ISSN: 2366-1313

Text Summarizer using NLP (Natural Language Processing)

G. Bhargavi [1], S. Sailaja [2]
[1] Associate professor, Associate Professor [2]
Department of Computer Science Engineering
RISE Krishna Sai Prakasam Group of Institutions

Abstract- Enormous amounts of information are available online on the World Wide Web. To access information from databases, search engines like Google and Yahoo were created. Because the amount of electronic information is growing every day, the real outcomes have not been reached. As a result, automated summarization is in high demand. Automatic summary takes several papers as input and outputs a condensed version, saving both information and time. The study was conducted in a single document and resulted in numerous publications. This report focuses on the frequency-based approach for text summarization.

Keywords: Automatic summarization, Extractive, Natural Language Processing, frequency-based

INTRODUCTION

Text summary is the way of selecting important points from the provided article or a document that can be reduced by a program. As the data overload problem increased, so did the interest in capturing the text as the amount of data increased. Summarizing a large document manually is challenging since it requires a lot of human effort and is time-consuming. There are mainly two methods for summarizing the text document that can be done by using extractive and abstractive techniques. Extractive summaries concentrate on selecting important passages, sentences, words, etc. from the primary text and connecting them into a concise form. The importance of critical sentences is concluded on the basis of analytical and semantic features of the sentences [1].

Summary systems are usually based on sentence delivery methods and for understanding the whole document properly as well as for extracting the important sentences from the document. The technique of generating a brief description that comprises a few phrases that describe the key concepts of an article or section is known as abstractive summarization. This function is also included to naturally map the input order of words in a source document to the target sequence of words called the summary [2].

LITERATURE SURVEY

The Internet is a vast source of electronic information. But the result of information acquisition becomes a tedious task for people. Therefore, automated summaries began the search for automatic retrieval of data from documents using our precious time. H.P. Luhn was the first to invent an

ISSN: 2366-1313

automatic summary of the text in 1958. There are helpful ways to produce a. summary-extraction and abstraction. Extraction is independent of the domain and takes key sentences and provides a summary on the other hand, abstracting depends on the domain and taking personal information by understanding the entire text and adjusting the policy to produce a summary. There are several methods that use different methods to obtain a summary of a text [3–10].

Frequency Based Approach Term Frequency (TF)

TF mainly determines that how often a word appears in a text document and it is considered to be an important factor. The paragraphs in the document are divided into sentences based on the punctuation marks that appears at the end of every sentence.

Keyword Frequency

The high frequency words in the sentence are known as keyword. It measures the frequency for every word once you've refined the content. Keywords are the terms that have the most important frequency. The word score is organized as a keyword, and the phrase is given some fixed points for each keyword found in the text based on this feature. Stop Words Filtering Any document will have a lot of words that appear regularly but do not give the document less or more meaning. Words like 'on', 'the', 'is' and 'and' appear frequently in the English language and there are many examples of many texts. While searching, these words do not add up value to the information when users submit a query. Clustering Approach K-means Clustering This approach aims to classify n observed in k groups where each recognition belongs to a category with a descriptive meaning, acting as a collective example. k-means can be applied to data with small size, is numerical, and continuous. The applications that can be benefited by the k-means algorithm are public transport data analysis, targeting crime hotspots, insurance fraud detection, customer segregation, document collection, etc.

PROBLEM STATEMENT

According to recent studies, each and every day there are about 18,000,000 pages read every day, that is 18 million. it may be a book you are reading online or may be the 10s of thousands of emails and papers that the researchers, scientist and people in admin have to read every day and well reading for you might be a pleasure but for the people that have to go through 1000s of email each day it can be quite tiring. These days, we are gaining instant access to more information. On the other hand, this information is not required nor relevant and therefore does not communicate the intended message. Suppose you're seeking specific information in an internet news item, for example. In that case, you should spend some time examining the material and removing irrelevant information before you locate what you're looking for.

PROPOSED SYSTEM

We have used NLP, which seeks to summarize articles by picking a collection of words that hold the most essential information, can address this problem with the help of extractive summarizer. This

ISSN: 2366-1313

approach takes a significant portion of a phrase and utilizes it to create a summary. To define sentence verbs and subsequently rank them in terms of significance and similarity, a variety of algorithms and approaches are utilized. There is a great need for text summary techniques to address the amount of text data available online to help people find the right information and use the right information quickly. In addition, the implementation of text summaries reduces reading time, speeds up the process of researching information, and increases the information that may not be in one field.

