TEST Engineering & Management

May-June 2020
ISSN: 0193-4120 Page No. 6875 - 6879

Snapshot of an Agreement based on Keywords

Sunil Bhutada, Professor, IT Department Sreenidhi Institute of Science and Technology Yamnampet, Hyderabad. (sunilb@sreenidhi.edu.in)
Lakmishetti Shreya, B. Tech IV year, IT Department Sreenidhi Institute of Science and Technology, Yamnampet, Hyderabad (shreyanani4@gmail.com)
Javvaji Varshini, B. Tech IV year, IT Department Sreenidhi Institute of Science and Technology, Yamnampet, Hyderabad (varshinijavvaji511@gmail.com)
Mudelli Uma Maheshwari, B. Tech IV year, IT Department Sreenidhi Institute of Science and Technology, Yamnampet, Hyderabad (uma2610maheshwari@gmail.com)
Sindhura Bandaru, B. Tech IV year, IT Department Sreenidhi Institute of Science and Technology, Yamnampet, Hyderabad (varoos99@gmail.com)

Article Info
Volume 83
Page Number: 6875 - 6879
Publication Issue:
May-June 2020

Article History
Article Received: *19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *18 May 2020*

Abstract:
In today's digital world everything is revolving around data. Now-a-days every individual require everything to be in a precise manner and wants the work to be done as fast as possible. So, our project is mainly based on "Summarization of Text", which is a process of deflating the text to help people save their time. If the data generated is enormous every individual do not have the time and patience to read and understand the given data. So, in this paper we provide a solution for the problem faced by the users to read and accept the terms and conditions before installing an application, or software or the user manuals of any electronic gadgets. Sometimes without reading the policies we agree upon them. But there might be some important points like the agreement policies say that our data may be used by some third-party. We don't want this to happen with our data. We want our data to be private. So, to overcome this problem we need someone who could go through policy and mention some important points which is depicted in this paper.

Keywords: *manual, service agreement, terms and conditions, refund policy, summary*

1. INTRODUCTION:

Summarization of text is demarcated into two types i.e. Extractive and abstractive summarization. In extractive summarization the important keywords are identified which cannot be modified further which are provided unchanged in the summary. Whereas the abstractive method is more complex than the other. Unlike in extractive method we can modify the mined important sentences and provide them in the summary. It involves huge effort but extractive method is still popular. We developed this text summarizer using python and a NLTK which is a natural language toolkit that works on human language. The problem faced by many individuals while reading and understanding the various privacy policies or user manuals is consumption of more time where the users lack of patience therefore it leads to the misuse of data. Accordingly, a solution to the problem is provided in this paper.

2. LITERATURE SURVEY:

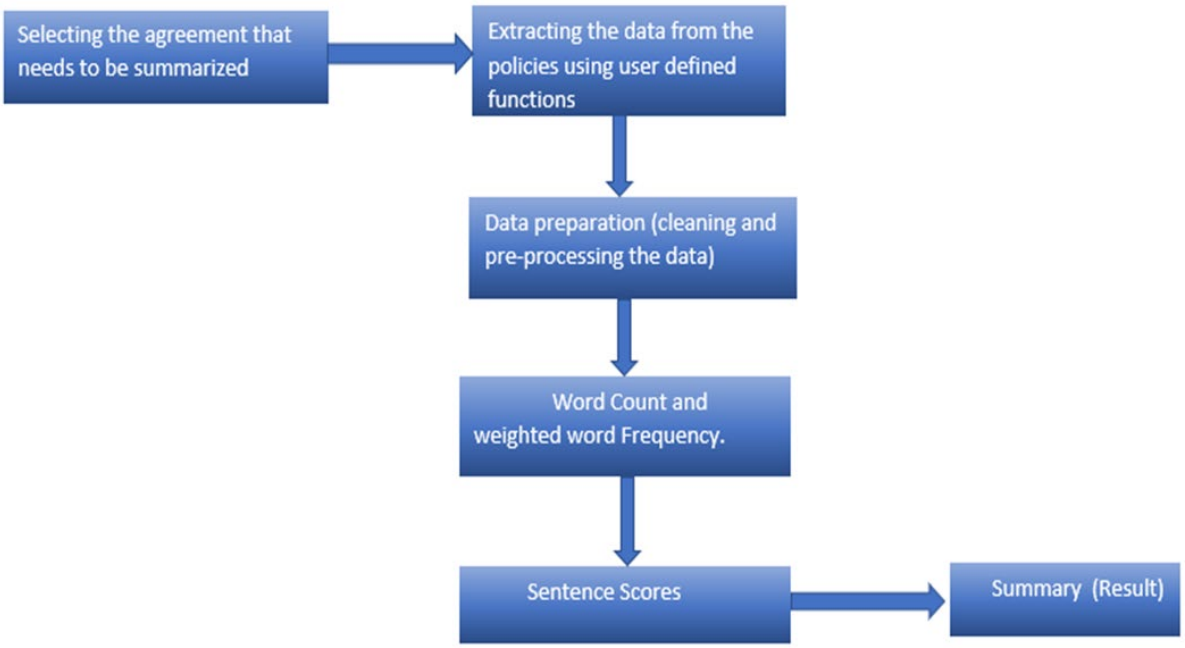
Author Inderjeet Mani [1] has explained about how important is the extraction of the summary or the important information from any given textual data.

