International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 05 | May 2021 www.irjet.net p-ISSN: 2395-0072

Splitting of Complete PDF Document into Individual Student Document

Pratiksha Gaikwad[1], Vishakha Fusate[2], Isha Gamare[3], Geeta Arwindekar[4]

[1]BE student, Information Technology, Datta Meghe College of Engineering, Navi Mumbai, India
[2]BE student, Information Technology, Datta Meghe College of Engineering, Navi Mumbai, India
[3]BE student, Information Technology, Datta Meghe College of Engineering, Navi Mumbai, India
[4]Ass.Professor, Information Technology, Datta Meghe College of Engineering, Navi Mumbai, India

---

Abstract- *This project splits one pdf file page wise which can be used for various application. It is build specially for our college ERP system (Enterprise Resource Planning) which a web application which contains all result and other modules of our college that can be handled by administrative and access by faculties , students for viewing , updating grade cards or other data related to college. As in our college the student's grade card of same semester are present in one pdf document so the faculty wants the grade card of every student in one individual pdf file with student id as the file name so that they can access every student's data uniquely. Our system will split that complete pdf document into separate file page wise and renamed as per student id. For splitting the pdf is converted to image format and then by using the pytesseract library we can get student id and split the pdf into separate documents. So using this system we can get all student's grade card into individual student pdf document so that they can search and access student by looking at that file name only.*

***Key Words*: PDF, ERP, grade card ,pytesseract, image.**

1. INTRODUCTION

As the project is based on ERP system of our college in which all student's grade card are present in pdf format so if any faculty wants to access particular student's grade card so our system will help them to split that complete pdf and save individual file pdf . we can directly convert the complete pdf document into individual document. So our project is about the splitting of complete pdf document into individual student document. In college or any educational institute we require lot of data of students like student's marks, personal information and all the details about each of the student.

It is an ERP based project on "Splitting of a complete pdf document into individual students documents". Grade cards of all students are generated into one pdf file. This project will split the complete pdf file into individual files and search for student id. Then save the individual file in concerned student id file. So the all the grade card of student are included into one complete document but we need to save the data of every student in different student folder. So using our system we can split the grade card of every student as per their student id as file name so we can search the required student's data.

1.1 Proposed System

The proposed system split the complete pdf document in which grade card of students are present into individual student document. In this we have taken input demo file in which some students grade card are combined as complete pdf. Then we have given a graphical user interface using python AppJar from which we can take input file from user and output directory where we have to save our individual files. In this we have converted pdf file document into images so that we can apply data extraction on it. Then using pytesseract libarary The system will extract all data from image and we get student id and then split the pdf into separate pdf files in desired folder as per student id as file name. So using this system the user can easily identify the unique student grade card from entire grade card pdf file.

1.2 Pytesseract

Pytesseract is an open-source text recognition engine. Tesseract was considered as one of the most accurate open-source OCR tool. we can use it directly or can use the API to extract the printed text from images. It supports an different variety of languages. It can also be used to recognize text from an image of a single text line. So it is called as Google's Tesseract-OCR Engine. It is also useful as a stand-alone invocation script to tesseract, as it can read all image types supported by the Pillow and jpeg, png, gif, bmp, tiff, and others. The text extraction from pdf pages is done using this library. So using image_to_string method we get all extracted data from images of pdf.

2. Approach

In this system it will consider all aspects for splitting of complete pdf document in which all students grade card are present. As this project is ERP based the system will specifically focuses on the pdf file which contains grade card of different student so it is more useful for teachers to access any student's data very quickly by just looking at their file name. The system will split the given input file into same number of output files and renamed as per student id of student from that grade card. First the graphical interface is done using appJar it is a cross platform python library for creating GUI in which the application will first ask the user for giving input file which contains combined grade cards of many students then it has option for taking path of output folder where we want to save our splitted data files. So after

