Text Summarization in Python Using Natural Language Processing

Sagar Mohite
Research Scholar
Dept of Computer Engineering Bharati Vidyapeeth (Deemed to be University) College of Engineering,
Pune-411043, India
email :sgmohite@bvucoep.edu.in

Sachin Wakurdekar
Research Scholar
Dept of Computer Engineering Bharati Vidyapeeth (Deemed to be University) College of Engineering,
Pune-411043, India
email :sbwakurdekar@bvucoep.edu.in

Shivam Kumar Srivastava
Research Student
Dept of Computer Engineering Bharati Vidyapeeth (Deemed to be University) College of Engineering,
Pune-411043, India
email : srivastavas168@gmail.com

Vishesh Rai
Research Student
Dept of Computer Engineering Bharati Vidyapeeth (Deemed to be University) College of Engineering,
Pune-411043, India
email :visheshrai87@gmail.com

Yash Srivastava
Research Student
Dept of Computer Engineering Bharati Vidyapeeth (Deemed to be University) College of Engineering,
Pune-411043, India
email :yashsri1510@gmail.com

Pratyush Ranjan
Research Student
Dept of Computer Engineering Bharati Vidyapeeth (Deemed to be University) College of Engineering,
Pune-411043, India
email:pratyushstunner@gmail.com

Shivaansh Agarwal
Research Student
Dept of Computer Engineering Bharati Vidyapeeth (Deemed to be University) College of Engineering,
Pune-411043, India
email :agarwal.shivansh620@gmail.com

Rahul Raj
Research Student
Dept of Computer Engineering Bharati Vidyapeeth (Deemed to be University) College of Engineering,
Pune-411043, India
email :rahulraj28121998@gmail.com

ABSTRACT

Broad amount of information is accessible online on the internet. The purpose of search engines like Google and Bing is to retrieve data from archives.The actual results have not yet been reached due to the rapid growth in number of computerized data. Subsequently, the default summary is much needed. The default summary takes several pages as insert and output an abridged version, saving both information and time.

In the current age, where immense amount of information is present online, it is much important to give the improved mechanism to draw out the information quickly and more efficiently. It is very tough that people can draw out a summary of large text documents on their own. So, there is a problem in inquiring for appropriate document from all the documents available, and consuming relevant information from it. In order to solve the problem above, the automatic text summarization is very much needed. The practice of extracting the most significant and meaningful information from a document or group of related papers and condensing it while retaining its overall meaning is known as text summarization.

Natural Language Processing is defined as the capability of a computer program to acknowledge human language when speaking and writing a language. Natural Language Processing is one of the key components of practical wisdom. It has many applications in various fields and we use it in many real-world applications such as Business Intelligence medical research and much more. We have been using Natural Language Processing for 50 years now.

The research was done in one volume and accumulated in many volumes. This research is centered on an idea build on the abbreviations of text summaries.

Keywords—Automatic Summarization, Extractive, NLP, frequency-based

I.INTRODUCTION

A text summary is a way to select key points in a given topic or document that can be reduced by program. As the problem of data overload grew, so did the interest in photography as the amount of data increased. Summarizing a large document by hand is a tough task as it needs a lot of human effort and time. There are mainly two methods for summarizing the text document that can be done by using extractive and abstractive techniques. Releasing abbreviations focus on selecting key passages, sentences, words, etc. in the native text and integrating it into a short description. The significance of critical sentences is determined on the basis of analytical and semantic features of sentences.

Summary systems are usually based on sentence delivery methods and comprehension of the whole document as well as extracting important sentences from the text.The Method of producing a brief description that includes a few phrases that describe the main ideas of an article or section is known as abstract summary.

This function is also included to automatically map the alphabetical order of the source document into a specific word order called an Abbreviation.

II. LITERATURE SURVEY

The World Wide Web is a great origin for electronic information. But the result of gaining knowledge turns out to be a daunting experience for humans. Therefore, the default summaries automatically began searching to retrieve information from documents saving our valuable time.

The first to create an automated version of the text was H.P. Luhn in 1958.There are practical ways to produce an abstract – background as well as abstraction. Extraction is not dependent on the domain and collating key sentences and providing a summary. Contrastingly, extraction depends on the domain and takes personal information by understanding the entire text and adjusting the policy to produce a summary. There are many methods that use different methods to get a summary of a text. The compression rate used in the extraction procedure determines the length of the summary and the source text. The summary is more detailed and contains more irrelevant material the higher the compression rate. The summary will be shorter and more info will be lost the lower the compression rate.Ideal vale for compression rate is 5-30%

II.A Frequency based approach:

- Term frequency (TF):

TF determines primarily how often a word appears in the text
and is considered important. Paragraphs are divided into sentences based on punctuation from the end of each 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 words that have a very important frequency. The word school is classified as a keyword, and a sentence is given a fixed point 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.