May-June 2020
ISSN: 0193-4120 Page No. 6875 - 6879

Here the method introduced by the author is focused on abstraction rather than sentence extraction. Author Giuliano Armano [2] tried to explain about the importance of the keywords for news and news aggregations. This type of methodology only suggests the keywords for the news and its aggregation. Author Aarti Patil [3] has developed the process based on the extractive summarization by using NLP. This methodology could be used only for the generation of the summaries for the given documents by calculating the indexing weights. Author Saiyed Saziabegum [4] has clearly explained the basics of automatic text summarization for the provided textual data or the document. Author Madhav Thaker [5] has used K-Means Clustering method for the retrieval of the summary from the provided textual data. In the above-mentioned papers and articles all the authors have concentrated on the summarization of text which is provided by the user to extract the smaller summary of the textual data produced. But no author has focused on the agreements, user manuals, policy terms and conditions list etc. which are commonly ignored by every individual as the points in an agreement are large in number. Finally, we say that our proposed methodology will overcome this problem which could not be done by the earlier authors.

3. PROPOSED METHODOLOGY

3.1. ARCHITECHTURE:



3.2. PSEUDO CODE:

Input:

Get the required data such as file or URL

Pre-processing is applied such as delimiter removal, stop word removal and stemming

Word count and weighted frequencies are applied on the data

for (all sentences x)

if (words in sentence x is less than 30)

Add those word frequencies to the sentence scores

Break out of for loop

end if

end for

add sentences with the highest scores to the summary

May-June 2020
ISSN: 0193-4120 Page No. 6875 - 6879

Output:

Summary of the data gathered is achieved with important data.

3.3. Steps:

- Gather the data from various sources which is needed to the user.
- Perform Pre-Processing
  - Delimiter removal - (, . ‘ “ ? ! { })
  - Stop word removal (which are mostly used by user) - the, is, an, a, like etc.
  - Stem word Removal (adjective or adverb of a word) - (ied, ed, mis, ary)
- Word Count
- Weighted word frequency
- Sentences scores
- Adding the highest sentence scores to the summary
- Final Result

4. Experimental setup:

First, we collect the data that needs to be summarised. The data can contain various formats such as text document, textual data or an URL. Now the collected data from any source needs to be fragmented into small units. They are called tokens. To understand it in a better way, it can be said as breaking of bigger sentences into individual words. Then we perform delimiter removal. Delimiters are the characters that are used to separate the text or data by specifying the boundary. These are present in plain text or other formats. Stop words are the most commonly used words which has less importance in performing the retrieval of summary so, they need to be removed. Stem word removal is known as stemming which is the process of obtaining the root word.

After the pre-processing of given data, we count the frequency of each word. This is known as word count. Later on, the weighted word frequencies of each word are calculated by using formula [Weighted word frequency= word frequency/maximum frequency]. Now we take the weighted word frequency of each word and then find them in sentences, add them together. Finally, the sentences with the highest scores are produced as summary. Now the retrieved data is the desired result.

