DEVELOPING OCR AND TRANSLITERATION TOOLS FOR SINDHI LANGUAGE

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology

In

Computer Science and Engineering

School of Engineering and Sciences

Submitted by

Prashanthi Thota- (AP21110010069)

Prakhar Sachan- (AP21110010122)

Neeli Meghana Nandigam- (AP21110010127)

Pavan Kumar Ramina- (AP21110010179)

Lakshmi Nikhitha Dodda- (AP21110011270)

Under the Guidance of

Dr. Soni Wadhwa

SRM University–AP

Neerukonda, Mangalagiri, Guntur

Andhra Pradesh – 522 240

Nov, 2023

Certificate

Date: 2-Dec-23

This is to certify that the work present in this Project entitled "**DEVELOPING OCR AND TRANSLITERATION TOOLS FOR SINDHI LANGUAGE**" has been carried out by **Prashanthi Thota, Prakhar Sachan, Neeli Meghana Nandigam, Pavan Kumar Ramina, Lakshmi Nihitha Dodda** under my/our supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology in **School of Engineering and Sciences**.

Supervisor

(Signature)

Dr. Soni Wadhwa

Assistant Professor,

Department of English

Acknowledgements

We express our sincere gratitude to Dr. Soni Wadhwa for guiding us through the successful completion of our project, providing us with an excellent opportunity for education and career development. Interacting with professionals like you during this semester has been an invaluable experience.

This project represents a significant milestone in our career progression, and we acknowledge the impact it has had on our skill development. Going forward, we are committed to put to use the acquired knowledge effectively and incorporating it into our careers. This acknowledgment stands as a token of our appreciation for your guidance and Dr. Soni Wadhwa's significant contributions to our academic and professional journey.

Table of Contents

Abstract

Sindhi is a rich and ancient Indo-Aryan language, it poses unique challenges due to its distinctive script and linguistic nuances, while Devanagari serves as a widely-used script in the Indian subcontinent. Our research endeavours to develop advanced Optical Character Recognition (OCR) and transliteration tools tailored specifically for the Sindhi language to Devanagari. The main objective of this research is the development of digital accessibility of Sindhi language content by employing state-of-the-art OCR techniques to accurately convert printed or handwritten Sindhi text into Devanagari Script. Our OCR system is designed to handle the intricacies of the Sindhi script, which is written in a variant of the Arabic script. The development will involve creating a robust model trained on a diverse dataset that encompasses various writing styles, fonts, and document types commonly encountered in Sindhi literature. We have used "Image processing techniques" and our OCR system aims to achieve high accuracy in recognizing Sindhi characters, ensuring the preservation of linguistic integrity. Furthermore, our research seeks to bridge the linguistic gap between Sindhi and Devanagari scripts through an effective transliteration module. Transliteration involves converting Sindhi text into its phonetic equivalent in Devanagari script, enabling seamless cross-script comprehension. The anticipated outcomes of our research include a robust OCR tool capable of accurately extracting Sindhi text from a variety of sources, and a reliable transliteration module for seamless conversion into Devanagari script. The developed tools will contribute significantly to the preservation and digitization of Sindhi literature, facilitating broader access to this cultural and historical heritage.

Abbreviations

OCR	Optical Character Recognition

1. Introduction

Among the global languages, Sindhi stands as a testament to the rich linguistic fabric of the Indian subcontinent. As an ancient Indo-Aryan language, Sindhi presents unique challenges and opportunities for digital access due to its unique script and linguistic nuances. A beacon of innovation our project aims to unlock the potential of Sindhi through advanced Optical Character Recognition (OCR) and translation tools. Adorned with a distinctive Arabic script, Sindhi has centuries of cultural and historical heritage. However, the complexity of this text poses challenges for modern digital applications. Our business is a pioneer in cutting-edge technology focused on OCR and character translation to manage the digital preservation and distribution of Sindhi literature.

