Student Helper App: NLP Based Text Summarization Android Application

Kavita Verma[1], Shaurya Chaubey[2]
[1]Dept. Of Computer Science and Engineering, KIET Group of Institutions, Ghaziabad

[2]Dept. Of Computer Science, KIET Group of Institutions, Ghaziabad

ABSTRACT

This digital era has simplified human life in every aspect. Technology is evolving continuously and with each going day, new tech is getting introduced to make our lives easier. With the same concept in mind, we are trying to develop a student helper app. A mobile-based application that would, as the title suggests, will help students to get a quick summary of any large document. In this application, we are using the technology of Natural Language Processing for text summarization. This application does summarization through extractive text summarization and helps to increase the productivity of students by saving their time and in general making their life better.

Keywords: Android, NLP, OCR, Text Algorithm, TextRank Summarization

1. INTRODUCTION

In this fast generation of mobile phones and google providing every needed information, most of the time students have a lot of content and less time to go through all the data. In today's world, technology has taken over everything and it has made space for itself in every aspect of human life. From waking up with a buzzing mobile alarm to sleeping with a last good night message on social media we are surrounded by technology. Every single day, the tech is evolving, and new things are coming out to ease the user experience. With this fast generation of mobile phones and google providing every needed information, most of the time students have a lot of content and less time to go through all the data.

The problem that we are aiming to solve through this project is the lack of time in students' lives and the abundance of content for them to read. Not only this we also aim to provide them with solutions to most of the problems related to student life and notes in one place. Hence, the Student Helper app is an application that will provide students with a summary of the content, which covers all the important parts of the given content. In this app, the students can provide their long notes either by copy-pasting the text or as a text file and the application will perform the needful processing on the given text and give out the summarized text without losing the important points in the text.

This application applies NLP to extract out the summary of a given large text, using JgraphT and Test Rank algorithm to make sure that important points in the text are included in the summary. TextRank is an extractive and unsupervised text summarization technique. The app will also be built in such a way that it provides the best user experience to the users so that they can stay and work on the app for a longer period.

The main objective of the project is to increase the productivity of the student and to facilitate them to save their time so that they can invest it into something else as they already have a tight schedule most of the time. So this app aims to make the life of students better and to keep most of the things required by the student at one single destination, this app.

2. LITERATURE REVIEW

2.1 NLP

[5] Natural language processing is a technology that enables humans to communicate with machines in their own natural human language. Natural Language Processing makes it much easier for humans to interact with computers and get their tasks completed. Natural Language Processing can be broadly said to comprise two components, these are, Natural Language Understanding or NLU and Natural Language Generation or NLG. Natural Language Understanding refers to the part of the process of processing natural, human language in which the computer tries to understand what is being conveyed to it through the process or algorithm that we have decided. Natural Language Generation is the next step to Natural Language Understanding. Once the model has understood what we are trying to convey to it, it prepares a proper

response for that and gives it to the humans in a form that they can understand, that is in the natural human language. The response can be generated in the form of phrases, paragraphs, or sentences.

There are several tasks that can be performed by harnessing the power of NLP, some of these are- text summarization, translation from one language to another, NER or Named Entity Recognition, POS tagging or Part of speech tagging, etc. NER is an entity detection mechanism, that is in this if we give our input, we get the known entities of the world, like- if we give in the name of a city, or a person it'll tag these entities. Translation from one language to another, also known as Machine translation, is the process in which the text is converted from one language to another through a model based on NLP. The difference that is made by NLP in machine translation is that it doesn't just translate the sentences word by word but also tries to keep the emotions of the sentence intact and convey them in the translated sentence. Text summarization is the ability to summarize a given long text into smaller and concise text that conveys the similar meaning by preserving all the important points as the original text.

