1. **PROJECT TITLE -** Smart OCR for Document Digitization
2. INTRODUCTION

2.1 Overview

Optical Character Recognition (OCR) is a process run by an OCR software. The software will open a digital image, e.g. a tiff file containing full text characters, and then attempt to read and translate the characters into recognizable full text and save them as a full text file.This is a very useful and cost efficient process for large scale digitisation projects for text based materials including books, journals and newspapers.

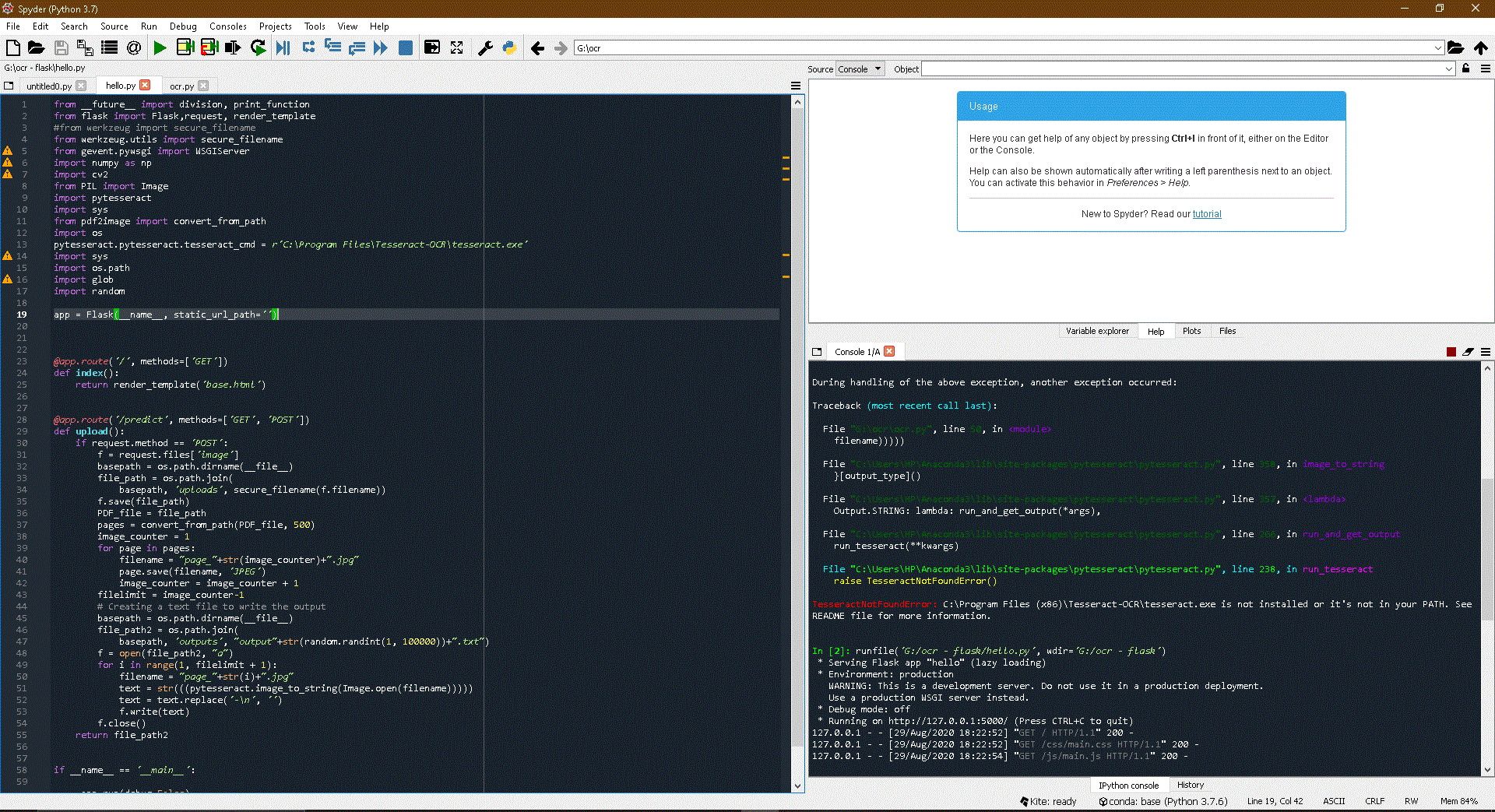
2.2 Purpose

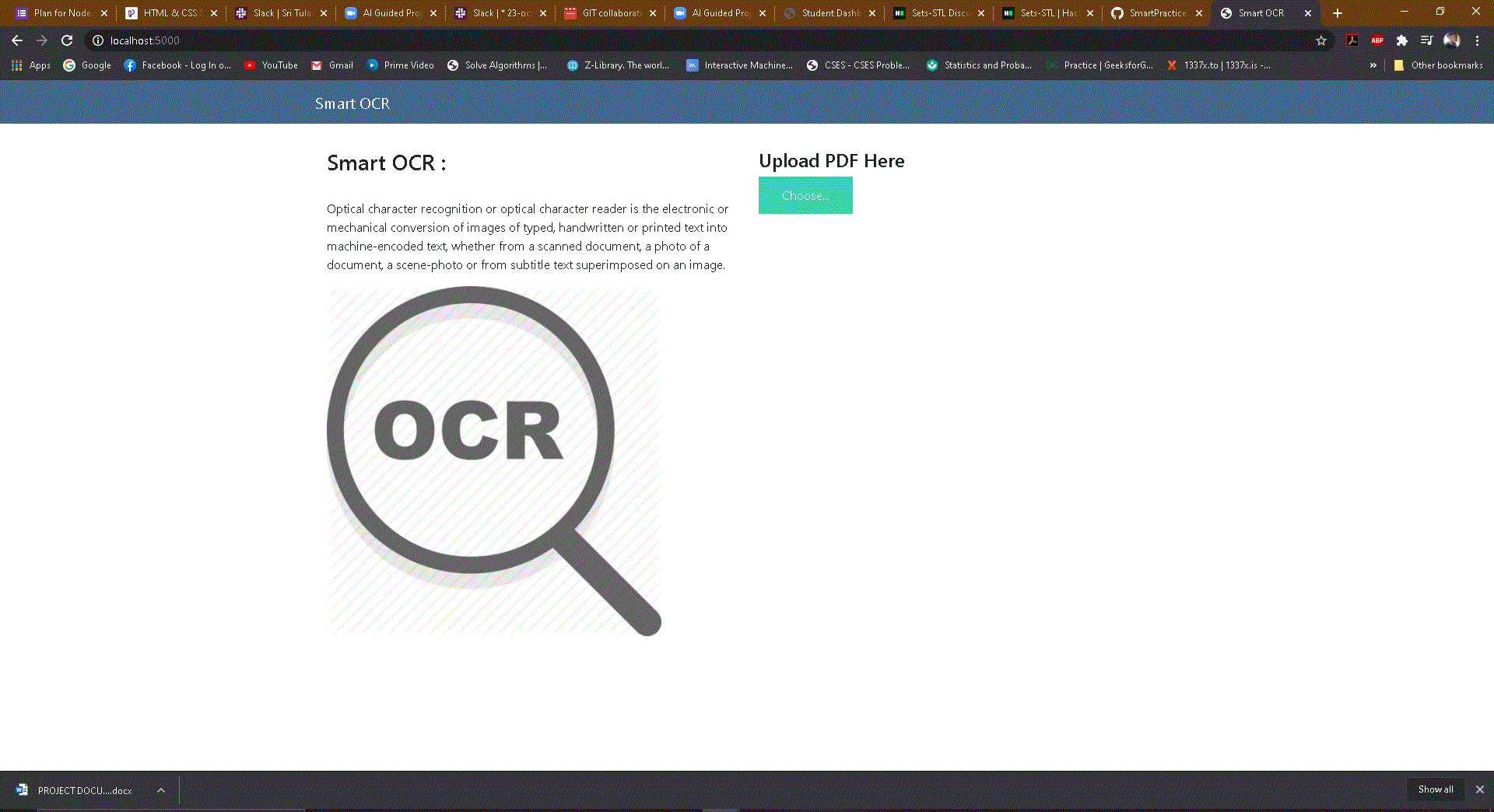
The project aims at creating an application from where the user can upload a pdf document, the document is analyzed by Optical character recognition package to extract text from it. The extracted text is again saved in a text document in the local drive.

