Introduction

This document provides an overview of the process of PDF text extraction and categorization. The goal is to illustrate how to extract text from PDF files, categorize it based on predefined tags, and store the categorized text in a structured format.

Methodology

The methodology involves several steps:

1. Extracting text from the PDF document using the PyPDF2 library.

2. Using regular expressions to identify and categorize sections of text based on keywords.

3. Storing the categorized text in a CSV file for easy analysis and manipulation.

Results

The results of this process show that text can be effectively extracted and categorized from PDF documents. The categorization is based on the occurrence of specific keywords that indicate different sections of the document.

Conclusion

In conclusion, PDF text extraction and categorization are useful techniques for processing and analyzing large volumes of text data from PDF documents. This approach can be used in various applications, including document management, information retrieval, and data analysis