**PROJECT TITLE:**

* Smart OCR for Document Digitization

**INTRODUCTION:**

**Overview:**

Stands for "Optical Character Recognition." OCR is a technology that recognizes text within a digital image. It is commonly used to recognize text in scanned documents, but it serves many other purposes as well.

OCR software processes a digital image by locating and recognizing characters, such as letters, numbers, and symbols. Some OCR software will simply export the text, while other programs can convert the characters to editable text directly in the image. Advanced OCR software can export the size and formatting of the text as well as the layout of the text found on a page.

**Purpose:**

With the advent of OCR techniques, much time has been saved by automatically extracting the text out of a digital image of any invoice or a document.

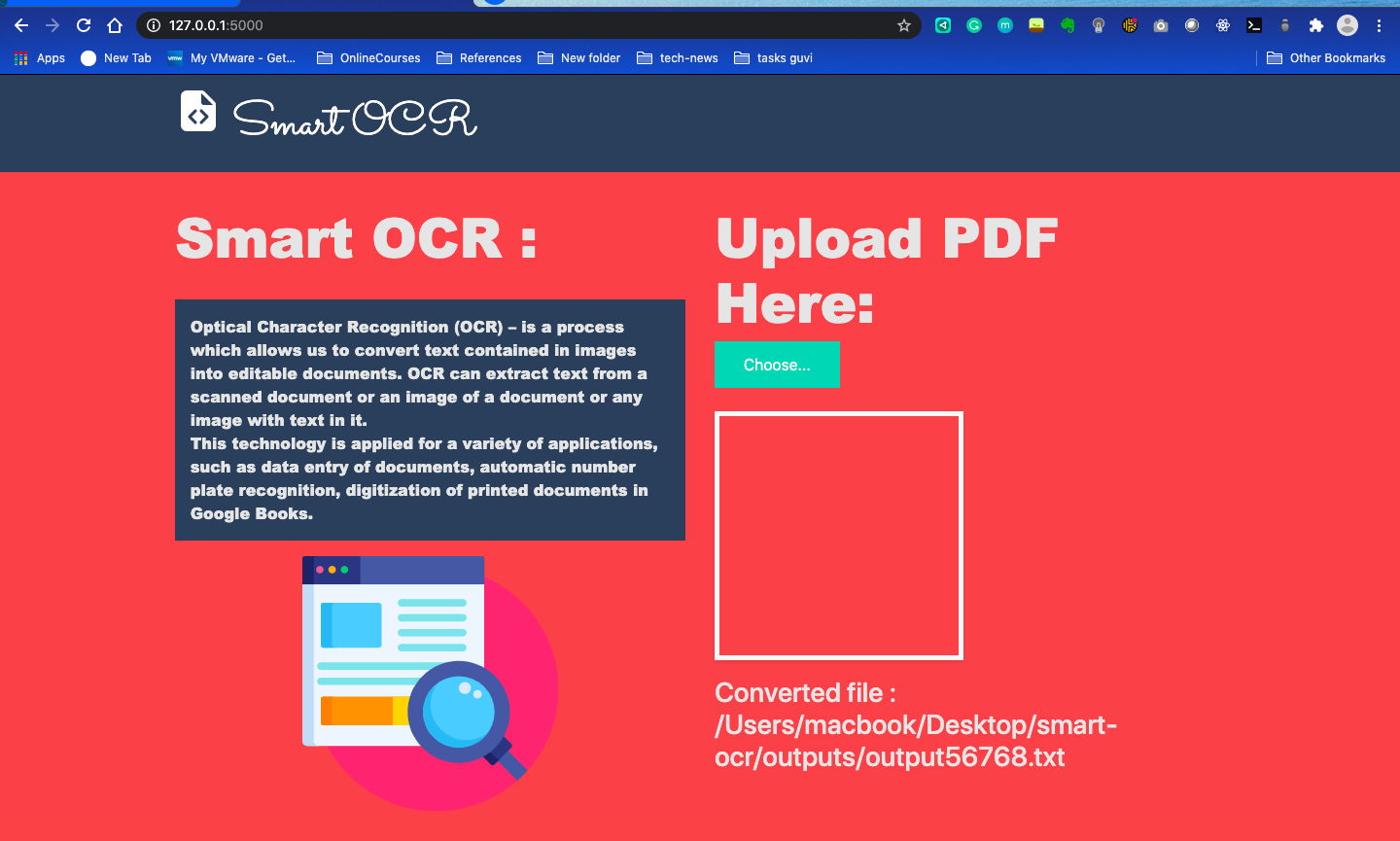
This project has been built to create an application from which the user can upload a pdf document, the document is analyzed by the Optical character recognition package to extract text from it. The extracted text is again saved in a text document in the local drive.

