International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 Impact Factor: 8.176 ISSN: 2582-3930

Palm leaf Digitization: A Blockchain-based approach for finding Hidden Treasures

Dr. R. PREMA[1], Jallepalli Aditya Sai [2], Kamireddy Hemanth [3]

[1]Assisstant professor, Department of Computer Science and Engineering, SCSVMV, Kanchipuram

[2] B.E graduate(IV year), Department of Computer Science and Engineering, SCSVMV, Kanchipuram

[3]B.E graduate(IV year), Department of Computer Science and Engineering, SCSVMV, Kanchipuram

---

Abstract - Many hidden treasures and useful information can be found in palm-leaf manuscripts. They contain information regarding medicine, astronomy, music, culture, etc. Manuscripts are written in different languages and different scripts. Identification of characters and linguistic scripts in palm-leaf manuscripts is a time-consuming and challenging process due to the manuscript's extreme fragility and susceptibility to deterioration from either natural or human-made activity. Characters, script, or language identification in palm leaf manuscripts is helpful for extracting useful information for knowledge retrieval and dissemination, and hence, this has created interest for many researchers in knowledge retrieval from palm leaf manuscripts since the last decade. Hence, it is necessary for the digitization and storing of manuscripts. The main aim of this work is to automatically identify, detect and recognize the metadata (Telugu and Tamil) from the digitized palm leaf manuscripts. Easy OCR is the step that helps in perceiving the text in pictures, tags, and so on. We used a Generative Pre-trained Transformer (GPT) for detecting or predicting the missing data from the text collected in the second stage. We relied upon trained language models based on image and character recognition. Each model is trained using GPT for the detection and prediction of the missing words or characters. For the translation of the extracted text from the existing language to the user's language, we have used Google Translator. Finally, all the metadata is secured in Blockchain for ensuring tamper-proof data. Metadata Blockchain Model (MDBC) ensures that the metadata, translated and transliterated data is not erased and secured in a trusted environment.

Key Words: *Blockchain Technology; Easy OCR; Generative Pre-trained Transformer; Google Translation, Optical Character Recognition.*

Introduction

Objectives

Manuscripts are written in a variety of languages and scripts. Because palm leaf manuscripts are exceedingly delicate and prone to damage caused by natural or man-made activity, identifying the characters and script of language in them is a tedious and difficult operation. It is required for manuscript digitalization and storage. The major goal of this project is to identify, detect, and recognize metadata (Telugu and Tamil) from digitized palm leaf manuscripts automatically. Using the Easy OCR Optical Character Recognition (OCR) device and Python, we conducted various experiments on over 3 lakh different types of digitized palm leaf manuscripts. Easy OCR aids in the recognition of text in images, tags, and other formats. At stage 1, noise reduction and background removal from palm-leaf manuscripts are done to improve the readability of the characters from scanned images. On incomplete projection profiles, text line segmentation is used. The Dropout technique is used to determine the accuracy. For detecting or predicting missing data from the text obtained via, we employed Generative Pre-trained Transformer (GPT-3) Easy OCR. We utilized MATLAB to recognize characters and text. We used images and character recognition-based trained language models. GPT is used to train each model for detecting and predicting missing words

or letters. We used Google Translator, CAT tools (computer-assisted translation), and Machine Translation for translation and transliteration (MT). Finally, to ensure tamper-proof data, all metadata is stored in Blockchain. The Metadata Blockchain Model (MDBC) ensures that metadata, translated, and transliterated material is not lost and are kept safe in a secure environment.

Scope of the Project

The main scope of the project is to extract valuable information from the multilingual palm-leaf manuscripts and automatically identify the missing words. In addition to translating it into the user's language and storing this precious information in a blockchain to safeguard it.

Existing System

In the late 1980s and early 1990s, the central ideas that underpin blockchain innovation emerged. Leslie Lamport founded the Pax OS convention in 1989, and he delivered The Part Time Parliament to ACM Transactions on Computer Systems in 1990; the work was later published and distributed in a 1998 edition. The paper portrays an agreement model for agreeing on this brings about an organization of PCs where the PCs or organisation itself might be inconsistent. In 1991, a marked chain of data was utilised as an electronic record for carefully marking the archives in a manner that could undoubtedly show none of the marked records in the assortment had been changed. These ideas were joined and applied to electronic money in 2008 and as depicted in the paper, Bitcoin: A Peer-to-Peer Electronic Cash System, which was distributed pseudonymously by Satoshi Nakamoto, and later in 2009, with the foundation of the Bitcoin digital currency blockchain network, Nakamoto paper contained the plan that most Modern digital money plans tend to follow this (albeit with variations and changes). Bitcoin It was only the first of numerous blockchain applications. In the initial stages, BCT (Blockchain Technology) was used in Bitcoin and cryptocurrency. But, in the coming days, blockchain will play a key role in many applications.

