

# **INTELLIGENT CONTENT AGGREGATOR AND KNOWLEDGE SYNTHESIZER**

**Vissarapu Srinath  
Pettugani Hotragn  
Eguturi Manjith  
WOXSEN UNIVERSITY**



# **INTELLIGENT CONTENT AGGREGATOR AND KNOWLEDGE SYNTHESIZER**

## **Capstone Report**

submitted to

School of Technology, Woxsen University

B. Tech

Artificial Intelligence and Data Science

By

Eguturi Manjith

Vissarapu Srinath

Pettugani Hotragn

Under the guidance of

Dr. Kiran Kumar Ravulakollu



WOXSEN UNIVERSITY, HYDERABAD

March 2024

## **CERTIFICATE**

This is to certify that the report entitled “INTELLIGENT CONTENT AGGREGATOR AND KNOWLEDGE SYNTHESIZER”, submitted by Vissarapu Srinath (20WU0101031), Pettugani Hotragn (20WU0101030) and Eguturi Manjith (20WU0101005) to the School of Technology, Woxsen University, for the successful completion of the capstone project - 2, is a record of bonafide research work carried out by them under my supervision and guidance. To the best of my knowledge, the work embodied in this Project have not been submitted to any other university or institute for the award of any other degree or diploma.

---

Prof. Dr. Jaswanth Nidamanuri  
(Program Chair)

---

Prof. Dr. Kiran Ravulakollu  
(Dean, School of Technology)

## **DECLARATION**

I certify that.

- ✓ The work contained in this report is original and has been done by me.
- ✓ I have followed the guidelines provided by the School of Technology while preparing the report.
- ✓ I have confirmed the norms and guidelines given in the Ethical Code of Conduct of the University.
- ✓ Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.



## **ACKNOWLEDGEMENTS**

First and foremost, we would like to express our deep sense of gratitude and indebtedness to my Supervisor Dr Kiran Kumar Ravulakollu for his valuable encouragement, suggestions, and support from an early edge of this project and providing us extraordinary experiences throughout the work. Above all, his priceless and meticulous experience supervision at each phase of work inspired me in innumerable ways.

We especially acknowledge his advice, supervision, and the vital contribution as and when required during this project. His involvement with originality has triggered and nourished my intellectual maturity that will help us for a long time to come. We are proud that we had an opportunity to work with an exceptionally experienced professor like him. We express our sincere thanks to all the faculty members of the School of Technology for their support.

V. Srinath – 20WU0101031

E. Manjith – 20WU0101005

P. Hotragn – 20WU0101030

## **List of abbreviations**

**ICAKS** – Intelligent Content Aggregator and Knowledge Synthesizer

**DOM** – Document Object Model

**API** - Application Programming Interface.



## **Abstract**

As browsing has become a daily routine, the time a person takes to analyze through multiple websites at a single shot and understand them is difficult. These findings had become tough while correlating the content from multiple websites which is leading to ambiguity in human brains.

Motivated by this, this project will focus on correlation between the information from various pages and formats. There are endless extensions like copy fish and vowel ai which can extract text or summarize wrt the feature of extension. Yet there's a lot of complications for us as humans to compare and thereby conclude a single concept.

Collective analysis of a single topic from various written styles would help users to increase productivity and management in terms of time invested. We will be having a single extension that can extract data from multiple active tabs with the utmost experience to users, which will improve their performance and work.

## Contents

Certificate	3
Declaration	5
Acknowledgement	7
List of Abbreviations	8
Abstract	9
Contents	10
List of Figures	11
1 Introduction	12
2. Literature Survey	13
2.1 Existing Word	13
2.2 Extension	14
2.3 Motivation and Problem statement	14
3. Proposed Methodology	15
4. Results and Discussions	19
5. Conclusion and Future Scope	24
References	26
Annexure	27

## List of Figures

Figure 1: Extension for extracting text from active tab .....	16
Figure 2: Summarized content for the active tab .....	17
Figure 3: API for receiving extracted content to the backend. ....	18
Figure 4: Extension to the browser. ....	20
Figure 5: summary dashboard .....	21
Figure 6: Summarized Content .....	21
Figure 7: Data Base Design .....	22
Figure 8: Extension Interface .....	23
Figure 9: Ablation Study .....	23

# 1.INTRODUCTION

In an era where browsing is ubiquitous, the challenge lies in efficiently analyzing multiple websites to grasp essential information. Current methods often lead to information overload and difficulty in synthesizing data across various sources. Our project presents the Integrated Content Aggregation and Knowledge Synthesis (ICAKS) extension, driven by these challenges. With the help of this browser extension, users will be able to extract and summarize data from various currently active web pages, which will improve their productivity and time management more efficiently.

Through the automated gathering and arrangement of real-time DOM data, ICAKS aims to deliver maximum effectiveness in terms of information perception and summarization. By utilising machine learning and natural language processing (NLP), the extension provides users with a unified overview of aggregated content from active websites.

## **2. LITERATURE SURVEY**

### **2.1Existing work [1]**

- An extension that has developed to help the users with poor English using the python dictionaries, google cloud firebase, and flask. This extension will generate pre-defined difficult words based on an article/website by parsing the content.
- With the information extracted, it will map the definition of words generated to a dictionary and check if the extension is able to accurately generate the words.
- In addition, this extension will have a track of the most difficult words for users and use this information for future users to have a seamless experience.
- In conclusion, this extension provides a simple understanding of difficult words to users in English.

### **2.2Extension [2]**

- The MaxAI extension has a variety of features that use most versions of ChatGPT and provide results for the user.
- It helps in summarizing a text from a single page, Generating the content for email, and even as an assistant chatbot.
- It works as an assistant and uses ChatGPT 4 to produce results only with the premium plan whereas the free plan only provides limited ChatGPT usage per day and a mini menu for a text selected.

## 2.3 Existing extension [3]

- Given any prompt with web access, the extension can give a response based on the latest news in ChatGPT. It can even scrape data from a URL by restricting responses to a specific URL and giving responses respectively.
- It can perform a one-click prompts wrt required language, Tone, Writing style and Topic for a specific selected article/website.
- Many extensions can replicate the same usage, but the mentioned extensions are the popular ones with most users.

## 2.4 Motivation and problem statement

As browsing has become a daily routine, the time a person takes to analyze through multiple websites in a single shot and understand them is difficult. These findings had become tough while relating the content from multiple websites which is leading to ambiguity in human brains. Motivated by this, this extension will focus on the correlation between the information from various pages and formats. Typical extensions provide us with a summary but there are a lot of complications for us as humans to compare multiple websites and thereby conclude a single concept. Collective analysis of a single topic from various written styles would help users increase productivity and management in terms of time invested. The ICAKS has a single extension that can extract data from multiple active tabs with the utmost experience for users, which will improve their performance and work. This can help in finding key insights within the information extracted, it will save time for visiting multiple tabs and find insights which reduces human effort.

The ICAKS mainly focuses on reducing redundancy and summarizing text without the need for an individual to revisit the pages again and again. Additionally, there is a chance that multiple people will try to visit the same/similar websites and it is hard to extract data each time. In this regard, there will be a tracker which not only visualizes the visited websites/content summarized, but also even stores the historical data concerning security measures. With the help of this data, we will suggest the users for the websites they are looking for and reduce the time even in terms of searching. In conclusion, this extension can help individuals find the summarization of data and let them track their progress with the help of an extension and website.

### 3 PROPOSED METHODOLOGY

**Step:1-** DOM Elements from the visited pages will be collected through an extension which will be built using JavaScript. Once the DOM content is accessed it will be sent to the backend using an API. Image tags will be extracted separately and will be sent to the backend in the form of URL or file format.

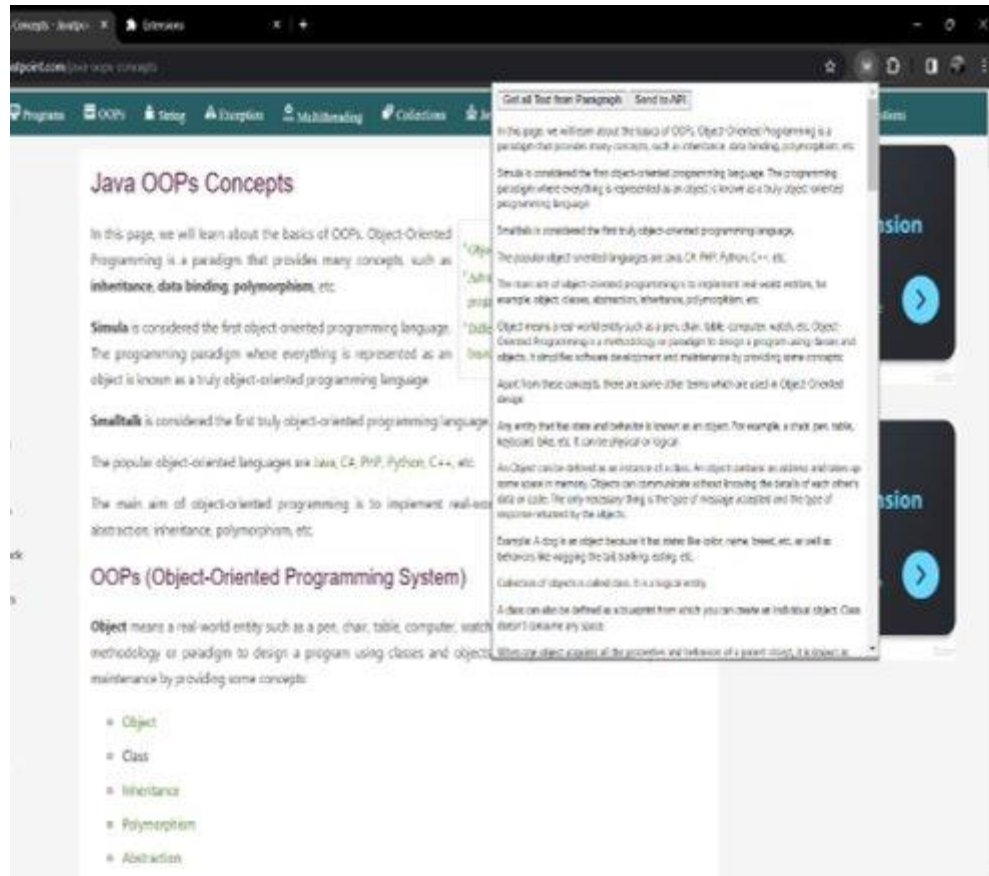


Figure 1: Extension for extracting text from active tab

## Step:2-

Model Based Extraction- As all websites do not follow the same structure, we try eliminating most of the unused content in the first step. We will try and build a machine learning model which should be able to separate tags based on their content.

