Text summarization of National Leader Online Article

Sanah Sayyed, Assistant professor, Department of C.S., Shivchhatrapti college Aurangabad, Maharashtra, India, Sanahsayyed92@gmail.com

Abstract The Internet is a great resource for learning about any subject you can think of. However, it becomes a difficult task to extract precise information due to the abundance of content. The primary goal of any text summarization system is to identify and present to users the essential information contained within a text. In today's world, there is a plethora of data at your fingertips. The accuracy of the data becomes problematic for the user. Reading everything and drawing conclusions about specific data is impossible. Text summarization is the process of condensing a large amount of data into a concise summary. If we're talking specifically about the data on Wikipedia, then yes, pretty much everything is accessible. With the right query, we can uncover a wealth of information. A text summarization service will condense long passages into digestible chunks without altering the meaning of the original text. For this study's demonstration purposes, we have taken information from Wikipedia and, using summarization techniques, condensed it without altering its meaning.

Keywords — ***Text summarization, Analyze the Data, POS tagging, OATS Model, summary generation.***

I. INTRODUCTION

To find the most pertinent information and treatment options, summarization has recently become evident in a variety of contexts, including media articles, financial records, and outcomes from search engines, medical summaries, and online portals. Even though there is an increasing amount of information on the internet, sub-branching has increased the capabilities of natural language processing. Information that is easy to understand is useful when searching [1][2][3].

You incorporate all pertinent information while summarizing key passages from the document using the machine learning method. Extractive and synthetizing are needed due to one of the main findings. The text summarization function is essential for advancing academic, research, teacher, and business marketing functions. It has been observed that the executive can process a small amount of data because they need a condensed summary of the details inside this time allotted. Text summarization produces a succinct, fluid description while trying to maintain the significance of key details and their overall content. Many automated text summarization techniques have been created and successfully used in a variety of fields. Text summarization is a technique for highlighting key details in a text document. The user is given concise reports and data that were gathered during this process. For humans, it's crucial to understand and be able to articulate the text's meaning. Due to its many applications in areas like book summaries, digests, financial institutions, media, showcases, basic science summaries, editorials, journals, etc., text summarization is crucial. A variety of tools exist for text summarization. The characteristics of scientific articles, such as their length and complexity, were not considered by these devices, which instead focused on news or essential documents. In this research work we have given emphasis on extractive text summarization.

II. LITERATURE REVIEW

Mishra et al. [4] looked at research from the years 2000-2013 and came across techniques like hybrid statistical and ML methods. The cognitive effects of ATS were not evaluated, and the researchers did not include this consideration. Processes like topic representation, frequency-driven, graph-based, and machine learning methods for ATS were all explored by Allahyari et al. [5]. Only the most common methods were included in this study.Ujjwal Rani [6] found that efficient collection, storage, and supervision require high rates of production in large data dimensions. Because there is so much data to store or search through, it can be challenging to find the right information at times. In this case, technology like big data is directly responsible for the proliferation of data in various formats, regardless of the value of mining information. According to Neelima G. et al. [7], the advent of the internet marked the beginning of the modern information technology era. There is a staggering daily increase in the amount of knowledge and data available.In their extensive survey for analysts, Saranyamol et al. [8] introduced many different facets of ATS, including its structure, strategies, datasets, evaluation metrics, etc. Two different text summarization techniques were combined in an attempt at analysis by Gambhir et al. [9].Many cutting-edge methods were not considered for this paper. Gholamrezazadeh et al[10] .'s study is the most thorough

DOI : 10.35291/2454-9150.2022.0488

and thorough comparison of extractive techniques used in ATS over the past decade. Several methods that can work in multiple languages have been mentioned.Research papers that use automated keyword extraction methods and techniques were surveyed by Bharti et al. [11]. Concepts related to the use of several data sources for text summarization are discussed.Text summarization from multiple documents and large amounts of web data was studied by Abualigah et al. [12]. The paper concludes with a detailed comparative table of recent studies.