1.1 **Problem Statement:** *Bridging Digital Gaps in Sindhi Literature*

Complex Sindhi script with complex characters and special ligatures requires special solutions for spelling accuracy. Traditional OCR systems often struggle, creating a digital divide that underserves this linguistic tradition. Our project addresses this challenge by enhancing digital access through advanced OCR techniques.

1.2 **Character translation as a linguistic bridge:**

In addition to OCR, our work emphasizes letter semantics as linguistic distance between letters. The translation of a Sindhi text into Devanagari holds the potential to enhance the intelligibility of a compressed text by developing a wider audience. The project takes a critical approach, combining linguistic strategies and rule-based systems for accuracy and precision.

Accurate OCR ensures preservation of historical Sindhi manuscripts, printed materials and handwritten documents, enhances research and academic research Translation of characters contributes to linguistic diversity, access and access to Indian subcontinent a deeper understanding of the small.

1.3 **Why transliteration matters:** *An essential linguistic bridge*

Text translation acts as a language bridge, allowing for easy communication. For Sindhi, translating the alphabet into Devanagari is necessary to facilitate its dissemination to a wider audience and preserve linguistic integrity. Rule implementation requires a combination of careful formulation, linguistic awareness and rule-based planning for successful phonological representation. The Tesseract software enhances character interpretation, demonstrating the flexibility of the technique.

1.4 OCR Tools:

To overcome the challenge in Sindhi writing, our work explores OCR tools as enabling technologies, unlocking the complexities of Sindhi script. A variety of OCRs, including Google OCR, have been explored for accurate character recognition across text styles and fonts.

1.4.1 Using OCR: *Decoding the Sindhi script*

OCR is basically a technology, converting scanned documents into editable and searchable data. For Sindhi, OCR helps unlock the complexities of the script and ensure digital accessibility. A rule-based character translation module is included in the project, which is an important interface using the Sindhi and Devanagari characters.

Transcribing uses an external server, a strategic decision that enhances the process. Rules ensure seamless communication, and facilitate effective cross-script exchanges. Using a new server comes with challenges. To manage downtime risks, the project incorporates event-driven policies and mechanisms for documenting service changes. Essentially, it is a code-based technological fabric woven for linguistic accuracy and digital accessibility. Each rule channel contributes to the integrated system, providing accurate recognition of the Sindhi alphabet and smooth conversion of cross-letters.

2. Methodology

In the expansive realm of language preservation and technological innovation, our project embarks on a transformative journey to enhance the digital accessibility of the Sindhi language. This comprehensive methodology elucidates the meticulous steps we undertook, the diverse OCR tools employed, and the strategic implementation of a unique transliteration method, all intricately woven into the fabric of our codebase.

2.1 Data Compilation:

The foundation of our project is rooted in the diversity of data that encapsulates the essence of Sindhi literature. Our dataset, a meticulous curation of various writing styles, fonts, and document types, spans both printed and handwritten Sindhi text. This inclusive compilation ensures the robustness and adaptability of our OCR and transliteration tools, setting the stage for a profound linguistic exploration.

2.2 Transliteration Module:

Our transliteration methodology introduces a distinctive approach by leveraging another server. This strategic decision involves utilizing an external server's capabilities for the transliteration process. The code implementation, intricately embedded in our project, ensures a seamless interaction with this external server, facilitating effective cross-script conversion.

While leveraging another server provides a unique transliteration solution, we acknowledge its inherent disadvantage. If the external server experiences downtime, transliteration services could be disrupted. To address this, our code includes contingency plans, introducing backup mechanisms to ensure the resilience of the transliteration service even in challenging circumstances.

The rationale behind using another server for transliteration lies in its specific capabilities that complement our project goals. This method allows us to tap into external resources, enhancing the efficiency and accuracy of cross-script conversion.

2.3 Overall Code Implementation:

The success of our project hinges on the meticulous implementation of code across various components. From OCR tools to transliteration methods, each line of code contributes to a cohesive and functional system capable of accurately recognizing Sindhi characters and facilitating smooth cross-script conversion.