- Real estate processing platform
- Secure sharing of medical data
- Music royalties tracking
- Real-time IOT operating systems
- Personal identity security

Drawbacks of Existing System

The main drawbacks of the existing work are that it can't automate the missing words in the extracted information. Also, it hasn't been safeguarded anywhere. So, in this project, we have to get over these drawbacks by using GPT-3 to automate the missing words in the extracted information and blockchain technology to safeguard the information.

Literature Survey

As we know, palm-leaf manuscripts are disappearing these days for many reasons, like man-made disasters and natural disasters. The information contained in these manuscripts is very valuable and also useful for present and future generations. So, we have to cherish these manuscripts. The only way to protect these manuscripts is to digitalize and store the information by using blockchain technology. Due to the interesting nature of palm-leaf manuscripts, the digitalization of palm leaves has a few difficulties. To conquer these difficulties, we have a few fundamental processes like "Natural Language Processing System, Google Translator, Metadata Generation, Optical Character Recognition (OCR), and digital image capture".

The strategic proposed a venture for National Digital Manuscripts Library. Without precedent for history, all the noteworthy abstract, aesthetic, and logical works of humanity can be carefully safeguarded and made effectively accessible for our way of life, research, training, and gratefulness and further more for all our people in the future. The aim of Digital Manuscripts Library of India is to foster creativity and easy access. NMM as a team with NIC has built up an easy to use web application for entering meta-information on the web and getting to the Manuscripts subtleties through a devoted web search tool. The Meta information will be transferred in the cloud server alongside the computerized pictures to be interlinked for giving simple

access to advanced Manuscripts for a wide open to be utilized for research and investigation of different subjects. NMM has formed programming considering the assorted variety of contents, material on which they are composed and subjects to which they allude. Such programming would empower proficient administration of information about custodial establishments, illustrative indexes, subject catalogs, safeguarding status of original copies and for getting to the computerized pictures through the web crawler and fusing the equivalent in the advanced library. It is being intended to set up a cutting edge Digital Manuscript Library at New Delhi with the limit with regards to 10 clients at first (to be extended up to 100 clients) to encourage access to this information base. Clients are relied upon to get to information from a Data Center through work area PCs. New information is required to be included to the capacity servers an everyday schedule. It is likewise necessitated that the Data Center holds fast to worldwide principles of information stockpiling, repetition, uptime, and accessibility.

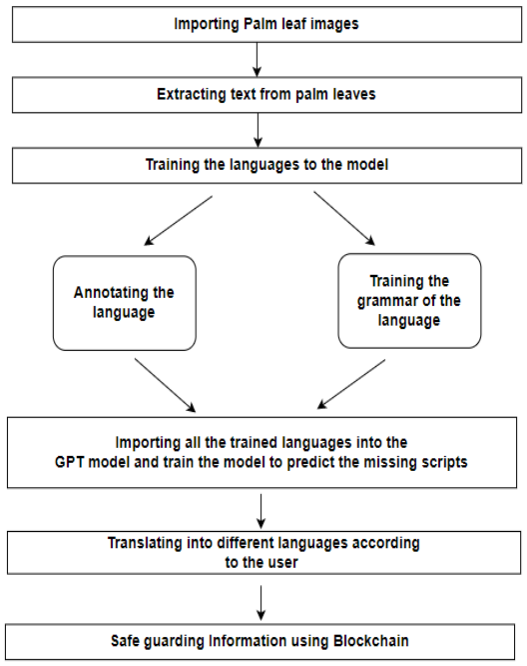
Digitization of manuscripts as means of protecting and documenting textual heritage is becoming popular in recent times. With the advancements in information technology, digitization provided easier way to access the manuscripts for scholars and researchers. In India, National Mission for Manuscripts was started in the year 2004 as a pilot study aiming at digitizing several caches of manuscripts across the country. In 2006, the Mission for setting standards and guidelines for digitization came into force.

