IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, June 2022
Impact Factor: 6.252

Abstraction Based Text Summarization using NLTK

Mr. Anand Tilagul[1], Santhosh T M[2], Prajwal M B[3], Nikhil R[4]
Assistant Professor, Department of Information Science and Engineering[1]
Student, Department of Information Science and Engineering[2,3,4]
S J C Institute of Technology, Chikkaballapur, Karnataka, India

Abstract: *As there is huge amount of content is produced in our day-to-day life using electronic devices in various form fields. The main problem arises from here huge form information once to analyzing and understanding the meaning of text become difficult and time taking, so the Text summarization is introduced. There are two types of text summarization types one is Extraction based text summarization and Abstraction based text summarization.*

Keywords: NLTK

I. INTRODUCTION

An abstraction-based text summarization is a method of generating summary for the huge input text given by the user. Basically, in this project user can give desired length of text to summarizer so the machine is trained such a way that it will produce summary for input text which will in reduced in length, reduced in words and meaningful sentences. By this summary it will make user to read the content of the huge text easily and also the main motive of the summarizer is to reduce the time of user without stuck in understanding inner meaning of the sentences. And summary should contain in the form human readable format with the most needed words which help to impact the meaning of the summary. Already many models are built to summarize the text using the neural form networks and using machine learning to get the meaning summary. Natural Language Toolkit (NLTK) which is toolkit build for purpose working with the Natural Language Processing with Python Programming language.

II. PROBLEM IDENTICATION

An Investigating Officer may sometimes be required to refer to online news articles to obtain further information about a case beyond what is already known through on-ground sources. Due to the proliferation of news websites on the internet, it is not uncommon for a simple search on a topic or suspect of interest to return thousands, and even lakhs, of relevant news articles. It would take an Investigating Officer hours and hours of manual effort to go through these news articles, understand them and assimilate key findings. Often information would be spread out and not available in a single article.

III. METHODOLOGY

STEPS INVOLVED TO GENERATE THE ABASTRACTION BASED SUMMARY

- **STEP 1:** Here the text is used collected from the user as input for summarizer.
- **STEP 2:** In this step collected text is cleaned, means deleting the stop words, special characters, numbers which is irrelevant to text and punctuations
- **STEP 3:** In this step word token and sentences token are created this process is called Tokenization
- **STEP 4:** In this step by those tokens created in pervious step, frequency is found for every word in the users input text.
- **STEP 5**: Here in this step weights are assigned to words.
- **STEP 6:** Based on the weights, most top rated 20% weighted sentences are called final summary.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, June 2022
Impact Factor: 6.252

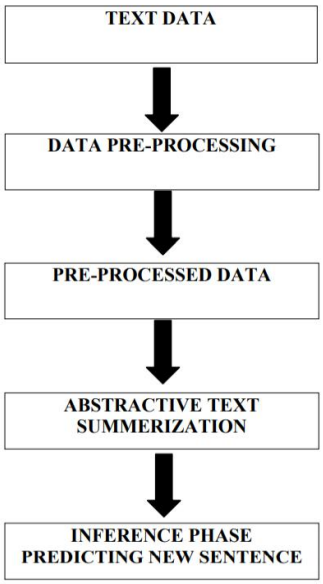


Figure 1: Summary Generation Process

IV. IMPLEMENTATION

The pseudocode of a NLTK TEXT- SUMMARIZER mechanism

Start

INPUT: Un-summarized Text

Output: Summary For given input Text

```
1. Import nltk , (import by typing " PIP INSTALL NLTK" command)
2.import Stopwords
3. Def nltk_summarizer(TEXT)
4. SW = set(stopword.word("English"))
5. words = word_tokenize(TEXT)
6. freqTable=dict( )
7. // Removing Stop Words
   for word in words
8.     word = word.lower( )
9. if word not in stopWords
10. if word in freqTable
11.       freqTable[word] += 1
12.  else
13.       freqTable[word] = 1
14.  end for
15. sentence_list = sent_tokenize(docx)
16. max_freq = max(freqTable.values( ))
17. for word in freqTable.keys( )
18.     freqTable[word] = (freqTable[word]/max_freq)
19. sentence_scores = { }
20. for sent in sentence_list
21. for word in nltk.word_tokenize(sent.lower( ))
22. if word in freqTable.keys( )
```

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, June 2022
Impact Factor: 6.252