III. PROPOSED MODEL:

OATS stands for Online Article Text Summarization. This model tries to identify the important sentences from the article so that the generated summary should not seem to have boundary information while the body is empty. The figure 1 below shows the work flow diagram of OATS

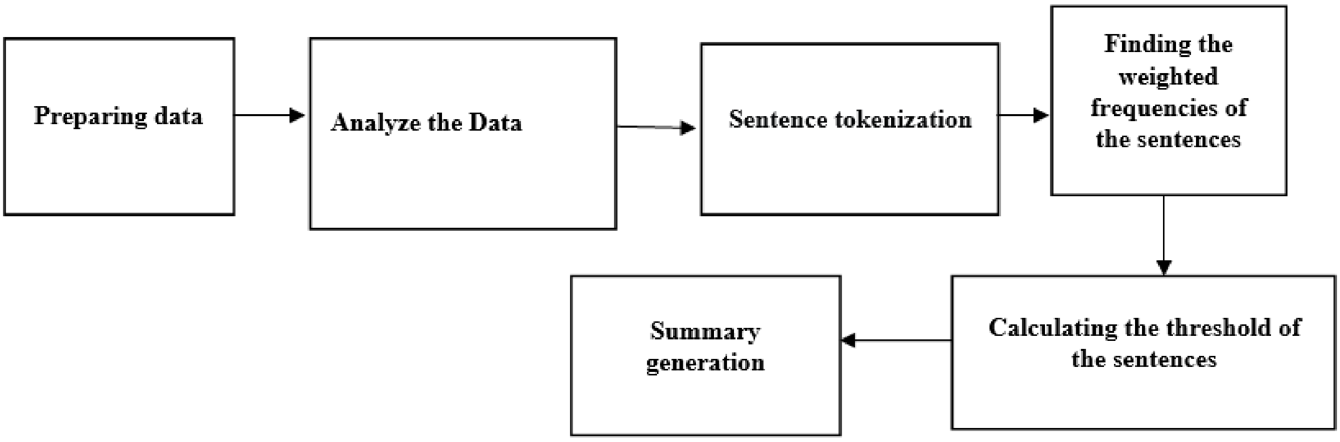
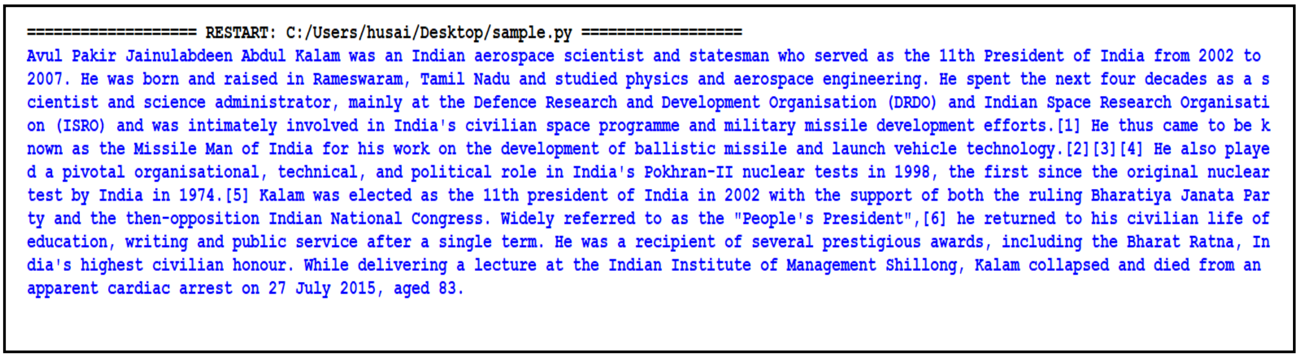

Preparing data
Analyze the Data
Sentence tokenization
Finding the weighted frequencies of the sentences
Calculating the threshold of the sentences
Summary generation

Figure 1. Workflow diagram of OATS Model

In this step, we want to summarize the information Wikipedia article such as 20 national leaders[13], The Beautiful Soup library will be used to retrieve the full text of the article. To ensure the highest quality of the textual data scraped, we will perform some basic text cleaning in the pre-processing step. To speed up the process, we will be using the nltk library's list of stop words. In addition, we'll be implementing PorterStemmer, a programme that deconstructs words into their individual morphemes. Tokenizing the article into sentences is the next step. The article will be broken down into sentences using the nltk library's built-in method. After that, you'll need to calculate the sentence weighted frequencies. Examining the frequency with which each term appears in the text will allow us to determine the overall grade for the essay. Specifically, we will award points to sentences based on how often specific keywords appear within them. Finally, we'll average the sentences to determine which types of sentences can be summed up. Now that we have collected all the data, we can generate a summary of the article.