APPROACH

This research paper focuses on the frequency-based approach for text summarization. The steps involved in text summarizer are Sentence and word tokenization and then calculating sentence score on the basis of TF-IDF score which is being used to select the most important sentences to retain the information and merge it to form a summary Figure 1 [11–15].

Step 1: Import all necessary libraries [5] NLTK (Natural Language toolkit) is a widely used library while we are working with text in python. Stop words contain a list of English stop words, which need to be removed during the preprocessing step shown in Figure 2.

Step 2: Generate Clean Sentences Text processing is the most important step in achieving a constant and positive approach result. The processing steps removes special digits, word, and characters as shown in Figure 3.

Step 3: Calculate TF-IDF and generate a matrix We'll find the TF and IDF for each word in a paragraph. TF (t) = (Frequency of t from document)/(total_no. of t in the document) IDF (t) = log_e (total_no. Of documents/No. of documents with t it) [4] Now, we will be generating a new matrix after multiplying the calculated TF and IDF values as shown in Figure 4.

Step 4: Score the sentences Here, we use TF-IDF word points in a sentence to give weight to a paragraph. However, Sentence scoring varies with different algorithms as shown Figure 5.

Step 5: Generate the summary This is the last stage of text summarization [16]. Top sentences are calculated based on the score and retention rate given to the user are included in the summary and finally, a summary is created as shown in Figure 6 [17–20].

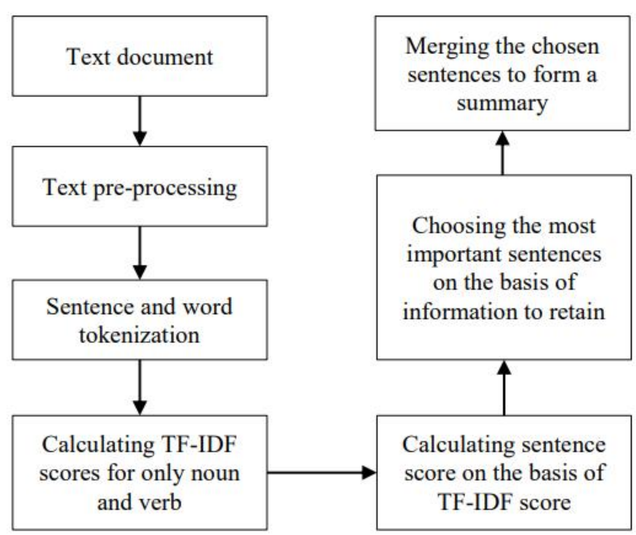

Figure 1. Frequency-based approach.

ISSN: 2366-1313

```
In [3]: from sklearn.feature_extraction.text import TfidfVectorizer
        from spacy.lang.en import English
        import numpy as np

In [4]: nlp = English()
        nlp.add_pipe(nlp.create_pipe('sentencizer'))

In [5]: text_corpus = """
        On the morning of Sept. 22, employees at Great Big Story received some good news. The digital video publis
        The announcement was especially encouraging because many employees had believed the automative brand would
        The feelings of relief, however, were short-lived.
        At 6:45 p.m. that evening, Great Big Story employees received an email from CNN vp of digital productions
        Despite being set for the usual meeting time, the invite featured a few irregularities. For starters, all-
        "Immediately everyone was texting everyone, and we all went to bed that night knowing we were going to get
        For many Great Big Story employees, the announcement on Sept. 23 that CNN was shutting down the company wa
        "A way that Great Big Story was explained to me in the early days was this is the place where you get to m
        """
```

