www.ijecs.in
International Journal Of Engineering And Computer Science
Volume 9 Issue 02 February 2020, Page No. 24940-24947
ISSN: 2319-7242 DOI: 10.18535/ijecs/v9i02.4437

Text Summarization

A. Vikas[1], G.V.N. Pradyumna[2], Tahir Ahmed Shaik[3]

[1]Student, Dept. of Computer Science and Engineering
[1]Anurag Group of Institutions
[1]Hyderabad, India
[2]Student, Dept. of Computer Science and Engineering
[2]Anurag Group of Institutions
[2]Hyderabad, India
[3]Student, Dept. of Computer Science and Engineering,
[3]Anurag Group of Institutions
[3]Hyderabad, India

Abstract:
In this new era, where tremendous information is available on the internet, it is most important to provide the improved mechanism to extract the information quickly and most efficiently. It is very difficult for human beings to manually extract the summary of a large documents of text. There are plenty of text material available on the internet. So, there is a problem of searching for relevant documents from the number of documents available and absorbing relevant information from it. In order to solve the above two problems, the automatic text summarization is very much necessary. Text summarization is the process of identifying the most important meaningful information in a document or set of related documents and compressing them into a shorter version preserving its overall meanings.
Keywords- information, automatic summarization, meaningful, shorter-version

1 Introduction

Text summarization is a process of extracting or collecting important information from original text and presents that information in the form of summary. In recent years, need for summarization can be seen in various purpose and in many domains such as news articles summary, email summary, short message of news on mobile etc. The automatic summarizer in general selects important sentences from the document and groups them together, it consumes less time or time saving to understand the content within the large document. The aim of automatic text summarization is to convert large document into shorter one and store important content.

The automatic summarization of text is a well-known task in the field of natural language processing (NLP). There are two general approaches to automatic summarization: extraction and abstraction. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. In contrast, abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might express. Such a summary might include verbal innovations.

1.1 Problem Definition

This paper describes a system for the summarization of single document. The system produces single document summaries using clustering techniques for identifying common terms across the set of documents. For each term, the system identifies representative passages that are included in the final summary. Results of our evaluation are also presented.

1.2. Application Areas

Text Summarization has been shown to be useful for Natural Language Processing tasks such as Question Answering or Text Classification and other related fields of computer science such as Information Retrieval. Since Geographical Information Retrieval can be considered as an extension of the Information Retrieval field, the generation of summaries could be integrated into these systems by acting as an intermediate stage, with the purpose of reducing the document length. These are some applications where text summarization can be used across the enterprise:

1 Newsletters: Many weekly newsletters take the form of an introduction followed by a curated selection of relevant articles. Summarization would allow organizations to further enrich newsletters with a stream of summaries (versus a list of links), which can be a particularly convenient format in mobile.

2. Search marketing and Search Engine Optimization (SEO): When evaluating search queries for SEO, it is critical to have a well-rounded understanding of what your competitors are talking about in their content. This has become particularly important since Google updated its algorithm and shifted focus towards topical authority (versus keywords). Multi-document summarization can be a powerful tool to quickly analyze dozens of search results, understand shared themes and skim the most important points.

3 Internal document workflow: Large companies are constantly producing internal knowledge, which frequently gets stored and under-used in databases as unstructured data. These companies should embrace tools that let them re-use already existing knowledge. Summarization can enable analysts to quickly understand everything the company has already done in a given subject, and quickly assemble reports that incorporate different points of view

4 Financial research: Investment banking firms spend large amounts of money acquiring information to drive their decision-making, including automated stock trading. When you are a financial analyst looking at market reports and news every day, you will inevitably hit a wall and won’t be able to read everything. Summarization systems tailored to financial documents like earning reports and financial news can help analysts quickly derive market signals from content.

5 Legal contract analysis: Related to point 4 (internal document workflow), more specific summarization systems could be developed to analyze legal documents. In this case, a summarizer might add value by condensing a

6 contract to the riskier clauses, or help you compare agreements.

