Jawaharlal Nehru University

Keyword Extractor

NLP Mini Project

Submitted to: Prof. Piyush Pratap Singh

Submitted By: Shikha Singh
Batch: 2022-24
MCA 3rd Sem

Introduction

The "Keyword Extractor" project is a mini Natural Language Processing (NLP) project that aims to extract keywords or key phrases from textual content. It provides a graphical user interface (GUI) for users to input text either by pasting it into a text area or by opening text files in various formats (TXT, PDF, DOCX). The project primarily utilizes Python and several libraries to achieve its objectives.

- Simplify Keyword extraction from text document.
- Utilizes Tkinter for the user friendly graphical interface.
- Process .txt, .pdf and .docx files.
- Used for S.E.O and text Summarization.

Objective

- The primary goal of the Keyword Extractor project is to assist users in automatically identifying the most significant keywords and key phrases within textual content.
- This can be especially useful for tasks such as document summarization, content indexing, or SEO (Search Engine Optimization).
- The project simplifies the process by providing an easy-to-use interface, making it accessible to individuals who may not have extensive programming or NLP knowledge.

Project Components

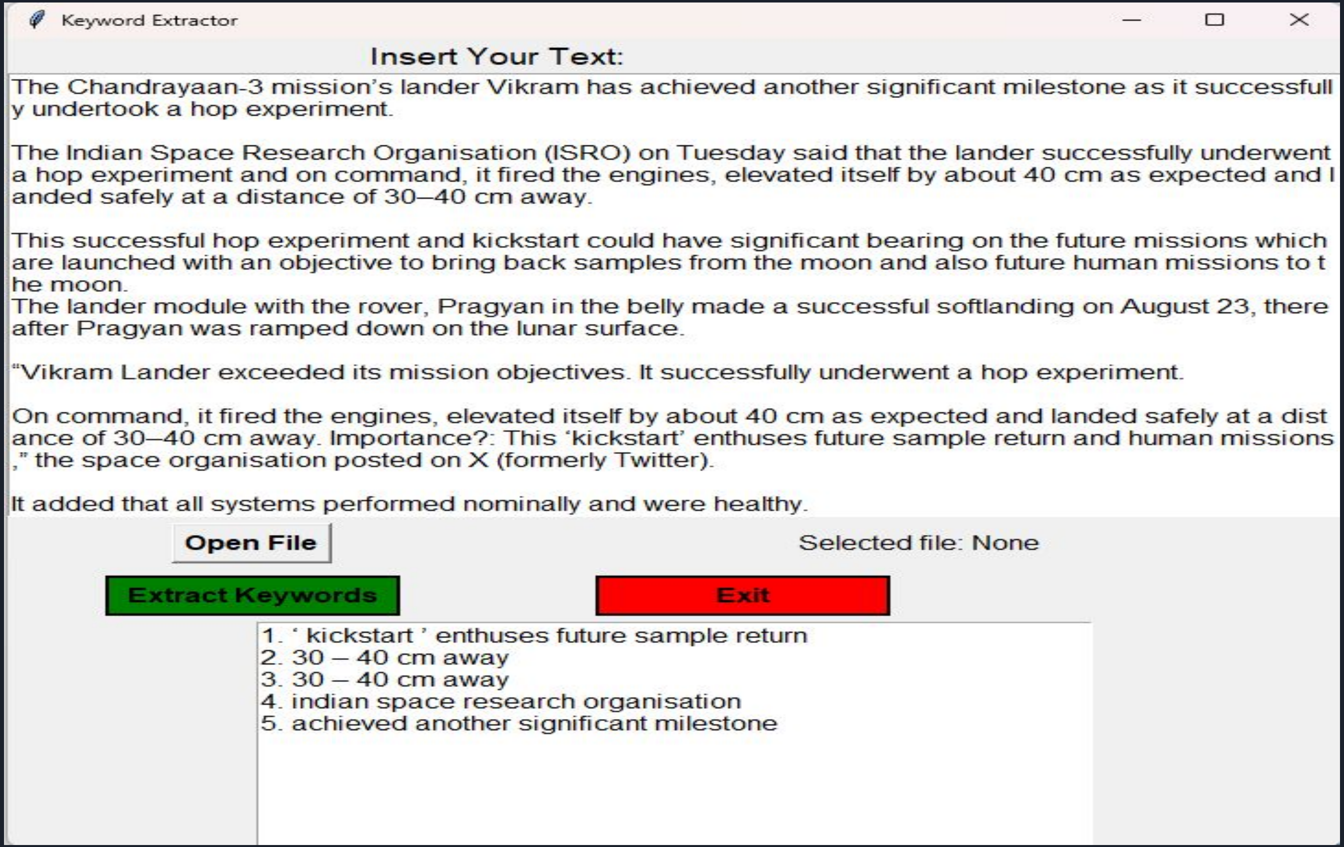
- Graphical User Interface (GUI) using Tkinter
- Text file input (TXT, PDF, DOCX)
- Text extraction from files
- Keyword extraction using Rake-NLTK
- Displaying extracted keywords

Technologies Used

List of technologies, libraries, and tools used in the project:

- Python
- Tkinter
- pdfplumber
- nltk
- Rake-NLTK
- docx

Demonstration - GUI



The Chandrayaan-3 mission's lander Vikram has achieved another significant milestone as it successfully undertook a hop experiment.

The Indian Space Research Organisation (ISRO) on Tuesday said that the lander successfully underwent a hop experiment and on command, it fired the engines, elevated itself by about 40 cm as expected and landed safely at a distance of 30–40 cm away.

This successful hop experiment and kickstart could have significant bearing on the future missions which are launched with an objective to bring back samples from the moon and also future human missions to the moon.
The lander module with the rover, Pragyan in the belly made a successful softlanding on August 23, thereafter Pragyan was ramped down on the lunar surface.

"Vikram Lander exceeded its mission objectives. It successfully underwent a hop experiment.

On command, it fired the engines, elevated itself by about 40 cm as expected and landed safely at a distance of 30–40 cm away. Importance?: This 'kickstart' enthuses future sample return and human missions," the space organisation posted on X (formerly Twitter).

It added that all systems performed nominally and were healthy.

```
1. ‘ kickstart ’ enthuses future sample return
2. 30 – 40 cm away
3. 30 – 40 cm away
4. indian space research organisation
5. achieved another significant milestone
```

Text Extraction

1. Text Extraction from TXT Files:

- For plain text (TXT) files, text extraction is straightforward. The project reads the content directly from the file without the need for additional libraries.
- It uses standard file I/O operations to open and read the contents of the TXT file.
- The extracted text is then available for keyword extraction and display.

2. Text Extraction from PDF Files:

- PDF files require a specialized library like "pdfplumber" to extract text. "pdfplumber" is a Python library for extracting text and other information from PDF documents.
- The project utilizes "pdfplumber" to open and process the PDF file, extracting text from each page. The text from all pages is concatenated to form the complete document text.

3. Text Extraction from DOCX Files:

- For Word documents in the DOCX format, the project relies on the "docx" library, which allows reading and extracting text from these files.
- "docx" provides methods to access the content of paragraphs in the document, and the text from each paragraph is collected to create the full document text.

Keyword Extraction

- The core functionality of the project is keyword extraction. It employs the "Rake-NLTK" library to identify and extract keywords from the provided text.
- Rake-NLTK is a Natural Language Processing library specifically designed for keyword extraction.
- Rake-NLTK uses heuristics, focusing on word co-occurrence and frequency, to identify potential keywords.
- It scores keywords based on their frequency and word proximity in the text.

THANK YOU