**Table of Contents**

1. Introduction………………………………………………….………………..1
2. Need for Text Summarization………………………………….………….….1
3. Main approaches to automatic summarization…………………….………….1
   1. Extraction-based summarization………………………………..……..1
   2. Abstraction-based summarization……………………………..………2
4. Methodology………………………………………………………….……….3
5. Results and Analysis…………………………………………………………..4
6. **Introduction**

Text Summarization is one of those applications of Natural Language Processing (NLP) which is bound to have a huge impact on our lives. With growing digital media and ever-growing publishing, no one has the time to go through entire articles/documents/ books to decide whether they are useful or not. Automatic Text Summarization is one of the most challenging and interesting problems in the field of NLP. Text summarization refers to the technique of shortening long pieces of text from multiple text resources such as books, news articles, blog posts, research papers, emails, and tweets. It deals with extracting summaries from huge chunks of texts. The intention is to create a coherent and fluent summary having only the main points outlined in the document. The demand for automatic text summarization systems is spiking these days thanks to the availability of large amounts of textual data.

There are two main types of techniques used for text summarization: NLP-based techniques and deep learning-based techniques. In this project, I have used a simple NLP-based technique for text summarization. I haven't used any machine learning library in this project. Rather I will simply use Python's NLTK library for summarizing texts.

1. **Need for text summarization**

Propelled by the modern technological innovations, data is to this century what oil was to the previous one. Today, our world is parachuted by the gathering and dissemination of huge amounts of data. In fact, the International Data Corporation (IDC) projects that the total amount of digital data circulating annually around the world would sprout from 4.4 zettabytes in 2013 to hit 180 zettabytes in 2025. That’s a lot of data.

With such a big amount of data circulating in the digital space, there is a need to develop some text summarization algorithms that can automatically shorten longer texts and deliver accurate summaries that can fluently pass the intended messages. Furthermore, applying text summarization reduces reading time, accelerates the process of researching for information, and increases the amount of information that can fit in an area.

1. **Main approaches to automatic summarization**

Text summarization in NLP can broadly be divided into two categories (Figure 1): abstraction-and extraction-based [1], [2].

* 1. **Extraction-based summarization**

The extractive text summarization technique involves pulling keyphrases from the source document and combining them to make a summary. The extraction is made according to the defined metric without making any changes to the texts. These methods rely on extracting several parts, such as phrases and sentences, from a piece of text and stack them together to create a summary. Therefore, identifying the right sentences for summarization is of utmost importance in an extractive method.

Here is an example:

*Source text:* Joseph and Mary rode on a donkey to attend the annual event in Jerusalem. In the city, Mary gave birth to a child named Jesus.

*Extractive summary:* **Joseph** and **Mary** attend event **Jerusalem**. **Mary** **birth** **Jesus**.

As you can see above, the words in bold have been extracted and joined to create a summary,  although sometimes the summary can be grammatically strange.