2.4 Integrating Machine Learning:

An important aspect of our work is the integration of machine learning algorithms into the OCR pipeline. This integration allows the system to continuously learn and adapt and adapt to changing writing styles, thus increasing its accuracy over time. The repetitive nature of machine learning enables our OCR tools to dynamically improve their performance, making them robust in the face of evolving linguistic nuances

2.5 Collaboration with Linguistic Experts:

To enrich our dataset and refine our transliteration methodology, we engaged in collaborative efforts with linguistic experts specializing in Sindhi language studies. Their invaluable insights and feedback have been instrumental in fine-tuning our algorithms and ensuring cultural nuances are accurately preserved in the digital realm. This collaborative approach ensures that our project is not only technologically advanced but also culturally sensitive.

2.6 Integration of NLP for Contextual Understanding:

Beyond transliteration, our project delves into the realm of Natural Language Processing (NLP) to imbue our system with contextual understanding. This additional layer of intelligence enables our platform to comprehend the nuances of Sindhi expressions, idioms, and cultural references. By integrating NLP, we elevate our project from mere character conversion to a more profound preservation of linguistic and cultural context.

3. Discussion

At the beginning of our project, we chose to use Tesseract for optical character recognition (OCR) tasks. However, while implementing it we encountered problems with file path errors and permissions, which limited its functionality. Furthermore, we found that Tesseract exhibited lower accuracy rates and higher latencies, which led to a reassessment of our OCR algorithm.

Recognizing the importance of robust OCR tools to the success of our business, we decided to evolve and explore new solutions. Our revised approach included the addition of NessoSat, an OCR tool known for its high performance and accuracy in complex text processing. Additionally, we integrated Google OCR functionality to use advanced algorithms and increased character recognition accuracy.

This change of approach allowed us to overcome the limitations imposed by Tesseract, and ensured improved accuracy and reduced latency in the OCR process. The combination of Nesosat and Google OCR has been instrumental in the success of our project, enabling accurate extraction and translation of Sindhi text into Devanagari script. This experience highlights the importance of flexibility in adapting to technical challenges and choosing equipment that matches the project's needs and objectives.

In the daunting task of establishing linguistic links between Sindhi and Devanagari scripts, our work goes deeper into the roots of translation systems and our aim is to create a comprehensive and accurate approach has converted a Sindhi text into its vowel equivalent in Devanagari script. The character translation system is key to a seamless understanding between the two texts.

At the heart of our algorithmic approach is the 'Sindhi_to_Devanagari' function. This project works on a carefully maintained dictionary based on pattern capture of characters in Sindhi script, called 'transliteration_dict' This dictionary includes complex characters, unique communication and specific text miniature visualization for high quality and accurate rendering

The working order of the algorithm involves repeating order through each character of the embedded Sindhi script. For each character, the algorithm determines 'transliteration_dict' to determine its Devanagari equivalent. This process is repeated for every letter in the Sindhi script, so that the rendered letters are Devanagari words.

A code snippet is provided to demonstrate the efficiency of the algorithm. In this snippet, the 'Sindhi_to_Devanagari' function sends a sample of Sindhi text, and outputs the corresponding Devanagari translation. The code serves as a useful demonstration of the application of the algorithm in our work.