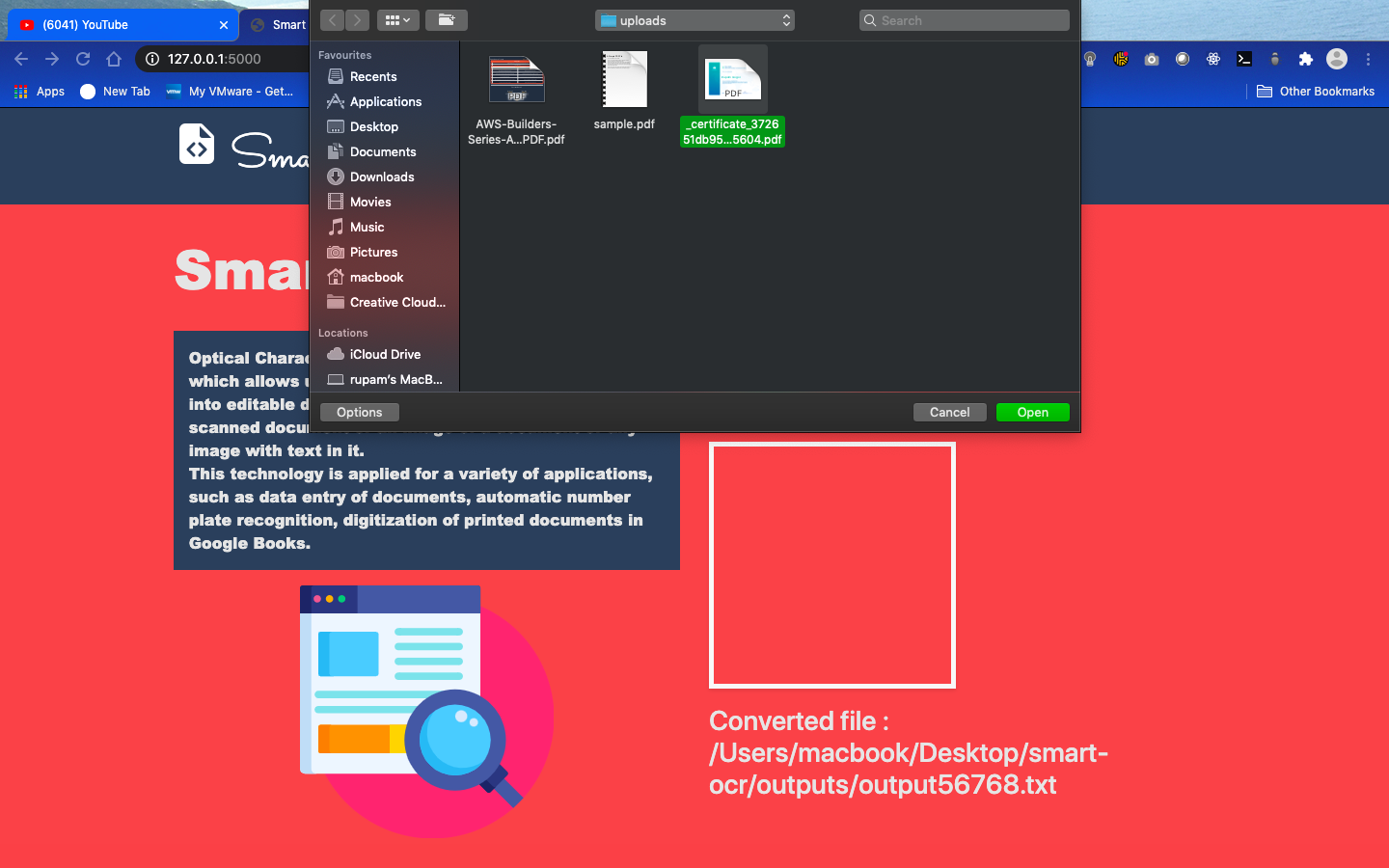
**Project Flow:**

1. Upload a pdf document
2. Convert PDF document to image
3. Extract the text from the image
4. Store the extracted text in the text document

