SCHOOL OF TECHNOLOGY AND APPLIED SCIENCES

CENTRE FOR PROFESSIONAL AND ADVANCED STUDIES

(Established by Government of Kerala)

Affiliated To Mahatma Gandhi University

Chuttippara, Pathanamthitta-689645

MINI PROJECT
ABSTRACT
ON
METADATA EXTRACTION TOOL USING PYTHON

SUBMITTED BY

Ashna S M

Ayishathul Adila K

SUBMITTED TO

Abdul Muhammed Rasheed

ABSTRACT

Metadata Extraction Tool Using Python

Metadata extraction is a crucial task in various domains, including information retrieval, data analysis, and content management. Python, being a versatile programming language, offers a wide range of libraries and tools that can be leveraged for extracting metadata from different types of files and documents. In this project, we aim to develop a metadata extraction tool using Python. The proposed tool will utilize various Python libraries such as `pandas`, `numpy`, `nltk`, and `beautifulsoup` to extract metadata from different file formats such as PDF, Word documents, images, and web pages. The tool will be designed to handle both structured and unstructured data, extracting relevant metadata fields like title, author, creation date, keywords, and more. The tool's workflow will involve parsing the input files using specific libraries tailored to each file format. For example, `PyPDF2` can be used to parse PDF documents, while `python-docx` can handle Word documents. Image metadata can be extracted using the `PIL` (Python Imaging Library) or `opencv-python` libraries. For web pages, the tool can utilize web scraping techniques with the help of `beautifulsoup` to extract metadata from HTML tags. The extracted metadata will then be stored and organized using the `pandas` library for further analysis and processing. The tool can be extended to support additional file formats and metadata fields based on specific requirements. Additionally, it can incorporate natural language processing techniques using the `nltk` library to extract relevant information from textual data, such as sentiment analysis, entity recognition, or topic extraction. Overall, the metadata extraction tool developed using Python will provide a flexible and efficient solution for extracting metadata from various file formats and enable further analysis and management of the extracted information.

Requirements:

1. Hardware Requirements:
   - Processor : Intel Core i3-9100/AMD Ryzen 3 3200G required
   - Hard Disc : A minimum of 100 MB to 500 MB of free disk
   - Main Memory : 4 GB or more recommended

2. Software Requirements:
   - Operating System : Windows 10 recommended
   - Programming Language : Python(With necessary libraries)