Step 3.1 Preparing the data: In this step, we want to summarize the information Wikipedia article such as 20 national leaders, example of it as we take gives an overview of the national leader APJ Abdul kalam. The Beautiful Soup library will be used to locate the article's text, the sample example of article is as shown in figure 2.

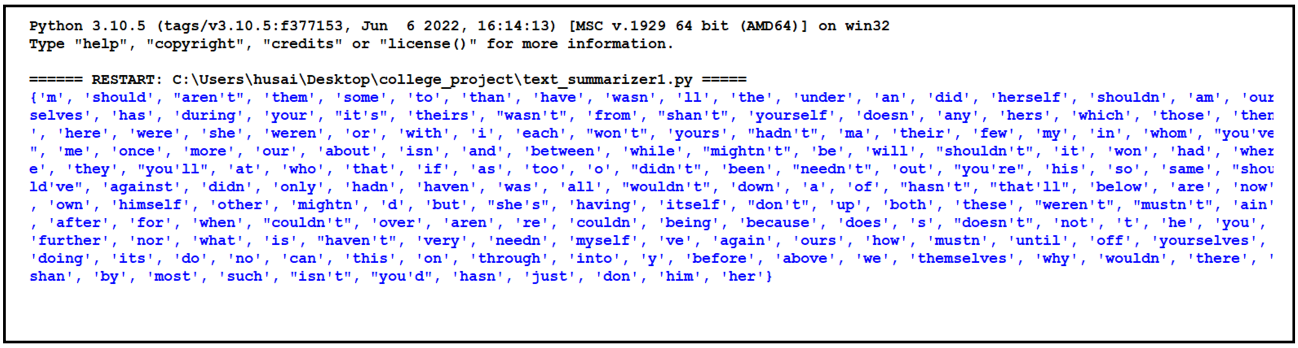


```
=================== RESTART: C:/Users/husai/Desktop/sample.py ==================
Avul Pakir Jainulabdeen Abdul Kalam was an Indian aerospace scientist and statesman who served as the 11th President of India from 2002 to
2007. He was born and raised in Rameswaram, Tamil Nadu and studied physics and aerospace engineering. He spent the next four decades as a s
cientist and science administrator, mainly at the Defence Research and Development Organisation (DRDO) and Indian Space Research Organisati
on (ISRO) and was intimately involved in India's civilian space programme and military missile development efforts.[1] He thus came to be k
nown as the Missile Man of India for his work on the development of ballistic missile and launch vehicle technology.[2][3][4] He also playe
d a pivotal organisational, technical, and political role in India's Pokhran-II nuclear tests in 1998, the first since the original nuclear
test by India in 1974.[5] Kalam was elected as the 11th president of India in 2002 with the support of both the ruling Bharatiya Janata Par
ty and the then-opposition Indian National Congress. Widely referred to as the "People's President",[6] he returned to his civilian life of
education, writing and public service after a single term. He was a recipient of several prestigious awards, including the Bharat Ratna, In
dia's highest civilian honour. While delivering a lecture at the Indian Institute of Management Shillong, Kalam collapsed and died from an
apparent cardiac arrest on 27 July 2015, aged 83.
```

Figure 2: Sample Example of Article

Step 3.2 Analyze the Data: To ensure the highest possible quality of the textual data we scraped, we will perform some basic text cleaning. To speed up the processing, we will use the nltk library's list of stopwords. In addition, we'll be implementing PorterStemmer, a programme that deconstructs words into their individual morphemes. The sample example of stop word removal is as shown in figure 3.

