# OCR & Entity Recognition for ID Cards

## 1. Project Title

**Project Name**: OCR & Entity Recognition for ID Cards

## 2. Project Overview

**Description**: This project aims to develop an Optical Character Recognition (OCR) and Named Entity Recognition (NER) system for extracting key details from various types of ID cards (such as PAN, Aadhaar, passport, and driving license). The system uses PaddleOCR for text extraction and spaCy for entity recognition. The output includes categorized information, allowing for easy data extraction and verification.

**Problem Statement**: Manual data entry from ID cards is time-consuming and prone to human error. This project addresses the need for an automated solution that can accurately read ID cards and extract essential information.

**Objectives**:  
- Accurately detect and classify text from multiple types of ID cards.  
- Extract specific information like names, dates, and identification numbers.  
- Implement a user-friendly interface using Streamlit for easy data upload and retrieval.

## 3. Team Member

Role:  
- OCR integration, entity extraction using NER Techniques using spacy and Streamlit setup.

## 4. Learning Objectives

**Skills Developed**:  
- OCR technology using PaddleOCR.  
- NLP and NER techniques using spaCy.  
- Streamlit for creating interactive interfaces.

**Learning Outcomes**:  
The team gained experience with integrating OCR and NLP for practical applications, learned to manage and preprocess ID card images, and developed skills in creating accessible data extraction interfaces.

## 5. Project Phases

**Phase 1**: Research and Planning  
**Activities**: Research on OCR methods and ID card text patterns; planning the integration of OCR with NLP.  
Insights: PaddleOCR and regex patterns provide effective text extraction for standard ID formats.

**Phase 2**: Development and Implementation  
**Technical Approach**:  
- Used PaddleOCR for extracting text from images.  
- Developed regex-based pattern matching to classify ID card types.  
- Created entity extraction functions for specific card types using spaCy.  
**Tools and Technologies**:  
- PaddleOCR for OCR.  
- spaCy for entity extraction.  
- Streamlit for the UI interface.

**Phase 3**: Testing and Validation  
**Testing Strategies**:  
- Tested with diverse ID card samples to ensure accuracy.  
- Validated entity extraction against known correct data.  
**Results**:  
The model accurately extracted and categorized data for recognized card types, achieving high reliability across all tested samples.

## 6. Challenges Faced

**Technical Challenges**:  
- Difficulty in handling low-quality or skewed images, addressed by preprocessing techniques.  
- Challenges in accurately classifying text patterns for multiple ID types.

**Learning Challenges**:  
- Understanding regex patterns for specific ID numbers.  
- Managing OCR inaccuracies for complex document layouts.

## 7. Results and Outcomes

**Final Results**:  
- A functional Streamlit web application capable of extracting ID information.  
- Downloadable text files for extracted information.

**Achievements**:  
Successfully created an efficient and user-friendly ID card processing system.

## 8. Feedback and Reflections

**Team Feedback**:  
Team members appreciated the integration between OCR and NLP and highlighted the ease of using Streamlit for the interface.

**Lessons Learned**:  
- Importance of effective text preprocessing in OCR.  
- Streamlit’s functionality in creating accessible user interfaces for machine learning applications.

## 9. Future Work

**Next Steps**:  
- Extend support for additional ID types.  
- Improve the OCR model to handle more complex document layouts.

**Areas for Further Exploration**:  
Further research into enhancing OCR for multilingual support and improving NER for region-specific entities.

## 10. Appendices

**References**:  
- PaddleOCR documentation, spacy documentation, Streamlit Documentation.