**Document Summarization**

Dharan Thaker , Paurav Thakkar

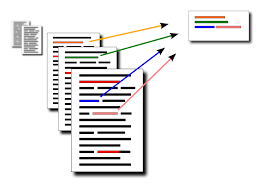
*Chandubhai S. Patel Institute of Technology, Changa , Gujarat ,India.*

**Abstract:**

Document Summarization is a part of text summarization in Natural Language processing (NLP) which is a concept used to summarize a document into lesser and simpler words. Text summarization is collecting all the necessary information from any original data and showing that information in the form od summary. Text summarization is a very important concept as the size data is constantly increasing in real world. Text summarization can be used various places such as search engines, sentimental analysis, etc. One can easily get the required amount of information in seemingly less time. This paper is a sincere effort to condense and epitomize the perspective of text summarization. The most fundamental technique which is used as a range for Summarization is Structured to Linguistic.

**Keywords:**

Document Summarization, Natural Language Processing, Text Summarization, Latent Semantic Analysis



**1.Introduction**

Before going to the Text summarization, first we, have to know that what a summary is. A summary is

a text that is produced from one or more texts, that conveys important information in the original text, and it

is of a shorter form. The goal of automatic text summarization is presenting the source text into a shorter

version with semantics.The most important advantage of using a summary is ,it reduces the reading time.

Text Summarization methods can be classified into extractive and abstractive summarization. An extractive

summarization method consists of selecting important sentences, paragraphs etc. from the original document

and concatenating them into shorter form. An Abstractive summarization is an understanding of the main

concepts in a document and then express those concepts in clear natural language

Before going to the Text summarization, first we, have to know that what a summary is. A summary is

a text that is produced from one or more texts, that conveys important information in the original text, and it

is of a shorter form. The goal of automatic text summarization is presenting the source text into a shorter

version with semantics.The most important advantage of using a summary is ,it reduces the reading time.

Text Summarization methods can be classified into extractive and abstractive summarization. An extractive

summarization method consists of selecting important sentences, paragraphs etc. from the original document

and concatenating them into shorter form. An Abstractive summarization is an understanding of the main

concepts in a document and then express those concepts in clear natural language

Before going to the Text summarization, first we, have to know that what a summary is. A summary is

a text that is produced from one or more texts, that conveys important information in the original text, and it

is of a shorter form. The goal of automatic text summarization is presenting the source text into a shorter

version with semantics.The most important advantage of using a summary is ,it reduces the reading time.

Text Summarization methods can be classified into extractive and abstractive summarization. An extractive

summarization method consists of selecting important sentences, paragraphs etc. from the original document

and concatenating them into shorter form. An Abstractive summarization is an understanding of the main

concepts in a document and then express those concepts in clear natural language

Before going to the Text summarization, first we, have to know that what a summary is. A summary is

a text that is produced from one or more texts, that conveys important information in the original text, and it

is of a shorter form. The goal of automatic text summarization is presenting the source text into a shorter

version with semantics.The most important advantage of using a summary is ,it reduces the reading time.

Text Summarization methods can be classified into extractive and abstractive summarization. An extractive

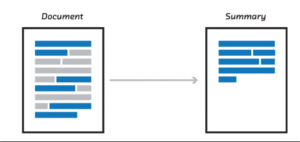
summarization method consists of selecting important sentences, paragraphs etc. from the original document

and concatenating them into shorter form. An Abstractive summarization is an understanding of the main

concepts in a document and then express those concepts in clear natural languag

Before diving into the world of Text Summarization, first we must know what a summary is. A summary is a text phrase which is generated from one or more texts, which transmits relevant and important information in the original text. The main objective behind automatic text summarization is to present the source text into a shorted version with semantics. This certainly helps in reducing the reading time.

