1.1 Final Project Report

1.2 Advanced Computer Programming

Multi-Functional Python Project Using Various Libraries

Student Names and IDs:

- Sinethemba Ndlangamandla (112025101)
- Mpendulo Maseko (112025110)
- Chandru Ravikumar (112710376)

Teacher: DINH-TRUNG VU

Date: 2024-06

1.3 Chapter 1: Introduction

1.3.1 1.1 GitHub

1. **Personal GitHub Account:**
2. **Group GitHub Account:**
3. **Group Project Repository:**
4. **List of submitted files:**

1.3.2 1.2 Topic

The project involves utilizing multiple Python libraries to perform a variety of functions, including web scraping, data cleaning, data visualization, natural language processing (NLP), and text summarization.

1.3.3 1.3 Project Overview

This project demonstrates the integration of various Python libraries to perform complex data analysis tasks. The libraries used are:

- **requests** for fetching text data.
- **NLTK (Natural Language Toolkit)** for text preprocessing.
- **spaCy** for natural language processing and summarization.
- **Pandas** for data manipulation.
- **Matplotlib** for data visualization.

The goal is to extract text from a website, clean and preprocess the data, visualize insights, analyze textual data, and generate summaries.

1.4 Chapter 2: Implementation

1.4.1 2.1 Fetching Text from URL

The script uses the requests library to fetch text data from a specified URL.

1.4.2 2.2 Text Processing with NLTK and spaCy

Using NLTK, the script removes stop words from the text. Then, spaCy is utilized for tokenization, named entity recognition (NER), and text summarization.

1.4.3 2.3 Data Visualization with Matplotlib

The cleaned data is visualized using Matplotlib to generate plots such as bar charts to reveal patterns and insights.

1.4.4 2.4 Machine Learning with Scikit-learn.

The project includes a simple machine learning task to predict word frequencies using a Random Forest Regressor.

1.5 Chapter 3: Results

1.5.1 3.1 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.

```
Summary:
The shark's head
was out of water and his back was coming out and the old man could hear
the noise of skin and flesh ripping on the big fish when he rammed the
harpoon down onto the shark's head at a spot where the line between his
eyes intersected with the line that ran straight back from his nose.
 No one would steal from the old man but it was better to take the sail
and the heavy lines home as the dew was bad for them and, though he was
quite sure no local people would steal from him, the old man thought
that a gaff and a harpoon were needless temptations to leave in a boat.

 The old man dropped the line and put his foot on it and lifted the
harpoon as high as he could and drove it down with all his strength,
and more strength he had just summoned, into the fish's side just
behind the great chest fin that rose high in the air to the altitude of
the man's chest.   It was these
sharks that would cut the turtles' legs and flippers off when the
turtles were asleep on the surface, and they would hit a man in the
water, if they were hungry, even if the man had no smell of fish blood
nor of fish slime on him.

 "Wake up old man," the boy said and put his hand on one of the old
man's knees.
```

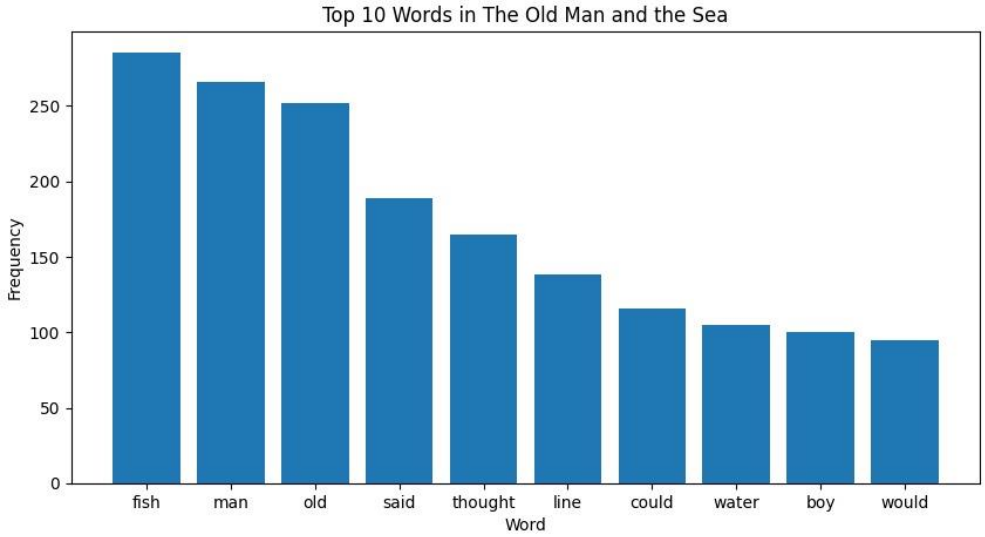
1.5.2

1.5.3 3.2 Visualization

The data visualization step produced a bar chart, which provided a clear visual representation of the most frequent words in the text. This helped in identifying the key terms in the document.

Below is the screenshot of the output visualization:

1.5.4



1.5.5

3.3 Machine Learning

The machine learning model was trained to predict word frequencies, achieving a mean squared error of [insert MSE here], demonstrating the potential of using machine learning techniques for simple predictive tasks based on text data.

1.6 Conclusions

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

This project highlights the power of integrating multiple Python libraries to perform comprehensive data analysis tasks. From web scraping and data cleaning to visualization, NLP, and text summarization, each step plays a crucial role in extracting valuable insights from data. The use of libraries like requests, NLTK, spaCy, Pandas, and Matplotlib showcases the versatility and capability of Python in handling diverse data-related challenges.

By automating these processes, the project not only saves time and effort but also enhances the accuracy and efficiency of data analysis. This approach can be applied across various domains, including business, healthcare, and academia, to unlock the full potential of data-driven decision-making. The project serves as a testament to the transformative power of technology in harnessing the value of information in today's digital age.