Keyword Extraction App Using TF-IDF

Introduction

This project is a Streamlit application that allows users to extract keywords from various types of documents (PDF, DOCX, PPTX, and TXT) using the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm. The application provides a user-friendly interface for uploading files, extracting text, and displaying the extracted keywords along with their corresponding scores.

Methodology

Text Extraction

The application supports extracting text from the following file formats:

- **PDF**: The `extract_text_from_pdf` function utilizes the `PyPDF2` library to extract text from PDF documents.
- **DOCX**: The `extract_text_from_docx` function uses the `docx` library to extract text from Microsoft Word documents.
- **PPTX**: The `extract_text_from_pptx` function employs the `pptx` library to extract text from Microsoft PowerPoint presentations.
- **TXT**: For plain text files, the `uploaded_file.read().decode('utf-8')` method is used to read the file contents.

Keyword Extraction

The `extract_keywords` function implements the TF-IDF algorithm using the `TfidfVectorizer` class from the `sklearn.feature_extraction.text` module. It takes the extracted text as input and returns a list of tuples, where each tuple consists of a keyword and its corresponding TF-IDF score.

The `clean_and_filter_keywords` function performs additional filtering and cleaning on the extracted keywords. It removes punctuation, converts keywords to lowercase, filters out stop words, and removes keywords with a length less than or equal to 4 characters. Additionally, it ensures that only unique keywords consisting of one or two words are included in the final list.

User Interface

The Streamlit application provides a user-friendly interface for interacting with the keyword extraction functionality. Users can upload files, view the file contents, and see the extracted keywords along with their scores. The application also offers additional features such as filtering keywords, displaying the top 100 keywords, and downloading the extracted keywords as a text file.

Dependencies

The following libraries are required to run the application:

- streamlit
- nltk
- PyPDF2
- docx
- pptx
- re
- sklearn

Usage

To run the application, execute `streamlit run app.py` in your terminal. The application will launch a Streamlit interface, where users can follow the on-screen instructions to upload files and extract keywords.

Conclusion

This Keyword Extraction App Using TF-IDF provides a convenient way to extract relevant keywords from various types of documents. The application's user-friendly interface and additional features make it a useful tool for text analysis and information retrieval tasks.