To achieve this process, we give the desired text document or textual data or an URL as the input in the input field then click on summarize button. If the n number of sentences defined by the user then we retrieve the summarized data. The important measure to be taken is that the mentioned URL or the path of the document need to be valid.

Examples:

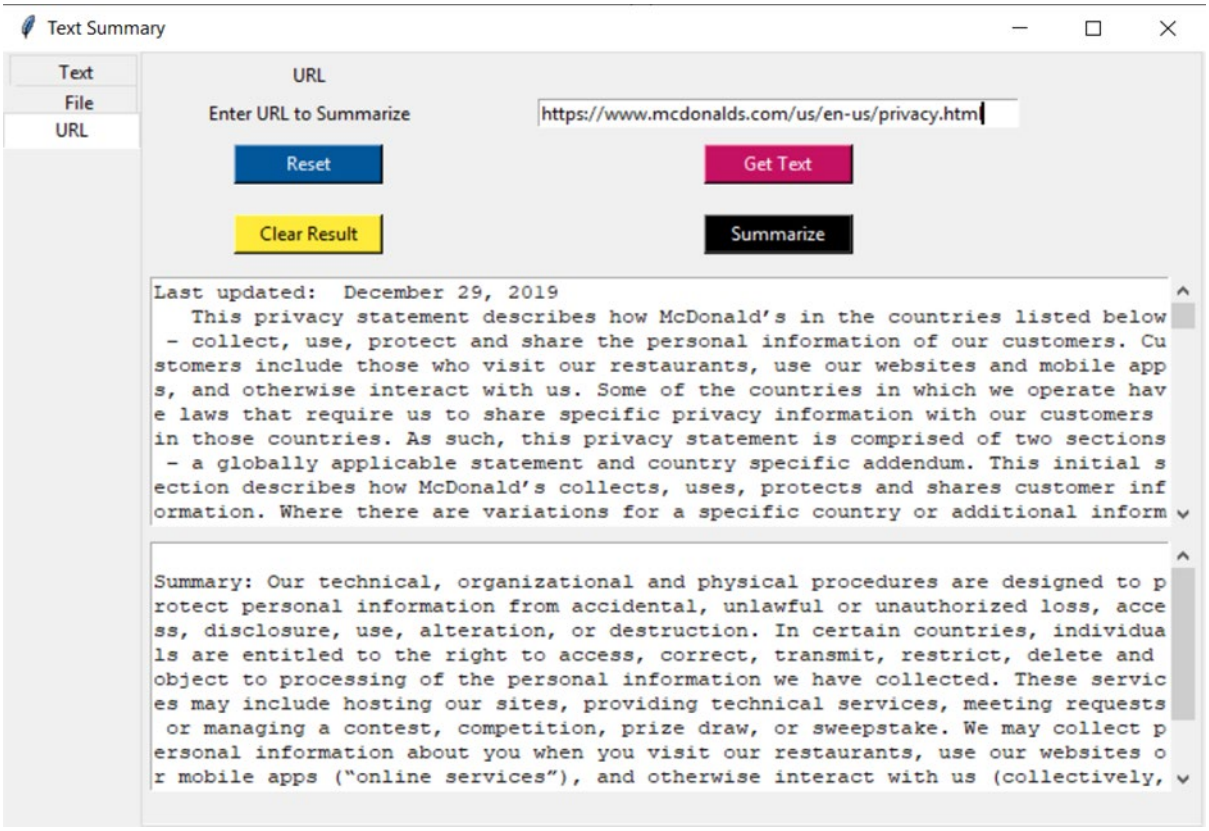
Now, we are generating summaries for 3 different scenarios like a Mc Donald’s privacy statement, an agreement of a car rental agency and a privacy policy of a gaming application which are in different formats like URL, text and a document**.**

Input type:

URL: https://www.mcdonalds.com/us/en-us/privacy.html

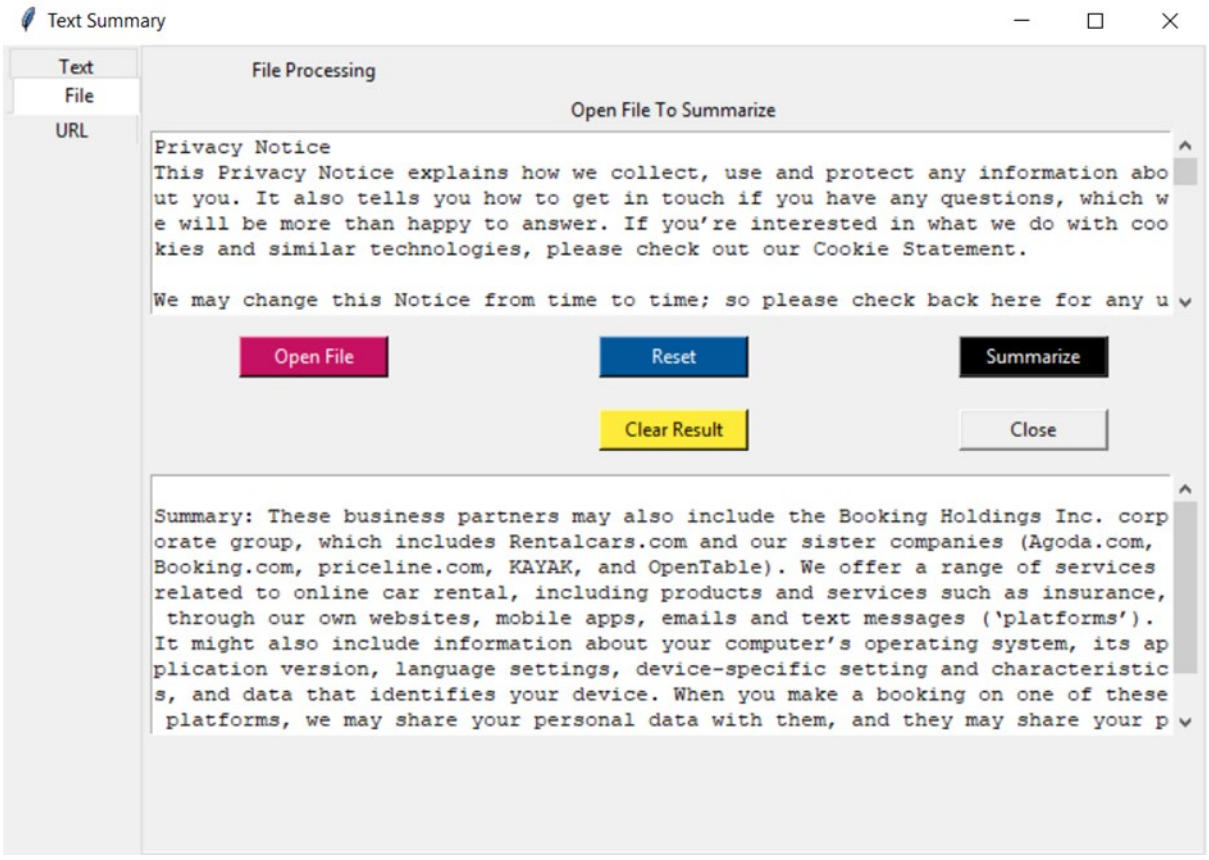
Output:

May-June 2020
ISSN: 0193-4120 Page No. 6875 - 6879



Input type: File

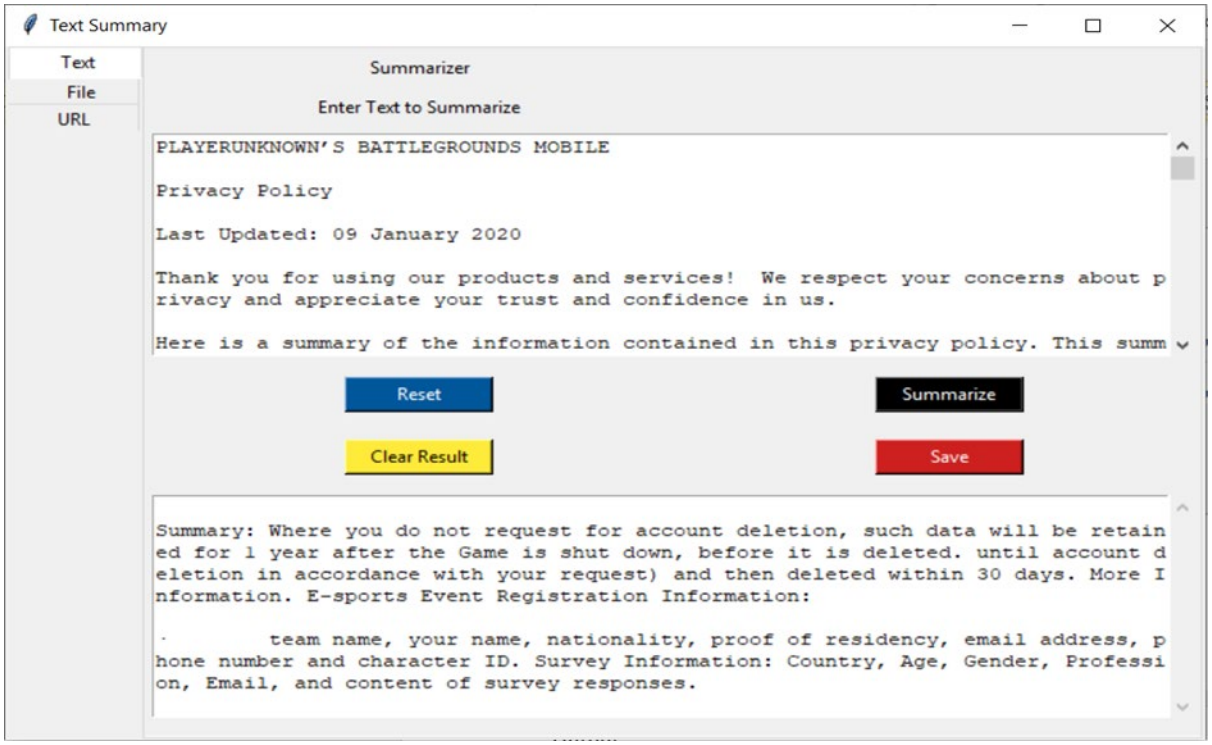
Output:



Input type: Text

Output:

May-June 2020
ISSN: 0193-4120 Page No. 6875 - 6879



5. Conclusion and Future Scope:

In day-to-day life no individual has the time or patience to read and agree for the terms and conditions for different works such as gaming applications, electronic gadgets user manuals etc. As the usage of textual data has been increased, reading and understanding the complete textual data of a document has become a complicated and time taking task to be performed. Many authors have discussed about retrieval of summaries from the provided textual data but no author has focused on the retrieval of important keywords or summary for an agreement or any kind of user manual etc. So, the main aim of this paper is to collect the important keywords from the provided agreements or the manuals and give the user a brief and short summary for the input. This paper provides the steps for the retrieve of important information from the agreements. This research area is done in the field of python by using a NLTK tool kit. It can be explored more by looking into the existing systems and by working on different techniques and algorithms.

REFERENCES

1. Inderjeet Mani, "Automatic summarization", MITRE corporation and Georgetown University, 2001.
2. Giuliano Armano, Alessandro Giuliani, "Experimenting Text Summarization on Multimodal Aggregation", Vol-771.
3. Dr. M. Suman, Tharun Maddu, "A New Approach for Text Summarizer, K.L University, COMPUSOFT", An international journal of advanced computer technology, 4 (4), April-2015 (Volume-IV, Issue-IV).
4. Aarti Patil, Komal Pharande, "Automatic Text Summarization", International journals of computer applications, Vol-107, January 2015.
5. Saiyed Saziabegum, "Review on Text Summarization Evaluation Methods", Indian Journal of Computer Science and Engineering (IJCSE), 2013.
6. https://towardsdatascience.com/comparing-text-summarization-techniques-d1e2e465584e.
7. https://becominghuman.ai/text-summarization-in-5-steps-using-nltk-65b21e352b65.
8. Therese Firmin Hand, "A proposal for Task-based Evaluation of Text Summarization Systems", Department of Defense, USA.