DOI : 10.35291/2454-9150.2022.0488

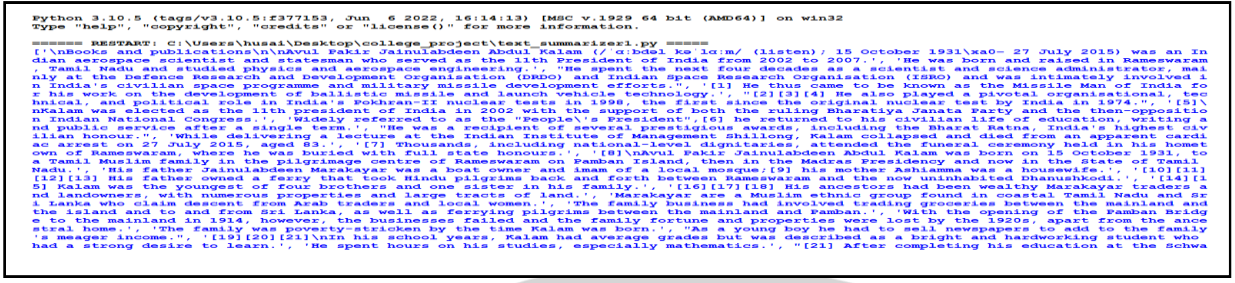


```
Python 3.10.5 (tags/v3.10.5:f377153, Jun  6 2022, 16:14:13) [MSC v.1929 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.

====== RESTART: C:\Users\husai\Desktop\college_project\text_summarizer1.py =====
{'m', 'should', "aren't", 'them', 'some', 'to', 'than', 'have', 'wasn', 'll', 'the', 'under', 'an', 'did', 'herself', 'shouldn', 'am', 'our
selves', 'has', 'during', 'your', "it's", 'theirs', "wasn't", 'from', "shan't", 'yourself', 'doesn', 'any', 'hers', 'which', 'those', 'then
', 'here', 'were', 'she', 'weren', 'or', 'with', 'i', 'each', "won't", 'yours', "hadn't", 'ma', 'their', 'few', 'my', 'in', 'whom', "you've
", 'me', 'once', 'more', 'our', 'about', 'isn', 'and', 'between', 'while', "mightn't", 'be', 'will', "shouldn't", 'it', 'won', 'had', 'wher
e', 'they', "you'll", 'at', 'who', 'that', 'if', 'as', 'too', 'o', "didn't", 'been', "needn't", 'out', "you're", 'his', 'so', 'same', "shou
ld've", 'against', 'didn', 'only', 'hadn', 'haven', 'was', 'all', "wouldn't", 'down', 'a', 'of', "hasn't", "that'll", 'below', 'are', 'now'
, 'own', 'himself', 'other', 'mightn', 'd', 'but', "she's", 'having', 'itself', "don't", 'up', 'both', 'these', "weren't", "mustn't", 'ain'
, 'after', 'for', 'when', "couldn't", 'over', 'aren', 're', 'couldn', 'being', 'because', 'does', 's', "doesn't", 'not', 't', 'he', 'you',
'further', 'nor', 'what', 'is', "haven't", 'very', 'needn', 'myself', 've', 'again', 'ours', 'how', 'mustn', 'until', 'off', 'yourselves',
'doing', 'its', 'do', 'no', 'can', 'this', 'on', 'through', 'into', 'y', 'before', 'above', 'we', 'themselves', 'why', 'wouldn', 'there', '
shan', 'by', 'most', 'such', "isn't", "you'd", 'hasn', 'just', 'don', 'him', 'her'}
```

Figure 3: Sample Example of Stop words Removal

Step 3.3 sentence tokenization: The article will be broken down into sentences using the nltk library's built-in method. The sample example of sentence tokenization is as shown in figure3.