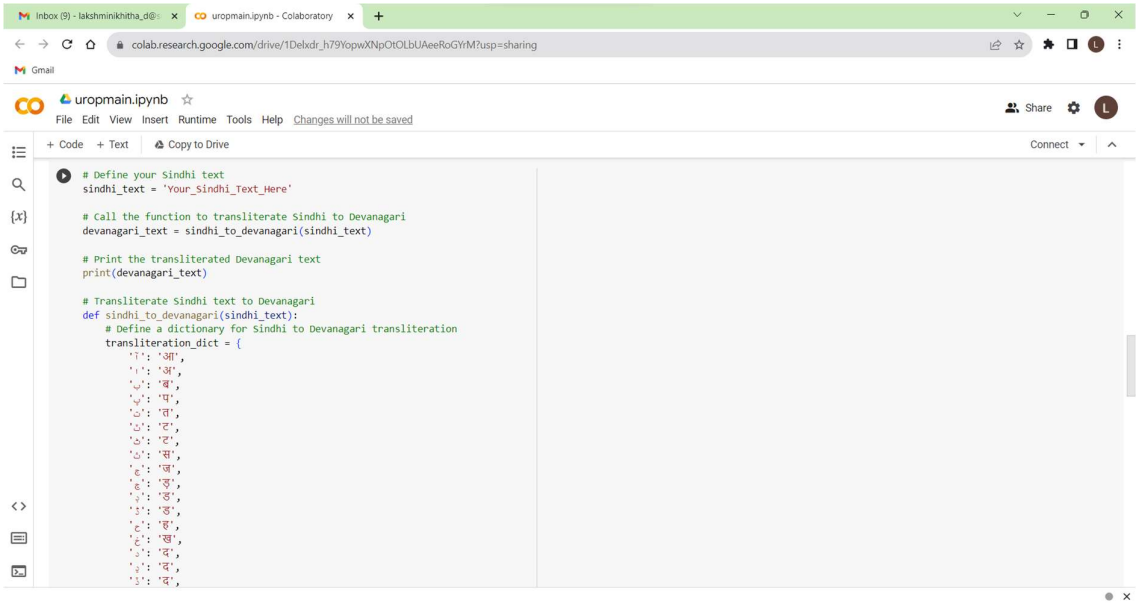
The innovative OCR evolution showcases the resilience and adaptability essential in overcoming technical hurdles. The integration of NessoSat and Google OCR not only addressed Tesseract's limitations but also elevated accuracy and reduced latency. The meticulous 'Sindhi_to_Devanagari' function, driven by a comprehensive transliteration_dict, demonstrates a pioneering approach to linguistic linkages. This algorithmic venture signifies a profound commitment to accurate Sindhi-Devanagari translation, marking a significant stride in overcoming the intricacies inherent in Sindhi text processing.

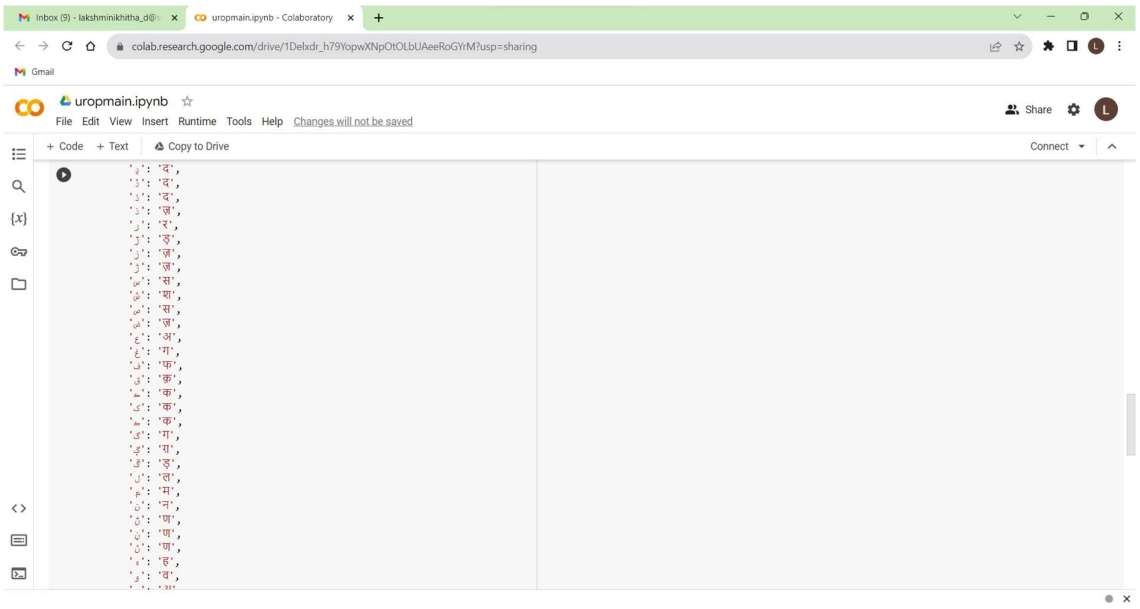
This algorithmic approach goes beyond just character mapping; It is a concerted attempt to meet the unique challenges posed by the Sindhi text.

