Extractive Text Summarization using NLP

[1] Mr.Parth Sachdeva, Mr. Vardaan Pruthi, Ms.Kanika Mittal, Ms. Mahima Harjani, [2]Ms.Sharanya

[1] Student,Dept. Of Computer Science, HMRITM, Hamidpur, New Delhi-110056, India

[2] Professor Dept. Of Computer Science, HMRITM, Hamidpur, New Delhi-110056, India

---***---

Abstract - In today's world, when data is of utmost importance, it becomes very important for us to retrieve information quickly and accurately. We cannot spend large amount of time to go through all the texts to extract meaningful information and spend a lot of time and effort. Instead, we formulated an automatic version of extracting useful information from enormous lines of text which works just like humans but saves us lot of time and effort. It is very quick too. Automatic Text Summarization is the one in which the contents of the text are reduced but the informational content of the text is preserved. There are two types of summarizations: - extractive and abstractive. In Extractive summarization involves extracting important sentences from the whole text provided as input and then putting them together to input the summary hence formed. In Abstractive summarization, the original sentences of the text are not preserved. It uses different terms to construct new sentences which convey the same meaning. In this paper, we have introduced extractive method of text summarization. Among many models, we chose word frequency as the indicator to define the importance of the word and hence the importance of sentence to be included in the summary.

1.INTRODUCTION

Text Summarization refers to extracting significant data from a large volume of data. The amount of data available on the internet grows every day, making it a matter of space and time to cope with such massive amounts of data. As a result, handling such a massive volume of data poses a significant challenge in a variety of real-world data management applications. The Automatic Text Summarization project makes it easier for users to use Natural Language applications such as data recovery, question answering, and content reduction. Automatic Text Summarization plays an unavoidable role in extracting relevant and unique material from a large amount of data. Filtering through a mountain of reports can be difficult and time-consuming. It can take minutes to make sense of what people will discuss in a paper or report without a summary or breakdown. As a result, the Automatic Text Summarization extracts a sentence from a content record, determines which are the most crucial, and returns them in a comprehensible and structured manner. Automatic Text Summarization is a subset of natural language processing, which is the process by which computers decipher and interpret human speech.

Automatic Text Summarization that looks through a large number of reports using the classifier structure and its rundown modules and delivers the phrases that are useful for constructing a summary. The most overlapping sentences are deemed high score words in the programmed outline of content, which uses overlapping sentences and synonyms or senses. The terms with the highest recurrence are given the most weight. And the most valuable terms are extracted from the content, ordered by frequency, and a summary is generated.

1.1 NATURAL LANGUAGE PROCESSING

Natural Language Processing is a branch of Artificial Intelligence (AI). It provides the computers the ability to understand text and spoken words just as human beings can. NLP focuses on interaction between data science and human language.

NLP Tools and Approaches - Python programming language provides various libraries and tools for performing NLP specific tasks. Many of these libraries and tools are found in NLTK (Natural Language Toolkit). NLTK includes libraries which are used to perform tasks such as word segmentation, sentence parsing, stemming, lemmatization, tokenization, semantic reasoning etc.

Following are the steps/phases involved in Natural Language Processing -

Morphological Processing - This phase involves breaking chunks of input data into sets of tokens such as words, sentences and paragraph.

Syntax Analysis - In this phase two functions are performed. Firstly, the sentences are checked whether they are well formed or not. Secondly, these sentences are broken up into a structure that shows syntactic relationships between the different words.

Semantic Analysis - This phase involves extracting exact meaning or dictionary meaning from the text and also the text is checked for meaningfulness.

Pragmatic Analysis - If a sentence has two semantic interpretations, then pragmatic analyzer will choose between these two possibilities.

1.2 ADVANTAGES OF NATURAL LANGUAGE PROCESSING

1) Provides objective and accurate analysis

While performing tasks like reading, analyzing text data etc., humans are prone to mistakes. NLP powered tools can be trained to the language and according the business criteria. And if once they are trained and set for running, they can perform much more efficiently than humans ever could.

2)Performs large scale analysis

NLP can process large amounts of data in just few seconds or minutes, that would take days or weeks of manual analysis. NLP tools can scale up or down according to our needs, so very little computation power is required.

3)Improves customer satisfaction

NLP enables us to analyze and sort customer service tickets by intent, sentiment, topic etc., and route them to the proper department. By analyzing customer satisfaction surveys, we can quickly know how happy customers are at every stage.

4)Streamlines processes and reduces cost

NLP tools can work at any scale according to our needs. To accomplish manual data analysis, a couple of employees are required to work full time. But, by using NLP SaaS tools, we can keep number of employees to minimum.

5)It helps in better understanding of market

NLP is having a huge impact on marketing. It provides you with better understanding of market segmentation, helps you to be better equipped to target the customers and also decreases customer churn.

6)Provides real and actionable insights

An extra level of analysis is required to analyze unstructured data such as open-ended survey responses, online reviews and comments etc. But NLP tools can make this analysis easier. NLP enables us to really dig into unstructured text for data driven, real world, and immediately actionable insights.