7 Social media marketing: Companies producing long-form content, like whitepapers, e-books and blogs, might be able to leverage summarization to break down this content and make it sharable on social media sites like Twitter or Facebook. This would allow companies to further re-use existing content.

8 Question answering and bots: Personal assistants are taking over the workplace and the smart home. However, most assistants are fairly limited to very specific tasks. Large-scale summarization could become a powerful question answering technique. By collecting the most relevant documents for a particular question, a summarizer could assemble a cohesive answer in the form of a multi-document summary.

2. Proposed Methodology

2.1. Objective

There are many aspects and research regarding automatic document summarization that, apart from their importance, cannot be investigated. Our system is subjected to have a wider scope and to build a document summarization using Feature extraction method and Clustering, providing the central idea behind a textual matter to the user that is easier to understand and quicker to read.

2.2. Proposed System

The aim is to auto summarize documents. The proposed system works on the extractive summarization based approach; the system calculates the frequency weightage of each word in the sentence in the entire document and also validates for the parts of speech of the words then assigns a total score for the sentences. System then incorporates the clustering technique for extracting the final summary sentences. The architecture of the proposed system is given in Fig 1.

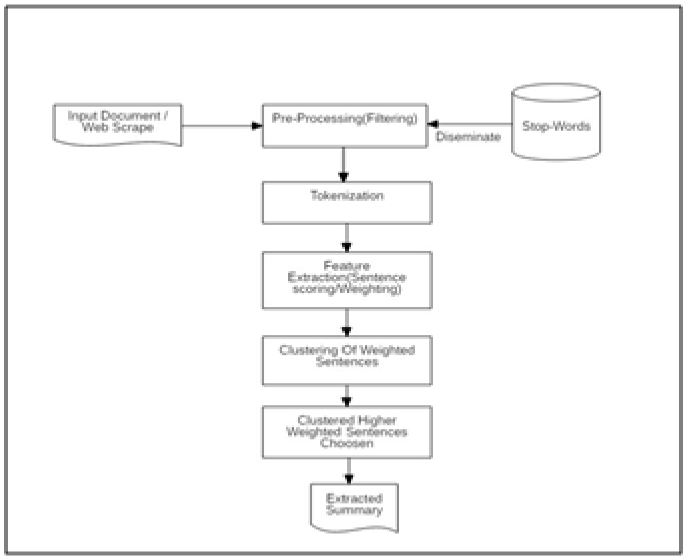

Input Document / Web Scrape
Pre-Processing(Filtering)
Diseminate
Stop-Words
Tokenization
Feature Extraction(Sentence scoring/Weighting)
Clustering Of Weighted Sentences
Clustered Higher Weighted Sentences Choosen
Extracted Summary

Figure 1: Proposed System Architecture

From the above Fig 1 then input process represents two formats such as a local document(docx, pdf ,txt) and web sources obtained from the URL. The input file may include certain common words such as root words or stop words (a, an, the, etc.), illegal characters and spaces; all these words are eliminated in the pre-process phase. Tokenization is the process of converting a sequence of characters into a sequence of tokens. In the feature extraction phase, the sentences are attached with the weights or scores which are used to prioritize them. In the Clustering phase the clusters are formed based upon the sentence scores and are segregated into lowest and highest weighted sentences from which the final phase provides the output based upon the highest scored clusters.

2.3. Scope

Extracting data from publication reports is a standard process in systematic review (SR) development. However, the data extraction process still relies too much on manual effort which is slow, costly, and subject to human error. In this study, we developed a text summarization system aimed at enhancing productivity and reducing errors in the traditional data extraction process.

3. Implementation

3.1. Software Libraries

The system uses the NLTK python module for achieving the NLP (Natural Language Processing) tasks. The NLTK provides tokenization for sentences and words, stop words and stemming of words to root forms. The Library also provides POS (Parts of Speech) tagger for identification of parts of speech in the data. Beautiful Soup module is used for HTML parsing. The pyFPDF is used for flushing the output to the pdf document. The communication with the HTML source document on web is provided by the urllib module.