The screenshot shows a web browser displaying the 'Java OOPs Concepts' page on the website 'javaTpoint.com'. The page has a dark green header with navigation links: Home, Java, Programs, OOPs, String, Exception, Multithreading, and Collections. A left sidebar contains a table of contents with sections like 'Basics of Java', 'Java Object Class', 'Java Inheritance', 'Java Polymorphism', 'Java Abstraction', and 'Java Encapsulation'. The main content area is titled 'Java OOPs Concepts' and contains introductory text about OOPs, a list of popular languages (Java, C#, PHP, Python, C++), and the main aim of OOPs. A 'Get Summary' popup is overlaid on the right side of the page, containing a detailed definition of Object-oriented programming (OOPs) and its key concepts. At the bottom of the page, there is a Windows taskbar showing the date as 08-02-2024 and the time as 23:01.

javaTpoint

Home Java Programs OOPs String Exception Multithreading Collections

Basics of Java

Java Object Class

- Java OOPs Concepts
- Naming Convention
- Object and Class
- Method
- Constructor
- static keyword
- this keyword

Java Inheritance

- Inheritance(IS-A)
- Aggregation(HAS-A)

Java Polymorphism

- Method Overloading
- Method Overriding
- Covariant Return Type
- super keyword
- Instance Initializer block
- final keyword
- Runtime Polymorphism
- Dynamic Binding
- instanceof operator

Java Abstraction

- Abstract class
- Interface
- Abstract vs Interface

Java Encapsulation

- Package
- Access Modifiers

### Java OOPs Concepts

In this page, we will learn about the basics of OOPs. Object-Oriented Programming is a paradigm that provides many concepts, such as **inheritance, data binding, polymorphism, etc.**

**Simula** is considered the first object-oriented programming language. The programming paradigm where everything is represented as an object is known as a truly object-oriented programming language.

**Smalltalk** is considered the first truly object-oriented programming language.

The popular object-oriented languages are Java, C#, PHP, Python, C++, etc.

The main aim of object-oriented programming is to implement real-world entities, for example, object, classes, abstraction, inheritance, polymorphism, etc.

### OOPs (Object-Oriented Programming System)

**Object** means a real-world entity such as a pen, chair, table, computer, watch, etc. **Object-Oriented Programming** is a methodology or paradigm to design a program using classes and objects. It simplifies software development and maintenance by providing some concepts:

- Object
- Class
- Inheritance

Object-oriented programming (OOPs) is a programming paradigm that uses the concept of "objects" to represent real-world entities, along with their attributes and behaviors. These objects can interact with each other to form complex systems. Some key concepts in OOPs include inheritance, data binding, polymorphism, abstraction, encapsulation, coupling, cohesion, association, aggregation, and composition. In OOPs, inheritance allows one object to acquire all the properties and behaviors of a parent object, promoting code reusability and runtime polymorphism. Polymorphism enables one task to be performed in different ways, while abstraction hides internal details and shows only the necessary functionality. Encapsulation binds code and data together into a single unit, enhancing security and reducing complexity. Objects can be associated with each other through relationships such as association, aggregation, and composition. Association represents a relationship between two objects, while aggregation and composition represent stronger forms of association where one object contains other objects as a part of its state. OOPs offers several advantages over procedural programming, including easier development and maintenance, data hiding, and improved simulation of real-world events. Popular object-oriented programming languages include Java, C#, PHP, Python, and C++. It's worth noting that object-based programming languages follow most of the features of OOPs, but they lack inheritance. Examples of object-based programming languages include JavaScript and VBScript. Additionally, while constructors in Java do not explicitly return a value, they play a crucial role in creating and initializing objects.

Get Summary

Questions

AD

SLEEP WORRY FREE on period nights

Get upto 100% leakage protection with **SafefreeNights** Cotton Soft Cloth

BUY NOW

Based on research in 2019 and consumer satisfaction survey for years in 2020, SafefreeNights is one of the most trusted brands for period care products. Choose SafefreeNights. The comfort of choice is always with you.

26°C Haze

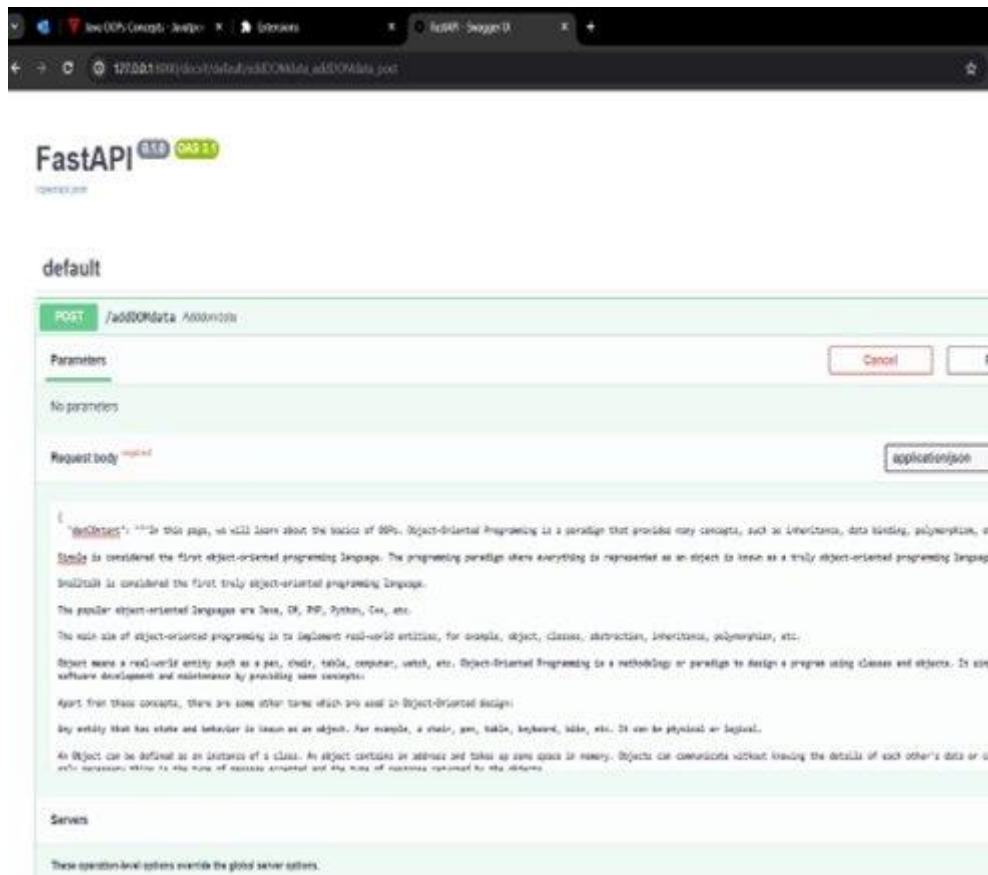
Search

ENG IN

23:01 08-02-2024

Figure 2: Summarized content for the active tab





*Figure 3: API for receiving extracted content to the backend.*

**Step:3-** Development of an algorithm that can do the below:

Based on the title of the DOM. Once all documents are segregated, the summarization of each title will be performed. Further, there will be a summary of all similar titles together and a separate summary considering all content.

**Step:4-**

A website will be developed to visually track the stats of a user and similar content using the data repository constructed & updated based on extension usage.

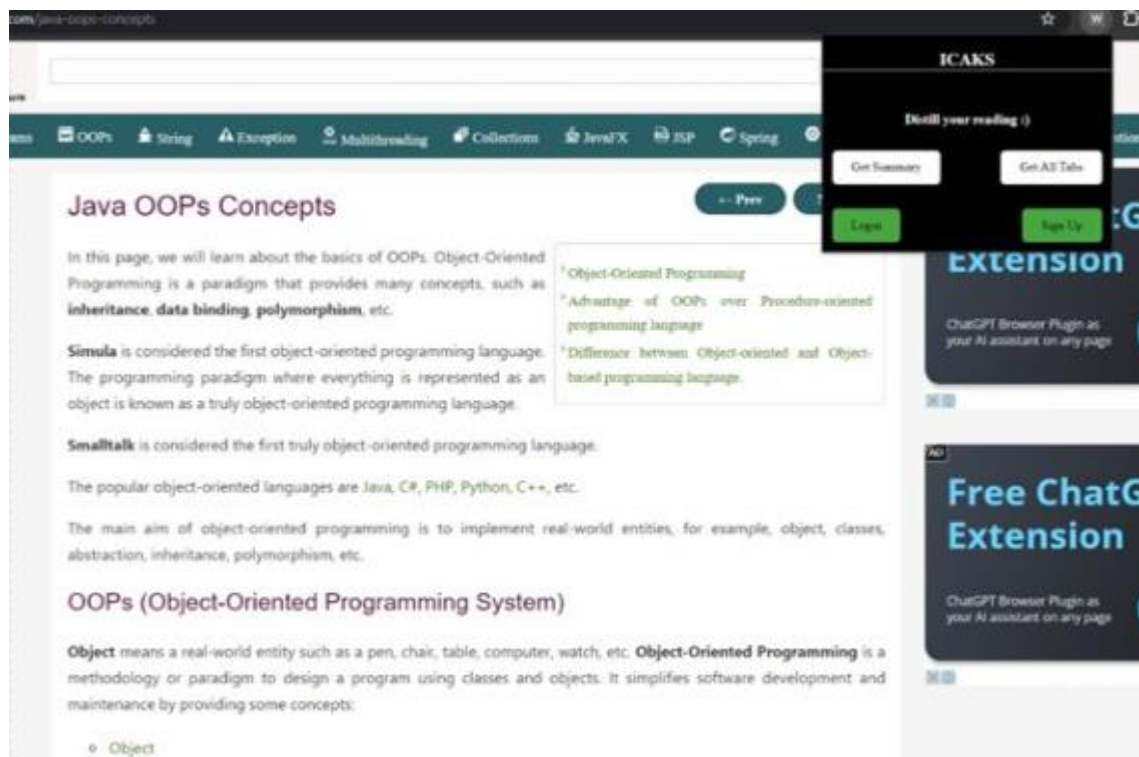
**Step:5- Tag Elimination**

As all websites do not follow the same structure, we try eliminating most of the unused content as the first step. We will try to eliminate the tags based on their content.

Once information is extracted it will be stored in the database along with the website name, topic searched, and timestamp. Summarized content is stored alongside DOM content. As the user stops browsing, the summarizing process will start in the backend which involves APIs and other text summarization models to summarize all the extracted content into one single chunk and store it. A user interface will be developed so users can login and view their searched topics and their summaries along with them. This website will be developed using react and the backend remains the same with additional fetch Apis added. As part of this development, we will be conducting regular reviews for the outputs and process to check the risks, working efficiency, and mitigate the risks as well as updating the methods or extraction accordingly which focuses on making the product for users to have the utmost experience. Furthermore, the approach will be proposed to incorporate this innovative idea in the form of a multilingual summarization extension and the Research/Content Recommendation Systems.