II.2 Clustering approach

- K-means clustering:

This method aims to classify the n marked in groups k where
each recognition belongs to a descriptive class, acting as a collective example.
k-methods can be applied to small, numerical, and continuous
data. Apps that can benefit from the k-means algorithm to analyze public transport data, targeted crime sites, insurance fraud detection, customer segregation, document collection, etc.

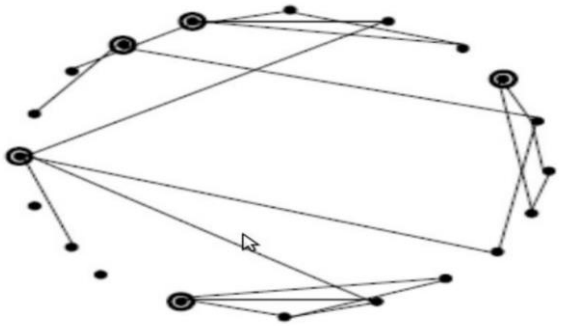
II.3 Graph theoretic approach

In this approach, for every input sentence there is a node. If different sentences share some common words, then they are said to be connected to the edge. Therefore, their similarly must be more than a minimum threshold. Two end results are provided by this approach.

First of these results is that, certain sub graphs are not connected to other sub graphs, known as partitions contained in the graph, different from topics covered in the given document. Second result is provided by graph-theoretic method. Here,

significant sentences in the document are identified. High cardinal nodes (nodes connected to high no. of edges) in a partition are considered important sentences, and thus are preferred to be included in the final summary.
An example of a graph report is shown in figure below. In the document there are about 3-4 topics and encircled nodes represents the informative sentences, as they exchange content and information with various sentences in the document. Thus, inter and intra document similarity can be easily visualized by graph theoretic method.

A.
1)
a)
B.



Graph Theoretic Approach

II.4 Machine Learning approach
In this approach, the educational dataset is used for testimonial and the process of briefing is modelled as a categorization problem: based on the attributes that they own sentences are categorized as (1) abstract sentences and (2) non-abstract sentences.

$$P(s \in <S \mid F1, F2, ..., FN) = P(F1, F2, ..., FN \mid s \in S) * P(s \in S) / P(F1, F2, ..., FN)$$

Applying Bayes' rule, the ranking possibilities are deduced analytically from the educational data, where s is a sentence from the document compilation and F1, F2, and FN are categorization-related factors. P (s∈S | F1, F2..., FN) is the likelihood that sentences will be chosen to create the summary given that they contain features F1, F2...FN. S is the summary that will be generated.

III.PROBLEM STATEMENT

According to recent studies, every day some 18,000,000 pages are read each day, which means 18 million. it can be a book you read online or 10 thousand emails and papers that researchers, scientists and people in charge should read every day and reading it well can be fun but for people who have to go over 1000 emails each day it can be tiring.

These days, we get quick access to more information. On the other hand, this information is neither necessary nor essential and therefore, does not address the intended message.

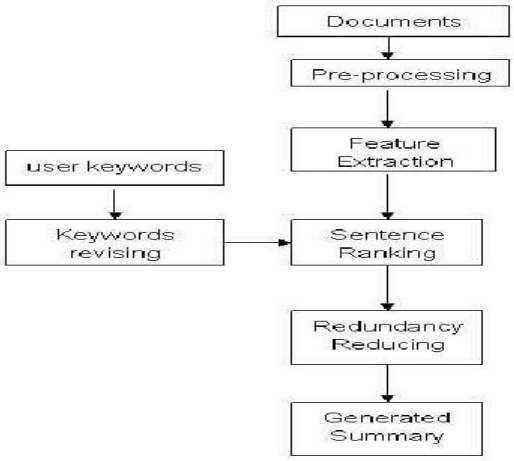
Suppose you were looking for information on an online news item, for example. If so, you should spend some time examining the equipment and deleting important information before you find what, you are looking for.

IV.PROPOSED SYSTEM

Using NLP, which seeks to summarize articles by selecting a collection of words that contain the most important information, we can address this issue with the help of an extracting summary. This method takes an important part of the phrase and uses it to form a summary. In order to define sentence verbs and later arrange them in order of importance and similarity, various algorithms and methods are used.

