Intelligent Content Aggregator And Knowledge Synthesizer

By

Manjith Eguturi(20WU0101005)

Hotragn Pettugani(20WU0101030)

Vissarapu Srinath(20WU0101031)

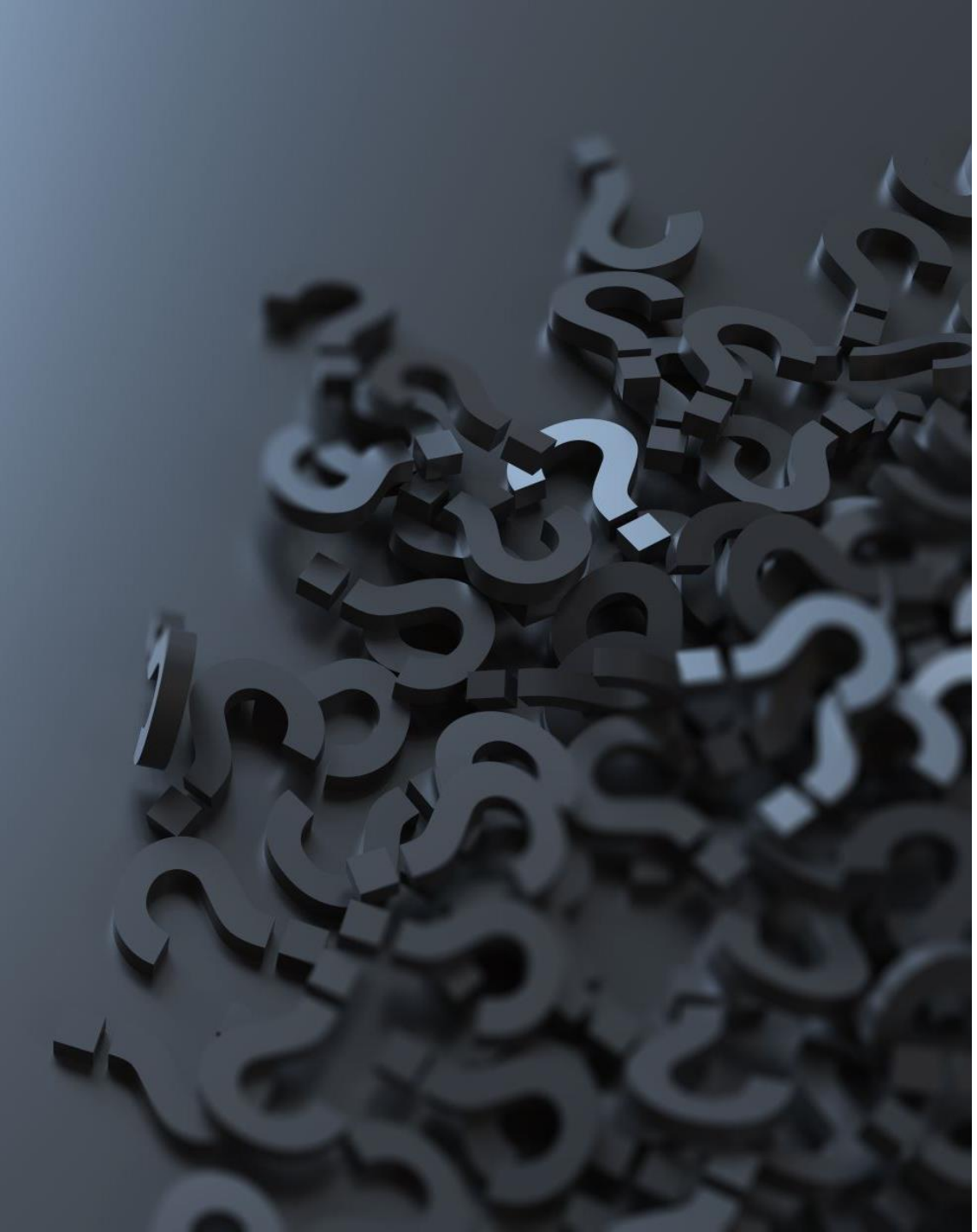
Mentor: Dr. Kiran Kumar Ravulakollu

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



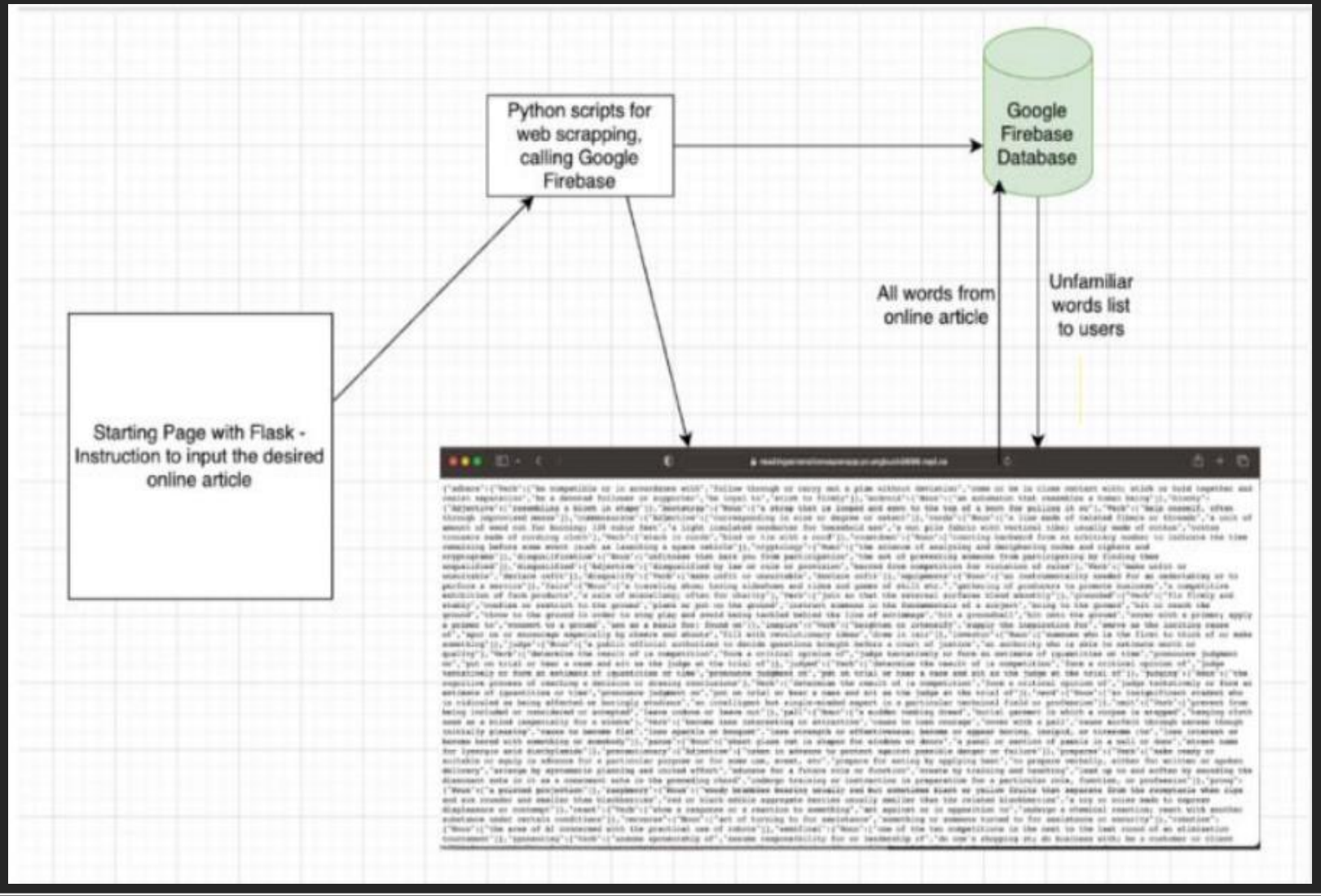