3.2. Class Diagram

Class diagram is a static diagram. It represents the static view of an application. Class diagram describes the attributes and operations of a class and also the constraints imposed on the system. The class diagrams are widely used in the modelling of object-oriented systems because they are the only UML diagrams, which can be mapped directly with object-oriented languages. Class diagram shows a collection of classes, interfaces, associations, collaborations, and constraints. It is also known as a structural diagram.

There are 5 main classes in our proposed system: User Interface, Preprocess, Tokenize Sentences, SentenceScoring, and Clusterize classes. Each of their fields and methods are going to be used in order to calculate the sentence features and extract the summary. The class diagram of the proposed method is shown in below Fig 2.

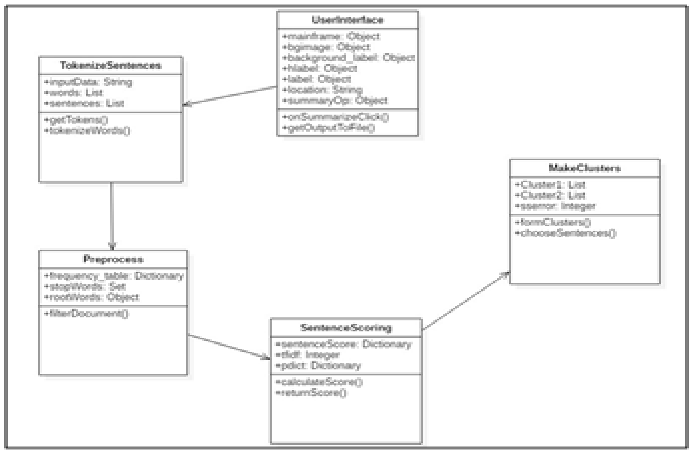


Figure 2: Class Diagram

3.3.Modules

The system is implemented using the libraries available in python programming language and divided into four modules.

The major four modules namely

1. UserInterface Module
2. Pre-process Module
3. Feature Extraction Module
4. Clustering Module

The description and usage of these modules is explained under the module description section. Each of these modules consists of defined classes, where each class performs a specific task or operation through its defined methods.

3.4. Module Description

1. User Interface Module**:** This module allows for the user interaction, providing the input for processing and generating the output to the user. This module uses the python tkinter library for generating the interaction components such as main window, buttons, text input. All the object instantiations and method invoking occurs in this module.
2. Pre-process Module: This module has the duty of filtering the input data through the processes such as removal of stop words, reduction of words to root forms, formatting of spaces and unrecognized symbols through the pre-process class. Then this module consists of the Tokenization class that performs the splitting of the data into sentences and words.
3. Feature Extraction Module: When the input data to an algorithm is too large to be processed and it is suspected to be redundant then it can be transformed into a reduced set of features (also named a feature vector). Determining a subset of the initial features is called feature selection. The core module for the application is calculating the sentence score and assigning the weights to the sentences. This process is done in Sentence Scoring class.
4. Clustering Module: K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes. This module clusters the weighted sentences to form a cluster of major sentences to be chosen for the summary.

4. Results

4.1 Testing

System testing is performed on the entire system in the context of either functional requirement specifications (FRS) or system requirement specification (SRS), or both. System testing tests not only the design, but also the behavior and even the believed expectations of the customer. The following tables depicts the testing done to the system:

1. Test Case-1**:** For test case-1 a local text document is provided as an input and the output file is generated and is described in table 1.

Table 1: Test Case-1

TEST CASE ID	#1
TEST CASE NAME	Local Text File Input
DESCRIPTION	Inputting a local text file to summarize
FREQENCY OF USE	Many Times
REMARKS	Success

2 Test Case-2: For test case-2 a web URL link is given as a source for input and the output file is generated and is described in table 2.