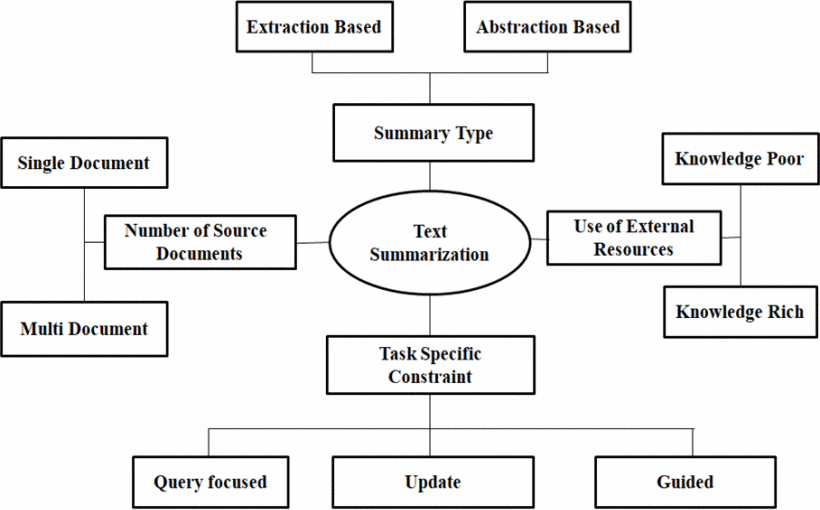


Figure 1. Types of text summarization

* 1. **Abstraction-based summarization**

The abstraction technique entails paraphrasing and shortening parts of the source document. When abstraction is applied for text summarization in deep learning problems, it can overcome the grammar inconsistencies of the extractive method. These methods use advanced NLP techniques to generate an entirely new summary. Some parts of this summary may not even appear in the original text. The abstractive text summarization algorithms create new phrases and sentences that relay the most useful information from the original text  just like humans do.

Here is an example:

*Abstractive summary*: Joseph and Mary came to Jerusalem where Jesus was born.

Therefore, abstraction performs better than extraction. However, the text summarization algorithms required to do abstraction are more difficult to develop; that’s why the use of extraction is still popular.

Summarization is also classified as single- or multi-document considering its source [3], [4]. In multi-document, the text overlaps on various texts and makes the task difficult. Summarizations depending on the resources can be knowledge-rich and poor where the first uses external corpus, e.g., Wikipedia, and wordnet. depending on the task of specific constraints, summarizations are of query-focused or query-oriented, updated, and guided. In query-focused, a summary is constructed with the information pertained to the query. The updated summary is used to identify new pieces of information from the document.

1. **Methodology**

In recent years, researchers have developed various types of extractive summarization techniques to generate a concise summary. The most common extractive text summarizing approaches are as follows:

1. Word frequency
2. Cluster
3. Graph theoretic
4. Neural networks
5. Machine learning, and
6. Fuzzy logic

In this project, I have used Frequency Based Summarization. In frequency summarizer, we initialize the values of min\_cut and max\_cut. The words that have the frequency lower than min\_cut or upper than max\_cut will be ignored. Then, we list the stop words which are not considered. After tokenization, each word in a sentence counts its frequency when not in the list of stop words. Then, we find the value of the max frequency of all words. After that, we find the importance of each word by Xi/y, where y indicates the max frequency and x the entire appearance of ith word. Then, crucial of a sentence is found by the summation of the value of important word belonging to the respective sentences. Finally, after sorting, the first k sentences are chosen [5].

Pre-processing Step:

1. *Removing Stopwords*: Stopwords like ‘the’, ‘a’, ‘in’ are ignored.
2. *Tokenization*: It is the way of splitting a flow of text into input characters, called tokens, which are fed to a form of processing. We have used two types of tokenization: sentence and word where the first is the way of extracting all the sentences from the text in which every word is extracted from every sentence.

Algorithm 1. Frequency Summarizer (input text T):

1. Tokenize sentences in T and saved to S.
2. Remove stop words from sentences
3. For each sentence in S
4. Count frequency of each word *W* in a sentence
5. For *Wi*
6. M = Maximum(freq [*Wi*])
7. freq[*Wi*]=freq[*Wi*]/M
8. If(fieq[*Wi*]>=max\_cut or freq[*Wi*]<=min\_cut)
9. Ignore Word *Wi*
10. For each *ith* sentence in S
11. For each word in S
12. If *Wi* in freq
13. rank[i] += freq [*Wi*]
14. TopK sentences are selected for the final summary

In line 8, two terms *max\_cut* and *min\_cut* are introduced. These variables contain two values which can be chosen using the observation between the ranges where the summarizer will produce a decent result. According to our observation, we use mincut = 0.1 and max cut = 0.9. Applying this approach, a single document can be summarized within a very short time. As there is a short task of mapping, counting and prioritizing the terms, the execution time is much less. However, there is a possibility to lose some information due to the range of max cut and min cut. Working with some random training data, it is seen that it returns a decent summary where redundant lines are easily eliminated in summary. The same approach can also be applied for multi-document summarizations.