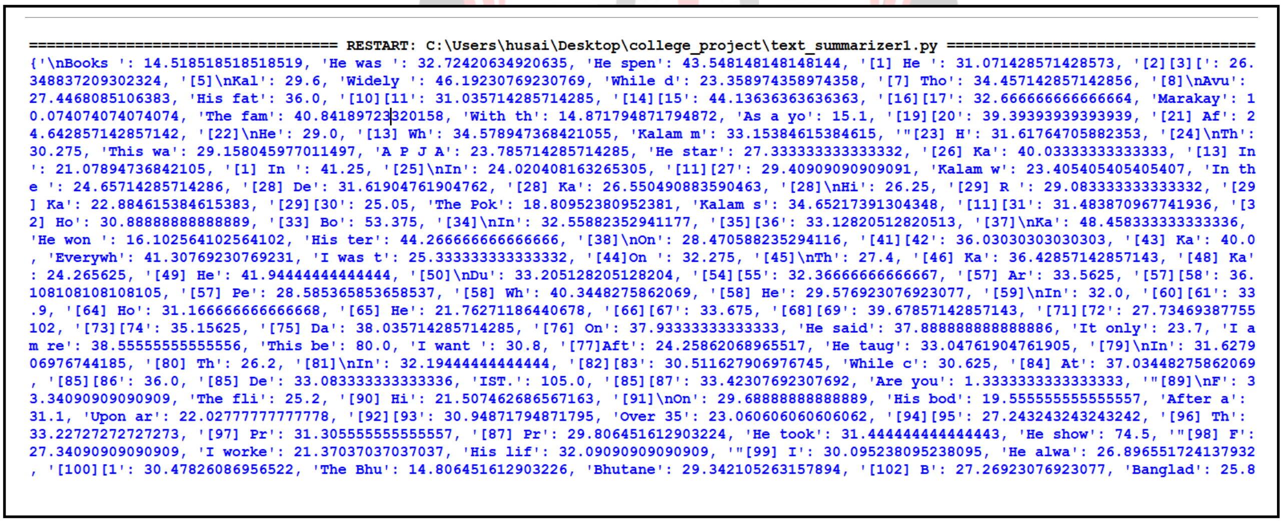
```
Python 3.10.5 (tags/v3.10.5:f377153, Jun  6 2022, 16:14:13) [MSC v.1929 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.

====== RESTART: C:\Users\husai\Desktop\college_project\text_summarizer1.py =====
['\nBooks and publications\n\nAvul Pakir Jainulabdeen Abdul Kalam (/ˈɑːbdəl kəˈlɑːm/ (listen); 15 October 1931\xa0– 27 July 2015) was an In
dian aerospace scientist and statesman who served as the 11th President of India from 2002 to 2007.', 'He was born and raised in Rameswaram
, Tamil Nadu and studied physics and aerospace engineering.', "He spent the next four decades as a scientist and science administrator, mai
nly at the Defence Research and Development Organisation (DRDO) and Indian Space Research Organisation (ISRO) and was intimately involved i
n India's civilian space programme and military missile development efforts.", '[1] He thus came to be known as the Missile Man of India fo
r his work on the development of ballistic missile and launch vehicle technology.', "[2][3][4] He also played a pivotal organisational, tec
hnical, and political role in India's Pokhran-II nuclear tests in 1998, the first since the original nuclear test by India in 1974.", '[5]\
nKalam was elected as the 11th president of India in 2002 with the support of both the ruling Bharatiya Janata Party and the then-oppositio
n Indian National Congress.', 'Widely referred to as the "People\'s President",[6] he returned to his civilian life of education, writing a
nd public service after a single term.', "He was a recipient of several prestigious awards, including the Bharat Ratna, India's highest civ
ilian honour.", 'While delivering a lecture at the Indian Institute of Management Shillong, Kalam collapsed and died from an apparent cardi
ac arrest on 27 July 2015, aged 83.', '[7] Thousands, including national-level dignitaries, attended the funeral ceremony held in his homet
own of Rameswaram, where he was buried with full state honours.', '[8]\nAvul Pakir Jainulabdeen Abdul Kalam was born on 15 October 1931, to
a Tamil Muslim family in the pilgrimage centre of Rameswaram on Pamban Island, then in the Madras Presidency and now in the State of Tamil
Nadu.', 'His father Jainulabdeen Marakayar was a boat owner and imam of a local mosque;[9] his mother Ashiamma was a housewife.', '[10][11]
[12][13] His father owned a ferry that took Hindu pilgrims back and forth between Rameswaram and the now uninhabited Dhanushkodi.', '[14][1
5] Kalam was the youngest of four brothers and one sister in his family.', '[16][17][18] His ancestors had been wealthy Marakayar traders a
nd landowners, with numerous properties and large tracts of land.', 'Marakayar are a Muslim ethnic group found in coastal Tamil Nadu and Sr
i Lanka who claim descent from Arab traders and local women.', 'The family business had involved trading groceries between the mainland and
the island and to and from Sri Lanka, as well as ferrying pilgrims between the mainland and Pamban.', 'With the opening of the Pamban Bridg
e to the mainland in 1914, however, the businesses failed and the family fortune and properties were lost by the 1920s, apart from the ance
stral home.', 'The family was poverty-stricken by the time Kalam was born.', "As a young boy he had to sell newspapers to add to the family
's meager income.", '[19][20][21]\nIn his school years, Kalam had average grades but was described as a bright and hardworking student who
had a strong desire to learn.', 'He spent hours on his studies, especially mathematics.', "[21] After completing his education at the Schwa
```