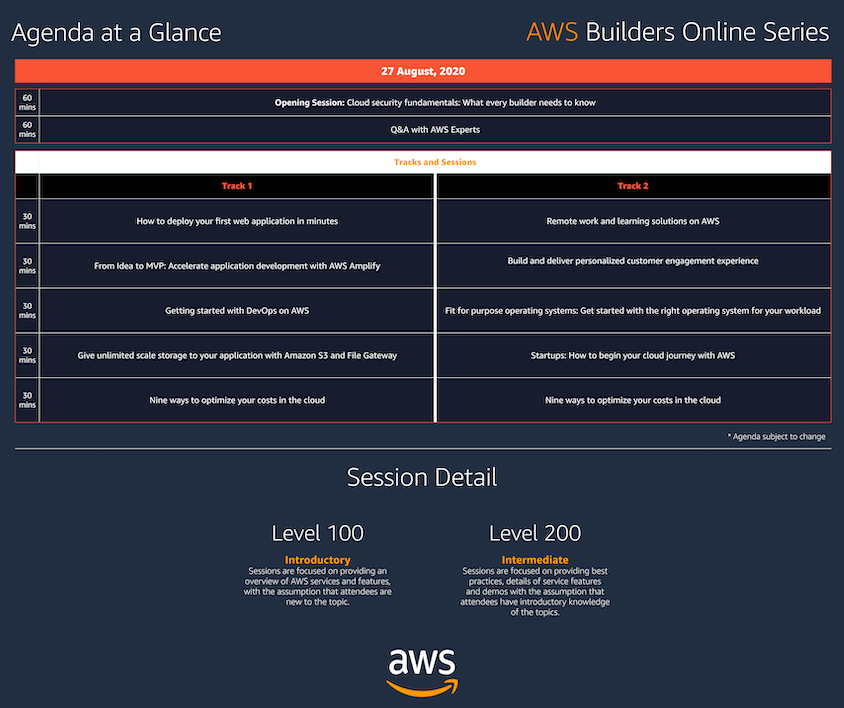
**RESULTS:**

1. **Screenshot of flask app:**

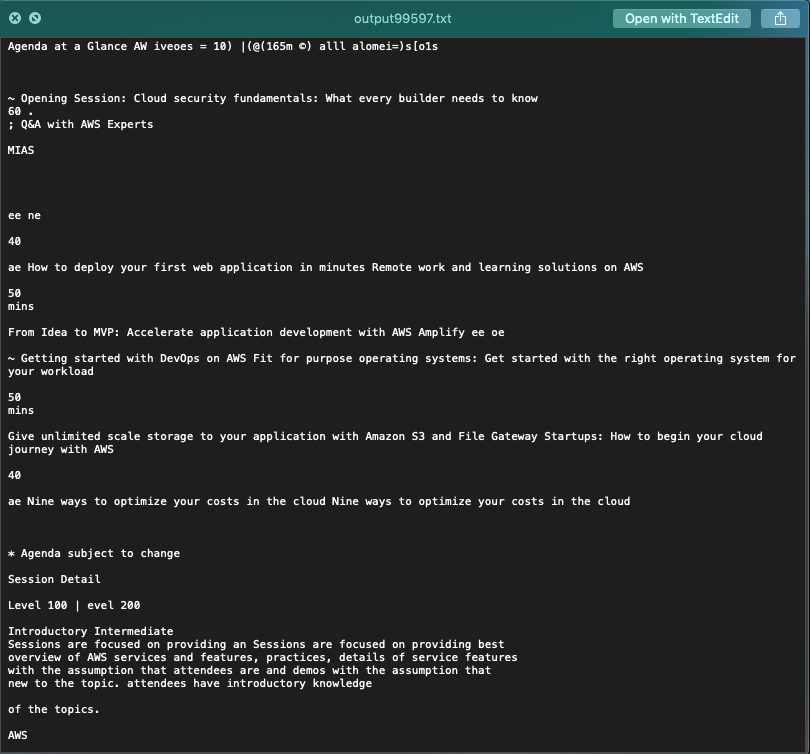
****

****

1. **Screenshot of input pdf file:**

****

1. **Screenshot of output .txt file:**

****

**APPLICATIONS:**

a) **Identity Document Analysis**

AI-backed OCR techniques enable officials to accelerate public data verification including identity documents, licenses, and contracts while handling variations. Unlike manual validation, OCR systems can easily scan through millions of copies to strengthen cybersecurity and examine threats.

b) **Policy Analysis**

The text and feature extraction processes under OCR can process lengthy policy documents to analyze major touchpoints and identify differences. It boosts the decision-making abilities of authorities and enables them to channelize human resources to other critical tasks.

c) **Healthcare**

Healthcare has also seen an increase in the use of OCR technology to process paperwork. Healthcare professionals always have to deal with large volumes of forms for each patient, including insurance forms as well as general health forms. To keep up with all of this information, it is useful to input relevant data into an electronic database that can be accessed as necessary. Form processing tools, powered by OCR, are able to extract information from forms and put it into databases, so that every patient's data is promptly recorded. As a result, healthcare providers can focus on delivering the best possible service to every patient.

**CONCLUSION:**

In conclusion, OCR is a very remarkable technology that holds a lot of potential. In this day and age, such tools are already quite advanced. However, Optical Character Recognition is going to look even better in the future. AI is on the way to becoming one of the most influential trends in the coming years, revolutionizing information as we know it.

**FUTURE SCOPE:**

* OCR technology is becoming very widespread in professional environments such as historical archives, museums, and libraries. This is a great way to preserve ancient texts or images in a digital format. More importantly, these documents can also be examined in the digital domain without disturbing the original physical materials.
* **Language Converter through OCR:** Once a complete OCR has been developed for two languages with font encoding, spell checker and grammatical sentence check, then a converter could be implemented to convert sentences from one language to another through a transliteration and translation scheme.
* **Speech recognition from OCR:** The most required application today is Speech recognition. The recognized Printed or Handwritten character could be recorded and through a voice synthesizer speech output could be generated. This would help the blind to send and receive information.

-----------------------------------\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*-----------------------------------