Table 2: Test Case-2

TEST CASE ID	#2
TEST CASE NAME	Web page source input
DESCRIPTION	Inputting a web page URL for summarizing
FREQUENCY OF USE	Many times
REMARKS	Success

3 Test Case-3: For test case-3 an unsupported file document is given as an input and the output file is not generated and is described in table 3.

Table 3 Test Case-3

TEST CASE ID	#4
TEST CASE NAME	unsupported files such as Docx, jpg, png, gif.
DESCRIPTION	Invalid document
FREQUENCY OF USE	-
REMARKS	Failure

4 Test Case-4**:** For test case-4 an invalid URL link or an error-based web page is given as an input and the output file is not generated and is described in table 4.

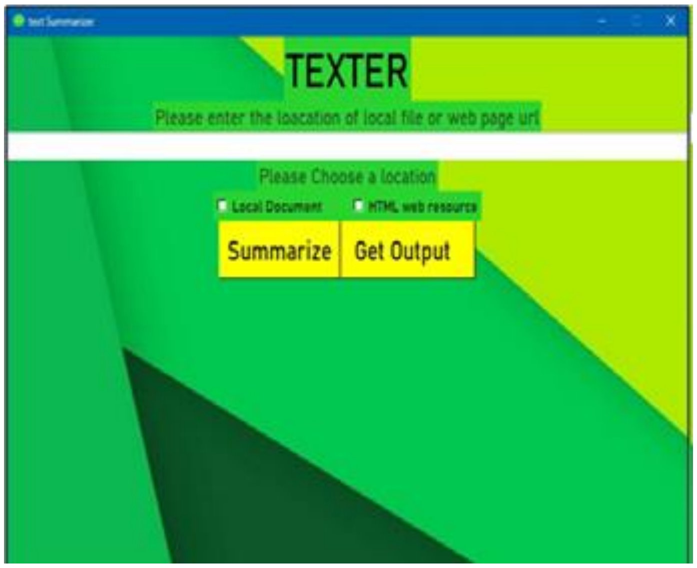
Table 4: Test Case-4

TEST CASE ID	#6
TEST CASE NAME	Inaccessible url
DESCRIPTION	Undefined web page

4.2 Screenshots

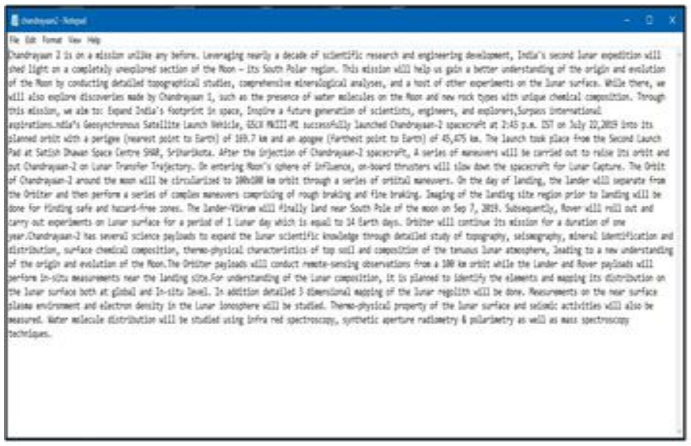
User Interface: This is the graphical user interface where the user provides a file location and gets a summarized output as shown in Fig 3

Figure 3: Main Screen



Sample Examples

A sample text document, Chandraayan II is given as an input in the Fig 4.