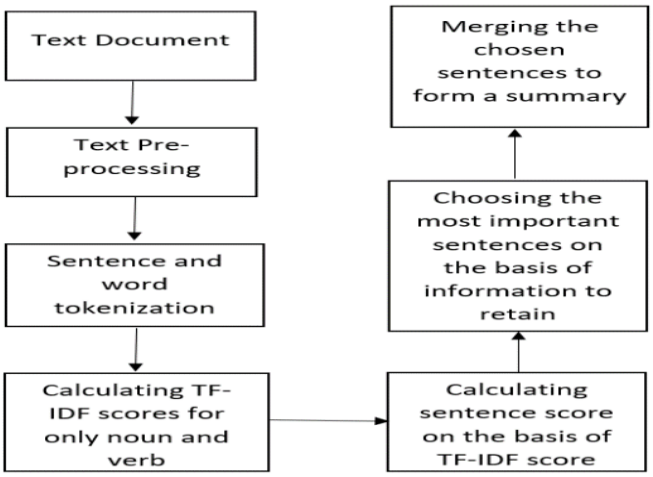
There is a great need for text summary techniques to address the amount of text data available online to help people find the right information and use the right information faster. Additionally, the use of text abbreviations reduces reading time, speeds up the process of researching information, and expands information that may not be in one field.

V. SYSTEM ARCHITECTURE



VI.APPROACH

This research paper focuses on a multidisciplinary approach to summarizing the text.Steps involved in abbreviating text to sentence and wording and then calculating sentence points based on TF-IDF points used to select the most important sentences to store information and combine it to form a summary.



STEP-1: Import all necessary libraries

Natural Language Toolkit (NLTK) is a library that is widely used when working with text in python. Terminals contain a list of English stop words, which need to be removed during the previous processing step.

```
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 publisher's biggest advertise

        The announcement was especially encouraging because many employees had believed the automotive brand would not be renewing its de

        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 Courtney Coupe, a founc

        Despite being set for the usual meeting time, the invite featured a few irregularities. For starters, all-hands meetings schedule

        "Immediately everyone was texting everyone, and we all went to bed that night knowing we were going to get let go that next day,"

        For many Great Big Story employees, the announcement on Sept. 23 that CNN was shutting down the company was a shock. Working at (

        "A way that Great Big Story was explained to me in the early days was this is the place where you get to make that story that you
        """
```

STEP-2: Generate clean sentences

Text processing is a very important step in achieving a continuous and positive performance result. Processing steps remove special digits, name, and letters.

```
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 biggest a
dvertiser, Hyundai-owned Genesis, had signed a new sponsorship deal worth more than $1 million, they were told during their reg
ular 9:30 a.m. morning meeting.', 'The announcement was especially encouraging because many employees had believed the automati
ve brand would not be renewing its deal and the lost revenue would put Great Big Story in an even more precarious financial sit
uation than they feared the CNN-owned company may already be in.', '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 Courtney Coupe, a founding
member of Great Big Story who oversaw the media company.', '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 irreg
ularities.', 'For starters, all-hands meetings scheduled with short notice are notorious within the media industry as a sign th
at bad news is imminent and layoffs are likely.', '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 o
f Great Big Story, who never attended the morning meeting, according to multiple former employees.', 'And the meeting was sched
uled to take place in Coupe's virtual meeting room, which was not the usual location.', '"Immediately everyone was texting ever
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 Big Sto
ry employees that Digiday spoke to for this article.', 'For many Great Big Story employees, the announcement on Sept. 23 that C
NN was shutting down the company was a shock.', 'Working at Great Big Story had been a dream job.', 'Editorial employees were a
ble to produce Emmy and Webby Award-winning short-form documentaries about subjects like an Oakland center for disabled artists
and an organization working to combat plastic pollution and poverty simultaneously.', 'And employees on the business side were
able to support content that they enjoyed watching themselves and were proud to show to sponsors and their own family member
s.', 'It's unclear how many people worked at Great Big Story at its peak, but a website built to showcase Great Big Story emplo
yees who lost their jobs lists 45 employees.', '"A way that Great Big Story was explained to me in the early days was this is t
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 tru
e," said a former employee.']

In [8]: # Let's create an organizer which will store the sentence ordering to later reorganize the
        # scored sentences in their correct order
        sentence_organizer = {v:k for v,k in enumerate(sentences)}
```