Figure 2. Preprocessing of data

```
In [6]: doc = nlp(text_corpus.replace("\n", ""))
        sentences = [sent.string.strip() for sent in doc.sents]

In [7]:
        print("Senetence are: \n", sentences)

        Senetence are:
         ['On the morning of Sept. 22, employees at Great Big Story received some good news.', 'The digital video publisher's bigge
        dvertiser, Hyundai-owned Genesis, had signed a new sponsorship deal worth more than $1 million, they were told during their
        ular 9:30 a.m. morning meeting.', 'The announcement was especially encouraging because many employees had believed the auto
        ve brand would not be renewing its deal and the lost revenue would put Great Big Story in an even more precarious financial
        uation than they feared the CNN-owned company may already be in.', 'The feelings of relief, however, were short-lived.', 'A
        45 p.m. that evening, Great Big Story employees received an email from CNN vp of digital productions Courtney Coupe, a four
        member of Great Big Story who oversaw the media company.', 'The email notified the employees that an all-hands meeting woul
        scheduled for the following morning at 9:30 a.m.  Despite being set for the usual meeting time, the invite featured a few i
        ularities.', 'For starters, all-hands meetings scheduled with short notice are notorious within the media industry as a si
        at bad news is imminent and layoffs are likely.', 'But this meeting was also set to be attended by Coupe, who had not atter
        Great Big Story's morning meetings for some time, and CNN evp and chief digital officer Andrew Morse, another founding memb
        f Great Big Story, who never attended the morning meeting, according to multiple former employees.', 'And the meeting was s
        uled to take place in Coupe's virtual meeting room, which was not the usual location.', '"Immediately everyone was texting
        yone, and we all went to bed that night knowing we were going to get let go that next day," said one of 11 former Great Bi
        ry employees that Digiday spoke to for this article.', 'For many Great Big Story employees, the announcement on Sept. 23 th
        NN was shutting down the company was a shock.', 'Working at Great Big Story had been a dream job.', 'Editorial employees we
        ble to produce Emmy and Webby Award-winning short-form documentaries about subjects like an Oakland center for disabled art
        and an organization working to combat plastic pollution and poverty simultaneously.', 'And employees on the business side w
        able to support content that they enjoyed watching themselves and were proud to show to sponsors and their own family membe
        s.', 'It's unclear how many people worked at Great Big Story at its peak, but a website built to showcase Great Big Story e
        yees who lost their jobs lists 45 employees.', '"A way that Great Big Story was explained to me in the early days was this
        he place where you get to make that story that you always wanted to make, but your boss said you couldn't.', 'And that was
        e," said a former employee.']
```

Figure 3. Result after preprocessing

```
In [10]: # Let's now create a tf-idf (Term frequncy Inverse Document Frequency) model
         tf_idf_vectorizer = TfidfVectorizer(min_df=2,  max_features=None,
                                             strip_accents='unicode',
                                             analyzer='word',
                                             token_pattern=r'\w{1,}',
                                             ngram_range=(1, 3),
                                             use_idf=1,smooth_idf=1,
                                             sublinear_tf=1,
                                             stop_words = 'english')

In [11]: # Passing our sentences treating each as one document to TF-IDF vectorizer
         tf_idf_vectorizer.fit(sentences)

Out[11]: TfidfVectorizer(analyzer='word', binary=False, decode_error='strict',
                         dtype=<class 'numpy.float64'>, encoding='utf-8',
                         input='content', lowercase=True, max_df=1.0, max_features=None,
                         min_df=2, ngram_range=(1, 3), norm='l2', preprocessor=None,
                         smooth_idf=1, stop_words='english', strip_accents='unicode',
                         sublinear_tf=1, token_pattern='\\w{1,}', tokenizer=None,
                         use_idf=1, vocabulary=None)
```

Figure 4. Calculation of TF-IDF

```
# Ordering our top-n sentences in their original ordering
mapped_top_n_sentences = sorted(mapped_top_n_sentences, key = lambda x: x[1])
ordered_scored_sentences = [element[0] for element in mapped_top_n_sentences]

# Our final summary
summary = " ".join(ordered_scored_sentences)

Our top_n_sentence with their index:

('At 6:45 p.m. that evening, Great Big Story employees received an email from CNN vp of digital productions Courtney Coupe, a f
ounding member of Great Big Story who oversaw the media company.', 4)
('But this meeting was also set to be attended by Coupe, who had not attended Great Big Story's morning meetings for some time,
and CNN evp and chief digital officer Andrew Morse, another founding member of Great Big Story, who never attended the morning
meeting, according to multiple former employees.', 7)
('The email notified the employees that an all-hands meeting would be scheduled for the following morning at 9:30 a.m.  Despite
being set for the usual meeting time, the invite featured a few irregularities.', 5)

In [17]: print("Summary: \n", summary)

Summary:
 At 6:45 p.m. that evening, Great Big Story employees received an email from CNN vp of digital productions Courtney Coupe, a fo
unding member of Great Big Story who oversaw the media company. The email notified the employees that an all-hands meeting woul
d be scheduled for the following morning at 9:30 a.m.  Despite being set for the usual meeting time, the invite featured a few
irregularities. But this meeting was also set to be attended by Coupe, who had not attended Great Big Story's morning meetings
for some time, and CNN evp and chief digital officer Andrew Morse, another founding member of Great Big Story, who never attend
ed the morning meeting, according to multiple former employees.
```

Figure 5. Result: Summarized Text

CONCLUSION

Text summaries have been shown to be useful for natural language processing tasks such as question and answer or other related fields of computer science such as text classification and data retrieval. And access time for information search will be improved. At the same time, sequencing enhances the effect and its algorithms are less biased than human creams. Using a text summary system, commercial capture services allow users to increase the number of texts they can process.