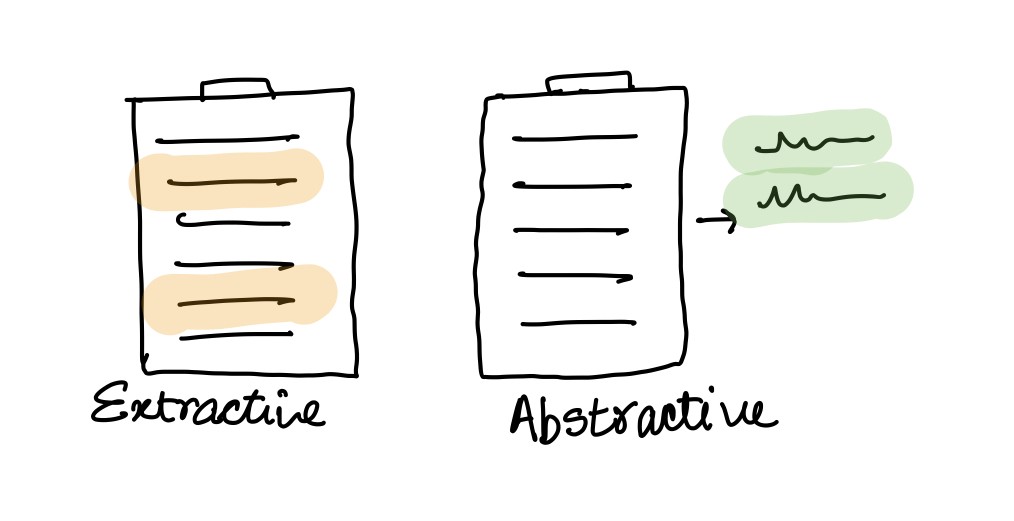
There are various models on the internet like news article summarizer such as Google, Microsoft News or Columbia Newsblaster. Also, there are numerous online summarization tools like Sumplify, Text Compacter, Tools4Noobs, FreeSummarizer, etc. Open Text Summarizer, NClassifier, CNGL Summarizer are few widely used open source summarization tools. There has been an immense need for presenting the information in an abstract form by the click of a button. Substantial conquests in the area of text summarization have been acquired using sentence extraction and statistical analysis.



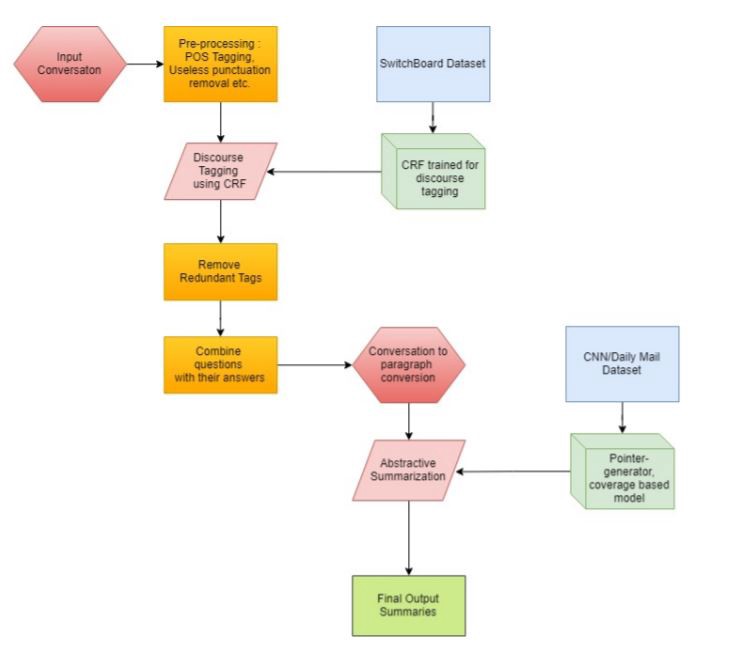
Text Summarization is broadly classified into two types Extractive and Abstractive Summarization. An Extractive Summarization method basically comprises of opting sentences, phrases etc from the original document and concatenating them into shorter understandable form, whereas Abstractive Summarization is an understanding of the important concepts in a document and then portray those concepts in clear natural language.