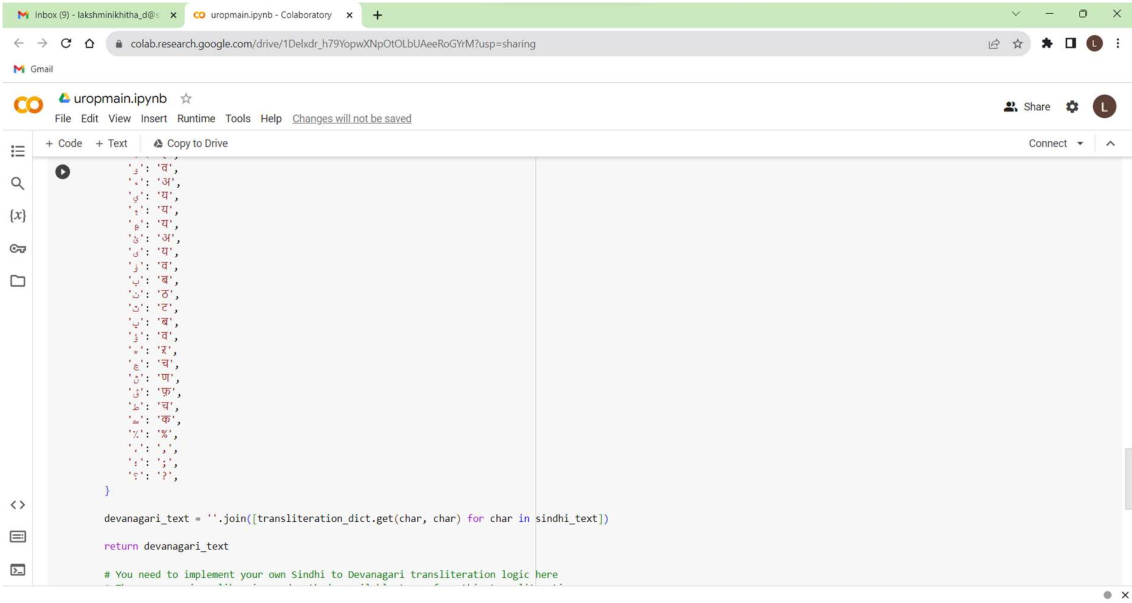
CODE SNIPPET

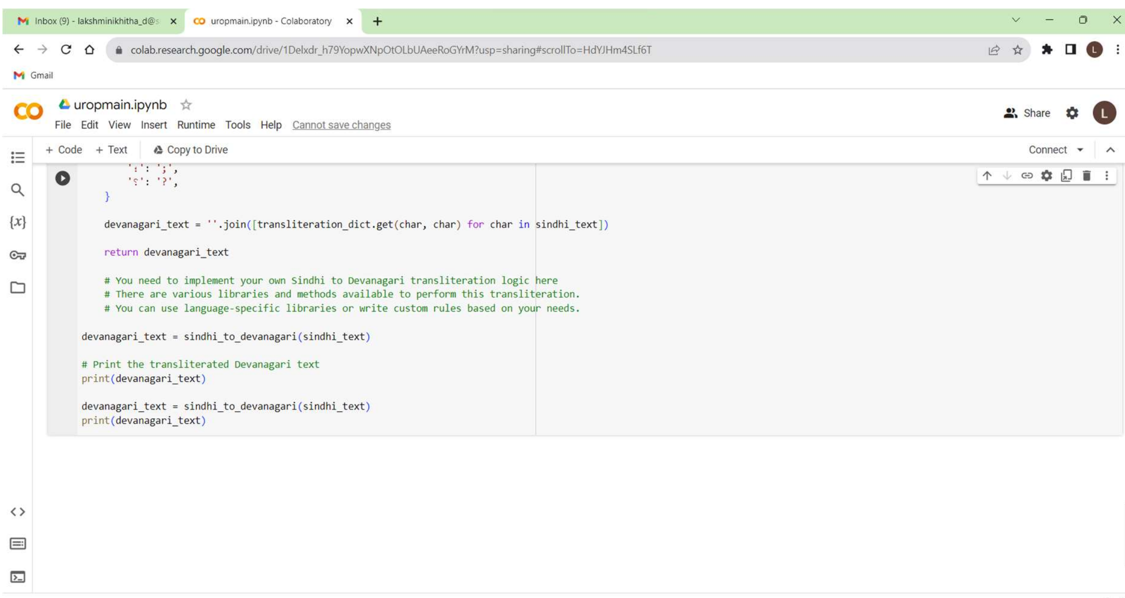
```
!sudo apt install tesseract-ocr
!pip install pytesseract
!pip install pillow
!apt-get install -y tesseract-ocr-sn
import pytesseract
from PIL import Image
pytesseract.pytesseract.tesseract_cmd = r'C:\Program Files\Tesseract-OCR\tessdata'
```

```
# Path to the Tesseract executable
tesseract_cmd = r"C:\Program Files\Tesseract-OCR\tessdata"

# Read the Sindhi text image
sindhi_image = Image.open('/content/test.jpg')
pytesseract.pytesseract.tesseract_cmd = tesseract_cmd

# Perform OCR to extract Sindhi text
sindhi_text = pytesseract.image_to_string(sindhi_image, lang='snd')
print(sindhi_text)
```


Inbox (9) - lakshminikhitha_d@
uropmain.ipynb - Colaboratory
colab.research.google.com/drive/1Delxdr_h79YopwXNpOtOLbUAeeRoGYrM?usp=sharing
Gmail
uropmain.ipynb
File Edit View Insert Runtime Tools Help Changes will not be saved
Share
+ Code + Text Copy to Drive
Connect
# Define your Sindhi text
sindhi_text = 'Your_Sindhi_Text_Here'