FUTURE SCOPE

In this section, we will list some of the future extensions for this study. In this article, we focused on summarizing news articles under the auspices of sports and technology. The strategies proposed here are flexible in some domains. One of the future plans would be to use an overview framework that focuses on the topic in news articles or blogs and to increase work on machine-dependent methods. Summaries focused on the headline article can be very accurate and very important for users. It would be even more interesting to work on topic modeling and summarizing in the future media domain.

REFERENCES

1. T. Kumar, "Automatic Text Summarization," Rourkela, 2014.

2. P.J. Patel, "https://machinelearningmastery.com/gentle-introduction-text-summarization/," Interna tional Journal Of Engineering And Computer Science, p. 5, 2015.

3. A. Jain, "Automatic Extractive Text Summarization using TF-IDF," 1 April 2019.

[Online]. Available: https://medium.com/voice-tech-podcast/automatic-extractive-text-summarization-using -tfidf-3fc9a7b26f5.

4. A. Panchal, "NLP—Text Summarization using NLTK: TF-IDF Algorithm," 10 June 2019. [Online]. Available: https://towardsdatascience.com/text-summarization-using-tf-idf-e64a0644ace3.

5. M. Mayo, "Getting Started with Automated Text Summarization," November 2019. [Online]. Available: https://www.kdnuggets.com/2019/11/getting-started-automated-text-summarization.html.

6. J. Brownlee, "A Gentle Introduction to Text Summarization," 7 August 2019. [Online]. Available: https://machinelearningmastery.com/gentle-introduction-text-summarization/.

7. A. Opidi, "A Gentle Introduction to Text Summarization in Machine Learning," 15 April 2019. [Online]. Available: https://blog.floydhub.com/gentle-introduction-to-text-summarization-inmachine-learning/.

8. H. Darji, "Text Summarization-Key Concepts," 8 January 2020. [Online]. Available: https://medium.com/@harshdarji_15896/text-summarization-key-concepts-23df617bfb3e.

9. J.M.a.O.D.P. Conroy, "Text summarization via hidden markov models," Proceedings of SIGIR '01, 2001.

10. D.M.D.W. Changjian Fanga, "Word-sentence co-ranking for automatic extractive text summarization," 5 March 2016. [Online]. Available: https://www.sciencedirect.com/science/ article/abs/pii/S0957417416306959?via%3Dihub.

11. "Recent automatic text summarization techniques: a survey," 29 March 2019. [Online]. Available: https://link.springer.com/article/10.1007/s10462-016-9475-9.

12. L.G. Gupta V, "A survey of text summarization extractive techniques," J Emerg Technol Web Intell, pp. 258–268, 2010.

13. T.U.K.M.B.C.Z. Chen, "Automatic Summarization Based on Sentence Extraction: A Statistical Approach," International Journal of Applied Electromagnetics and Mechanics, vol. 13, pp. 19–23, 2002. 14. R.a.R.D.R. McKeown, "Generating summaries of multiple news articles," in Generating summaries of multiple news articles, Seattle, 1995.

15. A.I.Z.A.a.Y.Y.C.S.P. Yong, "A Neural Based Text Summarization System," in 6th International Conference of Data Mining, 2005.

16. R.A. Rasim Alguliev, "Evolutionary Algorithm for Extractive Text Summarization," Intelligent Information Management, vol. 1, pp. 128–138, 2009.

17. V.H.L.S.&. H.R. Qazvinian, " Summarising text with a genetic algorithm," Int. J. Knowledge Management Studies, vol. 2, no. 4, pp. 426–444, 2008.

18. A.G. Sonali Behal, "Automatic Text Summarization using Natural Language," 2019.

19. G.S.L. Vishal gupta, "A Survey of Text Summarization Extractive Techniques," Journal of Emerging Technologies in Web Intelligence, vol. 2, no. 3, 2010.

20. R.S. P.a.U. Kulkarni, "Implementation and Evaluation of Evolutionary Connectionist," Journal of Computer Science 6, vol. 6, no. 11, pp. 1366–1376, 2010.