for creativity, intuition, and spatial awareness. However, the two hemispheres work together to perform most tasks.

The brain is constantly changing and adapting. It can learn new things and form new memories. The brain is also able to repair itself to some extent. However, the brain is also susceptible to damage, which can lead to a variety of neurological disorders.

Summary: (Word Count: 62)

The human brain is the most complex organ in the human body. The brain is divided into two halves, called the left and right hemispheres. The right hemisphere is responsible for creativity, intuition, and spatial awareness. However, the brain is also susceptible to damage, which can lead to a variety of neurological disorders.

9. Conclusion

In conclusion, extractive text summarization provides a powerful solution for distilling essential information from vast amounts of textual data. By selecting and assembling the most critical sentences or phrases, extractive summarization enables users to quickly grasp the main points of a document. It retains the original context and intent of the text while significantly reducing its length, making it invaluable for various applications such as information retrieval, content organization, and document skimming.

Through the utilization of data-driven techniques and advancements in Natural Language Processing (NLP), this project aims to develop an extractive summarization system. By harnessing the power of data-driven extraction and pre-processing techniques, this project aspires to unlock the essential insights hidden within lengthy documents. The combination of word frequency analysis, scoring sentences based on importance, and incorporating sentence ordering from the original text will contribute to the generation of informative and coherent summaries. Ultimately, this project seeks to empower users with the ability to efficiently navigate through the overwhelming abundance of textual data and extract the crucial knowledge encapsulated within.

We have also implemented the text summarizer in a web application to help the user to get a glimpse of the text summarization.

10. References

GitHub Link to the project - https://github.com/sayandas1302/Extractive-Text-Summarizer