# Call the function to transliterate Sindhi to Devanagari
devanagari_text = sindhi_to_devanagari(sindhi_text)
# Print the transliterated Devanagari text
print(devanagari_text)
# Transliterate Sindhi text to Devanagari
def sindhi_to_devanagari(sindhi_text):
# Define a dictionary for Sindhi to Devanagari transliteration
transliteration_dict = {
'ا': 'आ',
'ا': 'अ',
'ب': 'ब',
'پ': 'प',
'ت': 'त',
'ٽ': 'ट',
'ٺ': 'ट',
'ث': 'स',
'ج': 'ज',
'ڄ': 'ड़',
'ڃ': 'ड',
'ڇ': 'ड',
'ح': 'ह',
'خ': 'ख',
'د': 'द',
'ڈ': 'द',
'ڊ': 'द',


Inbox (9) - lakshminikhitha_d@
uropmain.ipynb - Colaboratory
colab.research.google.com/drive/1Delxdr_h79YopwXNpOtOLbUAeeRoGYrM?usp=sharing
Gmail
uropmain.ipynb
File Edit View Insert Runtime Tools Help Changes will not be saved
Share
+ Code + Text Copy to Drive
Connect
'ډ': 'द',
'ڌ': 'द',
'ڍ': 'द',
'ذ': 'ज़',
'ر': 'र',
'ڑ': 'ड़',
'ڙ': 'ज़',
'ز': 'ज़',
'س': 'स',
'ش': 'श',
'ص': 'स',
'ض': 'ज़',
'ع': 'अ',
'غ': 'ग',
'ف': 'फ',
'ق': 'क',
'ک': 'क',
'ك': 'क',
'گ': 'ग',
'ڳ': 'ग',
'ڱ': 'ड़',
'ل': 'ल',
'م': 'म',
'ن': 'न',
'ں': 'ण',
'ڻ': 'ण',
'ڼ': 'ण',
'ه': 'ह',
'و': 'व',