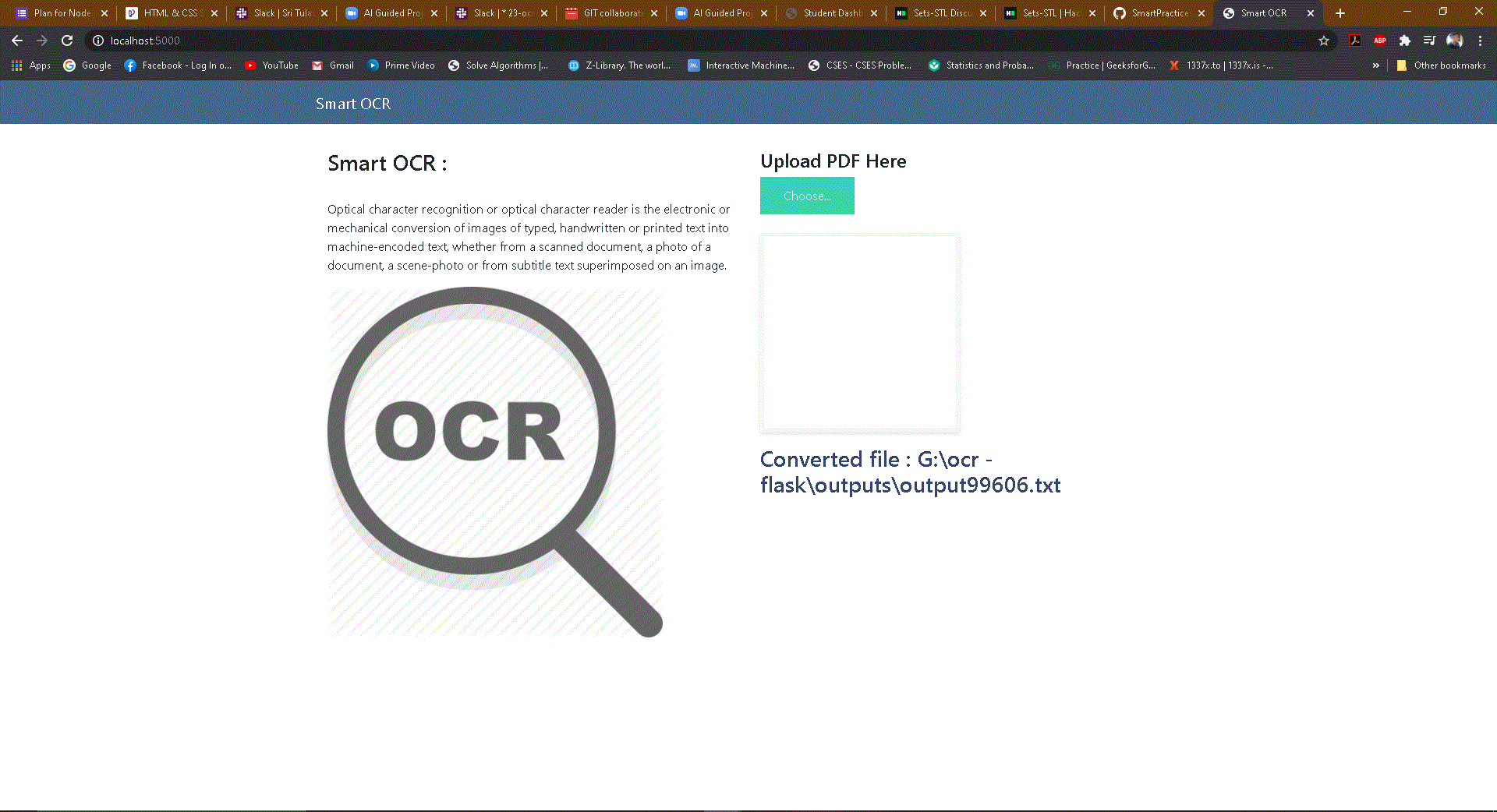
3. RESULT

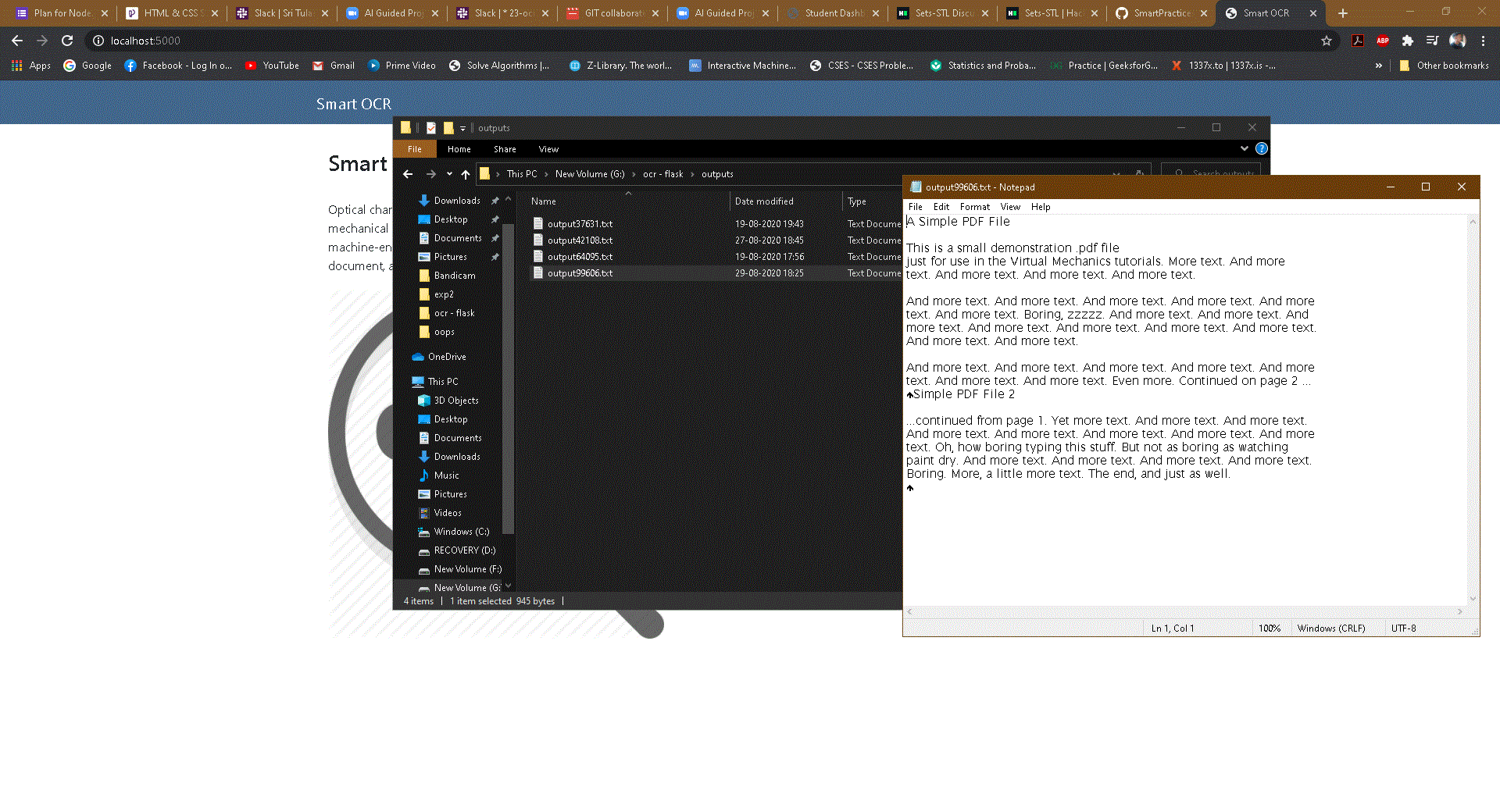
3.1 Screenshots of output

->With flask

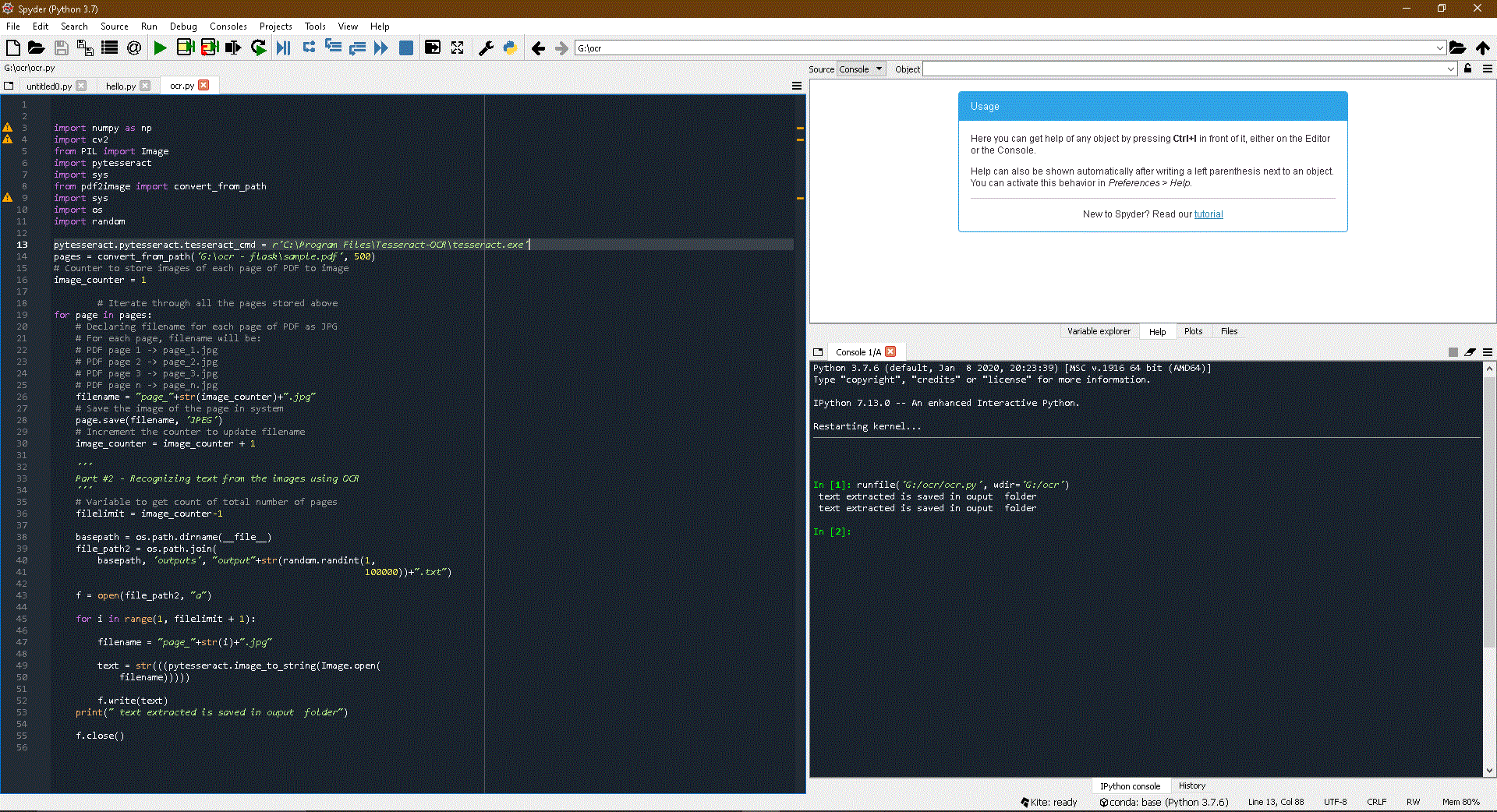


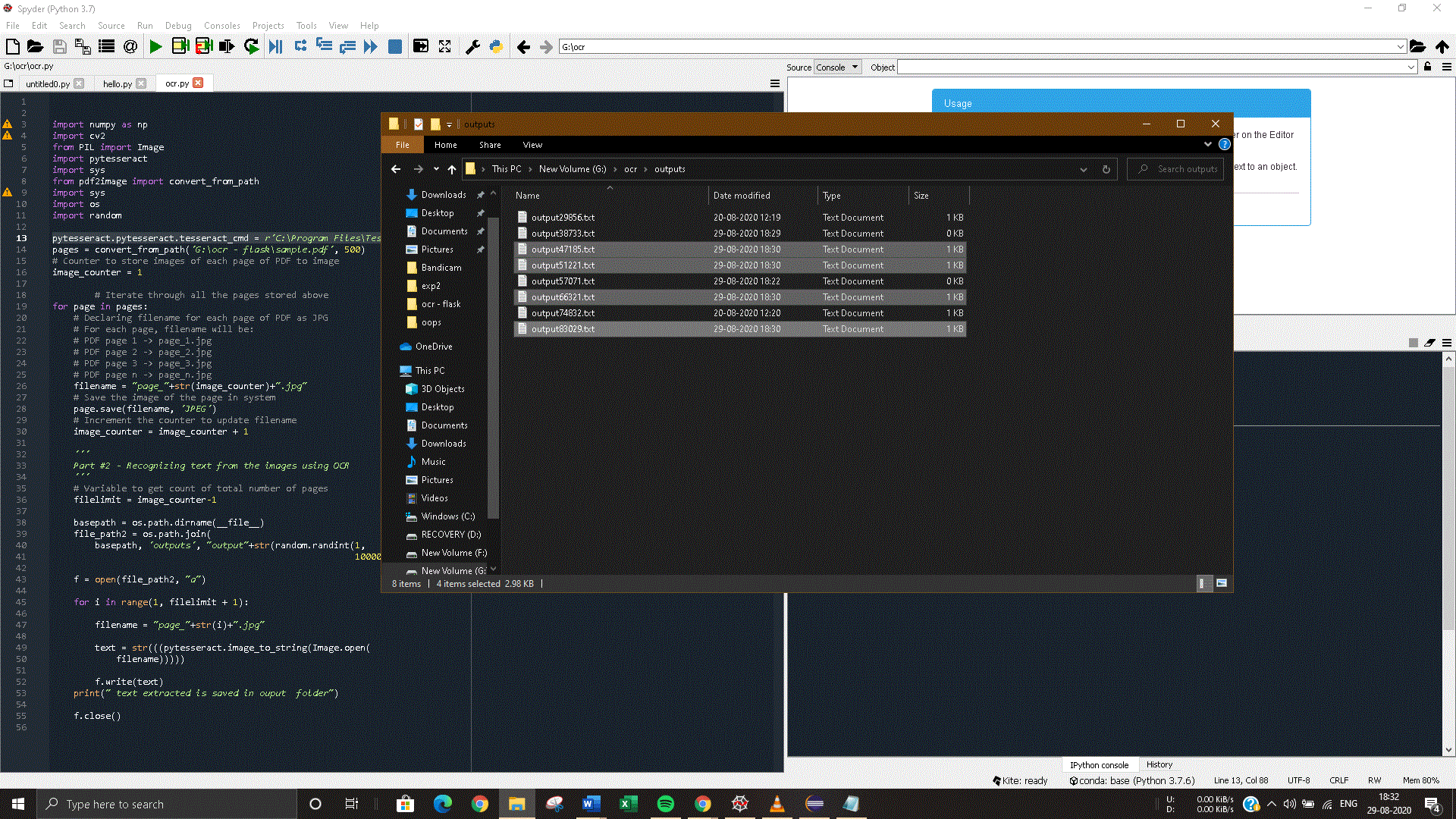






->Without flask





1. APPLICATIONS
2. Data entry for business documents, e.g. Cheque, passport, invoice, bank statement and receipt
3. Automatic number plate recognition.
4. In airports, for passport recognition and [information extraction](https://en.wikipedia.org/wiki/Information_extraction)
5. Automatic insurance documents key information extraction
6. [Traffic sign recognition](https://en.wikipedia.org/wiki/Traffic_sign_recognition)
7. Extracting business card information into a contact list
8. More quickly make textual versions of printed documents, e.g. book scanning for [Project Gutenberg](https://en.wikipedia.org/wiki/Project_Gutenberg)
9. Make electronic images of printed documents searchable, e.g. Google books.
10. Converting handwriting in real time to control a computer (pen computing)
11. Defeating CAPTCHA anti-bot systems, though these are specifically designed to prevent OCR. The purpose can also be to test the robustness of CAPTCHA anti-bot systems.
12. Assistive technology for blind and visually impaired users
13. Writing the instructions for vehicles by identifying CAD images in a database that are appropriate to the vehicle design as it changes in real time.
14. Making scanned documents searchable by converting them to searchable PDFs

1. CONCLUSION

The program was successfully executed.

1. FUTURE SCOPE

OCR is finally moving away from just seeing and matching. Driven by deep learning, it's entering a new phase where it first recognizes scanned text, then makes meaning of it. The competitive edge will be given to the software that provides the most powerful information extraction and highest-quality insights.