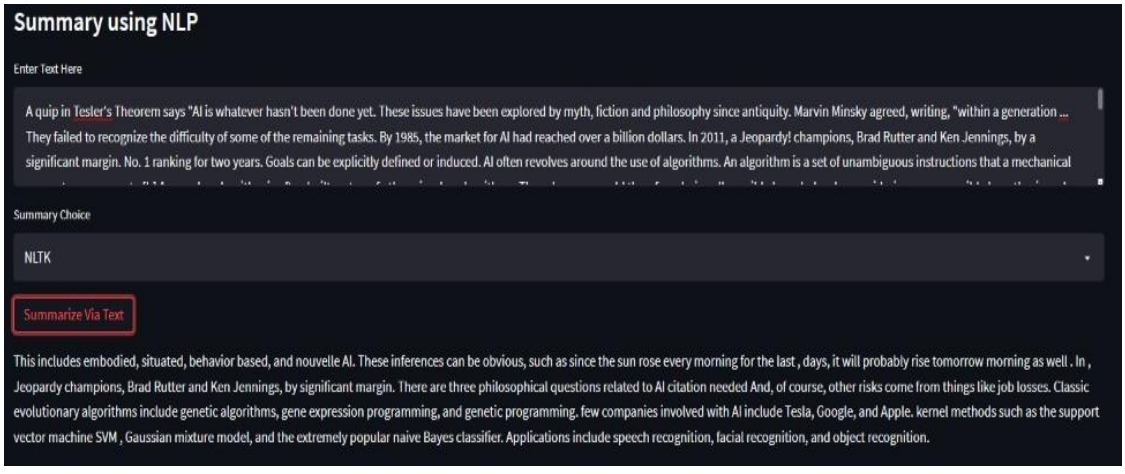
```
23.  if len(sent.split(‘    ‘)) < 30
24.     if sent not in sentence_scores.keys( )
25.          sentence_scores[sent] = freqTable[word]
26.      else
27.           sentence_scores[sent] += freqTable[word] //total number of length of words.
28. end for
29. end for
30. return summary
31. Stop
```

V. TESING.

C	TEST CASE	EXPECTED OUTPUT	OBTAINED OUTPUT	RESULT
1	Passing the text input	Text should be read and displayed in input area	Text should be read and displayed in input area	Pass
2	Passing the text input as null	Show alert messages “Enter the text”	Show alert messages “Enter the text”	Pass
3	Selecting the NLTK method	Ready to create summary	Ready to create summary	Pass
4	Without selecting the NLTK method	An error should be thrown specifying ”importing NLTK “	An error is thrown	Pass
5	The model should return reduced and meaningful summary	Abstractive summary will be displayed	Abstractive summary will be displayed	Pass

VI. RESULTS

1. The solution should take the desired length of summary from the user as an input should return summarized output.
2. The most important output of these Abstraction based text summarizer is to reduce the reading time.
3. Abstraction based text summarization produces meaningful sentences.
4. It makes the user to read the summarized output easily.
5. It gives short, exact and more content full summary without repetitive summary.



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, June 2022
Impact Factor: 6.252

REFERENCES

[1]. Mrs.Chetana Badgujar 'Abstractive Summarization using Graph Based Methods', Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018) IEEE Xplore Compliant - **Part Number**: CFP18BAC-ART; **ISBN**:978-1-5386-1974-2, 978-1-5386-1974-2/18/$31.00 ©**2018 IEEE**.

[2]. Siya Sadashiv, " Extractive text summarization by feature-based sentence"Extraction using rule-based.2017 2nd ieee international conference on recent trends in electronics information & communication technology (rteict), may 19-20, 2017, india", 978-1-5090-3704-9/17/$31.00 © 2017 IEEE.

[3]. Nidhi Patel, Prof,Nikhita Mangoakar, "Abstractive vs Extractive Text Summarization", 2020 IEEE International Conference for Innovation in Technology (INOCON) Bengaluru, India. Nov 6-8, 2020,

[4]. 978-1-7281-9744-9/20/$31.00 ©2020 IEEE

[5]. Lili Wan, " Extraction Algorithm of English Text Summarization For English Teaching", 2018 International Conference on Intelligent Transportation, Big Data & Smart City. 0-7695-6368-6/18/$31.00 ©2018 IEEE.

[6]. Leonhard Henning,Winfreid Umbarth "An Ontology-based Approach to Text Summarization", 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 978-0-7695-3496-1/08 $25.00 © 2008 IEEE.

[7]. Daksha Singhal, "Abstractive Summarization of Meeting Conversations", 2020 IEEE International Conference for Innovation in Technology (INOCON) Bengaluru, India. Nov 6-8,2020, 978-1-7281-9744-9/20/$31.00 ©2020 IEEE.

[8]. Michael T.Mills and Nikolas G, "Graph-Based Methods for Natural Language Processing", Ieee transactions on systems, man, and cybernetics: systems, vol. 44, no. 1, January 2014, 1094-6977 © 2013 IEEE.