

****

**Midterm Project Report**

**Advanced Computer Programming**

**Web Scraping with Python**

**Student Name: Maseko Mpendulo Mndeni**

**Student ID: 112025110**

**Teacher: DINH-TRUNG VU**

**2024-04**

# Introduction

## Github:

1. Personal Github Account: https://github.com/Superteam5
2. Group Github Account: https://github.com/Superteam5/2
3. Group Project Repository: https://github.com/Superteam5/2
4. List of submitted files:

**Topic**: Web Content Summarization using Text Analysis

**Project Overview:**

This project aimed to develop a Python script that scrapes text content from a website and generates a summary. The script utilizes the requests library to fetch the web content and BeautifulSoup to parse the HTML structure. The extracted text is then processed to create a summary based on sentence length.

**Implementation:**

Libraries:

requests: Used to fetch web content from a specified URL.

BeautifulSoup: Parses the downloaded HTML content to extract relevant text.

scrape\_and\_summarize Function:

Takes a URL and optional summary length as arguments.

Fetches the content using requests.get.

Parses the HTML with BeautifulSoup.

Extracts text content from paragraphs (<p>) using find\_all.

Tokenizes the text into sentences (assuming proper sentence structure).

Implements a simple summarization by selecting the top summary\_length sentences based on their word count (longer sentences are assumed to be more informative).

Returns the generated summary or None if an error occurs during scraping.

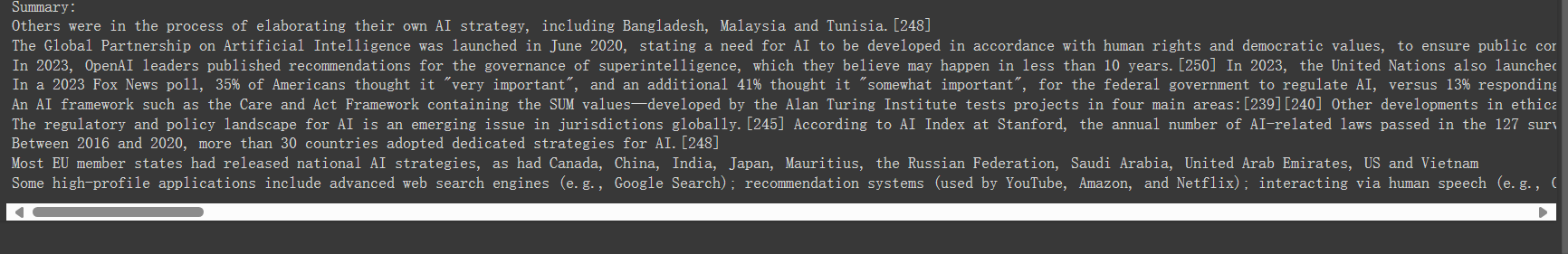
Error Handling:

The function includes basic error handling for cases where fetching the URL fails.

**Results:**

The script successfully scrapes text content from a provided URL and generates a summary based on the chosen parameters. By default, the summary includes the top three longest sentences.

Here is the screenshot of the summary.



**Conclusion:**

This Python script provides a basic solution for scraping text content from websites and generating summaries. While it currently uses a simple summarization approach, it offers a foundation for building more advanced text summarization techniques.