MBOX Data Extraction: Methodology and Implementation

This document details the structured approach employed to extract and process data from MBOX files, along with the implementation of a Python script utilizing CLI arguments.

Demo

Parsing Steps:

1. Message Iteration:
    - Utilized the mailbox.mbox library and its iteritems() method to iterate through each email message within the MBOX file.
2. Message Structure Analysis and Content Processing:
    - Inspected the key-value pairs of each message object to understand the structure and identifed relevant data fields.
    - Implemented charset handling to ensure accurate decoding of email headers and bodies.
    - Converted email body content to plain text to facilitate data analysis.
3. Dictionary Construction:
    - Iterated through the identified keys and constructed a dictionary representation for each email message, organizing the extracted data.

4. Attachment Extraction:
    - Iterated through message parts to identify and extract attachments, saving them to designated locations.
5. CSV Output:
    - Employed the Pandas library to organize the extracted data into a structured DataFrame.
    - Wrote the DataFrame to a CSV (Comma Separated Values) file for data storage and further analysis.

Python Script Implementation:

A Python script (mbox2csv.py) has been developed to automate the MBOX data extraction process. It takes the MBOX file path as a command-line argument.

Key Features:

- CLI Argument: The script accepts the MBOX file path and an optional output directory as command-line arguments, providing flexibility in processing different files and specifying output locations.
- Modular Design: The script is structured into functions for better organization and reusability.
- Error Handling: Basic error handling is included to manage potential issues during file processing.
- Output: The script generates a CSV file containing the extracted data and saves it, along with any extracted attachments, to the specified output directory.

How to Use the Script:

- Clone GitHub repo

```
git clone https://github.com/cchristion/mbox2csv
cd mbox2csv
```

- Download requirements

```
pip install -r requirements.txt
```

- Usage

```
usage: mbox2csv.py [-h] [-o OUTPUT_DIR] mbox_file

Mbox to CSV parser.

positional arguments:
mbox_file             Path to Mbox file

options:
-h, --help            show this help message and exit
```

```
  -o OUTPUT_DIR, --output_dir OUTPUT_DIR
                        Path to Output Directory. default
  "mbox_output"
```

- Execute the code

```
  python mbox2csv.py /path/to/mbox_file.mbox -o /path/to/output
```