1. **Results and Analysis**

To analyze and experiment the summarizer, I have applied the algorithm on the summary of the movie “Avengers: Endgame  (2019)” and got the desired respective output. Here, I only show a single input and output as follows (Figure 2):

***Input:***

After Thanos activates the completed Infinity Gauntlet, Clint Barton's family disintegrates, along with half of all life across the universe. Nebula and Tony Stark are stranded in space following their defeat by Thanos on Titan, but are returned to Earth by Carol Danvers and reunited with Natasha Romanoff, Bruce Banner, Steve Rogers, Rocket, Thor, and James Rhodes. The team formulates a plan to steal the Infinity Stones back from Thanos and use them to reverse his actions, but learn upon finding him that he had used the stones a second time to destroy them, preventing their further use. In anger, Thor cuts off Thanos's head.  
  
Five years later, Scott Lang escapes from the quantum realm to discover that his daughter Cassie is now a teenager and that Hope van Dyne, along with half of the population, has disappeared in the snap. Rogers has been leading grief counseling sessions for survivors still struggling with the effects of the snap, while Romanoff is tirelessly keeping watch over both Earth and the rest of the universe with the help of Rhodes, Danvers, Okoye, Rocket, and Nebula. Lang goes to Romanoff and Rogers, and explains that while five years had passed for them, only five hours had passed for him in the quantum realm. The three go to Stark, who is now raising a child with Pepper Potts, and explain their theory that the quantum realm can be used to go back in time and steal the Infinity Stones before Thanos is able to collect them. Stark initially rejects their proposal with concern about risking his family and the peace he has found, but after reflecting upon the loss of Peter Parker decides to test theoretical models that would work with Lang's quantum tunnel, eventually finding one that works.

With Stark now on board with the plan, the remaining Avengers set out to reassemble their team. Bruce Banner has now embraced the Hulk as a part of him, and has melded his own consciousness and the Hulk's together into one. Romanoff, after hearing reports from Rhodes of an assassin that operates with similar methods to Barton, goes the Japan to track him down. Barton, consumed with grief after the loss of his family, has been operating under the mantle "Ronin" while brutally massacring cartels and gangs in order to try and improve the world that's still left. After some convincing from Natasha, he agrees to rejoin the team in order to try and bring his family back. Banner and Rocket go to the small town of New Asgard, where Valkyrie and the last survivors of Asgard have settled. They there find Thor, who has become overwhelmed by guilt over not killing Thanos the first time, in Wakanda. Thor has become majorly overweight, his hair and beard are overgrown, and he spends his free time eating junk food, getting drunk, and playing Fortnite with his friends Korg and Miek. Thor begrudgingly agrees to return to the Avengers after some convincing from Rocket and Banner.