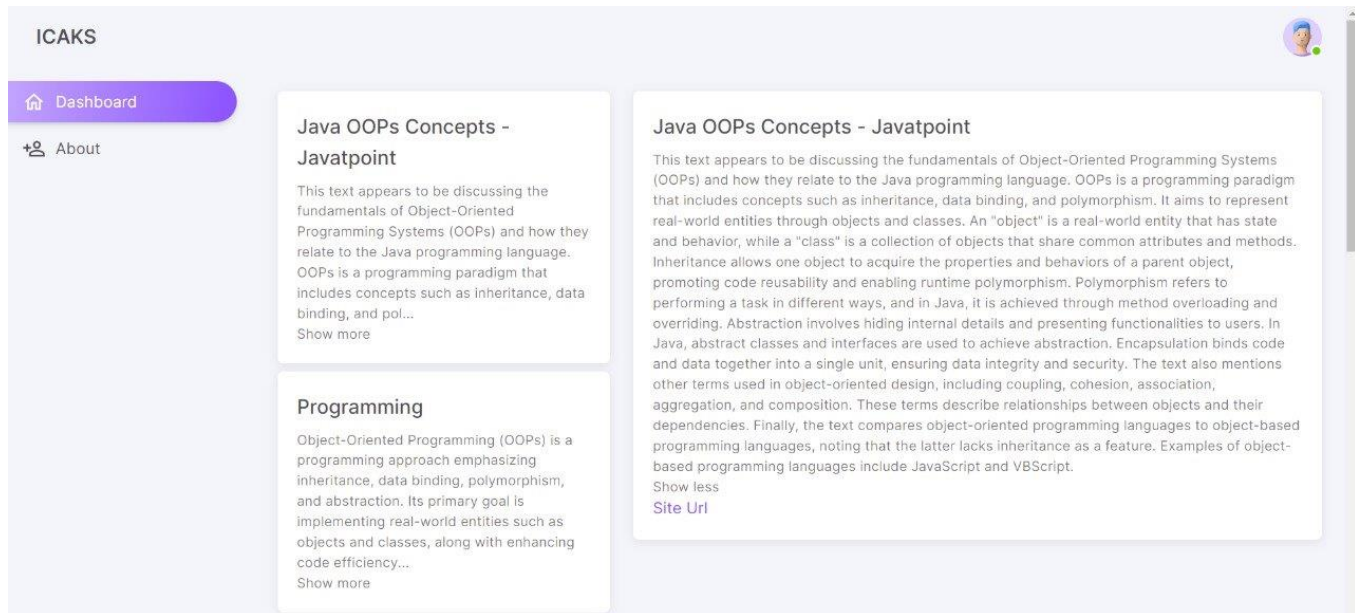
## 4 Results and Discussions

The ICAKS extension has been developed, enabling the extraction of data from multiple active tabs, and providing users with succinct summaries of the content. This functionality streamlines information processing, facilitating enhanced productivity and efficiency in accessing and comprehending information from diverse sources.



*Figure 4 Extension to the browser.*

From the above image, we can observe the interface of the extension, which includes options for “Get Summary” and “Get All Tabs.” Upon clicking the “Get All Tabs” button, the extension will store information from all open tabs. Subsequently, clicking the “Get Summary” button will display a summary of the contents from all open tabs.



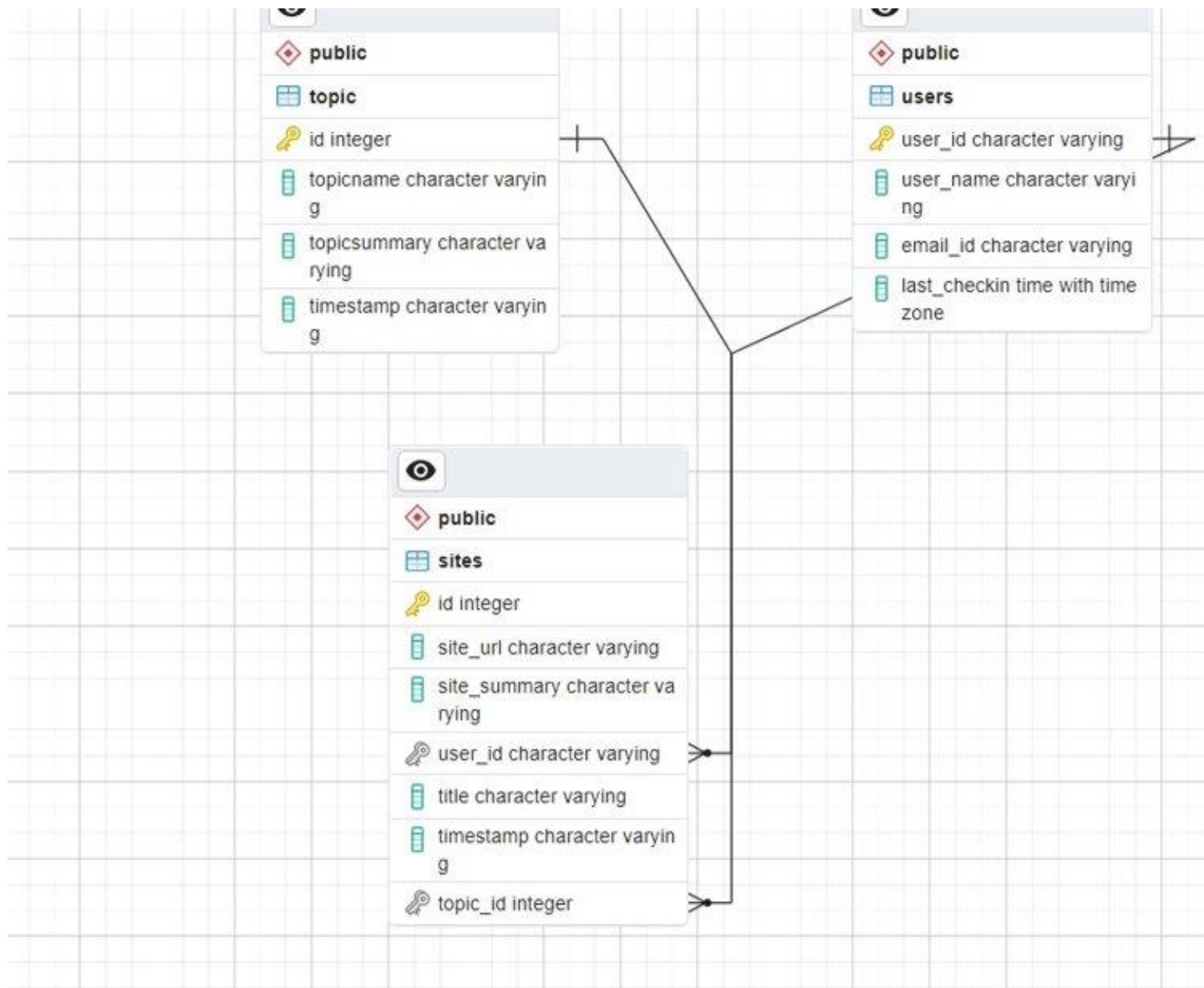
*Figure 5 summary dashboard*

## FastAPI - Swagger UI

The text describes three topics obtained from GET/getalltopics API endpoint in FastAPI 0.1.0 with OpenAPI 3.1 specification. The first topic is about Object-Oriented Programming System (OOPS) concepts in Java. It explains fundamental concepts of OOPS like inheritance, data binding, polymorphism, encapsulation, and abstraction. It also covers other related terms such as coupling, cohesion, association, aggregation, and composition. The second topic introduces Programming, specifically focusing on Object-Oriented Programming (OOP) and Python. OOP is a programming approach that utilizes inheritance, data binding, polymorphism, and abstraction to enhance code efficiency, maintainability, and scalability. Keywords and constructs in Java, such as abstract class, interface, methods, constructors, static keyword, this keyword, instance initializer blocks, and final keyword, are discussed. Python is described as a simple, readable, and widely-used programming language with extensive community assistance and powerful libraries. The third topic pertains to Beverages but contains only code snippets, styling rules, and configurations for building a webpage, making it unsuitable for summarization. However, the code relates to a template script defining a div structure for menu items with server-side rendering, JavaScript functions handling postbacks and page request management, style declarations for specific DOM elements, interaction with third-party services or plugins, and configuration and initialization of analytics tracking. This response also provides information about response headers and controls accept header with example value schema string and no links schemas. Lastly, there is an embedded Swagger UI bundle code block showing the client-side representation of the API documentation generated using OpenAPI Specification.

*Figure 6 Summarized content*

From figures 5 and 6 above, we can observe the dashboard displaying the summarized content, where various websites are summarized. The extension includes a “Get Summary” button that redirects users to this dashboard. Within the summarized content, users can find the web URL. From which the content was extracted, located at the end of each summary.



*Figure 7 Data Base Design*

Figure 9 shows the DB design of extension where it has topic, sites, and Users. In topic it Has id, topic summary and timestamp. In user DB has User id, User name, email address and last\_check in time. And in sites DB has id, site URL, site summary.



Figure 8 Extension Interface

Models	Bleu	Rouge	METEOR	BERT	Word Mover's Distance	Self-Bleu	Perplexity vp8=
MPT(MosaicML_84kTokens)	0.38	0.36	0.22	0.51	0.46	0.30	80
FLAN-T5	0.43	0.38	0.25	0.55	0.50	0.26	76
<a href="#">philschmid/bart-large-cnn-samsum</a>	0.42	0.35	0.26	0.49	0.48	0.29	79
<a href="#">Hugchat</a>	0.46	0.41	0.31	0.60	0.52	0.34	68
<a href="#">Falconsai/text_summarization</a>	0.39	0.32	0.25	0.52	0.44	0.29	84
<a href="#">ARTElab/it5-summarization-mlsum</a>	0.40	0.37	0.29	0.54	0.47	0.28	77

Figure 9 Ablation study

- Figure 11 shows the ablation study of the models, encompassing evaluations based on metrics such as Bleu, Rouge, METEOR, BERT, Word Mover's Distance, self-Bleu, and perplexity. Utilizing Hugging Face yields superior scores compared to other models, demonstrating its effectiveness in the summarization task.
- Hugging Face's includes varied pre-trained models such as BART and T5, preparing to different summarization needs across various content types found in active tabs.
- Models on Hugging Face are finely tuned and consistently achieve top scores across evaluation metrics like BLEU, ROUGE, and METEOR, ensuring the generation of accurate and informative summaries for users.
- Our extension can tailor the summarization process to suit different types of content, ensuring that summaries extracted from active tabs are relevant and comprehensive.