Background And Introduction

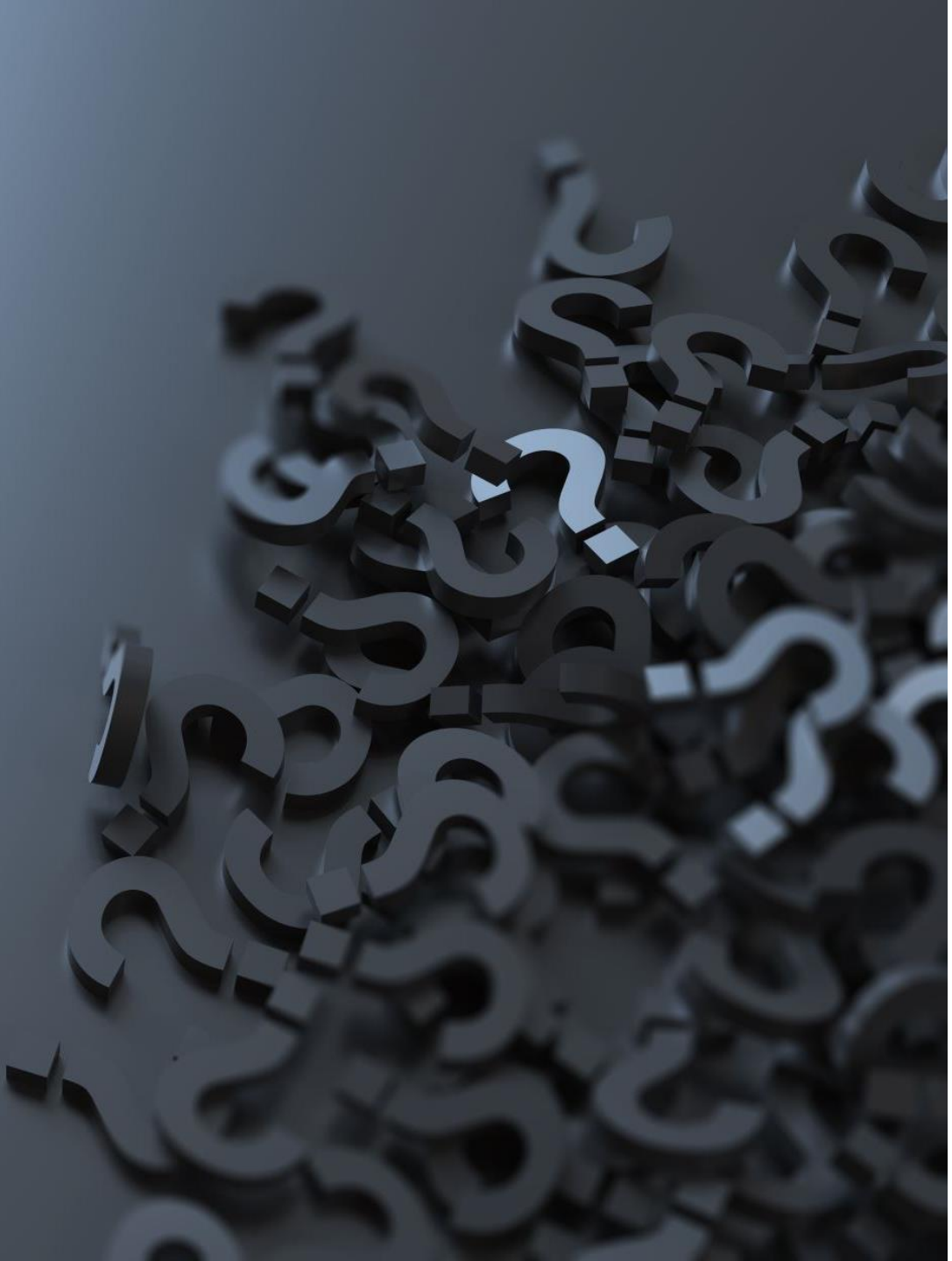
- Copy fish[4], MaxAI[2], and WebChatgpt[3] which can extract text or summarize w.r.t the feature of the extension
- Multiple Written styles from different websites
- Extension that can increase productivity and performance

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.