this the pdf file is open using fitz library which is used for accessing or manipulating the images in pdf. As the pdf is portable document format file which is encrypted, the system convert the pdf pages into image format. And then next step is to iterate over the pdf pages so that it gives the output that how many images are available on that pdf page.so It show how many number of images found on first page or second page etc. After this using xref of all images it has to convert it to image bytes and then save as image extension. After getting xref of image using PIL which is pillow library for accessing image files, open it as image and then by applying pytesseract for extracting data from images, all image data is converted to string and save. The next step is for searching the desired student id in that extracted data, so for that using regular expression the proper condition is given such as the student id is of 11 character number so the condition is [0-9]{4}[A-Z]{4}[0-9]{3}. So after it will findall the matches found for that condition and the user get all student id present on every image. The last step is whenever the student id is found on every image the system will save that image in PDF format and renamed that file as particular student id. So after clicking on split button the user will get all the splitted grade cards in desired output folder as per their student id. For example if student id is 2017FHIT009 then the file will get 2017FHIT009 as pdf file name so any user can identify and access unique student's grade card very easily.

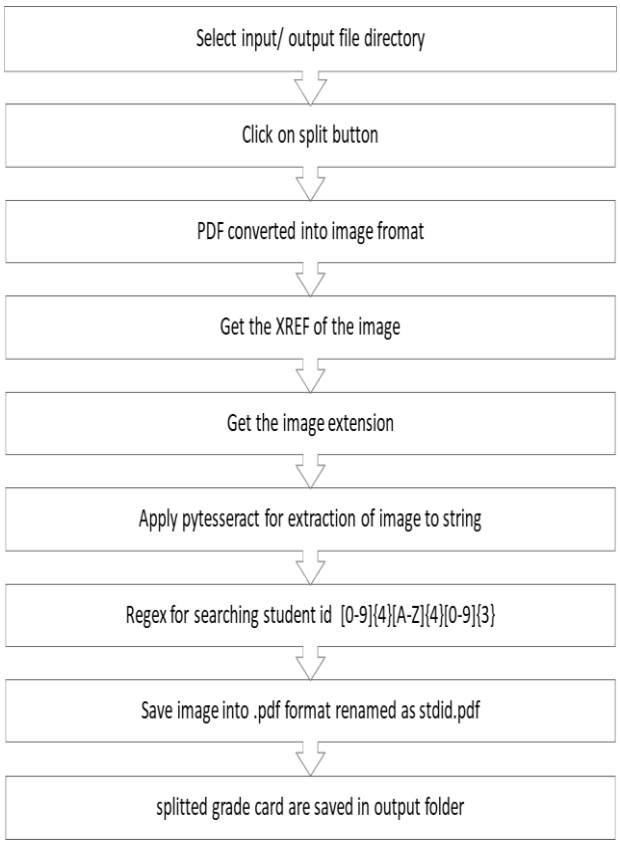

Fig -1: Flow chart of project

3. CONCLUSIONS

We have developed a python application system using which the teacher's or user can get the output of pdf grade card files as per student id of student. As it is mainly focuses on Engineering college data files like grade cards of students are in form of complete pdf file so the system can extract all student's data and get the desired student id and renamed it as per unique student id. So it splits one pdf file into multiple student pdf file as per their unique id. Therefore this system is mainly useful for all college's teachers or staffs for searching and accessing particular student's grade card very easily.

ACKNOWLEDGEMENT

Working on the project on "Splitting of complete pdf document into individual student document" was a source of immense knowledge to us. However, this would not have been possible without the equal support of all the members. We would like to express our sincere gratitude to Asst. Prof. Geeta Arwindekar for her guidance and valuable support. We would like to extend a sincere thanks to her.

REFERENCES

[1] https://pbpython.com/pdf-splitter-gui.html

[2] https://www.tutorialexample.com/python-convert-pdf-to-images-with-given-scale-using-pymupdf-python-tutorial/

[3] https://www.geeksforgeeks.org/python-reading-contents-of-pdf-using-ocr-optical-character-recognition/