## 5 CONCLUSION

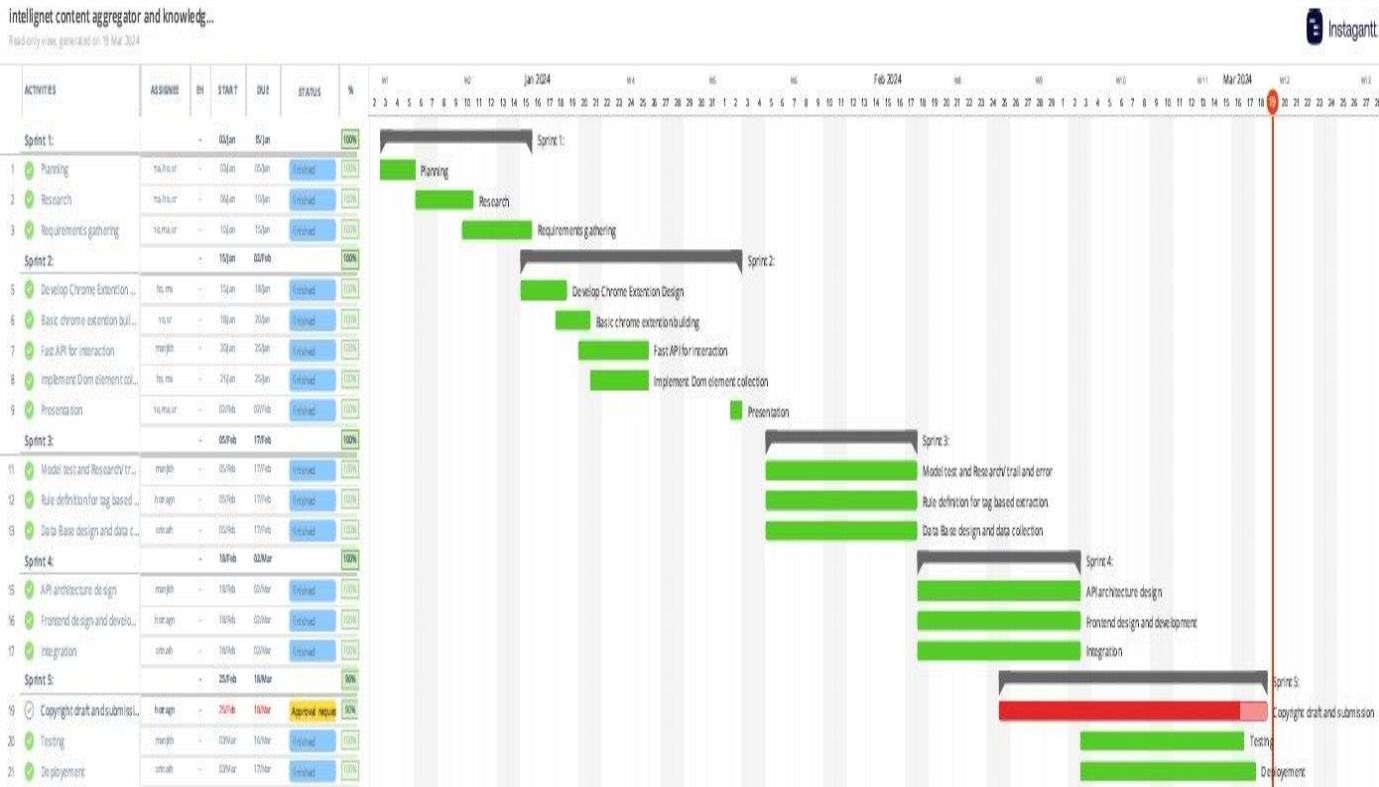
In conclusion, the development of the ICAKS extension, or Intelligent Content Aggregator and Knowledge Synthesiser, we have successfully developed an addon that can take data from several open tabs and synthesise it into summaries, increasing user productivity and efficiency through careful design and execution. We have protected our creative solution and made it possible for it to be widely used by filing the extension for copyright protection.

### **Future Scope:**

Moving forward, several avenues for further enhancement and expansion of the ICAKS extension present themselves. Firstly, efforts should be directed towards ensuring universal support across all major web browsers, thereby maximizing accessibility and usability. Additionally, continued refinement of the summarization algorithms is necessary to optimize the maximum length of summaries while maintaining coherence and relevance. Moreover, stringent measures must be implemented to prioritize data privacy and security throughout the data extraction and storage processes, addressing concerns regarding user confidentiality and trust. Furthermore, the adaptation of algorithms to accommodate the diverse formats of webpages is crucial for maintaining flexibility and accuracy across various online sources. By addressing these constraints and pursuing future developments, the ICAKS extension is poised to further revolutionize the way users interact with and synthesize.



GANTT CHART



## REFERENCES

- [1] Zhao, Tongde, and Khoa Tran. “A powerful CHROME EXTENSION: TRANSLATION PROGRAM USING PYTHON, WEBSITE ANALYSIS AND GOOGLE FIREBASE SERVICES.” *CS & IT Conference Proceedings*. Vol. 13. No. 7. CS & IT Conference Proceedings, 2023.
- [2] <https://maxai.me/>
- [3] <https://chromewebstore.google.com/detail/webchatgpt-chatgpt-with-i/pfemeioodjbpieminkklglpmhlnghcn>
- [4] <https://chrome.google.com/webstore/detail/copyfish-%F0%9F%90%9F-free-ocr-soft/eenjdnjldapjajjofmldgmkjaienebbj>
- [5] <https://huggingface.co/mosaicml/mpt-7b>
- [6] <https://huggingface.co/philschmid/bart-large-cnn-samsum>
- [7] <https://huggingface.co/chat/>
- [8] [https://huggingface.co/Falconsai/text\\_summarization](https://huggingface.co/Falconsai/text_summarization)
- [9] <https://huggingface.co/ARTELab/it5-summarization-mlsum>

# **ANNEXURE**

# **Software Requirements Specification**

for

## **Intelligent content Aggregator and Knowledge Synthesizer**

Eguturi Manjith  
Vissarapu Srinath  
Pettugani Hotragn

**WOXSEN UNIVERSITY**

**19<sup>th</sup> March 2024**

## Table of Contents

1.0	Introduction.....	3
1.1	Purpose.....	3
1.2	Scope.....	3
1.3	References.....	3
1.4	Abbreviations / Acronyms / Definitions.....	4
2.0	Project Overview: .....	4
3.0	Constraints/directives: .....	4
4.0	Project Approach .....	4
5.0	Functional requirements.....	5
6.0	Technical specifications.....	5
7.0	Recommended solutions .....	6
7.1	Systems Architecture Diagram .....	6
7.2	Software Architecture Diagram (optional) .....	6
8.0	Acceptance Criteria.....	7
8.1	Usability .....	7
8.2	Reliability.....	7
8.3	Performance .....	7
8.4	Supportability.....	7
9.0	Supplementary Specifications.....	7
9.1	Online User Documentation and Help System Requirements.....	7
9.2	Purchased Components.....	7
9.3	Interfaces.....	7
9.4	User Interfaces .....	7
9.5	Software Interfaces .....	7
9.6	Communication Interfaces .....	7
10.0	Critical Success Factors and Assumptions .....	7

## 1.0 Introduction

In an era where browsing is ubiquitous, the challenge lies in efficiently analyzing multiple websites to grasp essential information. Current methods often lead to information overload and difficulty in synthesizing data across various sources. Our project presents the Integrated Content Aggregation and Knowledge Synthesis (ICAKS) extension, driven by these challenges. With the help of this browser extension, users will be able to extract and summarize data from various currently active web pages, which will improve their productivity and time management more efficiently.

Through the automated gathering and arrangement of real-time DOM data, ICAKS aims to deliver maximum effectiveness in terms of information perception and summarization. By utilising machine learning and natural language processing (NLP), the extension provides users with a unified overview of aggregated content from active websites.

### 1.1 Purpose

The goal of this project is to develop the Integrated Content Aggregation and Knowledge Synthesis (ICAKS) extension to automate the extraction and correlation of data from multiple web pages, as stated in the (SRS).

### 1.2 Scope

The project's scope includes developing the Integrated Content Aggregation and Knowledge Synthesis (ICAKS) extension to use natural language processing (NLP) to automate the extraction of data from numerous web pages. This covers backend development for data processing as well as text summarization. The project also intends to create an intuitive user interface for a better browsing experience.

### 1.3 References

- [1] Zhao, Tongde, and Khoa Tran. "A powerful CHROME EXTENSION: TRANSLATION PROGRAM USING PYTHON, WEBSITE ANALYSIS AND GOOGLE FIREBASE SERVICES." *CS & IT Conference Proceedings*. Vol. 13. No. 7. CS & IT Conference Proceedings, 2023.
- [2] <https://maxai.me/>
- [3] <https://chromewebstore.google.com/detail/webchatgpt-chatgpt-with-i/lpfemeioodjbpieminkklglpmhngfcn>
- [4] <https://chrome.google.com/webstore/detail/copyfish-%F0%9F%90%9F-free-ocr-soft/eenjdnjldapjajjofmldgmkjaienebbj>

## **1.4 Abbreviations / Acronyms / Definitions**

**ICAKS** – Intelligent Content Aggregator and Knowledge Synthesizer

**DOM** – Document Object Model

**API** - Application Programming Interface.

## **2.0 Project Overview:**

The main goal of the extension is to use natural language processing techniques to automate the extraction of data from different web pages. Its goal is to improve content summarization and streamline information synthesis, which will enable effective data retrieval and analysis by cutting down on redundancy.

## **3.0 Constraints/directives:**

ICAKS generally includes: -

- All browsers must support the extension.
- Maximum length of the summarization
- Data privacy and security during data extraction and storage is imperative.
- Various webpage formats require algorithms that are flexible.

## **4.0 Project Approach:**

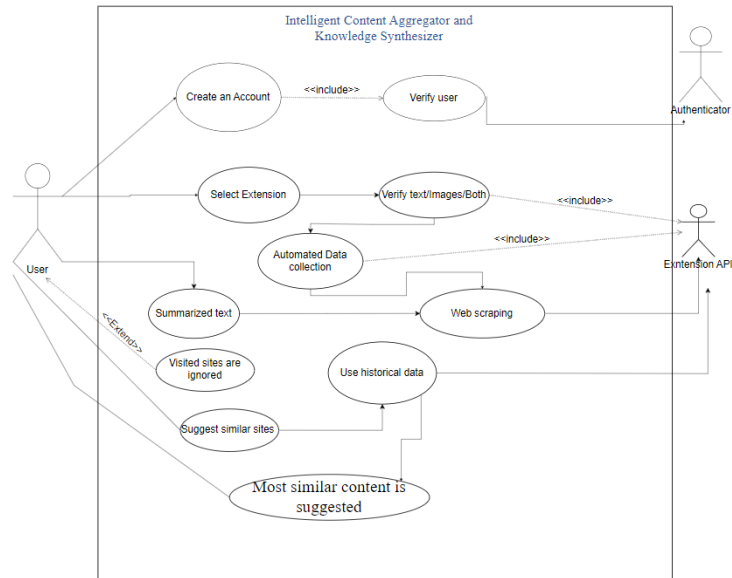
By streamlining the processes of information retrieval and summarization, this extension hopes to improve knowledge management and user productivity.