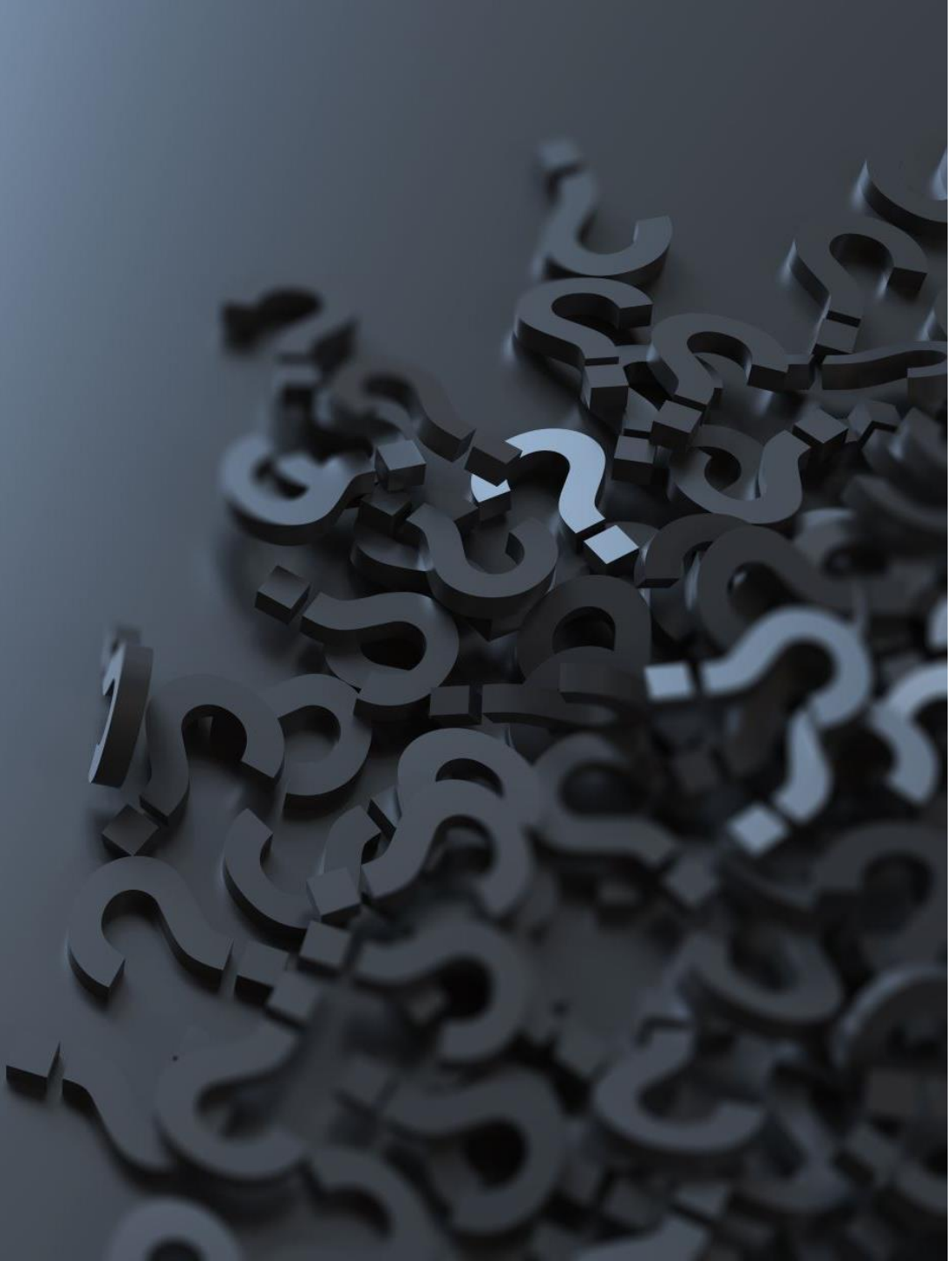
Inferences from the Work [1]





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]

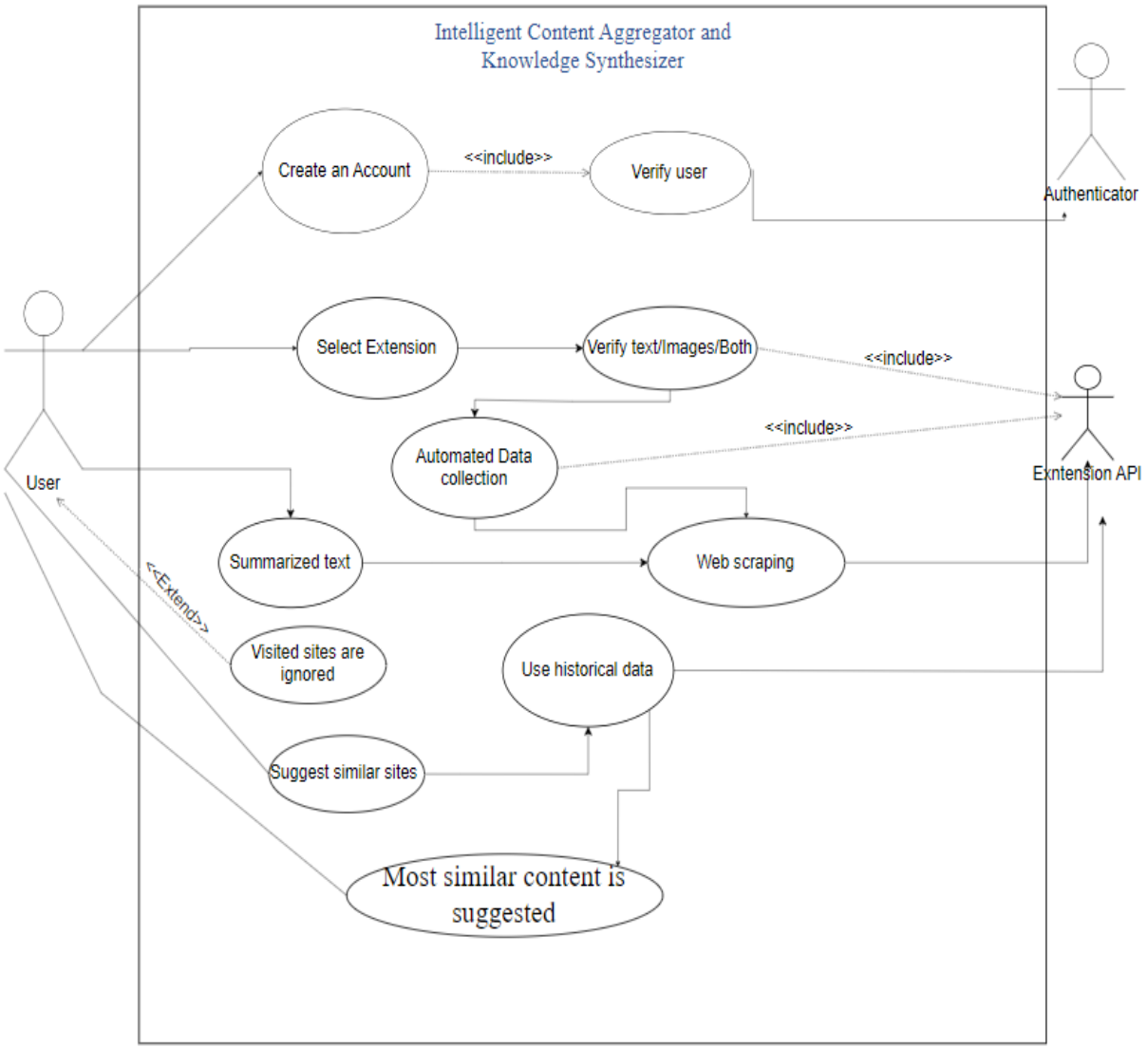
- Given any prompt with web access, the extension can give a response based on the latest news in ChatGPT. It can even scrape data from a URL by restricting responses to a specific URL and giving responses respectively.
- It can perform a one-click prompts wrt required language, Tone, Writing style and 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