**2. Techniques Used for Text Summarization**

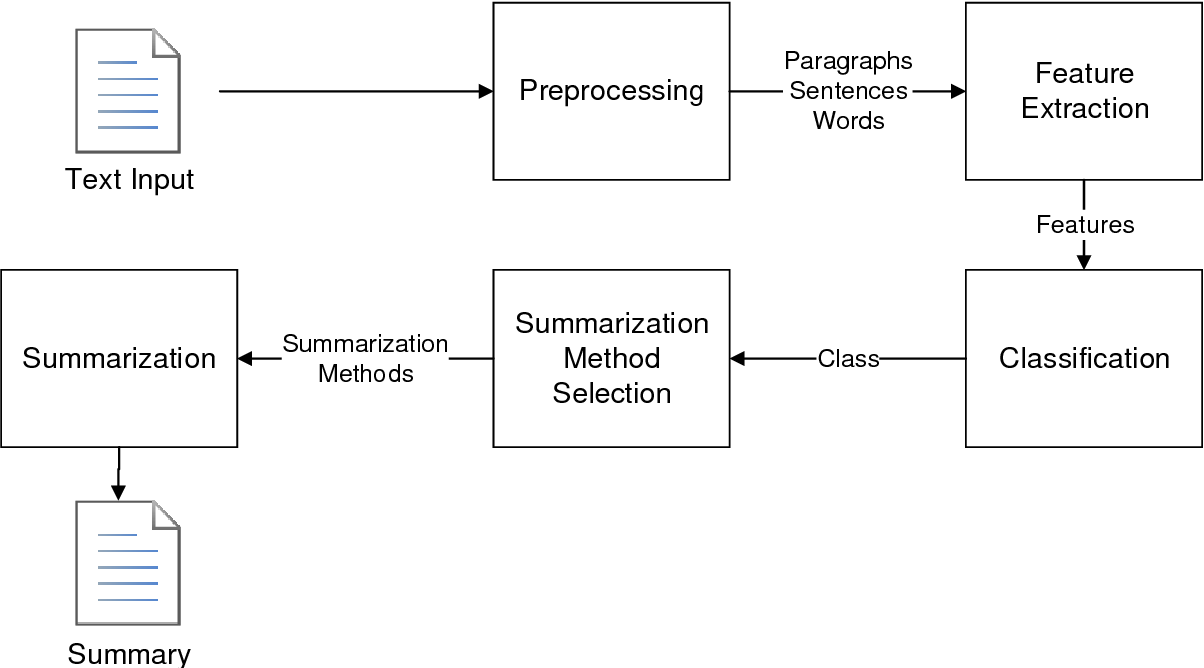
Abstractive and Extractive Summarization are the two types of Text Summarization. Following section shows the brief description on both the approaches:



1. Abstractive Summarization: This uses Abstractive techniques which again are classified into two categories: Structured Based Approach and Semantic Based Approach.

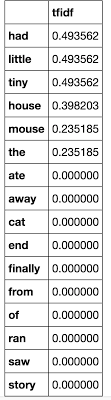


1. Structured Based Approach: Structured based Approach conceals the most important information through subjective schemes such as templates, extraction rules and other structures such as ontology, body phrase structure and tree.
2. Semantic Based Approach: In this approach, the representation of the document is done semantically and indicatively, by feeding it into the Natural Language Generation (NLG) system. Here, linguistic data is processed by focusing on recognizing the noun phrase and verb phrase.
3. Extractive Summarization: In this approach, the summarization is done by selecting essential sentences from the original document and chaining the document into a better and shorter form. The Importance which is to be given to the sentences is decided based on the linguistic and statistical features of sentences.

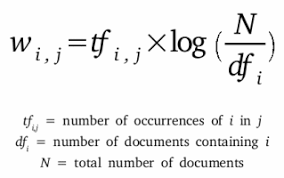


**3. Term Frequency-Inverse Document Frequency (TF-IDF) method**

The whole idea behind the TFIDF model is to calculate the numeric statistical value which shows how important the word is in the entire document. There is a proportionate increase in the TF-IDF value with respect to the number of occurrences of a word in the sentence. Weighted term frequency and inverse sentence frequency are the two paradigms through which the TF-IDF value is calculated in which sentence frequency marks the number of sentences in the document that contain that particular term. The vector values of sentences are obtained by the analogy to the query and amongst that the highest scoring sentences are picked to be part of the summary. Summarization is query-specific.