STEP-3: Create a matrix by calculating the TF-IDF

Each word in a paragraph will have its TF and IDF determined.
The formula for TF (t) is (Frequency of t from document) / (Total Number of t in document).
IDF (t) is equal to log e (total number of documents / number of documents containing t it).
Now, we will be producing a new matrix after multiplying the calculated TF and IDF values.

```
In [10]: # Let's now create a tf-idf (Term frequnecy 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)
```

STEP-4: Score the sentences

Here, we use TF-IDF word points in a sentence to give a paragraph weight. However, sentence points vary by different algorithms.

```
In [13]: # Getting sentence scores for each sentences
         sentence_scores = np.array(sentence_vectors.sum(axis=1)).ravel()

         # Sanity checkup
         print(len(sentences) == len(sentence_scores))

         True
```

STEP-5: Generate the summary

This is the last stage of text summarization. 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.

```
# 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.
```

VII.CONCLUSION

Text summaries have been demonstrated to be helpful in tasks involving natural language processing, such as question-and-answer exchanges, as well as in areas of computer science that are closely related, such text fragmentation and data retrieval.And access to information search time will be improved. At the same time, sequence enhances the effect and its algorithms are slightly biased than human creams. By using a text summary system, commercial recording services allow users to enhance the number of texts they can process.

VIII.FUTURE SCOPE

In this section, we will list some of the future extensions for this study. In this article, we focus on summarizing the news headlines under the auspices of sports and technology. The strategies proposed here are adapted to the conditions in other domains. One of the future plans will be to use an overview framework focused on the topic in news articles or blogs and to expand the work on machine-based approaches. Summaries focused on the article can be very accurate and very important for users. It would be even more interesting to work on modelling the title and summarizing the future media domain.

IX.REFERENCES

[1] Adhika Widyassari, S. R. (2020). "Review of automatic text summarization techniques & methods. *Journal of King Saud University - Computer and Information Sciences*, 18."

[2] Amigó E, G. J. (2005)." A framework for the evaluation of text summarization systems. *proceedings of the 43rd annual meeting on association for computational linguistics."* ACL '05.

[3] Antiqueira L, O. O. (2007)."A complex network approach to text summarization. *Information Sciences*."

[4] B. Cretu, Z. C. (2002). "Automatic summarization based on sentence extraction. *International Journal of Applied Electromagnetic and mechanics*."

[5] Changjian Fanga, D. M. (2016, March 5). "*Word-sentence co-ranking for automatic extractive text summarization*."

[6] Conroy, J. M. (2001). "Text summarization via hidden markov models. *Proceedings of SIGIR '01*."

[7] D. Gillick, K. R. (2009). "A global optimization network for meeting summarization. *Proc. IEEE Int. Conf. Acoust*, 1-4."

[8] Darji, H. (2020, January 8). *Text Summarization-Key Concepts*.

[9] Evans, D. K. (2005). "*Similarity-based multilingual multidocument summarization.* Technical Report CUCS-014-05."

[10] Gupta V, L. G. (2010). "A survey of text summarization extractive techniques. *J Emerg Technol Web Intell*, 258-268."

[11] J.Patel,P. (2015). *International Journal Of Engineering And Computer Science*, 5.

[12] Jain, A. (2019, April 1). "*Automatic Extractive Text Summarization using TF-IDF*."

[13] KS, J. (2007). "Automatic summarising: the state of the art. *Inf Process Manag 43*, 1449-1487."

[14] Kumar, T. (2014). *Automatic Text Summarization.* Rourkela.

[15] Mayo, M. (2019, November). "*Getting Started with Automated Text Summarization."*

[16] Mr.Vikrant Gupta, M. P. (2012)."An Statiscal Tool for Multi-Document Summarization. *International Journal of Scientific and Research (ISSN 2250-3153)*."

[17] Neelima Bhatia, A. J. (2015)." Literature Review on Automatic Text Summarization: Single and Multiple Summarizations. *International Journal of Computer Applications*, 1-5."

[18] Okumura, H. T. (2009). “Text Summarization Model based on the budgeted median problem . *Proc. 18th ACM Conf. Inf. Knowledge*, 1-4.”

[19] Opidi, A. (2019, April 15).” *A Gentle Introduction to Text Summarization in Machine Learning*.”

[20] Panchal, A. (2019, June 10). “*NLP — Text Summarization using NLTK: TF-IDF Algorithm*.”

[21] *Recent automatic text summarization techniques: a survey*. (2019, March 29).

[22] S, S. (2011). “Automatic Text Summarization: The current state of the art. *International Journal of”*.

[23] YLLIAS CHALI, S. A. (2011). “Query-focused multi-document summarization: automatic data annotations and supervised learning approaches. *Cambirdge University Press*.”