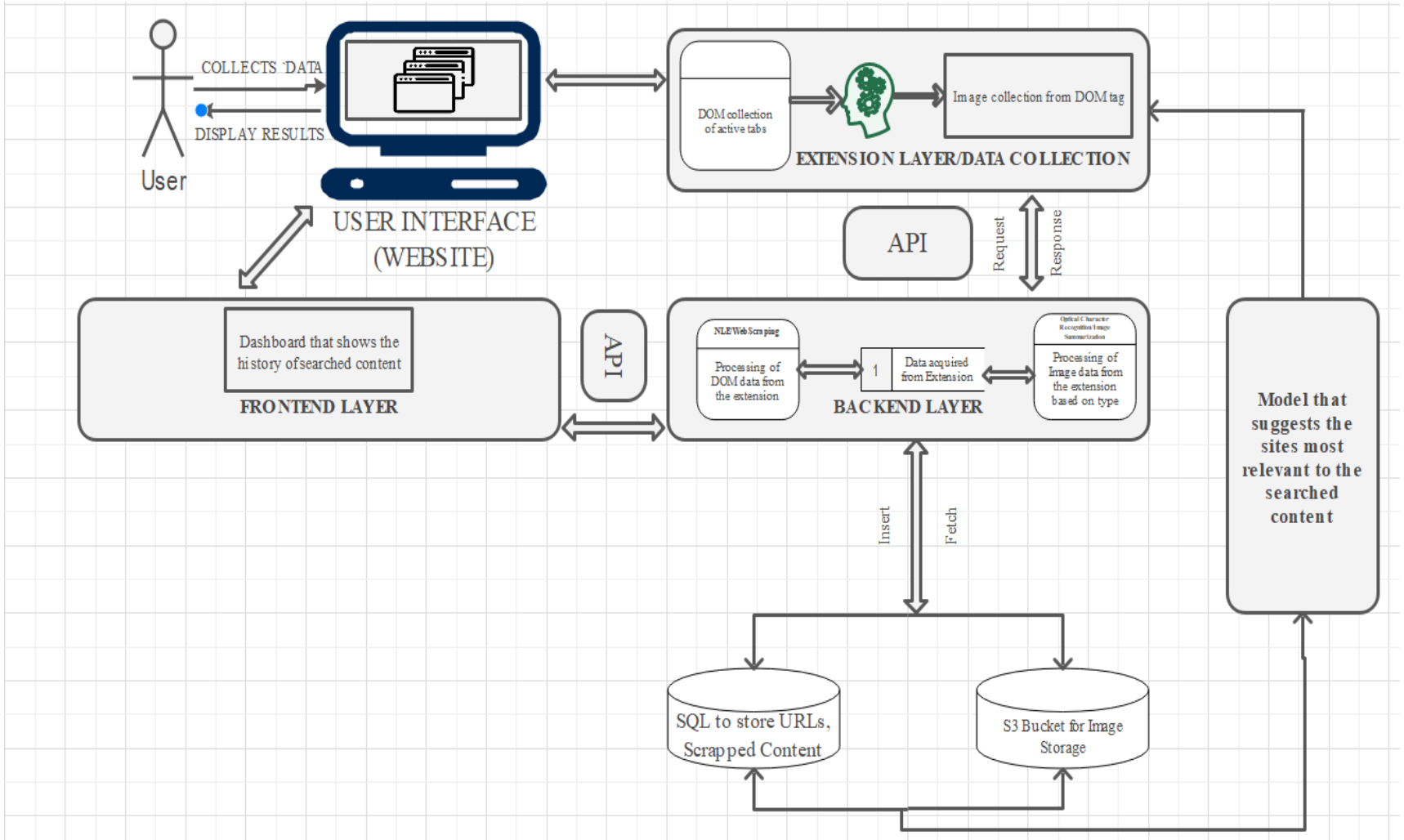
Help users to gather and summarize information automatically from visited websites.