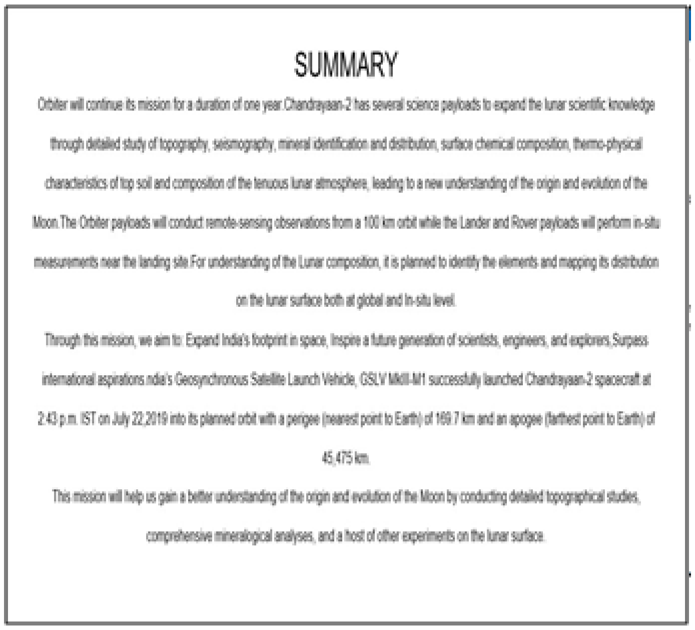


Chandrayaan 2 is on a mission unlike any before. Leveraging nearly a decade of scientific research and engineering development, India's second lunar expedition will shed light on a completely unexplored section of the Moon – its South Polar region. This mission will help us gain a better understanding of the origin and evolution of the Moon by conducting detailed topographical studies, comprehensive mineralogical analyses, and a host of other experiments on the lunar surface. While there, we will also explore discoveries made by Chandrayaan 1, such as the presence of water molecules on the Moon and new rock types with unique chemical composition. Through this mission, we aim to: Expand India's footprint in space, Inspire a future generation of scientists, engineers, and explorers,Surpass international aspirations.ndia's Geosynchronous Satellite Launch Vehicle, GSLV MkIII-M1 successfully launched Chandrayaan-2 spacecraft at 2:43 p.m. IST on July 22,2019 into its planned orbit with a perigee (nearest point to Earth) of 169.7 km and an apogee (farthest point to Earth) of 45,475 km. The launch took place from the Second Launch Pad at Satish Dhawan Space Centre SHAR, Sriharikota. After the injection of Chandrayaan-2 spacecraft, A series of maneuvers will be carried out to raise its orbit and put Chandrayaan-2 on Lunar Transfer Trajectory. On entering Moon's sphere of influence, on-board thrusters will slow down the spacecraft for Lunar Capture. The Orbit of Chandrayaan-2 around the moon will be circularized to 100x100 km orbit through a series of orbital maneuvers. On the day of landing, the lander will separate from the Orbiter and then perform a series of complex maneuvers comprising of rough braking and fine braking. Imaging of the landing site region prior to landing will be done for finding safe and hazard-free zones. The lander-Vikram will finally land near South Pole of the moon on Sep 7, 2019. Subsequently, Rover will roll out and carry out experiments on Lunar surface for a period of 1 Lunar day which is equal to 14 Earth days. Orbiter will continue its mission for a duration of one year.Chandrayaan-2 has several science payloads to expand the lunar scientific knowledge through detailed study of topography, seismography, mineral identification and distribution, surface chemical composition, thermo-physical characteristics of top soil and composition of the tenuous lunar atmosphere, leading to a new understanding of the origin and evolution of the Moon.The Orbiter payloads will conduct remote-sensing observations from a 100 km orbit while the Lander and Rover payloads will perform in-situ measurements near the landing site.For understanding of the lunar composition, It is planned to identify the elements and mapping its distribution on the lunar surface both at global and in-situ level. In addition detailed 3 dimensional mapping of the lunar regolith will be done. Measurements on the near surface plasma environment and electron density in the Lunar Ionosphere will be studied. Thermo-physical property of the lunar surface and seismic activities will also be measured. Water molecule distribution will be studied using infra red spectroscopy, synthetic aperture radiometry & polarimetry as well as mass spectroscopy techniques.

Figure 4: A Sample Document: Chandrayaan-II.txt

The summarized output of the text document is shown in the Fig 5.

The accuracy obtained for the output is 77.8.