Figure 4: Sample Example of Sentence Tokenization

Step 3.Finding the weighted frequencies of the sentences:

Examining the frequency with which each term appears in the text will allow us to determine the overall grade for the article. Specifically, we will award points to sentences based on how often specific keywords appear within them. The sample example of weighted frequencies of the sentences is as shown in figure 5.



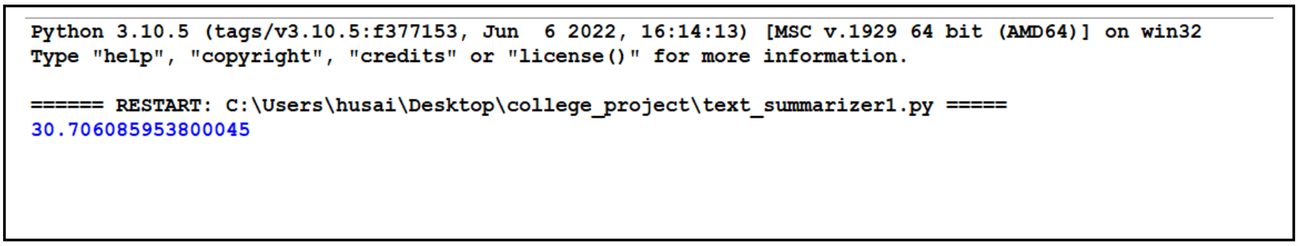
```
================================== RESTART: C:\Users\husai\Desktop\college_project\text_summarizer1.py ==================================
{'\nBooks ': 14.518518518518519, 'He was ': 32.72420634920635, 'He spen': 43.548148148148144, '[1] He ': 31.071428571428573, '[2][3][': 26.
348837209302324, '[5]\nKal': 29.6, 'Widely ': 46.19230769230769, 'While d': 23.358974358974358, '[7] Tho': 34.457142857142856, '[8]\nAvu':
27.4468085106383, 'His fat': 36.0, '[10][11': 31.035714285714285, '[14][15': 44.13636363636363, '[16][17': 32.666666666666664, 'Marakay': 1
0.074074074074074, 'The fam': 40.84189723320158, 'With th': 14.871794871794872, 'As a yo': 15.1, '[19][20': 39.39393939393939, '[21] Af': 2
4.642857142857142, '[22]\nHe': 29.0, '[13] Wh': 34.578947368421055, 'Kalam m': 33.15384615384615, '"[23] H': 31.61764705882353, '[24]\nTh':
30.275, 'This wa': 29.158045977011497, 'A P J A': 23.785714285714285, 'He star': 27.333333333333332, '[26] Ka': 40.03333333333333, '[13] In
': 21.07894736842105, '[1] In ': 41.25, '[25]\nIn': 24.020408163265305, '[11][27': 29.40909090909091, 'Kalam w': 23.405405405405407, 'In th
e ': 24.65714285714286, '[28] De': 31.61904761904762, '[28] Ka': 26.550490883590463, '[28]\nHi': 26.25, '[29] R ': 29.083333333333332, '[29
] Ka': 22.884615384615383, '[29][30': 25.05, 'The Pok': 18.80952380952381, 'Kalam s': 34.65217391304348, '[11][31': 31.483870967741936, '[3
2] Ho': 30.88888888888889, '[33] Bo': 53.375, '[34]\nIn': 32.55882352941177, '[35][36': 33.12820512820513, '[37]\nKa': 48.458333333333336,
'He won ': 16.102564102564102, 'His ter': 44.266666666666666, '[38]\nOn': 28.470588235294116, '[41][42': 36.03030303030303, '[43] Ka': 40.0
, 'Everywh': 41.30769230769231, 'I was t': 25.333333333333332, '[44]On ': 32.275, '[45]\nTh': 27.4, '[46] Ka': 36.42857142857143, '[48] Ka'
: 24.265625, '[49] He': 41.94444444444444, '[50]\nDu': 33.205128205128204, '[54][55': 32.36666666666667, '[57] Ar': 33.5625, '[57][58': 36.
108108108108105, '[57] Pe': 28.585365853658537, '[58] Wh': 40.3448275862069, '[58] He': 29.576923076923077, '[59]\nIn': 32.0, '[60][61': 33
.9, '[64] Ho': 31.166666666666668, '[65] He': 21.76271186440678, '[66][67': 33.675, '[68][69': 39.67857142857143, '[71][72': 27.73469387755
102, '[73][74': 35.15625, '[75] Da': 38.035714285714285, '[76] On': 37.93333333333333, 'He said': 37.888888888888886, 'It only': 23.7, 'I a
m re': 38.55555555555556, 'This be': 80.0, 'I want ': 30.8, '[77]Aft': 24.25862068965517, 'He taug': 33.04761904761905, '[79]\nIn': 31.6279
06976744185, '[80] Th': 26.2, '[81]\nIn': 32.19444444444444, '[82][83': 30.511627906976745, 'While c': 30.625, '[84] At': 37.03448275862069
, '[85][86': 36.0, '[85] De': 33.083333333333336, 'IST.': 105.0, '[85][87': 33.42307692307692, 'Are you': 1.3333333333333333, '"[89]\nF': 3
3.34090909090909, 'The fli': 25.2, '[90] Hi': 21.507462686567163, '[91]\nOn': 29.68888888888889, 'His bod': 19.555555555555557, 'After a':
31.1, 'Upon ar': 22.027777777777778, '[92][93': 30.94871794871795, 'Over 35': 23.060606060606062, '[94][95': 27.243243243243242, '[96] Th':
33.22727272727273, '[97] Pr': 31.305555555555557, '[87] Pr': 29.806451612903224, 'He took': 31.444444444444443, 'He show': 74.5, '"[98] F':
27.34090909090909, 'I worke': 21.37037037037037, 'His lif': 32.09090909090909, '"[99] I': 30.095238095238095, 'He alwa': 26.896551724137932
, '[100][1': 30.47826086956522, 'The Bhu': 14.806451612903226, 'Bhutane': 29.342105263157894, '[102] B': 27.26923076923077, 'Banglad': 25.8
```