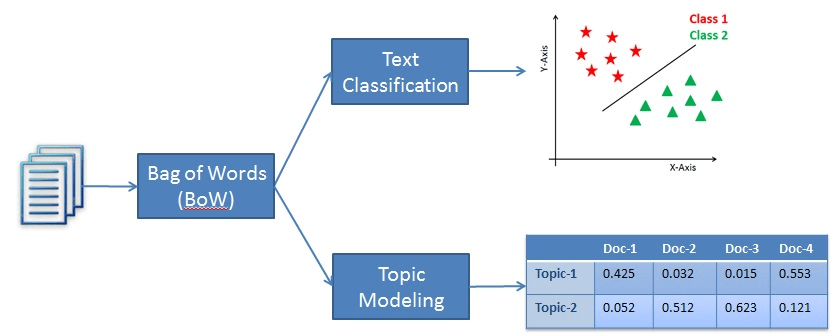


If the number of occurrences of certain specific words in a given sentence is more then that particular sentence is considered to be relatively more important. The method actually works by comparing the term frequency (tf) values in a given document. The targeted words mostly nouns. The TF/IDF score is calculated as follows:

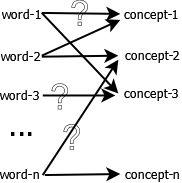


**4. Latent Semantic Analysis (LSA)**

Latent semantic analysis (LSA) is a method to analyse text semantics based on observed keywords. Initially a method was proposed by Gong and Liu in which LSA was used to select highly ranked sentences and use that sentences to create single and multi-document summarization. In the LSA method, a term-sentence matrix is built by giving each and every word as the row input and all the sentences present in the document as the column input. Inside the matrix, weight of words in corresponding sentences are stored which are calculated using TF-IDF model and words that do not belong to the corresponding sentence are given zero score. Then single value decomposition (SVD) is used to transform the matrix into three matrices: A = U ΣV T.



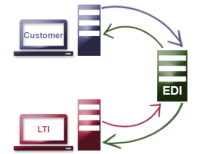
Here U represents Matrix U(number of words X number of sentences) known as term-topic matrix. Σ represents Matrix Σ where Matrix Σ is a diagonal matrix (number of sentences X number of words) and each row corresponds to the weight of a topic. V T stands for topic sentence matrix. The matrix D = ΣV T is tells us how a sentence represents a topic and so dij shows the weight of topic I in sentence j.



Gong and Liu’s method was to choose one sentence per each topic, therefore, based on the length of summary in terms of sentences, they retained the number of topics. This strategy has a drawback due to the fact that a topic may need more than one sentence to convey its information. Therefore, substitutive solutions were suggested to enhance the performance of LSA-based techniques for summarization. One such improvement that was decided was to influence the weight of each topic to determine the relative size of the summary that should cover the topic. This in turn would give the flexibility of having a variable number of sentences. Another advancement is described Steinberger introduced a LSA-based method which achieves a significantly better performance than the original work.

**5. Bayesian Topic Models**

There are certain limitations to the existing models when it comes to Text Summarization methods. 1) The sentences are thought to be autonomous to each other, as a matter of fact topics that are present in the document are disregarded. 2) Many scores of the sentences are just calculated abruptly using the heuristics and they do not signify a clear probabilistic interpretation.

.

Bayesian topic model is the solution to the above limitations. It is severely used for uncovering and representing the topics of the model. As to which is the main idea or the theme of the document. It helps in identifying the main highlight of the entire document. This approach helps us in seeking the right similarities and differences between various documents, ultimately giving a hand to more robust text summarization model.