The strategic proposed a venture for National Digital Manuscripts Library. Without precedent for history, all the noteworthy abstract, aesthetic, and logical works of humanity can be carefully safeguarded and made effectively accessible for our way of life, research, training, and gratefulness and furthermore for all our people in the future. NMM as a team with NIC has built up an easy to use web application for entering meta-information on the web and getting to the Manuscripts subtleties through a devoted web search tool. The Meta information will be transferred in the cloud server alongside the computerized pictures to be interlinked for giving simple access to advanced Manuscripts for a wide open to be utilized for research and investigation of different subjects. NMM has formed programming considering the assorted variety of contents, material on which they are composed and subjects to which they allude. Such programming would empower proficient administration of information about custodial establishments, illustrative indexes, subject catalogs, safeguarding status of original copies and for getting to the computerized pictures through the web crawler and fusing the equivalent in the advanced library. It is being intended to set up a cutting edge Digital Manuscript Library at New Delhi with the limit with regards to 10 clients at first (to be extended up to 100 clients) to encourage access to this information base. Clients are relied upon to get to information from a Data Center through work area PCs. New information is required to be included to the capacity servers an everyday schedule. It is likewise necessitated that the Data Center holds fast to worldwide principles of information stockpiling, repetition, uptime, and accessibility.

Problem Statement

As of now, the palm-leaf manuscripts have only been digitized. By taking a further step, we are going to extract the data from the digitized palm-leaf manuscripts with the maximum accuracy rate and secure them safely. In the process of extracting the data from digitized pal-leaf manuscripts, we have to overcome some sorts of challenges. As the palm-leaf manuscripts belong to the late centuries, the words or letters in those manuscripts will fade away. So, while we are extracting the data, we have to predict the faded ones accurately. We have to convert the regional language to the various user languages.

Methodology



Flowchart for the proposed method

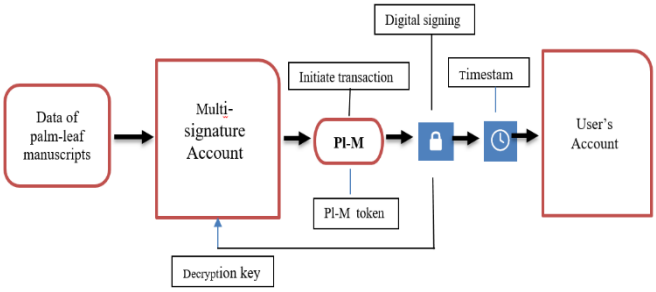
Optical character recognition methods involve a series of pre-processing processes in which the inspected document is cleaned and rendered noise-free. The material is then transformed into binary data to assist in the contour identification of lines and columns a description of the conventional OCR model. Finally, using machine learning techniques like K-nearest neighbors and support vector machines, the characters that make up the lines are extracted, segmented, and recognized. They perform well on straightforward OCR datasets, such as easily recognizable printed material and handwritten MNIST data. The pre-processing stage comes first in the text line segmentation process. The pre-processing changes the color pictures of the palm leaf text to binary images since it is simpler to execute an algorithm to recognize an obstruction using image enhancement techniques.

Here we use Easy OCR, which is an open source to extract the data from the palm-leaf manuscripts. Easy OCR consists of different inbuilt languages, which could be useful in the case of multilingual palm-leaf manuscripts.

Name entity recognition is a data processing task where it can train a language's PoS (noun, verb, adverb, adjective, etc.). For this, we have to collect a lot of text data to annotate and get the output. We have to give the annotated output as an input to the model which we have to train.

The task of sentence infilling is formally defined as follows. Consider a dataset of n paragraph $\{p^{(k)}\}\ N_{k=1}$. Each paragraph $p^{(k)}$ = ($S_1^{(k)}$,$S2^{(k)}$ …….$S_M^{(k)}$) Consists of M_k consecutive sentences. For each k, we are given a positive integer $m_k \leq M_k$ and the context $S_1^{(k)}$,S2(k), …….$S_M^{(k)}$. The task is to generate a sentence $s\hat{}_{mk}^{(k)}$ in the missing position such that it fits the context. For simplicity and without any confusion, we drop the index k from now on (note that M and m may depend on k).

An auto encoder implants sentences into vectors in the idle space. We trust that the implanting is smooth as in semantically comparative sentences are planned to vectors that are near one another. Specifically, interjection between two focuses in the dormant space ought to compare to a



smooth semantic change in the text area. To this end, we utilize the accompanying two stunts.