1.3 ADVANTAGES

1)Manual Text summarization is a tedious task which requires a lot of human effort of reading the whole document, extracting the important points from large amount of information and then presenting the same to the user. Whereas, automatic text summarizers can do the same amount of work, even more in few seconds.

2)Automatic text summarizers are not biased like human text summarizer.

3)They also improve the efficacy of indexing.

4)They can also summarize large texts in any language.

5)It also helps in increasing the productivity of the user by reducing the time required for doing the summarization manually.

2.EXTRACTIVE SUMMARIZATION

The process of extractive summarization includes picking out salient sections of the text and generating them line by line resulting in subset of sentences from the original text.

Following are the three independent tasks which are performed by an extractive summarizer –

1) Formation of an intermediate representation of the input text

Indicator representation and topic representation are the two types of representation-based approaches. In topic representation process, the text is transformed into an intermediate representation and the topics present in the text are interpreted. The techniques used for this process are divided into topic word approaches, frequency driven approaches, Bayesian topic models and latent semantic analysis. In indication representation process, every sentence is described as a list of formal features such as position in the document, having certain phrases, sentence length etc.

2) Scoring the sentences based on the representation

An importance score is assigned to each sentence after the intermediate representation of the text. In topic representation process, score is assigned to a sentence based on how well it demonstrates important topics of the text. In indicator representation process, the score is calculated by aggregating the evidences from various weighted indicators.

3) Selection of a summary

The top k most important sentences are selected by the summarizer system in order to produce a summary. To select the important sentences, some approaches use greedy algorithms while some approaches convert the process of sentence selection into an optimization problem in which sentences are selected, considering the criteria that these sentences should maximize coherency and overall importance minimize the redundancy.

3.IMPLEMENTATION

Step 1: Importing required libraries.

In order to build an efficient text summarizer, we need to import two NLTK libraries.

from nltk.corpus import stopwords

from nltk.tokenize import word_tokenize

from nltk.stem import PorterStemmer

Corpus - corpus is a collection of text. It can be anything such as writings by an author, poems by a particular poet etc.

Tokenizer - Tokenizer divides a text into a series of tokens. Word, sentence and regex tokenizer are the three main tokens in a tokenizer. But we will be using only the sentence and the word tokenizer.

PorterStemmer - Its function is to separate the common endings from words in English.

Step 2: Text Tokenizing and Preprocessing – In this step, we perform word_tokenize(text) which returns a list of syllables by applying Legality Principle in combination with Onset Maximization. Then, we input our stream of tokenized words into the stemmer i.e.,

PorterStemmer. After which, we remove the stop words from the output stem. Stop words are the words that have no role in adding value to the meaning of a sentence. For e.g. - the, a, for, an, is etc. We can reduce the number of words and still can preserve the meaning of the sentence after removing the stop words.

Step 3: Creating a frequency table of the words.

In order to keep a record of number of times a word appears in the text, a python dictionary is used. Then this dictionary is used over each stemmed word to store the resulting words left after removing stop words. sentence to know which sentences have the most overall content in the overall text.

Step 4: Finding Sentence score - The text is provided as input in sent_tokenize (), which further divides the sentences into more fragments. A score is assigned to each sentence depending on the words it contains and the frequency table A dictionary is also used to keep track of the score of each sentence. Later, this dictionary is used to create a summary.

Step 5: Comparison of sentences within the text.

To compare the scores, average score of a particular sentence is calculated. This average score is a good threshold. After applying this threshold value, the sentences are stored in an order into the summary.

4.OUTPUT

Step 1: Run Text-sum-core.py file. Then, ignite the Text Summarizer Web application i.e., on http://localhost:5000/text-summarizer.

Step 2: Copy the content that you wish to summarize from any website and paste it on the app (shown in Fig 1).

Step 3: Summarized document will be given as output. It can be viewed after saving it on local computer.