SUMMARY

Orbiter will continue its mission for a duration of one year Chandrayaan-2 has several science payloads to expand the lunar scientific knowledge through detailed study of topography, seismography, mineral identification and distribution, surface chemical composition, thermo-physical characteristics of top soil and composition of the tenuous lunar atmosphere, leading to a new understanding of the origin and evolution of the Moon The Orbiter payloads will conduct remote-sensing observations from a 100 km orbit while the Lander and Rover payloads will perform in-situ measurements near the landing site For understanding of the Lunar composition, it is planned to identify the elements and mapping its distribution on the lunar surface both at global and In-situ level.

Through this mission, we aim to: Expand India's footprint in space, Inspire a future generation of scientists, engineers, and explorers,Surpass international aspirations ndia's Geosynchronous Satellite Launch Vehicle, GSLV MkIII-M1 successfully launched Chandrayaan-2 spacecraft at 2:43 p.m. IST on July 22,2019 into its planned orbit with a perigee (nearest point to Earth) of 169.7 km and an apogee (farthest point to Earth) of 45,475 km.

This mission will help us gain a better understanding of the origin and evolution of the Moon by conducting detailed topographical studies, comprehensive mineralogical analyses, and a host of other experiments on the lunar surface.

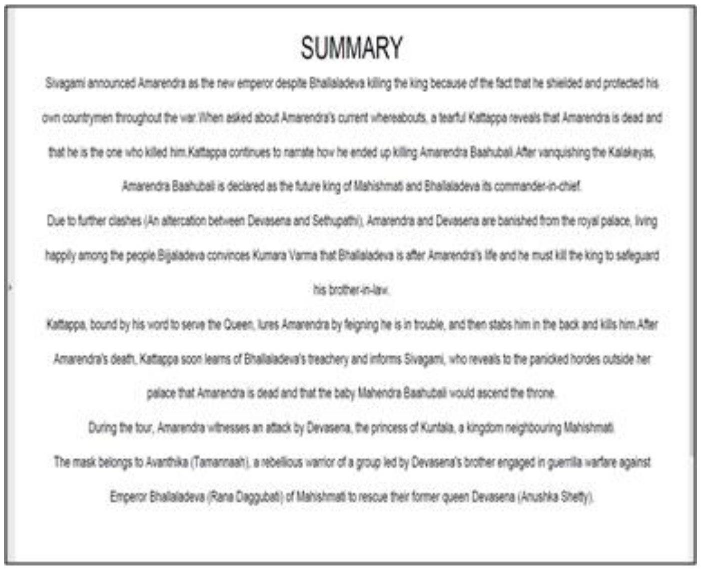
Figure 5: The output file of Chandrayaan-II

1 A sample text document, baahubhali is given as an input, shown in Fig 6.

[illegible]

Figure 6: A Sample Document: baahubhali.txt

The summarized output of the text document is given below in Fig 7.



SUMMARY

Sivagami announced Amarendra as the new emperor despite Bhallaladeva killing the king because of the fact that he shielded and protected his own countrymen throughout the war.When asked about Amarendra's current whereabouts, a tearful Kattappa reveals that Amarendra is dead and that he is the one who killed him.Kattappa continues to narrate how he ended up killing Amarendra Baahubali.After vanquishing the Kalakeyas, Amarendra Baahubali is declared as the future king of Mahishmati and Bhallaladeva its commander-in-chief.

Due to further clashes (An altercation between Devasena and Sethupathi), Amarendra and Devasena are banished from the royal palace, living happily among the people.Bijjaladeva convinces Kumara Varma that Bhallaladeva is after Amarendra's life and he must kill the king to safeguard his brother-in-law.

Kattappa, bound by his word to serve the Queen, lures Amarendra by feigning he is in trouble, and then stabs him in the back and kills him.After Amarendra's death, Kattappa soon learns of Bhallaladeva's treachery and informs Sivagami, who reveals to the panicked hordes outside her palace that Amarendra is dead and that the baby Mahendra Baahubali would ascend the throne.