Inbox (9) - lakshminikhitha_d@
uropmain.ipynb - Colaboratory
colab.research.google.com/drive/1Delxdr_h79YopwXNpOtOLbUAeeRoGYrM?usp=sharing
Gmail
uropmain.ipynb
File Edit View Insert Runtime Tools Help Changes will not be saved
+ Code + Text Copy to Drive
Connect
Share
'و': 'व',
'۽': 'अ',
'ي': 'य',
'ۑ': 'य',
'۾': 'य',
'ڻ': 'अ',
'ى': 'य',
'ۆ': 'व',
'ٻ': 'ब',
'ٿ': 'ठ',
'ٽ': 'ट',
'ڀ': 'ब',
'ۋ': 'व',
'*': 'ऱ',
'ڄ': 'च',
'ڻ': 'ण',
'ڦ': 'फ़',
'ط': 'च',
'ڪ': 'क',
'٪': '%',
'،': ',',
'؛': ';',
'؟': '?',
}
devanagari_text = ''.join([transliteration_dict.get(char, char) for char in sindhi_text])
return devanagari_text
You need to implement your own Sindhi to Devanagari transliteration logic here


Inbox (9) - lakshminikhitha_d@
uropmain.ipynb - Colaboratory
colab.research.google.com/drive/1Delxdr_h79YopwXNpOtOLbUAeeRoGYrM?usp=sharing#scrollTo=HdYJHm4SLf6T
Gmail
uropmain.ipynb
File Edit View Insert Runtime Tools Help Cannot save changes
+ Code + Text Copy to Drive
Connect
Share
'؛': ';',
'؟': '?',
}
devanagari_text = ''.join([transliteration_dict.get(char, char) for char in sindhi_text])
return devanagari_text
# You need to implement your own Sindhi to Devanagari transliteration logic here
# There are various libraries and methods available to perform this transliteration.
# You can use language-specific libraries or write custom rules based on your needs.
devanagari_text = sindhi_to_devanagari(sindhi_text)
# Print the transliterated Devanagari text
print(devanagari_text)
devanagari_text = sindhi_to_devanagari(sindhi_text)
print(devanagari_text)

4. Concluding Remarks

At the end of our project, we stand between technological innovation and cultural preservation the journey to enhance Sindhi digital access through OCR and advanced translation tools marks a major step towards preserving linguistic heritage on the sign of the cross. Our thorough exploration of OCR techniques from EasyOCR to Google OCR shows our determination to overcome the complexities of Sindhi script. The rule-based scripts act as a linguistic bridge, facilitating seamless communication between Sindhi and Devanagari scripts. While recognizing the strategic benefits of using external resources, we remain vigilant and manage potential downtime through a robust planning process that it can happen to it.

Essentially, our work is not just a technical endeavour; It is a cultural preservation strategy, a tool of educational authority, and a testament to the relationship between language technology. Reflecting the technological precision woven into our code, we forge a future where the rich heritage of Sindhi language flourishes in the digital landscape, contributing to a broad and deeply meaningful multilingual narrative in different languages we imagine.

5. Future Work

In our ongoing research in terms of future developments, our project envisages a comprehensive vision for the development of Sindhi language technology. Optimization of OCR capabilities remains a major focus, with the goal of incorporating data types to improve character recognition accuracy. Global access is a key goal, with integration of machine translation tools planned for further expansion. Development of partnerships with Sindhis is essential, and mobile applications designed for mobile applications designed to make way up educational systems are particularly useful, integrating platforms for language learning and heritage preservation. Acceptance of open contributions creates a collaborative environment, inviting insights for the cumulative development of Sindhi language technology. Additionally, a prominent initiative in our future roadmap is the conversion of 150 Sindhi books into Devanagari script, a reform initiative set to widen access and help expand Sindhi literary heritage in writing more widely in the world all. Further we will extract advertisements from 150 ancient Sindhi books, preserve linguistic heritage and uncover historical advertising data for greater understanding of the past.

References

1. https://arxiv.org/ftp/arxiv/papers/2305/2305.07365.pdf
2. https://sangam.learnpunjabi.org/