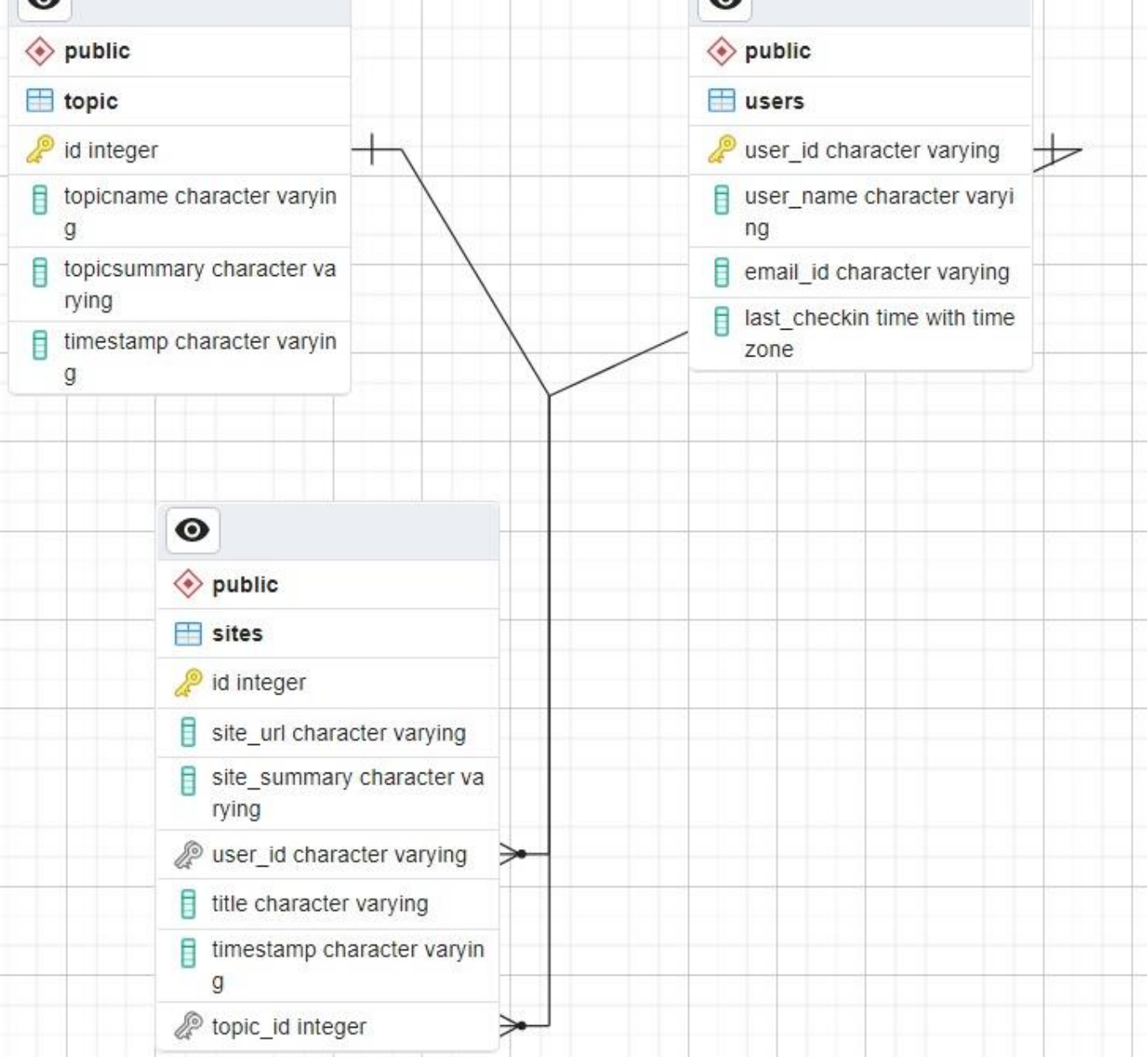
Minimal time to Research and site suggestions.

Overcome Challenges: Redundancy.



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:

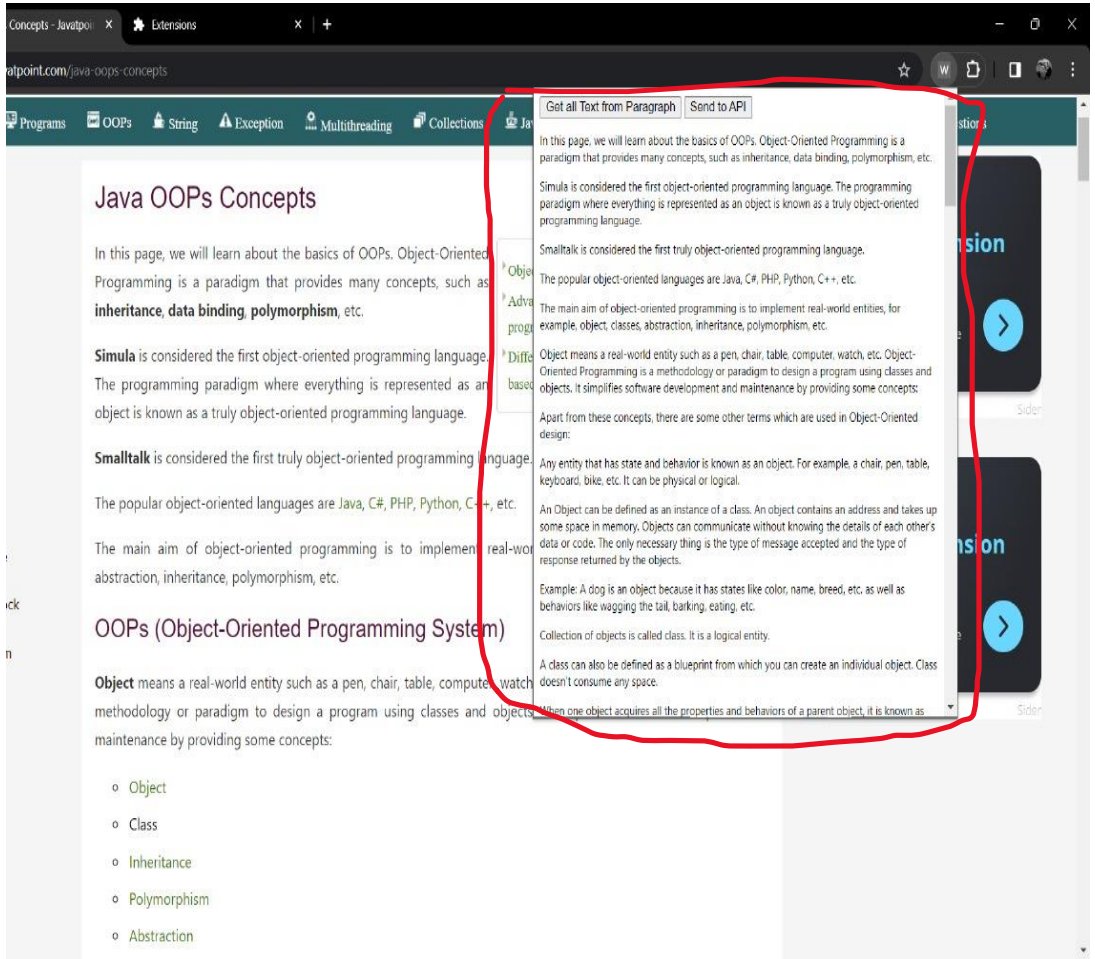
Based on the title of the DOM. Once all documents are segregated, the summarization of each title will be performed. Further, there will be a summary of all similar titles together and a separate summary considering all content.