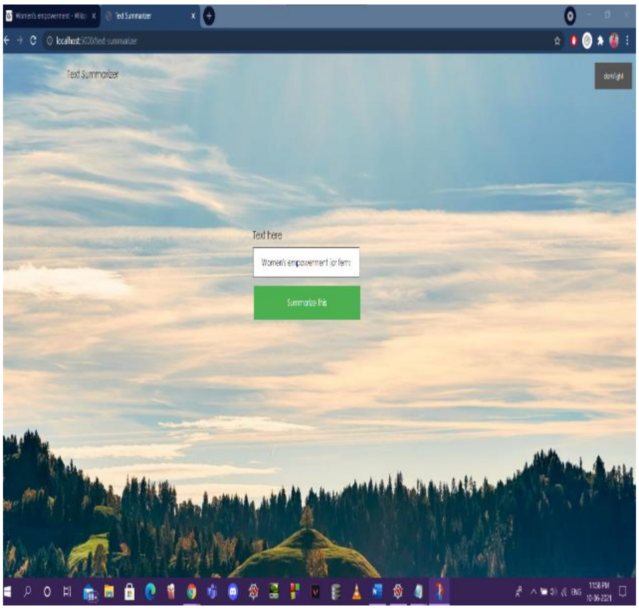
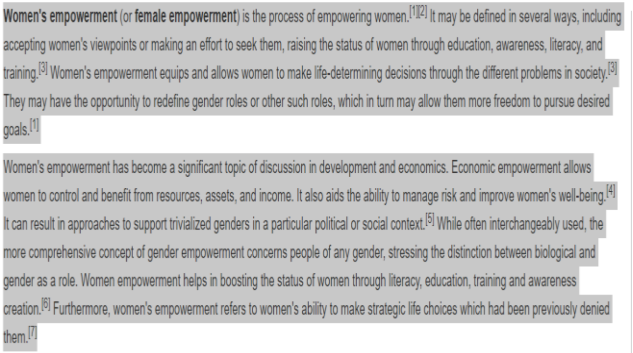


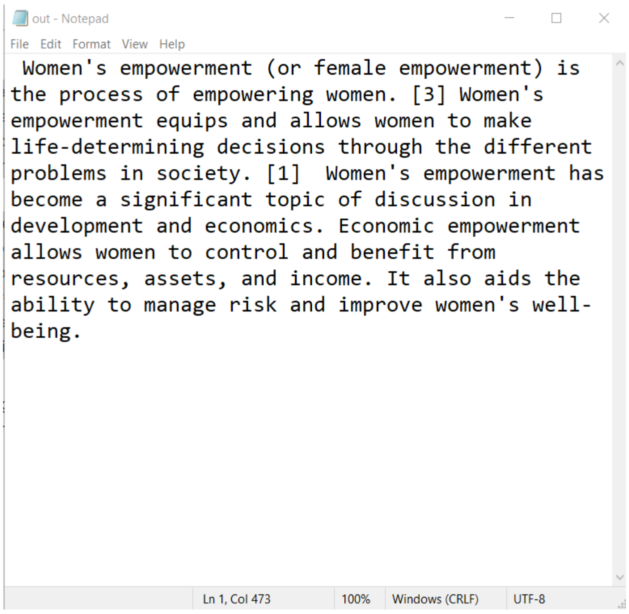
Fig 1: Text summarizer



Women's empowerment (or female empowerment) is the process of empowering women.[1][2] It may be defined in several ways, including accepting women's viewpoints or making an effort to seek them, raising the status of women through education, awareness, literacy, and training.[3] Women's empowerment equips and allows women to make life-determining decisions through the different problems in society.[3] They may have the opportunity to redefine gender roles or other such roles, which in turn may allow them more freedom to pursue desired goals.[1]

Women's empowerment has become a significant topic of discussion in development and economics. Economic empowerment allows women to control and benefit from resources, assets, and income. It also aids the ability to manage risk and improve women's well-being.[4] It can result in approaches to support trivialized genders in a particular political or social context.[5] While often interchangeably used, the more comprehensive concept of gender empowerment concerns people of any gender, stressing the distinction between biological and gender as a role. Women empowerment helps in boosting the status of women through literacy, education, training and awareness creation.[6] Furthermore, women's empowerment refers to women's ability to make strategic life choices which had been previously denied them.[7]

Fig 2: Input to Text summarizer



```
 Women's empowerment (or female empowerment) is
the process of empowering women. [3] Women's
empowerment equips and allows women to make
life-determining decisions through the different
problems in society. [1]  Women's empowerment has
become a significant topic of discussion in
development and economics. Economic empowerment
allows women to control and benefit from
resources, assets, and income. It also aids the
ability to manage risk and improve women's well-
being.
```

Fig 3: Output from Text summarizer

5.CONCLUSION

In this paper, we presented an extractive text summarizer that can help condense large reports or paragraphs into easily understandable chunks with key points. This model is made up of two main components i.e., sentence ranking and finding sentence score. In sentence ranking, we constructed a frequency table to keep the count of important words. In the second phase, we performed key points extraction based on the calculated average score of each sentence from the original text. Since text summarization when done by humans is a tedious task which is why this laborious task of reading the whole text to extract just the key points is done by machines with the help of NLP. Further research is required to incorporate it into a larger system with a valuable user interface. Also, some functionalities like summarizing the whole pdf or website will be much more demanding in the future due to the explosion in the size of data.

REFERENCES

[1] https://arxiv.org/abs/1910.14142
[2]https://www.sciencedirect.com/science/article/pii/S1532046416301514
[3] https://www.scirp.org/html/902.html
[4] https://towardsdatascience.com/text-summarization-using-tf-idf-e64a0644ace3
[5] https://www.mygreatlearning.com/blog/text-summarization-in-python/
[6]https://media.neliti.com/media/publications/166330-EN-single-document-automatic-text-summariza.pdf