Figure 5: sample example of weighted frequencies of the sentences.

Step 3.5 calculating the threshold of the sentences:

Finally, we'll average the sentences to determine which types of sentences can be summed up. We can eliminate sentences that did not meet our minimum standard by scoring them. The sample example of calculating threshold of the sentences is as shown in figure 6.

DOI : 10.35291/2454-9150.2022.0488

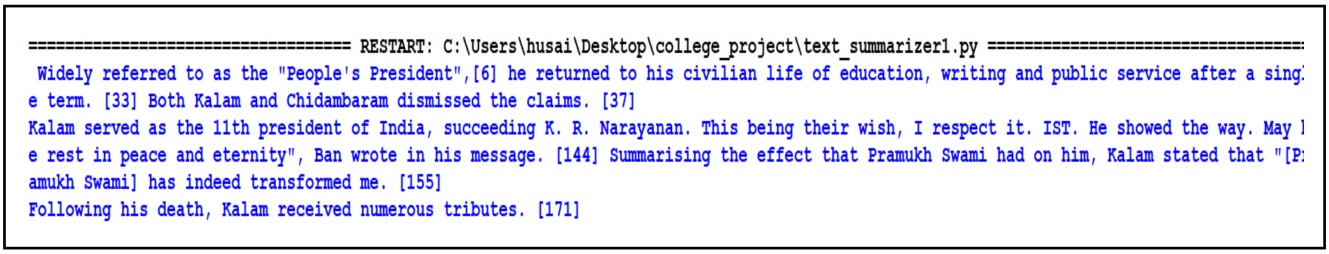


```
Python 3.10.5 (tags/v3.10.5:f377153, Jun  6 2022, 16:14:13) [MSC v.1929 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.

====== RESTART: C:\Users\husai\Desktop\college_project\text_summarizer1.py =====
30.706085953800045
```

Figure 6: Sample Example of Calculating Threshold of Sentences.