2.2 Text Summarization

[4] Text Summarization refers to the process of converting a given text to a much smaller summary of the text, and this summary has all the main and important points of the original text and reduces the redundancy in the data. [9] According to Edward Hovy and others, the summary is a text that is almost half of the original text and contains the significant parts of the original text(s) from which it is derived from.[11] There are mainly two techniques for text summarization, and these are- Extractive Text Summarization and Abstractive Text Summarization. In Abstractive Text Summarization, the new summary is generated by understanding the whole text and then producing new sentences that might not be present in the original text. While in Extractive Text Summarization, the summary of the text is extracted by finding out the most important sentences from the text and then combining all of them together to give the final summary of the text. [5] A summary can be produced either for a single document or for multiple documents. Several techniques have been developed to get a summary out of a text and some of these are- TensorFlow Model, TF-IDF, Clustering and Classification, Neural Network, Graph-based approach, etc. All these methods give varying results and are suitable for different languages and situations.

TF-IDF stands for the term frequency-inverse document frequency. In this approach the documents are represented using their TF-IDF scores of the words in the document. Term frequency refers to the number of appearances of a word in each sentence in a document. Inverse document frequency is the values computed from the whole of the document. Each weight is calculated as a combination of the term frequency and inverse document frequency. Sentences are ranked using a score called SumBasic score. The SumBasic score is an approach to find out the significance of any given sentence based on the frequency of words.

An algorithm has been developed by Google Brain Team which creates a summary by extracting interesting parts of the text and rephrasing them to create an abstractive summarization.

In the Graph-based approach, the importance of each sentence is calculated by comparing the semantic similarities between the various sentences that we have in our text. An algorithm is followed to get the most important sentences in the document and then they are combined to give out the summary. In this approach, all the sentences are represented as nodes in a graph, and they are scored and sorted, and then it is decided whether a sentence will be included or not.

2.3 Android

[3] Android is an open-source, operating system developed by Google in the year 2007. It is a Linux-based OS. In the android operating system, each app is assigned a unique user ID, so the app runs with a distinct identity. This feature also helps in isolating one resource from another, hence acting as a multiuser OS. The Linux kernel provides a variety of features related to security as a base for the computing environment of mobile phones.

Android applications can be written in Java, Kotlin, and C++. It is designed in such a way that the communication between software and hardware can be done easily with a user interface. It provides different API libraries that consist of everything needed to build an application including source code, manifest, and resource files.

2.4 JGraphT

[2] JGraphT is an open-source programming library written in Java. It is interoperable, stable, and provides efficient performance. JGraphT contains generic graph data structures along with a proper collection of algorithms. It was first released in 2003 and has widened its scope over the years. It now provides an implementation of simple graphs, multigraphs, pseudo graphs, etc, and consists of many graph algorithms like routing, planning(path), combinatorial optimization, network analysis, and other applications in computational biology. Besides containing basic graph algorithms like MST or shortest path, the library also consists of many complex algorithms like subgraph isomorphism, NP-hard problems, approximation algorithm, etc.

Nowadays, JGraphT is used widely for commercial, non-commercial, and research purposes.