During the tour, Amarendra witnesses an attack by Devasena, the princess of Kuntala, a kingdom neighbouring Mahishmati.

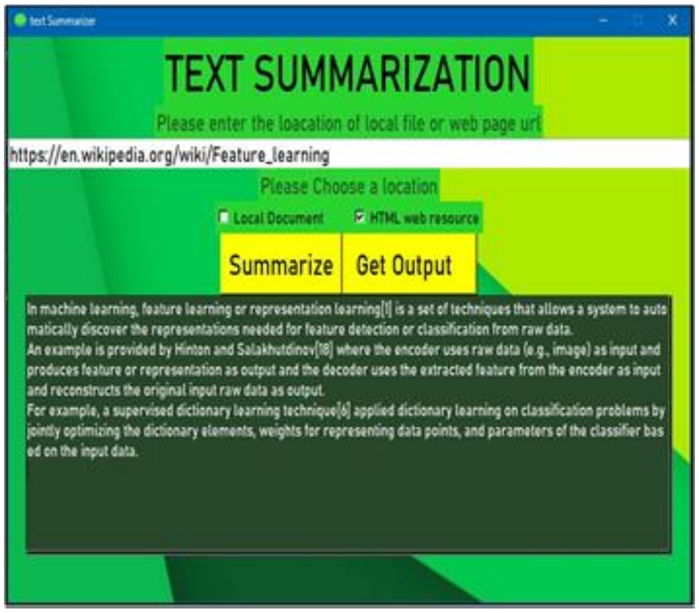
The mask belongs to Avanthika (Tamannaah), a rebellious warrior of a group led by Devasena's brother engaged in guerilla warfare against Emperor Bhallaladeva (Rana Daggubati) of Mahishmati to rescue their former queen Devasena (Anushka Shetty).

Figure 7: The output file of baahubhali.txt

The accuracy obtained for the output summary generated is 82.1.

2 We have also taken input from the specified URL on internet. The below Fig 8 shows the sample output.

The accuracy obtained for the above output summary is 47.0.



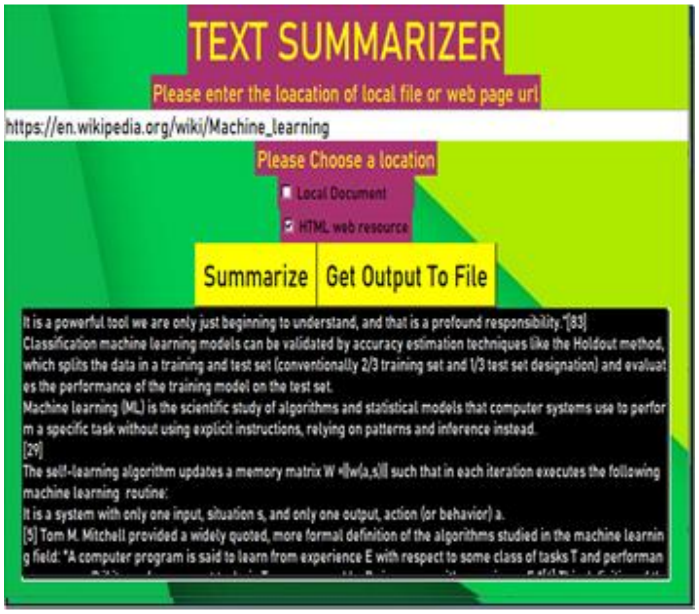
https://en.wikipedia.org/wiki/Feature_learning

[ ] Local Document [✓ HTML web resource]

In machine learning, feature learning or representation learning[1] is a set of techniques that allows a system to auto matically discover the representations needed for feature detection or classification from raw data.
An example is provided by Hinton and Salakhutdinov[18] where the encoder uses raw data (e.g., image) as input and produces feature or representation as output and the decoder uses the extracted feature from the encoder as input and reconstructs the original input raw data as output.
For example, a supervised dictionary learning technique[6] applied dictionary learning on classification problems by jointly optimizing the dictionary elements, weights for representing data points, and parameters of the classifier bas ed on the input data.