The Formula that we use to for Bayesian Topic Models is as follows:-

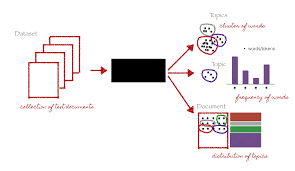
DK L ( P || Q ) = ∑w P(w) log P(w)/Q(w)

Where, KL = measure of divergence

P and Q = two probabilistic distributions

P(w) and Q(w) = Probabilities

Probabilistic topic models have really became these days. LDA ( Latent Dirichlet Allocation ) is used for digging out the thematic intel of the collection of documents. Here the main idea is that each and every topic resembles to the probability distributions over words and that the documents are nothing but the random mixture of latent topics.

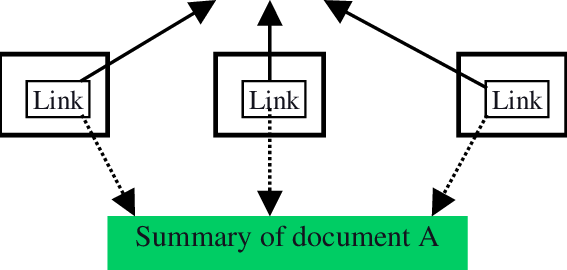


**6. The Impact of Context in Summarization**

Summarization systems has a wide variety of usage in different areas such as Web Summarization, Scientific Articles Summarization, Email Summarization etc. There are even blog posts readers who wants to know which parts of information inside the blogs are critical and which needs to be paid attention to. Following the areas where Summarization systems are used extensively.

**6.1 Web Summarization Field**

Web pages contains a lot of information which has only images which cannot be summarized. The Written information is very scarce, thus summarization gets very limited. The summarization of web pages is very useful as it might cover a lot of topics in a short easy to read paragraph.



**6.2 Scientific Article Summarization**

As we all know that scientific and research based articles contains tons of valuable information and it probably has huge data incorporated in it. Thus each and every piece of information in the scientific article is at most important to the researchers. Also, because of the huge amount of data it becomes very difficult to read everything. Article summarization helps to extract only those aspects in paper which conveys valuable information. KL divergence method is used to score the sentences in the original paper.

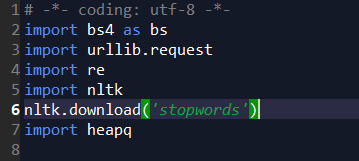


**6.3 Email Summarization**

Emails as we all know are the best approach to professional communication on the internet till date. Emails has both the aspects like spoken conversation and written text. There can be multiple conversations between two or more peoples. In Email summarization there is a machine learning technique which resembles to the position of the sentences in the thread, recipients, etc. Thus the idea is to cluster the messages into groups and then digging out the summarized form of those clusters.

**7. Implementation of Document Summarization Model**

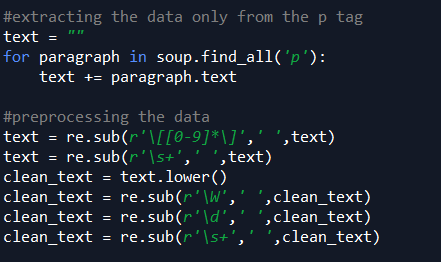
Step-1:- Importing the requires Libraries



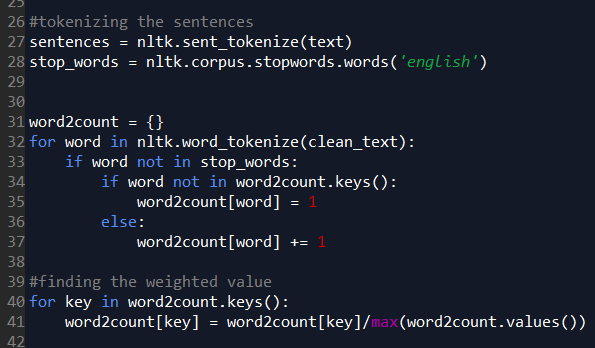
Step-2:- Reading the URL and parsing it with Beautiful Soup