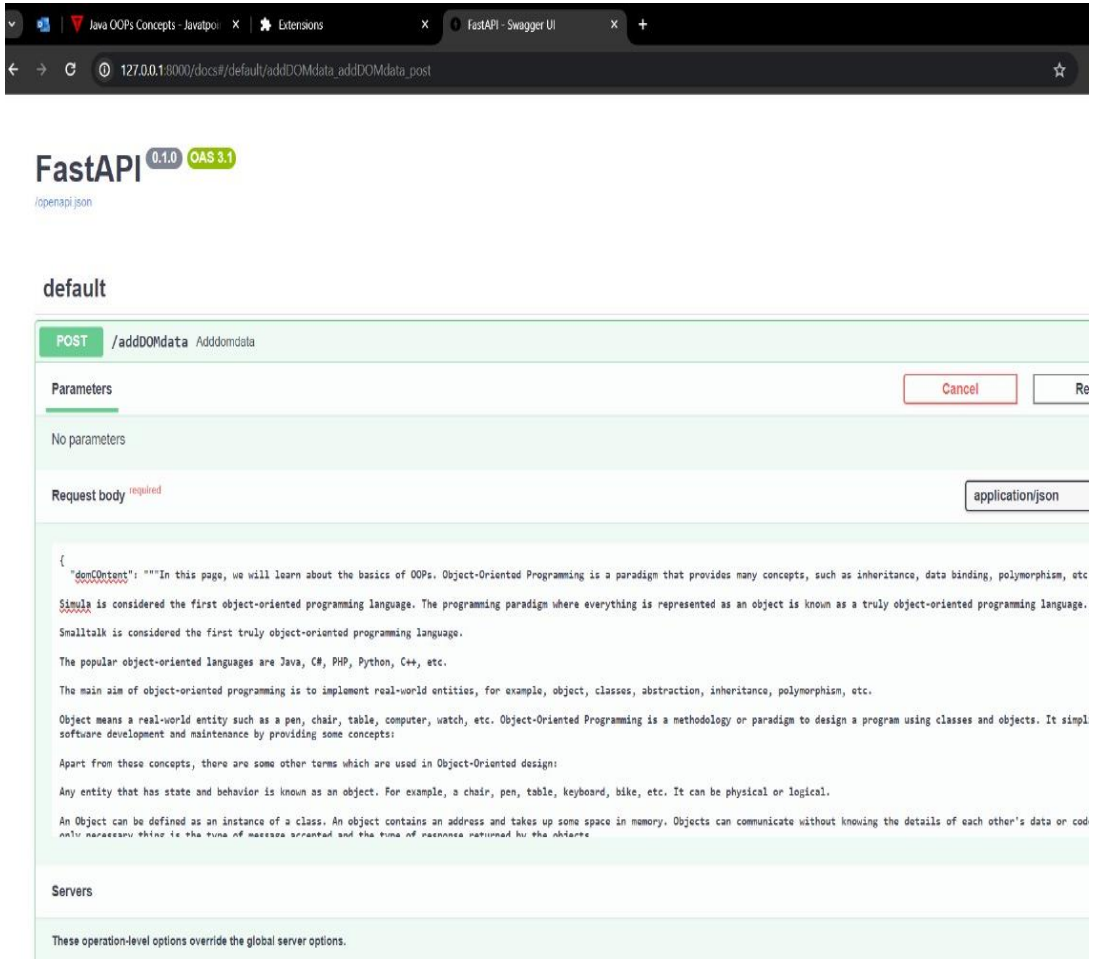
4

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

1



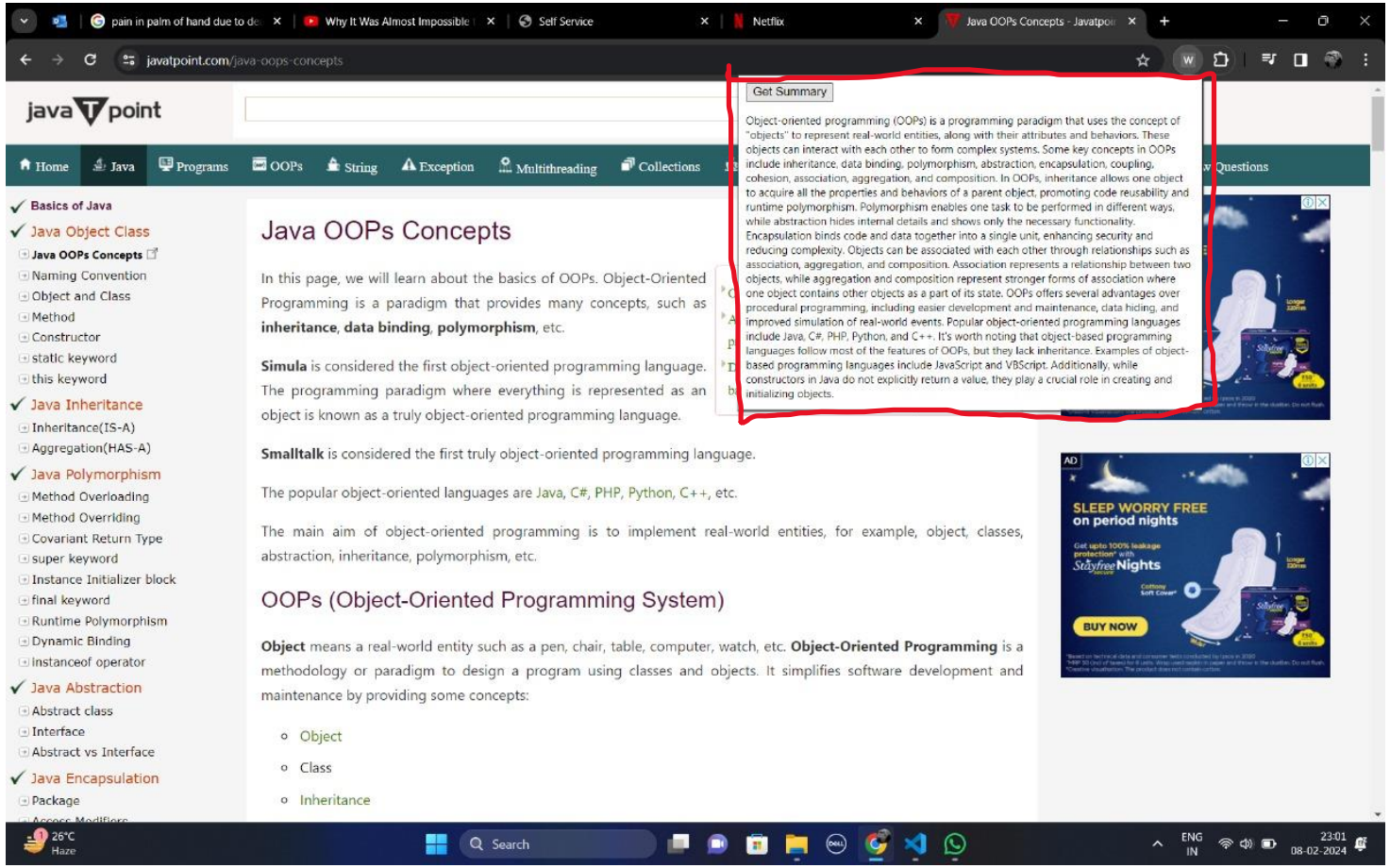


Extension for extracting text from active tab

API for receiving extracted content to the backend

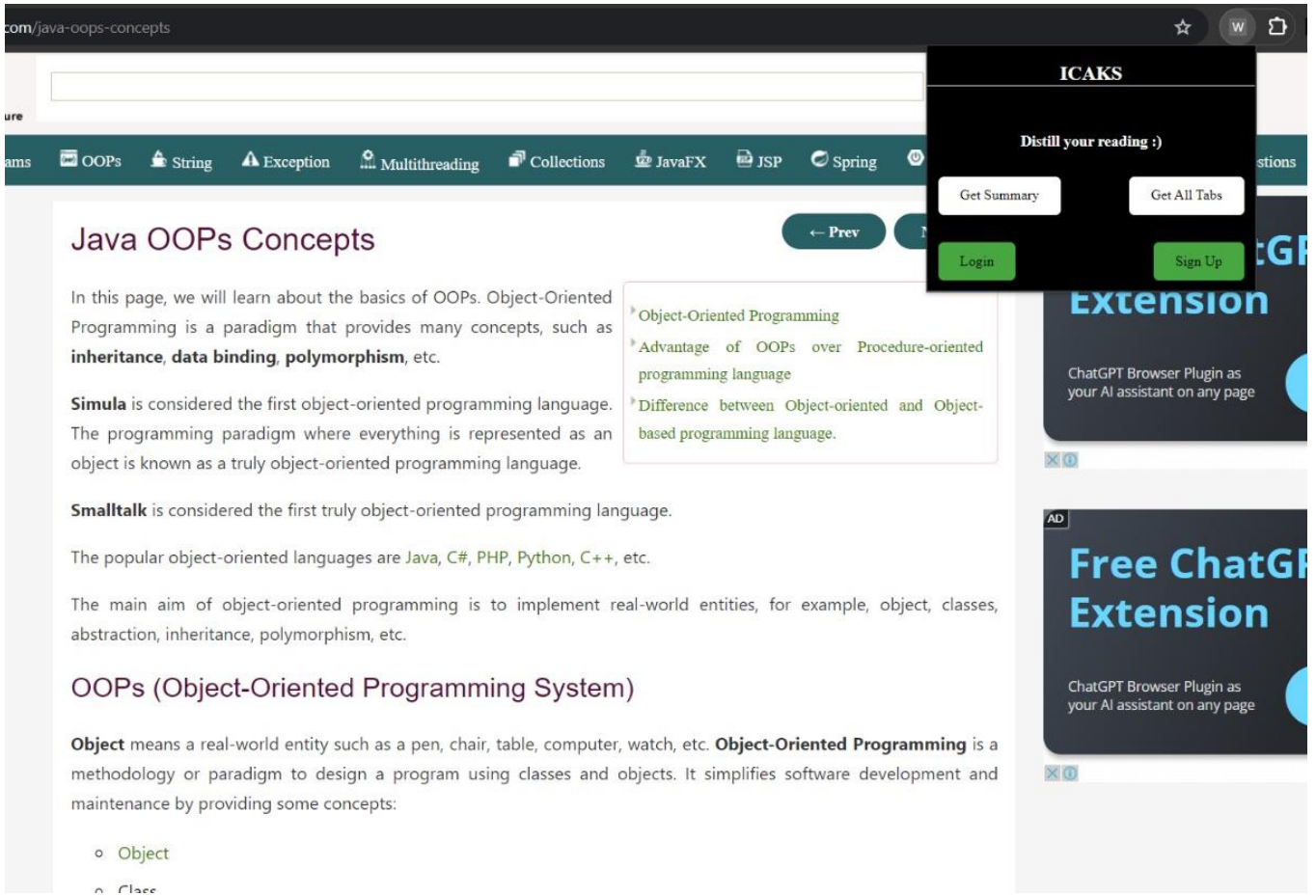
Preliminary Results

2

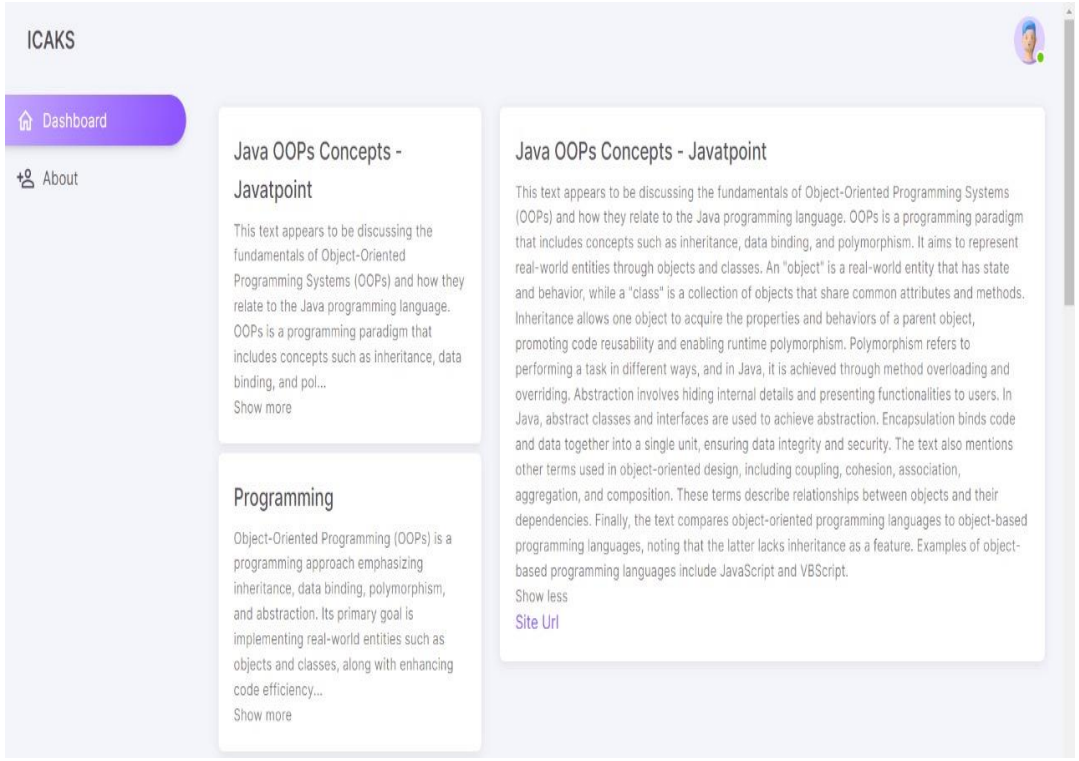


Summarized content for the active tab

Results


ICAKS
Distill your reading :)
Get Summary
Get All Tabs
Login
Sign Up
OOPs
String
Exception
Multithreading
Collections
JavaFX
JSP
Spring
Java OOPs Concepts
← Prev
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.
Object-Oriented Programming
Advantage of OOPs over Procedure-oriented programming language
Difference between Object-oriented and Object-based 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