Figure 8: Web page given as input for summarization

3 Another specified URL on the internet is taken as the input for the system, given below in Fig 9 is the output generated for the specified URL.


TEXT SUMMARIZER
Please enter the loacation of local file or web page url
https://en.wikipedia.org/wiki/Machine_learning
Please Choose a location
Local Document
HTML web resource
Summarize
Get Output To File

```
It is a powerful tool we are only just beginning to understand, and that is a profound responsibility."[83]
Classification machine learning models can be validated by accuracy estimation techniques like the Holdout method,
which splits the data in a training and test set (conventionally 2/3 training set and 1/3 test set designation) and evaluat
es the performance of the training model on the test set.
Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perfor
m a specific task without using explicit instructions, relying on patterns and inference instead.
[29]
The self-learning algorithm updates a memory matrix W =||w(a,s)|| such that in each iteration executes the following
machine learning  routine:
It is a system with only one input, situation s, and only one output, action (or behavior) a.
[5] Tom M. Mitchell provided a widely quoted, more formal definition of the algorithms studied in the machine learnin
g field: "A computer program is said to learn from experience E with respect to some class of tasks T and performan
```

Figure 9: Web page given as input for summarization.

The accuracy obtained for the above output is 53.5.

5. Conclusion

The rate of information growth due to the World Wide Web has called for a need to develop efficient and accurate summarization systems. Although research on summarization started about 50 years ago, there is still a long trail to walk in this field. Over time, attention has drifted from summarizing scientific articles to news articles, electronic mail messages, advertisements, and blogs. Both abstractive and extractive approaches have been attempted, depending on the application at hand. Majority of the work was traditionally focused on extractive approaches due to the ease of defining hard-coded rules to select important sentences than generate new ones. Also, it promises grammatically correct and coherent summary but they often don't summarize long and complex texts well as they are very restrictive. Usually, abstractive summarization requires heavy machinery for language generation and is difficult to replicate or extend to broader domains. In contrast, simple extraction of sentences has produced satisfactory results in large-scale applications. The application has achieved its objectives thereby reducing the input textual matter or data to a more concise reduced summarized output.

6. Future Enhancements

Currently the paper aims to produce summaries for single documents on the extractive approach. However, we can focus on a much more effective application that is capable to generate more precise summaries based on contexts using a wide range high

level technology such as deep learning, neural networks. Abstractive summarization systems generate new phrases, possibly rephrasing or using words that were not in the original text. Naturally abstractive approaches are harder. For perfect abstractive summary, the model has to first truly understand the document and then try to express that understanding in short possibly using new words and phrases. Much harder than extractive. Has a complex capability like generalization, paraphrasing and incorporating real-world knowledge? The traditional rule-based AI did poorly on Abstractive Text Summarization. But the recent advances in Deep Learning changed that for the good. The text summary with deep learning is still under research.

References

[1] https://pdfs.semanticscholar.org/8120/02cd523d65adf20fd9f8f2683f73b70672f7.pdf

[2] https://www.tutorialspoint.com/uml/uml_class_diagram.htm

[3] https://blog.frase.io/20-applications-of-automatic-summarization-in-the-enterprise/

[4] Auto Text Summarization R. Shelke, Marathwada Mitra, P. Sarode, Ashwini, S. Satyan, International Journal of IT, Engineering and Applied Sciences Research

[5] (IJIEASR) ISSN: 2319-4413 Volume 4, No. 4, April 2015

[6] A Survey on Automatic Text Summarization Dipanjan Das Andr´e

F.T. Martins Language Technologies Institute Carnegie Mellon University, November 21, 2007.

[7] Vishal Gupta and Gurpreet Singh Lehal, “A Survey of Text Summarization

[8] Extractive Techniques”, Journal of Emerging Technologies in Web Intelligence, Vol. 2, No. 3, August 2010.

[9] https://en.m.wikipedia.org/wiki/System_testing