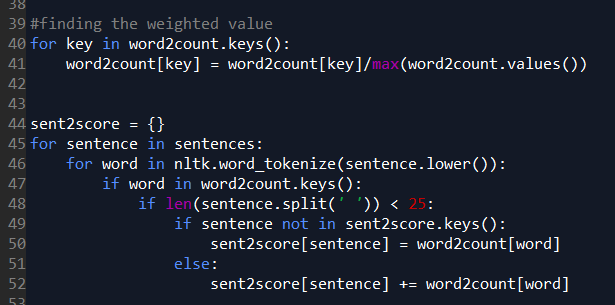
Step-3:- Extracting the data from the ‘p’ tag and Pre processing it with regex



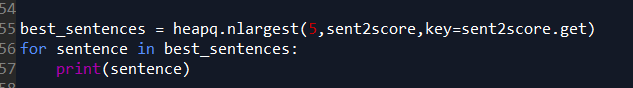
Step-4:- Tokenizing the words



Step-5:- Finding the weighted Value and storing it according to the Priority



Step-6 :- Displaying the Summary



**8. References**

[1] Mehdi Allahyari and Seyedamin Pouriyeh , 2017 - Text Summarization Techniques – A Brief Survey.

[2] Shai Erera and Michal Shmuli-Scheuer, 2019 – A Summarization System for Scientific Documents.

[3] Samrat Babar, 2013 – Text Summarization: An Overview

[4] Deepali K. Gaikwad and C. Namrata Mahender, 2016 – A Review Paper on Text Summarization

[5] Xiaodan Xu, 2009 – Research on Automatic Summarization System Based on Topic Partition

[6] Nedunchelian Ramanujan , 2016 – An Automatic Multidocument Text Summarization Approach based on Naïve Bayesian Classifier using Timestamp Strategy.

[7] Chintan Shah and Anjali G. Jivani, 2016 – Literature Study on Multi-document Text Summarization Techniques

[8] Arpita Sahoo and Dr. Ajit Kumar Nayak, 2018 – Review Paper on Extractive Text Summarization

[9] Vahed Qazvinian and Dragomir R. Radev, 2016 – Scientific Paper Summarization Using Citation Summary Networks

[10] Ruchika Aggarwal and Latika Gupta, 2017 – Automatic Text Summarization

[11] Sandeep Sripada, Venu Gopal Kasturi and Gautam Kumar Parai , 2015 – Multi-document extraction based Summarization

**9. Conclusion**

The increasing growth of Internet and data consumption is increasing day by day. It is difficult for humans to summarize such large amounts of texts and so in this document we have discussed a way of single or multi-document summarization using techniques such as abstract summarization(structural based approach and semantic based approach) and extractive summarization. Along with that we have also talked about the TFIDF model which proves to be one of the best models for finding the numerical value of words in the document and Latent Semantic Analysis and Bayesian Topic modelling. With we have describes the impact of context in summarization followed by the Implementation process of the Document summarization.

It is a numerical statistic which reflects how important a word is in a given document.The TF-IDF

value increases proportionally to the number of times a word appears in the document.This method mainly

works in the weighted term-frequency and inverse sentence frequency paradigm .where sentence-frequency is

the number of sentences in the document that contain that term. These sentence vectors are then scored by

similarity to the query and the highest scoring sentences are picked to be part of the summary.Summarization

is query-specific .

The hypothesis assumed by this approach is that if there are ‘‘more specific words’’ in a given

sentence, then the sentence is relatively more important. The target words are usually nouns .This method

performs a comparison between the term frequency (tf) in a document -in this case each sentence is treated

It is a numerical statistic which reflects how important a word is in a given document.The TF-IDF

value increases proportionally to the number of times a word appears in the document.This method mainly

works in the weighted term-frequency and inverse sentence frequency paradigm .where sentence-frequency is

the number of sentences in the document that contain that term. These sentence vectors are then scored by

similarity to the query and the highest scoring sentences are picked to be part of the summary.Summarization

is query-specific .

The hypothesis assumed by this approach is that if there are ‘‘more specific words’’ in a given

sentence, then the sentence is relatively more important. The target words are usually nouns .This method

performs a comparison between the term frequency (tf) in a document -in this case each sentence is treated