The project's goal is to gather DOM elements from visited web pages using browser extensions that are based on JavaScript. This will make it easier to extract text and image tags. Then, using rule-based extraction and tag elimination techniques, this data is processed and stored in a backend created with FastAPI (Python), where content extraction and summarization are accomplished. When a user closes their browser, the backend starts text summarising with APIs and other models, compiling the content that has been extracted for quick storage and retrieval.

## 5.0 Functional requirements:

- Content summarization
- Data segregation and storing
- Interface enables easy content access.

## Use Case Diagram-



## 6.0 Technical specifications:

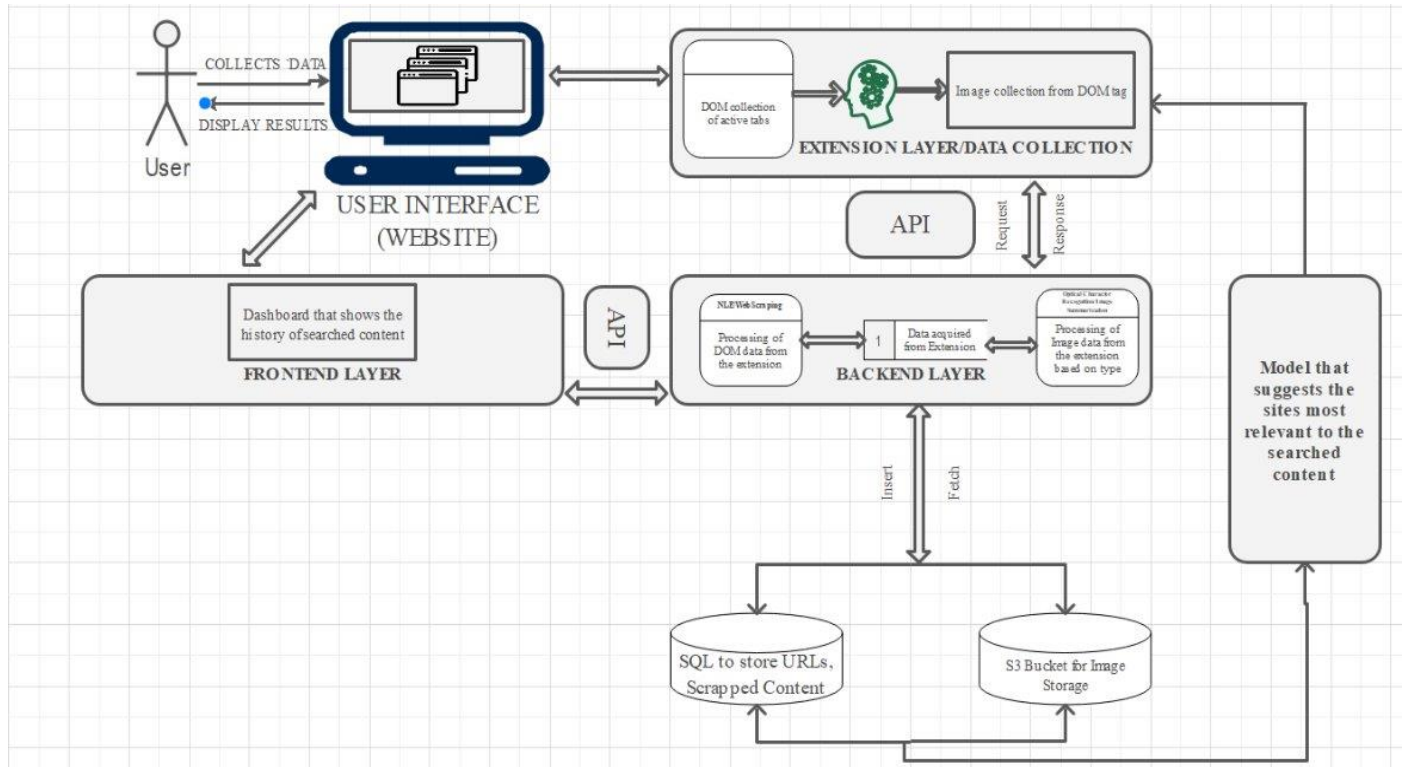
These are the prerequisites for website to access and thereby use it for summarization:

- Web application
- Chrome extension
- Data storage
- Text summarization



## 7.0 Recommended solutions:

### 7.1 Systems Architecture Diagram



## **8.0 Acceptance Criteria**

The extension extracts and summarizes content from visited web pages, providing concise and relevant summaries and the user interface allows seamless navigation and access to stored content.

### **8.1 Usability**

Analyze and understand content from various formats within the minimal time.

### **8.2 Reliability**

Exact parameters like Bleu, Rouge, Word Movers Distance, Perplexity scores of the summarization vs input are to be evaluated and taken in further implementations.

### **8.3 Performance**

- Assigning content such that the model can summarize in a less amount of time.
- The solution provided by the algorithm is evaluated and found if it's the best one.
- Handling the complexity in a faster way by integrating FastAPI.

### **8.4 Supportability**

Chrome Web Store and Website

## **9.0 Supplementary Specifications**

### **9.1 Online User Documentation and Help System Requirements**

User manuals or support will be provided.

### **9.2 Purchased Components**

Not Applicable

### **9.3 Interfaces**

User Interface  
Extension Interface

## **10.0 Critical Success Factors and Assumptions:**

- The Implementation of algorithm which can manage the summarization with respect to contents and tabs segregated.
- Add-ons are updated in further if there is any usage of tools/software provided by third-party.



# Intelligent Content Aggregator And Knowledge Synthesizer

**By**

Manjith Eguturi(20WU0101005)

Hotragn Pettugani(20WU0101030)

Vissarapu Srinath(20WU0101031)

**Mentor:** Dr. Kiran Kumar Ravulakollu



# AGENDA

---

Problem  
Statement

Background  
and  
Introduction

Literature  
Review

Use Case

Architecture  
Diagram

Methodology

Preliminary  
Results

Expected  
Outcomes

Gantt Chart

# Problem Statement

—

The problem lies in the time-consuming and often unsatisfactory process of retrieving relevant information from websites due to the vast amount of available content



# Background And Introduction

- Copy fish[4], MaxAI[2], and WebChatgpt[3] which can extract text or summarize w.r.t the feature of the extension
- Multiple Written styles from different websites
- Extension that can increase productivity and performance

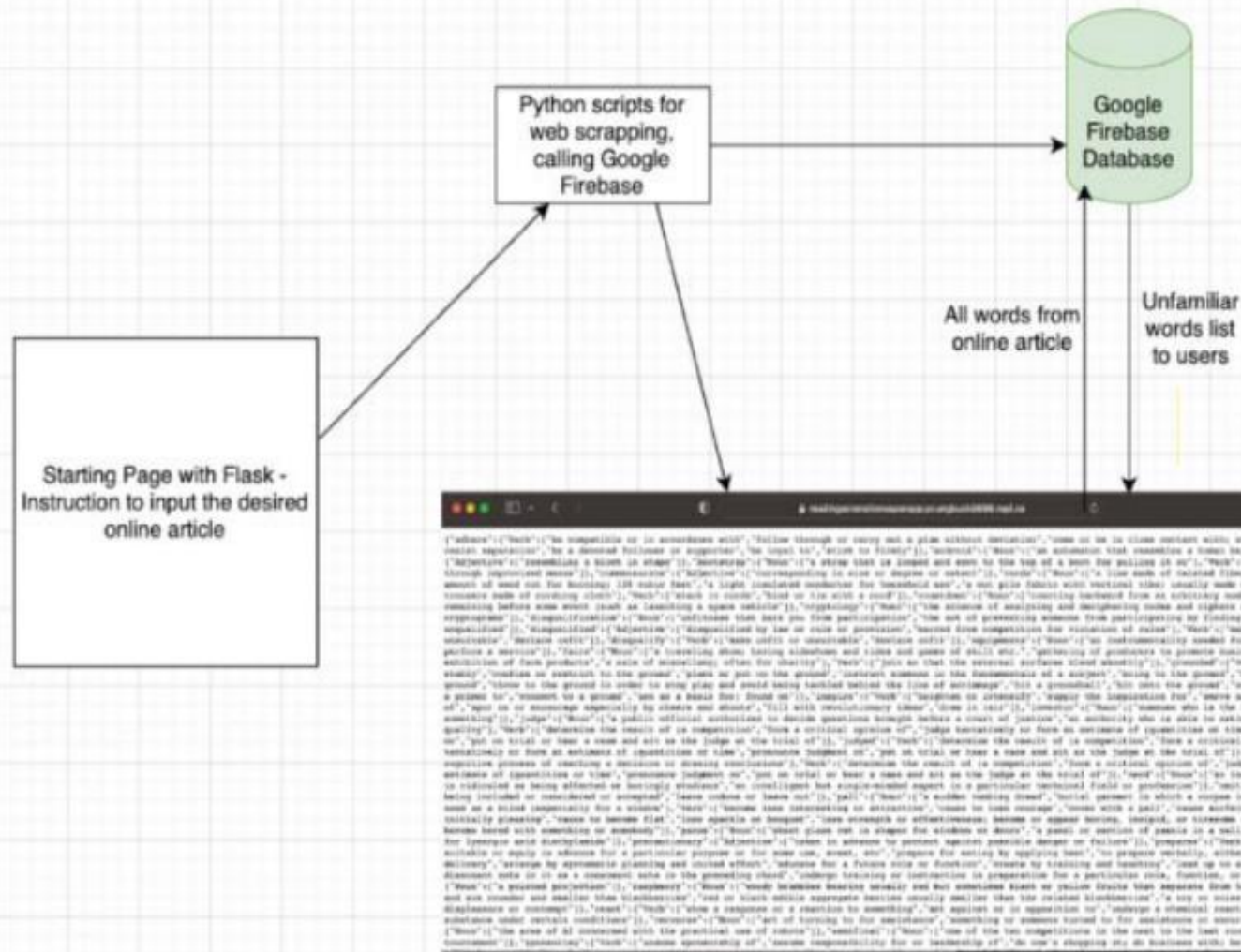


# Literature Review [1]

- An extension that has been developed to help users with poor English using the Python dictionaries, Google Cloud Firebase, and Flask. This extension will generate pre-defined difficult words based on an article/website by parsing the content.
- With the information extracted, it will map the definition of words generated to a dictionary and check if the extension can accurately generate the words.
- In addition, this extension will have a track of the most difficult words for users and use this information for future users to have a seamless experience.
- In conclusion, this extension provides a simple understanding of difficult words to users in English.



## Inferences from the Work [1]





---

## Existing Extension [2]

- The MaxAI extension has a variety of features that use most versions of ChatGPT and provide results for the user.
- It helps in summarizing a text from a single page, Generating the content for email, and even as an assistant chatbot.
- It works as an assistant and uses the ChatGPT 4 to produce results only with the premium plan whereas the free plan only provides limited ChatGPT usage per day and a mini menu for a text selected.