Extension
ChatGPT Browser Plugin as your AI assistant on any page
AD
Free ChatGP
Extension
ChatGPT Browser Plugin as your AI assistant on any page

Summarization of content from different websites



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
philschmid/bart-large-cnn-samsum	0.42	0.35	0.26	0.49	0.48	0.29	79
Hugchat	0.46	0.41	0.31	0.60	0.52	0.34	68
Falconsai/text_summarization	0.39	0.32	0.25	0.52	0.44	0.29	84
ARTeLab/it5-summarization-mlsum	0.40	0.37	0.29	0.54	0.47	0.28	77

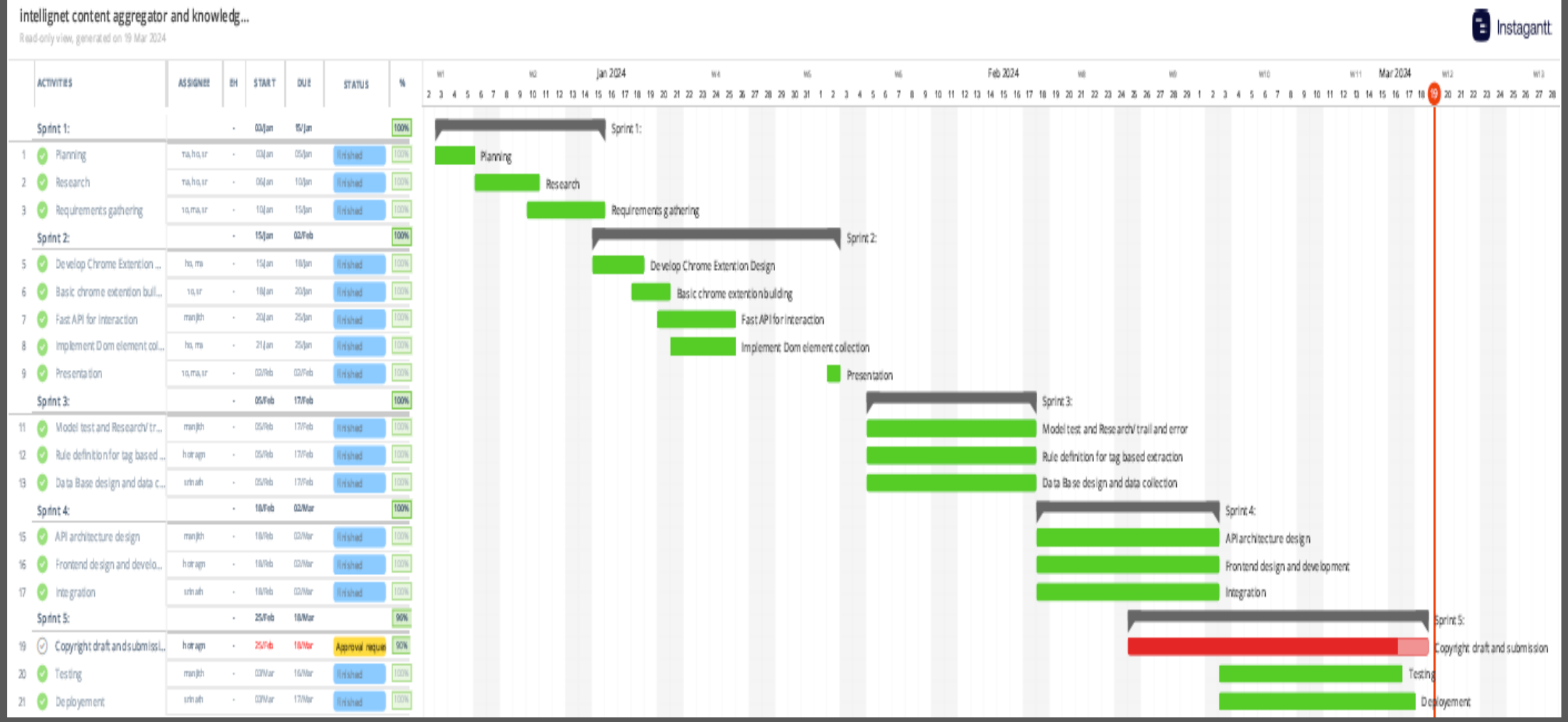
Hugging Face

- 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


intellignet content aggregator and knowledg...
Read-only view, generated on 19 Mar 2024
Instagantt
ACTIVITIES
ASSIGNEE
EH
START
DUE
STATUS
%
Jan 2024
Feb 2024
Mar 2024
Sprint 1:
Planning
Research
Requirements gathering
Sprint 2:
Develop Chrome Extention Design
Basic chrome extention building
Fast API for interaction
Implement Dom element collection
Presentation
Sprint 3:
Model test and Research/ trail and error
Rule definition for tag based extraction
Data Base design and data collection
Sprint 4:
API architecture design
Frontend design and development
Integration
Sprint 5:
Copyright draft and submission
Testing
Deployement

References

[1] Zhao, Tongde, and Khoa Tran. "APowerful 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/lpfemeioodjbpieminkklglpmhlngfcn

[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

[9] https://huggingface.co/ARTeLab/it5-summarization-mlsum

Thank You