

****

**Midterm Project Report**

**Advanced Computer Programming**

**Web Scraping with Python**

**Student Name: Sinethemba Bhekinkosi Ndlangamandla**

**Student ID: 112025108**

**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:

-summary

## Topic

The project centers around the application of text processing and natural language processing (NLP) methodologies to automate the summarization of textual content.

## Project Overview

In today's information-rich environment, the ability to quickly extract key insights from large volumes of text is paramount. This Python script aims to address this challenge by implementing text summarization techniques using the spaCy library. By summarizing textual content, this project facilitates efficient information retrieval and aids in decision-making processes across various domains. The link for book: *https://gutenberg.ca/ebooks/hemingwaye-oldmanandthesea/hemingwaye-oldmanandthesea-00-t.txt.*

# Implementation

The implementation section delves into the technical aspects of the Python script execution:

## Fetching Text from URL:

## The script employs the requests library to retrieve textual content from a specified URL. This content may include articles, documents, or any other textual resource accessible via a URL.

## Text Processing with spaCy:

Leveraging the spaCy library, the script undertakes a series of text processing tasks:

* **Tokenization**: The fetched text is tokenized into individual words and sentences, enabling granular analysis of the content.
* **Word Frequency Calculation:** The script computes the frequency of occurrence for each word in the document, shedding light on the prominence of specific terms.
* **Stop Word Removal:** To focus on the essence of the text, common stop words (e.g., "the," "is," "and") are filtered out, leaving only meaningful content for analysis.
* **Sentence Score Computation:** Based on the normalized word frequencies, the script assigns scores to each sentence, capturing their relative importance within the document.

# Results

## Summary

The results section showcases the summarized content generated by the Python script. The summary comprises the most significant sentences identified by the algorithm, reflecting the essence of the original text.

Below is the screenshot for the output summary:

A screenshot of a computer

Description automatically generated

Conclusions

The conclusion section encapsulates the project's significance and implications:

Amidst the deluge of textual information available today, effective text summarization techniques offer a lifeline for navigating through vast volumes of content. By automating the summarization process, this project streamlines information retrieval, enhances decision-making, and facilitates knowledge discovery across diverse domains. Moreover, the utilization of NLP libraries like spaCy underscores the potential of computational linguistics in transforming unstructured text into actionable insights. Moving forward, advancements in text processing and NLP hold promise for revolutionizing information management and fostering innovation in numerous sectors, from academia and business to healthcare and beyond. Thus, this project serves as a testament to the transformative power of technology in unlocking the value embedded within textual data.