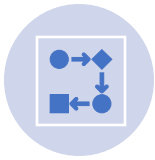


---

## Existing Extension [3]

- Given any prompt with web access, the extension can give a response based on the latest news in ChatGPT. It can even scrape data from a URL by restricting responses to a specific URL and giving responses respectively.
- It can perform a one-click prompts wrt required language, Tone, Writing style and Topic for a specific selected article/website.
- Many extensions can replicate the same usage, but the mentioned extensions are the popular ones with most users.

# Objectives



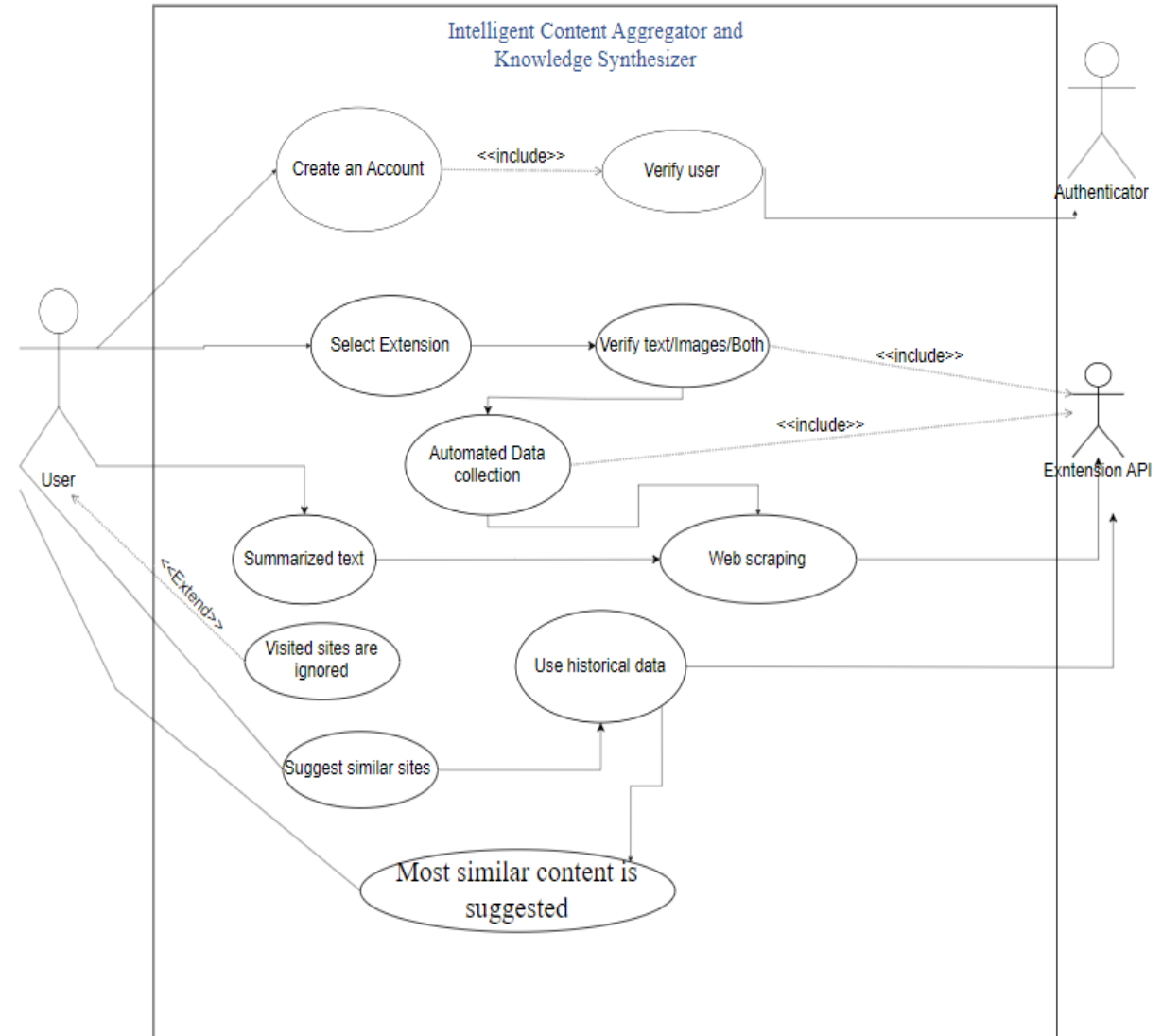
Help users to gather and summarize information automatically from visited websites.



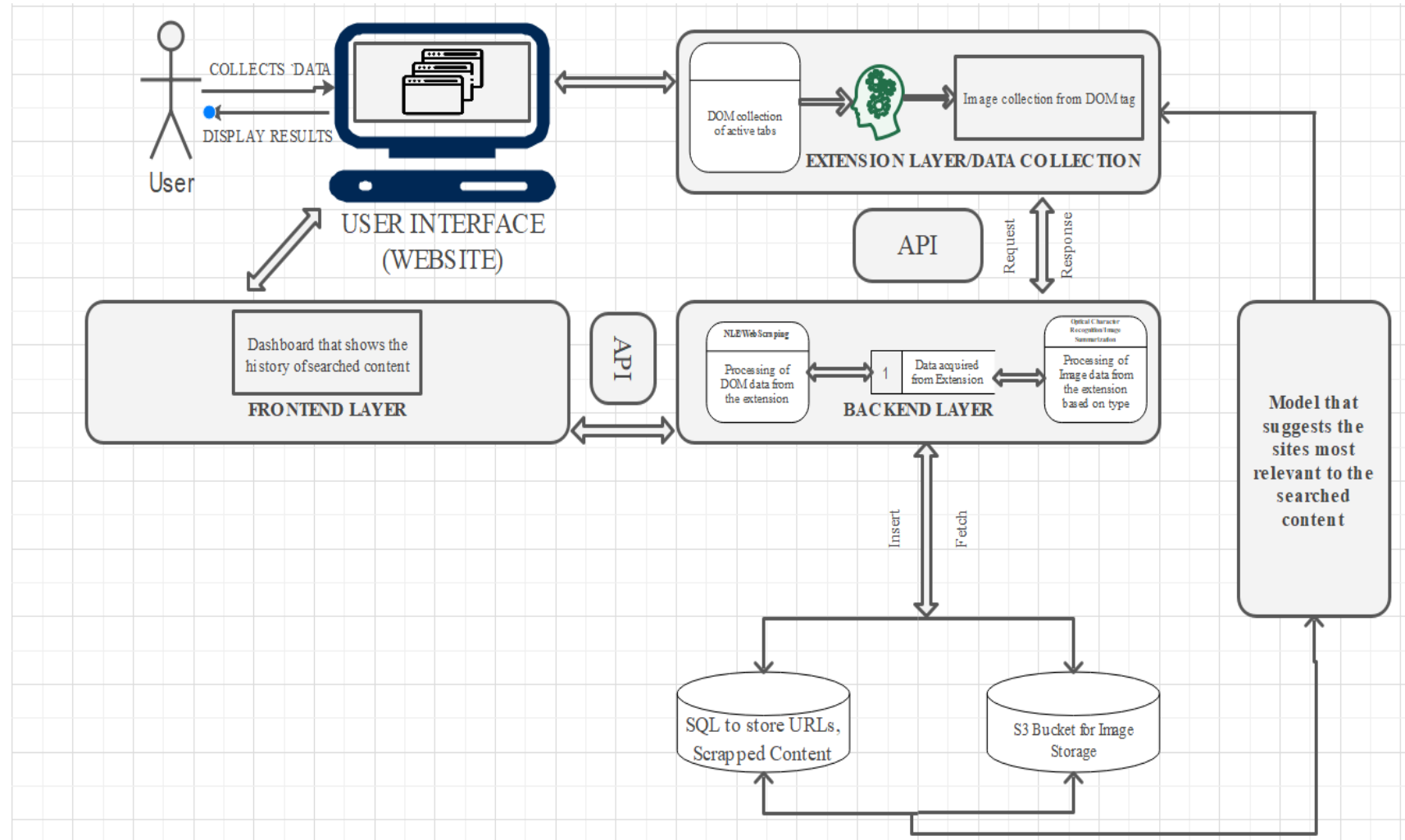
Overcome Challenges: Redundancy.

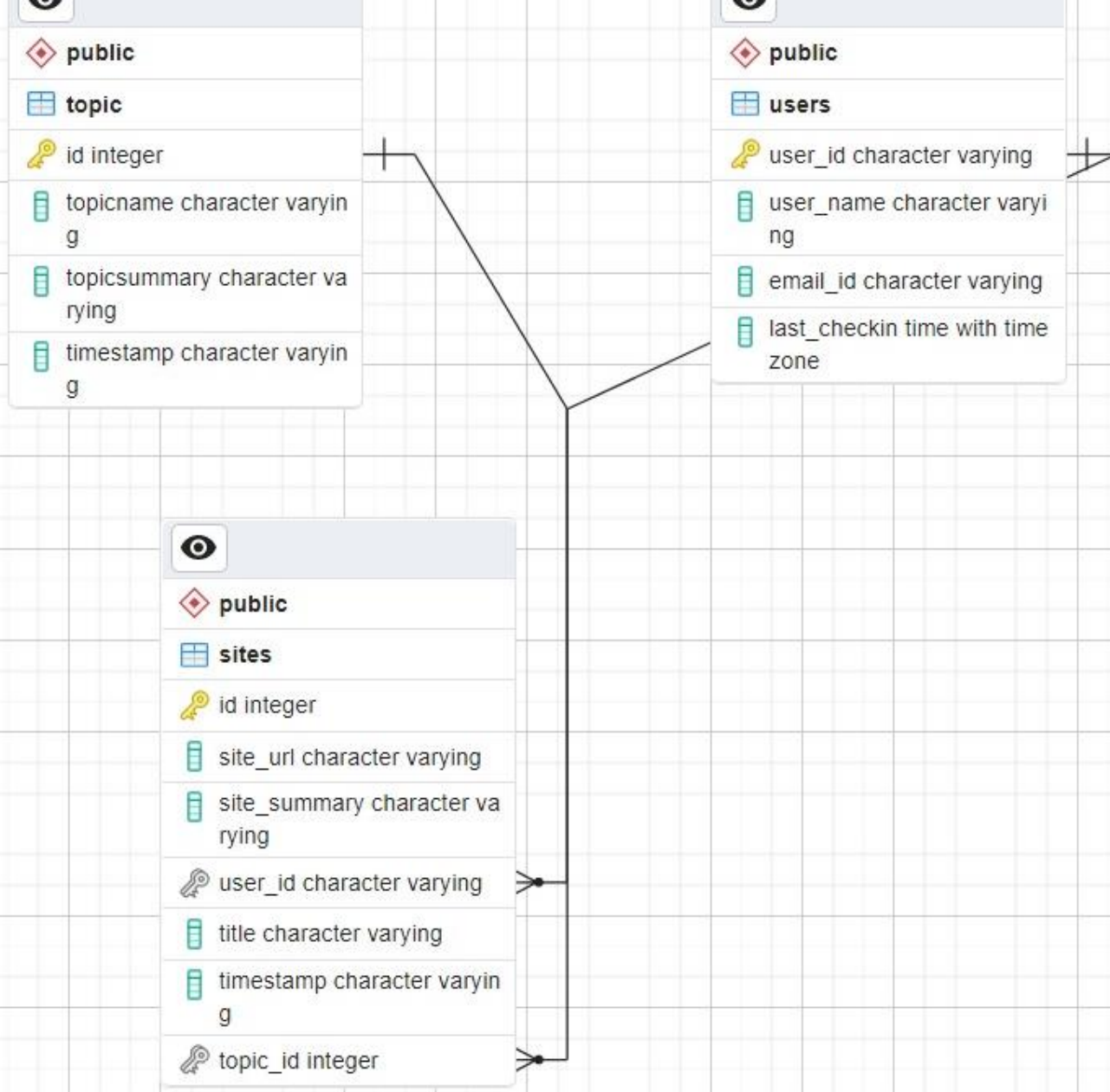


Minimal time to Research and site suggestions.