Sentences are consequently encoded and put away as vectors in the idle space. We trust that the implanting is smooth as in vectors close to one another are doled out to phrases with comparable semantic substance. Specifically, a consistent semantic progress in the text space ought to relate to the addition between two areas in the dormant space. We utilize the two strategies recorded beneath to do this.

A shrewd blockchain will be utilized to defend the data for better use by people in the future. We utilize the NEM public or private blockchain to store the information extricated from palm-leaf compositions, both on and off the chain. The thought is to utilize information encryption to guarantee security and command over our removed information as well as multi-signature contracts for access control in information the board. Executions could likewise utilize the other blockchain frameworks recorded, Ethereum and NEO.

The client will have unlimited authority over who sees the information and admittance to it. The client can indicate who can peruse and compose the substance of the information to the record and characterize access privileges for them. Applications off-chain could use this entrance control the executives. The client would have the option to allot new licenses and remote access through the application, as well as distinguish which substances approached the individual information. Concerning time periods and information types, the consents could be adaptable. The off-chain organization program would ensure that time-subordinate agreements were completed and grant a simple to-utilize blockchain interface.

Process of storing the data in blockchain

The stage's multi-signature contracting usefulness and cryptography structure the underpinning of the design. Numerous gatherings can deal with a record's exercises, claim its resources, or structure new agreements on account of multi-signature contracts. The multi-signature capability of NEM makes an editable agreement that moves the power and freedoms of one record to different records. A multi-signature understanding can be made so a significant number of the imprints are expected to sign an exchange; this is called an "m-of-n" multi-signature understanding, where m can be any number comparable to or not quite so much as n. In the plan of the clinical benefits blockchain, a 2-of-n multi-signature understanding would be applied. The client would control two keys in the arrangement, and absolute information would be remembered for the understanding connected with the client record. This suggests that the client will remain accountable for the record by having the vital number of keys to modify the understanding.

New data would be placed away on the blockchain in a going with way: the client, with something like one key connected with the client's multi-signature account, would begin a trade, sending a data mosaic containing the data as a message. The message would be encoded, and the translating key would be transported off the multi-signature account in an alternate trade. Exactly when the data mosaic trade is checked, either genuinely or thus, the trade is transported off a record associated with that specific data type. This ensures that the principal components with permission to the multi-signature record can scrutinize the client's data using the disentangling key.

Result

Palm-leaf manuscripts contain a wealth of information and undiscovered treasures. They contain knowledge on astronomy, music, culture, and other subjects. Different languages and scripts are used to write manuscripts. Due to the manuscript's extreme fragility and susceptibility to degradation from either natural or human-made activity, identifying characters and linguistic scripts in palm-leaf manuscripts is a time-consuming and difficult process. Identification of characters, scripts, or languages in palm leaf manuscripts is helpful for extracting relevant information for knowledge retrieval and dissemination; as a result, this has sparked interest in knowledge retrieval from palm leaf manuscripts among many researchers over the past ten years.

Conclusion

On a major new assortment of reports, this examination gave an extensive trial of three general OCR processors. It recommends that the server-based motors Archive simulated intelligence and Simple OCR, particularly on loud records, give fundamentally higher out-of-the-container precision than the independent Simple OCR library. It additionally shows that "incorporated" commotions like haze and salt and

pepper cause more error than "superimposed" clamors like watermarks, scrawls, and even ink stains. Moreover, it shows that the "OCR language hole" actually exists, regardless of the way that Report artificial intelligence seems to somewhat have connected it. For the sociologies and humanities, the main message is that high-exactness OCR is presently more open than any other time. Scientists who were recently put off by the chance of significant report pre-handling or corpus-explicit model preparation currently approach easy to use advances that produce astounding outcomes from the outset. More researchers will surely utilize OCR innovation, and more verifiable reports will be digitized subsequently. The respectability of records saved in a data set can be guaranteed utilizing blockchain innovation. It tends to be achieved utilizing the blockchain's all-around framed exchanges, validation, and reviewing highlights. Diminishing the quantity of likely dangers to information integrity is attainable. When one of the three essential security highlights has been gotten, blockchain might be utilized to guarantee the excess two characteristics of information. Palm-leaf compositions contain an abundance of data and unseen fortunes. They contain information on cosmology, music, culture, and different subjects. Various dialects and contents are utilized to compose original copies. Because of the composition's outrageous delicacy and vulnerability to debasement from one or the other normal or human-made action, recognizing characters and phonetic contents in palm-leaf compositions is a tedious and troublesome cycle. ID of characters, contents, or dialects in palm leaf compositions is useful for separating significant data for information recovery and scattering; subsequently, this has ignited interest in information recovery from palm leaf compositions among numerous specialists throughout the course of recent years.