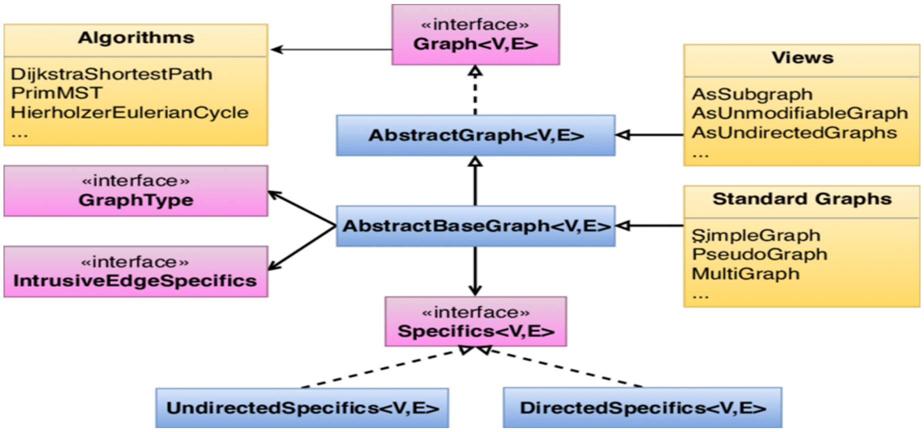

Figure 1. Functions in JgraphT library

2.5 TextRank Algorithm

[1] [6] TextRank is an unsupervised algorithm, which is used to get a summary of texts that are written in natural language. It uses extractive text summarization techniques, in which those sentences or words which are most important or have occurred more times than the other sentences or the words in the original text are extracted and used to construct a summary. [8][10] This algorithm follows a graph-based approach and the sentences, or the words are taken as the graph's vertices. It uses a weighted graph in which weight is directly proportional to the importance of a word or sentence in the text. The TextRank algorithm is unsupervised, language-independent, well-defined, and developed hence, used in extractive text summarization. TextRank algorithm is an extension of the PageRank algorithm which is used to measure the importance of web pages by Google [11]. PageRank algorithm is based on the concept of voting. PageRank value of each page is calculated by considering the connections between the page and other pages. The page having higher PageRank value is considered as more significant and then it is presented above the other web pages in web search results. The following diagram represents the whole process of how a summary is extracted from the given text by the TextRank algorithm.

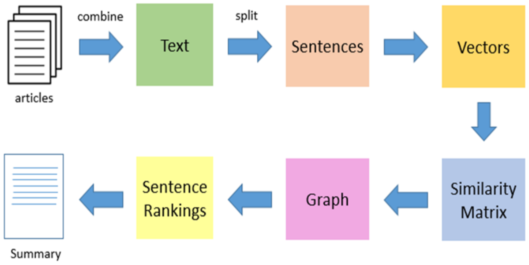

Figure 2. Steps followed in TextRank algorithm for text summarization

3. RELATED WORK

There are some already existing android applications based on text summarization. Here is a short explanation of their existing features and how our app will add or modify some features and is better than these existing ones.

3.1 SumIt: An android app that provides a summary of the provided text.

Features of the app:

- Free to use
- Simple and easy way to summarize the text
- You can also share a summary through various outlets (E-mail, Twitter, Facebook, Whatsapp, text message, etc.)
- If you find pasting the URL not working, simply copy all the text and paste it to summarize.

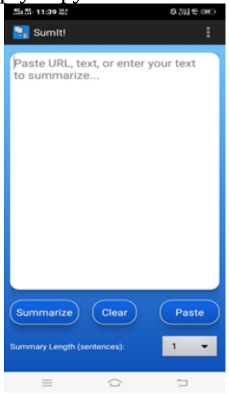
11:39
SumIt!
Paste URL, text, or enter your text to summarize...
Summarize
Clear
Paste
Summary Length (sentences):
1

Figure 3. Home Screen of SumIt application

3.2 Text Summary: An android app that is capable of summarizing texts of all sizes and text from photos and web pages.

Features of the app:

- Allows copy and paste text
- Allows copy and paste the link to the web page
- Allows either to click a picture of text with your camera or select images of text from your phone gallery

Figure 4. Splash Screen of Text Summary application

3.3 Improvement in our approach over these applications

In the approach that we propose here uses, the application can be used to extract text from images also. In some of the other existing applications, this feature is there but the results provided by them are not up to the mark. In our approach, we are using Optical Character Recognition (OCR) for text recognition. OCR or Optical Character Recognition is an approach to extract text from a given text image which can have either hand-written or printed text on it. In this approach, the text extracted is converted into machine readable format so that it can be further used for various purposes. Using this approach our application can give us better results for text recognition.

We are also providing a sharing feature in our application that will enable the user to share the extracted summary of their notes with others via email, WhatsApp, or any other social platform.

The UI and UX provided by the application are much more user-friendly that'll enable more users to use the app easily and conveniently. The application also provides the feature to extract text from document files, which will help users to extract text from their documents.

4. FUTURE WORK

The study shows that a lot of work is possible and required in this field. In our future work, we are planning to expand the application in such a way that it can help the students in many different co-curricular activities. Along with getting the content summary, students can share the summary on different social sites. To make the application useful different activities from other fields can be added to the app like games like brain teasers, word of the day, etc. This will engage the students, and make them learn new things in an interesting way. The feature of setting daily reminders can also be helpful to students to remain disciplined and follow their defined schedules in their day-to-day life.