# System Architecture





# Data Base Design





# Methodology

1

DOM Elements from the visited pages will be collected through an extension which will be built using JavaScript. Once the DOM content is accessed it will be sent to the backend using an API. image tags will be extracted separately and will be sent to the backend in the form of URL or file format.

2

**Model Based Extraction-** As all websites do not follow the same structure, we try eliminating most of the unused content in the first step. We will try and build a machine learning model which should be able to separate tags based on their content.

3

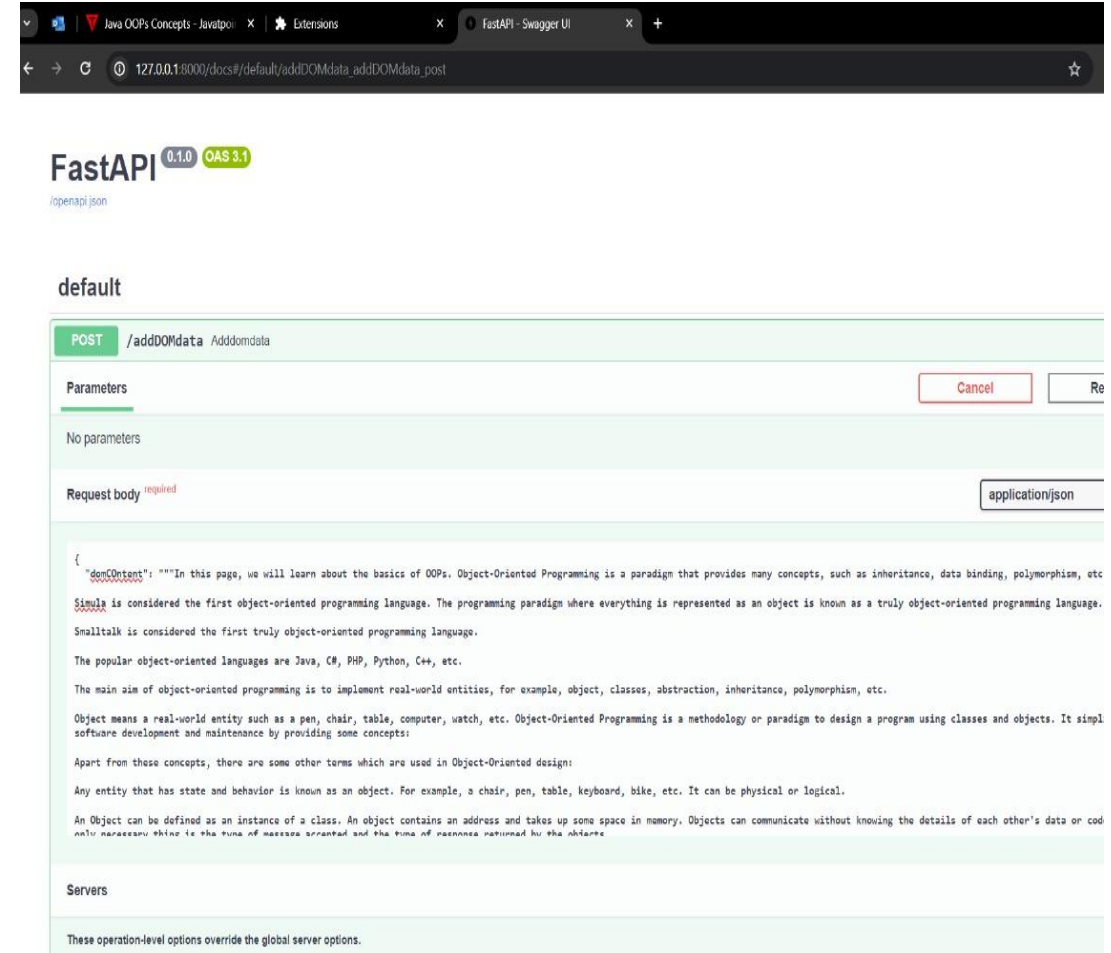
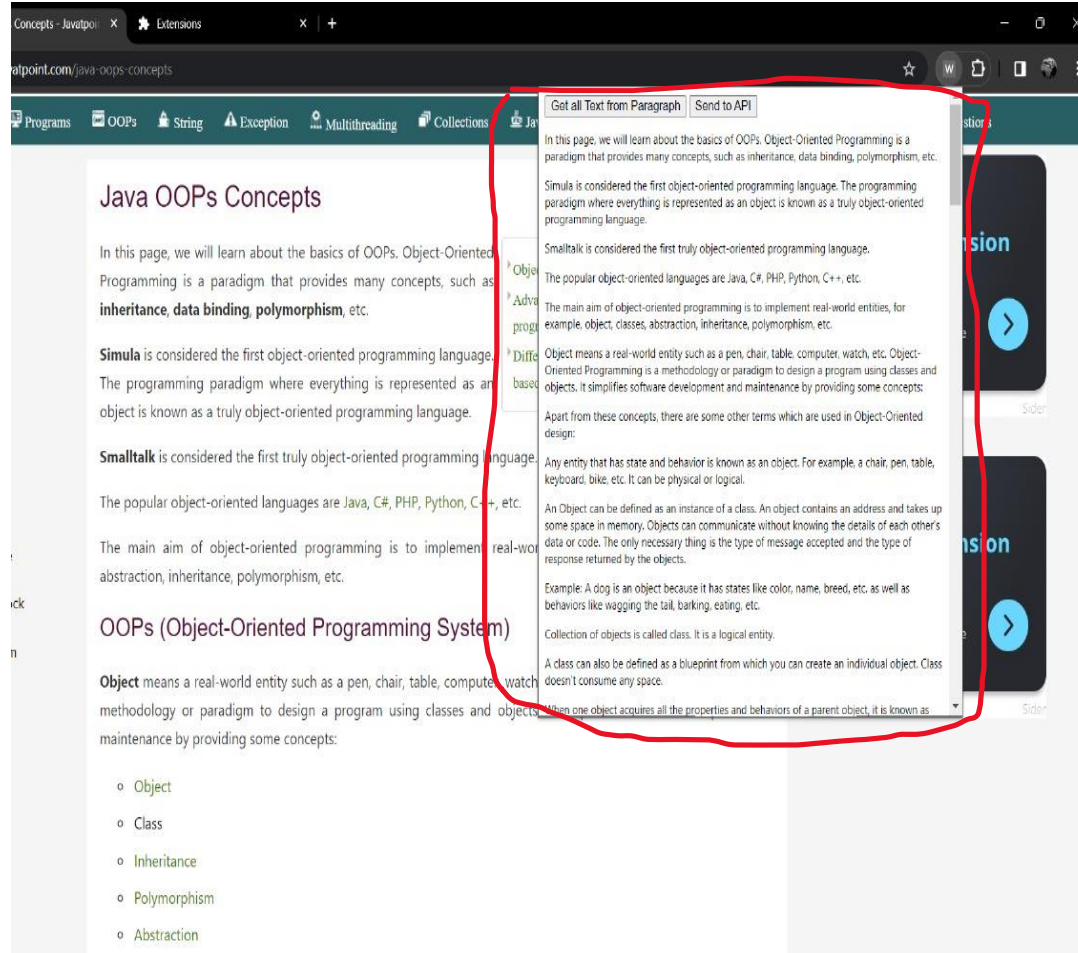
**Development of an algorithm that can do the below:**

Based on the title of the DOM. Once all documents are segregated, the summarization of each title will be performed. Further, there will be a summary of all similar titles together and a separate summary considering all content.

4

A website will be developed to visually track the stats of a user and similar content using the data repository constructed & updated based on extension usage

# Preliminary Results



**Extension for extracting text  
from active tab**

**API for receiving extracted  
content to the backend**



# Preliminary Results

2

The screenshot shows a web browser displaying the 'Java OOPs Concepts' page on Javatpoint.com. The page has a dark green header with navigation links: Home, Java, Programs, OOPs, String, Exception, Multithreading, and Collections. A left sidebar lists topics under 'Basics of Java', including 'Java Object Class', 'Java Inheritance', 'Java Polymorphism', 'Java Abstraction', and 'Java Encapsulation'. The main content area is titled 'Java OOPs Concepts' and contains introductory text about OOPs, mentioning Simula and Smalltalk as early languages, and listing popular languages like Java, C#, PHP, Python, and C++. It also defines OOPs as an Object-Oriented Programming System and lists core concepts: Object, Class, and Inheritance. A red box highlights a 'Get Summary' button and a text box containing a concise summary of OOPs: 'Object-oriented programming (OOPs) is a programming paradigm that uses the concept of "objects" to represent real-world entities, along with their attributes and behaviors. These objects can interact with each other to form complex systems. Some key concepts in OOPs include inheritance, data binding, polymorphism, abstraction, encapsulation, coupling, cohesion, association, aggregation, and composition. In OOPs, inheritance allows one object to acquire all the properties and behaviors of a parent object, promoting code reusability and runtime polymorphism. Polymorphism enables one task to be performed in different ways, while abstraction hides internal details and shows only the necessary functionality. Encapsulation binds code and data together into a single unit, enhancing security and reducing complexity. Objects can be associated with each other through relationships such as association, aggregation, and composition. Association represents a relationship between two objects, while aggregation and composition represent stronger forms of association where one object contains other objects as a part of its state. OOPs offers several advantages over procedural programming, including easier development and maintenance, data hiding, and improved simulation of real-world events. Popular object-oriented programming languages include Java, C#, PHP, Python, and C++. It's worth noting that object-based programming languages follow most of the features of OOPs, but they lack inheritance. Examples of object-based programming languages include JavaScript and VBScript. Additionally, while constructors in Java do not explicitly return a value, they play a crucial role in creating and initializing objects.'