References

[1] Dylan Yaga, Peter Mell, Nik Roby, Karen Scarfone; “Blockchain Technology Overview”

[2] Jannice Kall; “Blockchain Control”; Elsevier; 30 May 2018

[3] Chung-Shan Yang “Maritime shipping digitalization: Blockchain-based technology applications, future Improvements, and intention to use” PLOS One; 2019

[4] Jesse Yli-Huumo, Deokyoon Ko, Sujin Choi, Sooyong Park, Kari Smolander; “Where is Current Research on Blockchain Technology?”;Springer; October 3 2016;

[5] Piyadasa Ranasinghe, W.M. Tharanga Dilruk Ranasinghe “Preserving Sri Lanka’s Traditional Manuscript Culture: Role of the Palm Leaf Digitization Project of the Faculty of Social Sciences, University of Kelaniya”; International Federation of Library Associations (IFLA); 2019

[6] R. Vasanth Kumar Mehta, Nagendra Panini Challa “Facilitating enhanced user access through Palm-leaf Manuscript Digitization – Challenges and solutions”; IEEE; 2017

[7] Narahari sastry panyam, Vijaya lakshmi T.R, Ramakrishnan krishnan, koteswara rao N.V "Modeling of palm leaf character recognition system using transform based techniques"; ELSEVIER; 2016.

[8] Dr. M. Mohamed sathik, R. Spurgen ratheash "Text line segmentation in tamil language palm leaf manuscripts-a novel approach"; Journal of tianjin university science and technology,2021.

[9] T. Jerry alexander, S. Suresh kumar "A novel binarization technique based on whale optimization algorithm for better restoration of palm leaf manuscript"; Journal of ambient intelligence and humanized computing, 2020

[10] T.S. Suganya, S. Muruguvalli "A hybrid group search optimization: firefly algorithm-based big data framework for ancient script recognition"; springer, 2019.

[11] Rajalakshmi Krishnamurthi, Tuhina Shree "A brief analysis of blockchain algorithms and its challenges"; Research gate, 2021.

[12] Ashok Kumar Pundir, Jadhav Devpriya, Mrinmoy Chakraborty, L Ganpathy "Technology integration for improved performance: A case study in digitization of

supply chain with integration of internet of things and blockchain technology"; Springer, 2017

[13] Igor Zikratov, Alexander Kuzmin, Vladislav Akimenko, Viktor Niculichev, Lucas Yalansky "Ensuring Data Integrity Using Blockchain Technology" Proceeding Of The 20th Conference Of Fruct Association.

[14] Michael Piotrowski "Natural Language Processing for Historical Texts"; Morgan & Claypool Publishers series, 2012

[15] Piyadasa Ranasinghe, W.M. Tharanga Dilruk Ranasinghe "Preserving Sri Lanka's Traditional Manuscript Culture: Role of the Palm Leaf Digitization Project of the Faculty of Social Sciences, University of Kelaniya"; Rare Books and Manuscripts Section & Asia and Oceania Section, 2013.

[16] Lewis Rinaudo Cohen, Lee Samuelson and Hali Katz" How Securitization Can Benefit from Blockchain Technology"; Institutional Investor Journals, 2018.

[17] Thomas Hegghammer "OCR with Tesseract, Amazon Textract, and Google Document AI: a benchmarking experiment"; Springer, 2021.

[18] Simon Lebech Cichosz, PhD1, Mads Nibe Stausholm, MSc1, Thomas Kronborg, MSc1, Peter Vestergaard, PhD2,3,4, and Ole Hejlesen, PhD1; "How to Use Blockchain for Diabetes Health Care Data and Access Management: An Operational Concept"; Original Article; 2019.

[19] Yichen Huang, Yizhe Zhang, Oussama Elachqar, Yu Cheng; INSET: Sentence Infilling with INter-SEntential Transformer; "INSET: Sentence Infilling with INter-SEntential Transformer"; arXiv; 2020