***Output***:

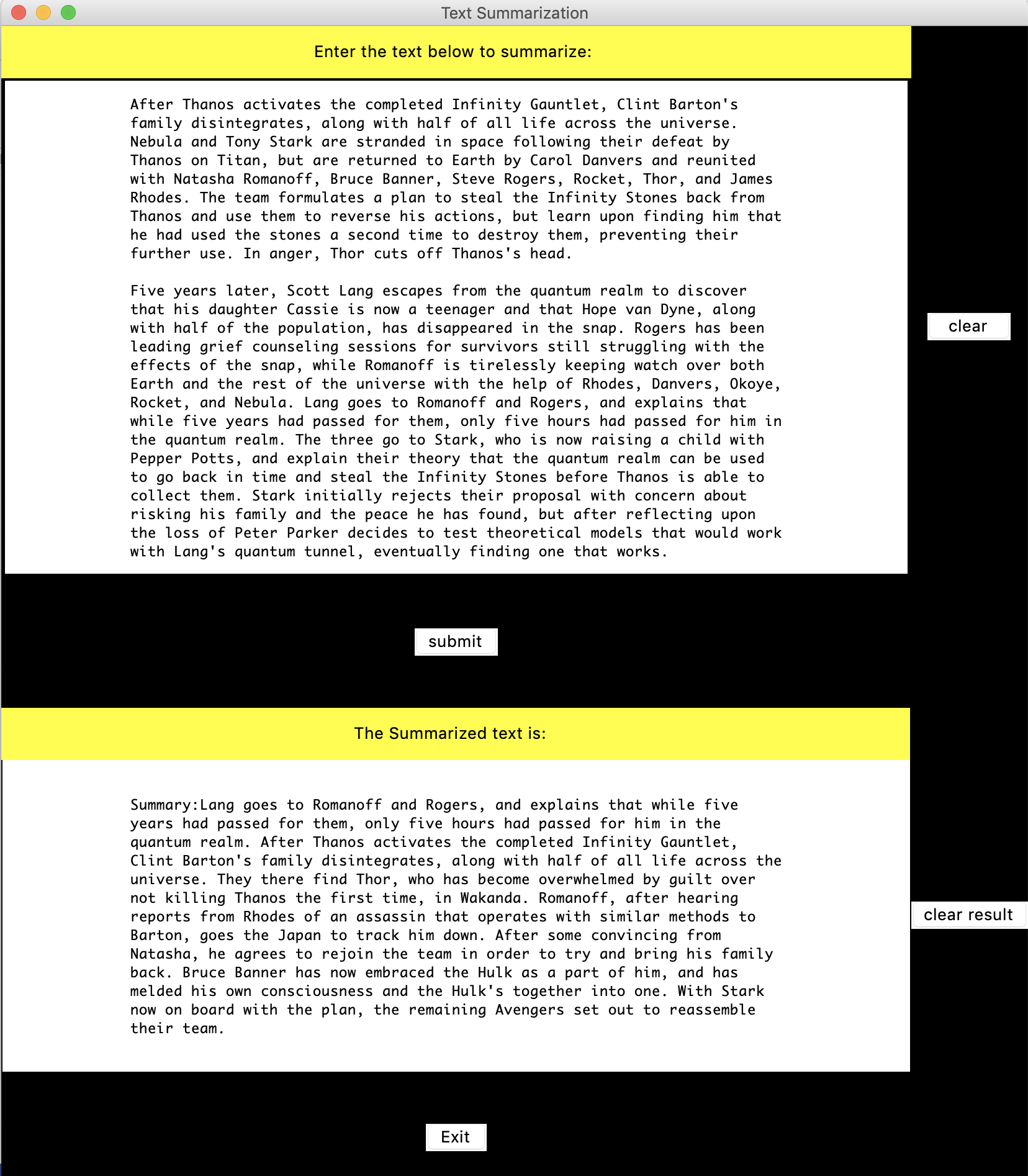
Lang goes to Romanoff and Rogers, and explains that while five years had passed for them, only five hours had passed for him in the quantum realm. After Thanos activates the completed Infinity Gauntlet, Clint Barton's family disintegrates, along with half of all life across the universe. They there find Thor, who has become overwhelmed by guilt over not killing Thanos the first time, in Wakanda. Romanoff, after hearing reports from Rhodes of an assassin that operates with similar methods to Barton, goes the Japan to track him down. After some convincing from Natasha, he agrees to rejoin the team in order to try and bring his family back. Bruce Banner has now embraced the Hulk as a part of him, and has melded his own consciousness and the Hulk's together into one. With Stark now on board with the plan, the remaining Avengers set out to reassemble their team.

Figure 2. Output of the text summarization

1. **References**

1. V. Gupta, G. S. LehaL, "A Survey of Text Summarization Extractive Techniques", *J. of Emerging Technologies in Web Intelligence*, vol. 2, no. 3, 2010.

2. G. Erkan, D. R. Radev, "LexRank: Graph based Lexical Centrality as Salience in Text Summarization." in J. of Artificial Intelligence Research, pp. 457-479, 2004.

3. Y. J. Knmar et al., "A Review on Automatic Text Summanzaticn Approaches", *J. of Computer Science*, vol. 12, no. 4, pp. 178-190, 2016.

4. Y. A. AL-Khassawnch, N. Salim, M. Jairah, "Inproving Triangle-Graph Based Text Summarization using Hybrid Similarity-Function", *Indian J. of Science and Technology*, vol. 10, no. 8, 2017.

5. P. C. Karmaker and M. S. Hossen, "Performance Analysis of Frequency and Graph Theoretic Based Text Summarization," *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, Cox'sBazar, Bangladesh, 2019, pp. 1-5.  
doi: 10.1109/ECACE.2019.8679452.