Summarized content for the active tab

# Results

The screenshot shows a web browser window with the address bar displaying 'com/java-oops-concepts'. The page title is 'Java OOPs Concepts'. A navigation bar at the top contains links for 'ams', 'OOPs', 'String', 'Exception', 'Multithreading', 'Collections', 'JavaFX', 'JSP', and 'Spring'. The main content area has a heading 'Java OOPs Concepts' and a subheading 'In this page, we will learn about the basics of OOPs. Object-Oriented Programming is a paradigm that provides many concepts, such as inheritance, data binding, polymorphism, etc.' Below this, it states 'Simula is considered the first object-oriented programming language. The programming paradigm where everything is represented as an object is known as a truly object-oriented programming language.' It then mentions 'Smalltalk is considered the first truly object-oriented programming language.' and 'The popular object-oriented languages are Java, C#, PHP, Python, C++, etc.' The main aim of object-oriented programming is to implement real-world entities, for example, object, classes, abstraction, inheritance, polymorphism, etc. The section 'OOPs (Object-Oriented Programming System)' defines 'Object' as a real-world entity such as a pen, chair, table, computer, watch, etc. and 'Object-Oriented Programming' as a methodology or paradigm to design a program using classes and objects. It simplifies software development and maintenance by providing some concepts: 

- Object
- Class

 An overlay titled 'ICAPS' is visible on the right side of the page, containing the text 'Distill your reading :)' and buttons for 'Get Summary', 'Get All Tabs', 'Login', and 'Sign Up'. Two advertisements for 'Free ChatGPT Extension' are also visible on the right side of the page.

# Summarization of content from different websites

ICAHS

Dashboard

About

Java OOPs Concepts - Javatpoint

This text appears to be discussing the fundamentals of Object-Oriented Programming Systems (OOPs) and how they relate to the Java programming language. OOPs is a programming paradigm that includes concepts such as inheritance, data binding, and pol...

Show more

Programming

Object-Oriented Programming (OOPs) is a programming approach emphasizing inheritance, data binding, polymorphism, and abstraction. Its primary goal is implementing real-world entities such as objects and classes, along with enhancing code efficiency...

Show more

Java OOPs Concepts - Javatpoint

This text appears to be discussing the fundamentals of Object-Oriented Programming Systems (OOPs) and how they relate to the Java programming language. OOPs is a programming paradigm that includes concepts such as inheritance, data binding, and polymorphism. It aims to represent real-world entities through objects and classes. An "object" is a real-world entity that has state and behavior, while a "class" is a collection of objects that share common attributes and methods. Inheritance allows one object to acquire the properties and behaviors of a parent object, promoting code reusability and enabling runtime polymorphism. Polymorphism refers to performing a task in different ways, and in Java, it is achieved through method overloading and overriding. Abstraction involves hiding internal details and presenting functionalities to users. In Java, abstract classes and interfaces are used to achieve abstraction. Encapsulation binds code and data together into a single unit, ensuring data integrity and security. The text also mentions other terms used in object-oriented design, including coupling, cohesion, association, aggregation, and composition. These terms describe relationships between objects and their dependencies. Finally, the text compares object-oriented programming languages to object-based programming languages, noting that the latter lacks inheritance as a feature. Examples of object-based programming languages include JavaScript and VBScript.

Show less

Site Url

## FastAPI - Swagger UI

The text describes three topics obtained from GET/getalltopics API endpoint in FastAPI 0.1.0 with OpenAPI 3.1 specification. The first topic is about Object-Oriented Programming System (OOPS) concepts in Java. It explains fundamental concepts of OOPS like inheritance, data binding, polymorphism, encapsulation, and abstraction. It also covers other related terms such as coupling, cohesion, association, aggregation, and composition. The second topic introduces Programming, specifically focusing on Object-Oriented Programming (OOP) and Python. OOP is a programming approach that utilizes inheritance, data binding, polymorphism, and abstraction to enhance code efficiency, maintainability, and scalability. Keywords and constructs in Java, such as abstract class, interface, methods, constructors, static keyword, this keyword, instance initializer blocks, and final keyword, are discussed. Python is described as a simple, readable, and widely-used programming language with extensive community assistance and powerful libraries. The third topic pertains to Beverages but contains only code snippets, styling rules, and configurations for building a webpage, making it unsuitable for summarization. However, the code relates to a template script defining a div structure for menu items with server-side rendering, JavaScript functions handling postbacks and page request management, style declarations for specific DOM elements, interaction with third-party services or plugins, and configuration and initialization of analytics tracking. This response also provides information about response headers and controls accept header with example value schema string and no links schemas. Lastly, there is an embedded Swagger UI bundle code block showing the client-side representation of the API documentation generated using OpenAPI Specification.

## Extension Interface




# Ablation Study

Models	Bleu	Rouge	METEOR	BERT	Word Mover's Distance	Self-Bleu	Perplexity vp8-=/
MPT(MosaicML_84kTokens)	0.38	0.36	0.22	0.51	0.46	0.30	80
FLAN-T5	0.43	0.38	0.25	0.55	0.50	0.26	76
<u>philschmid/bart-large-cnn-samsum</u>	0.42	0.35	0.26	0.49	0.48	0.29	79
<u>Hugchat</u>	0.46	0.41	0.31	0.60	0.52	0.34	68
<u>Falconsai/text_summarization</u>	0.39	0.32	0.25	0.52	0.44	0.29	84
<u>ARTELab/it5-summarization-mlsum</u>	0.40	0.37	0.29	0.54	0.47	0.28	77





# Hugging Face

- Hugging Face's includes varied pre-trained models such as BART and T5, preparing to different summarization needs across various content types found in active tabs.
  - Models on Hugging Face are finely-tuned and consistently achieve top scores across evaluation metrics like BLEU, ROUGE, and METEOR, ensuring the generation of accurate and informative summaries for users.
  - Our extension can tailor the summarization process to suit different types of content, ensuring that summaries extracted from active tabs are relevant and comprehensive.
- 

# Expected Outcomes



Copy right  
registration

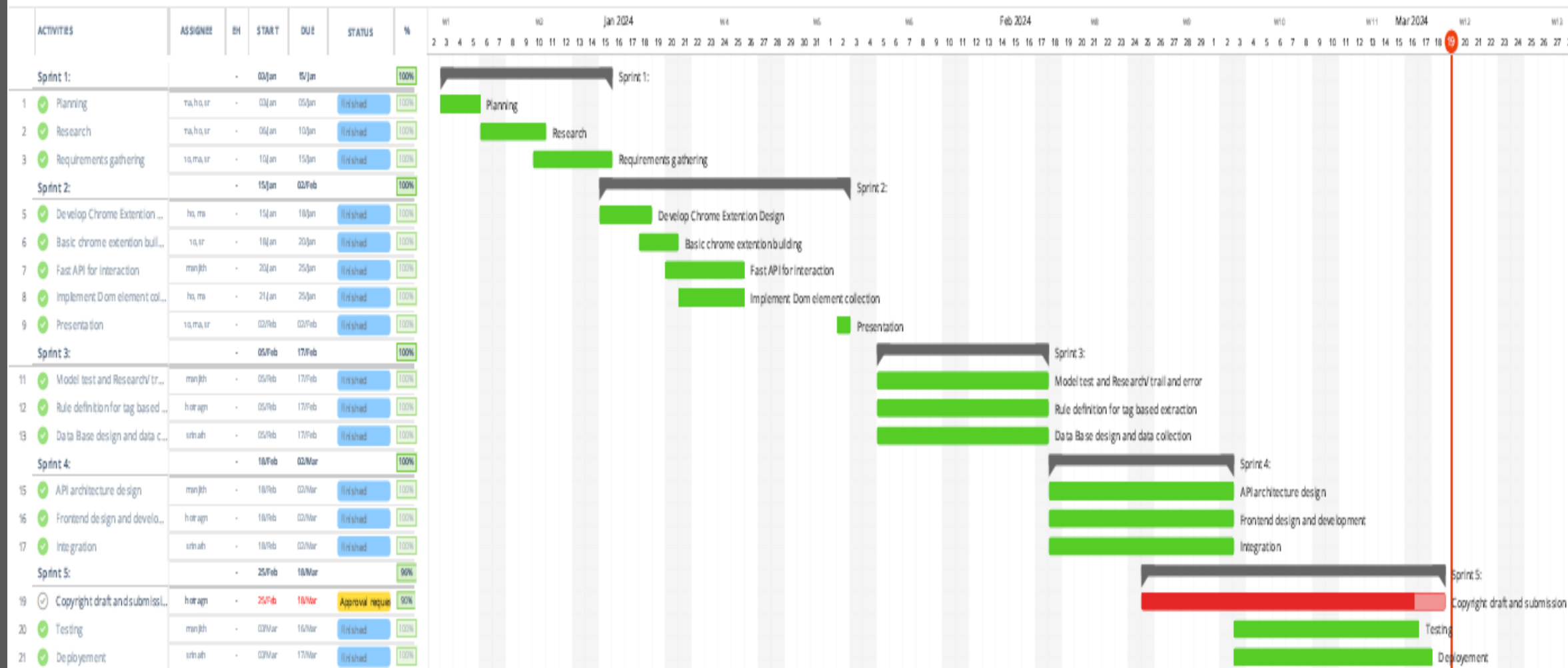
Chrome  
Extension

Platform for  
management of  
researched  
content

# Gantt Chart

intelligent content aggregator and knowledg...

Read-only view, generated on 19 Mar 2024





# References

- [1] Zhao, Tongde, and Khoa Tran. "A Powerful CHROME EXTENSION: TRANSLATION PROGRAM USING PYTHON, WEBSITE ANALYSIS AND GOOGLE FIREBASE SERVICES." *CS & IT Conference Proceedings*. Vol. 13. No. 7. CS & IT Conference Proceedings, 2023.
- [2] <https://maxai.me/>
- [3] <https://chromewebstore.google.com/detail/webchatgpt-chatgpt-with-i/lpfemeioodjbpieminkklglpmhlnghcn>
- [4] <https://chrome.google.com/webstore/detail/copyfish-%F0%9F%90%9F-free-ocr-soft/eenjdnjldapjajjofmldgmkjaienebbj>
- [5] <https://huggingface.co/mosaicml/mpt-7b>
- [6] <https://huggingface.co/philschmid/bart-large-cnn-samsum>
- [7] <https://huggingface.co/chat/>
- [8] [https://huggingface.co/Falconsai/text\\_summarization](https://huggingface.co/Falconsai/text_summarization)
- [9] <https://huggingface.co/ARTELab/it5-summarization-mlsum>

**Thank You**