Step 3.6 summary generation: We can create an article summary now that we have all the required inputs. The sample example of generated summary is as shown in figure 7.



```
================================== RESTART: C:\Users\husai\Desktop\college_project\text_summarizer1.py ==================================
 Widely referred to as the "People's President",[6] he returned to his civilian life of education, writing and public service after a sing
e term. [33] Both Kalam and Chidambaram dismissed the claims. [37]
Kalam served as the 11th president of India, succeeding K. R. Narayanan. This being their wish, I respect it. IST. He showed the way. May
e rest in peace and eternity", Ban wrote in his message. [144] Summarising the effect that Pramukh Swami had on him, Kalam stated that "[P
amukh Swami] has indeed transformed me. [155]
Following his death, Kalam received numerous tributes. [171]
```

Figure 7: Sample Example of generated summary

IV. CONCLUSION

Even though text summarization is an old field of study, new studies in the field are constantly being published. However, text summarization produces mediocre results in general, and the summaries it generates are not always satisfactory. As a result, researchers are working to improve current text summarization methods. It is also critical to develop new summarization methods that can meet human standards and produce more reliable summaries. As a result, ATS should be integrated with other intelligent systems to improve its effectiveness. Automatic text summarization is a prominent area of research that has been widely implemented and integrated into a wide range of applications to summaries and reduce text volume. In this research work we are giving emphasis on the article summarization on national leaders. We got satisfactory result.

REFERANCES

[1]Ishitva Awasthi et.al., "Natural Language Processing (NLP) based Text Summarization - A Survey", *Proceedings of the Sixth International Conference on Inventive Computation Technologies IEEE Xplore* .2021.

[2]Pradeepika Verma et.al., "A Review on Text Summarization Techniques", *Journal of Scientific Research*, Volume 64, Issue 1, 2020.

[3] Deepali K. Gaikwad et.al., "A Review Paper on Text Summarization", *International Journal of Advanced Research in Computer and Communication Engineering* Vol. 5, Issue 3, 2016

[4] R. Mishra, J. Bian, M. Fiszman, C. R. Weir, S. Jonnalagadda, J. Mostafa, and G. Del Fiol, "Text summarization in the biomedical domain: a systematic review of recent research*," Journal of biomedical informatics*, vol. 52, pp. 457–467, 2014.

[5]] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, "Text summarization techniques: a brief survey*," arXiv preprint arXiv:*1707.02268, 2017

[6] Ujjwal Rani, "Review paper on automatic text summarization"., *Int. Res. J. Eng. Technol*. 3349–3354,2019.

[7] Neelima, G.V.M., "Extractive text summarization using deep natural language fuzzy processing", *Int. J. Innov. Tech. Explor. Eng.*, 990–993, 2019.

[8] C. Saranyamol and L. Sindhu, "A survey on automatic text summarization," *International Journal of Computer Science and Information Technologies*, vol. 5, no. 6, pp. 7889–7893, 2014.

[9] M. Gambhir and V. Gupta, "Recent automatic text summarization techniques: a survey," *Artificial Intelligence Review*, vol. 47, no. 1, pp. 1–66, 2017.

[10]] S. Gholamrezazadeh, M. A. Salehi, and B. Gholamzadeh, "A comprehensive survey on text summarization systems",*2nd International Conference on Computer Science and its Applications. IEEE*, 2009, pp. 1–6.

[11] S. K. Bharti and K. S. Babu, "Automatic keyword extraction for text summarization: A survey," *arXiv preprint arXiv*:1704.03242, 2017.

[12] L. Abualigah, M. Q. Bashabsheh, H. Alabool, and M. Shehab, "Text summarization: a brief review," *Recent Advances in NLP: The Case of Arabic Language*, pp. 1–15, 2020

[13] Article accessed from: https://www.edsys.in/freedom-fighters-of-india/

DOI : 10.35291/2454-9150.2022.0488