The algorithms like TextRank, OCR, etc, applied in the paper can used in a more effective way to get the best and expected results.

5. EXPERIMENTAL DATA

For the testing purpose, we have taken data samples from different web pages and book sources that can be possibly used as content by students. The sample data could be classified into groups as:

English Essays
History
Social Science
Rural Development

We have tested the application with text from various subjects that a student can have in their curriculum, the texts consist of varying lengths of content. Some text samples were several pages while some were not too large to test the reliability of the summarization algorithm.

6. CONCLUSION

This research paper proposes a way to combine the work in the field of text summarization and Android to develop an application that can be used daily by students to save their time. Several approaches to extract a summary from the text have been devised. Mainly there are two approaches, extractive text summarization and abstractive text summarization. We have used an extractive approach in our application. There can be a variety of real-life scenarios where text summarization can be used, some of these are- summarization of news, articles, monitoring social media, etc. The paper here proposes one of the applications of text summarization, that is using it for the benefit of students helping them save their time from reading long notes and thus increase their productivity. Further research can be done in this field to find different ways to summarize notes.

REFERENCES

[1] Lu Yao, Zhang Pengzhou, Zhang Chi "Research on News Keyword Extraction Technology Based on TF-IDF and TextRank." School of Computer Science, Communication University of China, Beijing, China, 2019

[2] Dimitrios Michail, Joris Kinable, Barak Naveh, and John V Sichi, "JGraphT - A Java library for graph data structures and algorithms", Dept. of Informatics and Telematics, Harokopio University of Athens, Greecemichail@hua.gr; Dept. of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, The Netherlandsj.kinable@tue.nl; Robotics InstituteCarnegie Mellon University, Pittsburgh, USA4barak@3pq.com; The JGraphT projectjsichi@gmail.com (2019)

[3] Vijay Deshmane, Sarita Sawale, Krishna Bharambe, Pratik Lahudkar, "APPLICATION DEVELOPMENT WITH ANDROID: A REVIEW", Final year IT Student, Department of Information Technology, Anuradha Engineering College, Chikhli, MH, India; Assistant Professor, Department of Information Technology, Anuradha Engineering College, Chikhli, MH, India (2018)

[4] Dhanya P.M, Sreekumar A, Jathavedan M “A SURVEY OF RECENT TECHNIQUES IN AUTOMATIC TEXT SUMMARIZATION”, Volume 9, Issue 2, March-April 2018, pp. 74–85, Article ID: IJCET_09_02_007

[5] Diksha Khurana, Aditya Koli, Kiran Khatter and Sukhdev Singh “Natural Language Processing: State of The Art, Current Trends and Challenges”, August 2017

[6] Bartłomiej Balcerzak, Wojciech Jaworski, and Adam Wierzbicki. "Application of TextRank algorithm for credibility assessment".Polish-Japanese Institute of Information Technology, Koszykowa 86, 02-008 Warsaw, Poland, Institute of Informatics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland, wjaworski@mimuw.edu.pl (2014)

[7] Deepali K. Gaikwad and C. Namrata Mahender “A Review Paper on Text Summarization”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 3, March 2016

[8] Mihalcea, Rada.” Language independent extractive summarization.” Proceedings of the ACL 2005 on Interactive poster and demonstration sessions. Association for Computational Linguistics, (2005)

[9] E. H. Hovy, Automated Text Summarization, The Ox- ford Handbook of Computational Linguistics, chapter 32, Oxford University Press, Oxford, 2005, pp. 583-598.

[10] Mihalcea, Rada. ” Graph-based ranking algorithms for sentence extraction, applied to text summarization.” Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. Association for Computational Linguistics, (2004)

[11] L Page, S Brin, R Motwani, T Winograd. The PageRank Citation Ranking: Bringing Order to the Web[R]. Stanford InfooLab,1999