**PDF Entity Recognition**

**RATIONALE OF WORK**

* Organizations like hospitals and retail shops still depend on their prescriptions or bills in handwritten forms which have countless disadvantages.
* For example, it becomes difficult to store these data later on for the organizations and after some time, it becomes very difficult to understand the written data. It is also vulnerable to climate and storage conditions.
* A solution for this is if we can store this data in a database based on their data in a structured manner just by scanning them.
* This will make the storage of all their present and past data very easy. Also, by using this as a additional tool for their registration processes, they can continue with the present manner of their process and just use this as an additional task.
* We will be developing our software for this purpose.

**OBJECTIVES OF WORK**

We need to develop software to:-

* Extract data from pdf file
* Differentiate between data and its headings
* Compare the headings with columns of a database for similarity
* Give output as structured tables of data
* To achieve this, we first use a pdf scanner to convert any document into a virtual pdf document.
* Then read the pdf document and convert it into a text file
* Extract headings from the text file
* Compare the headings with the column headings of a database to check for similarity
* Store data in a database as structured data

**METHODOLOGY**

The methodology that we are going to follow is as follows:-

1. First, we are going to take a PDF file as input for our software.
2. Next, we will use OCR for recognition of characters in the input file.
3. This will give us the text file for our input PDF file.
4. In this output text file, we will extract the headings and data under that heading in the form of key-value pairs.
5. Then, we will examine the similarity of the headings using NLP, with the column titles of our database.
6. If the two are similar or same, then we store the data under that column.
7. This will create a structured database for the organization with all its data stored at one digital location.

**TOOLS USED**

* Pdfminer: contains a command line tool called “pdf2txt.py” that extracts text contents from a PDF file
* Subprocess: a standard library module that allows us to invoke the “pdf2txt.py” command line tool within our code
* NLTK: the Natural Language tool-kit, or NLTK, serves as one of Python’s leading platforms to analyze natural language data
* String: provides variable substitutions and value formatting to strip non-printable characters from the output of the text extracted from our journal article PDFs
* Unicode: allows Latin Unicode characters to degrade gracefully into ASCII. This is an important feature because some Unicode characters won’t extract nicely

**MODULES**

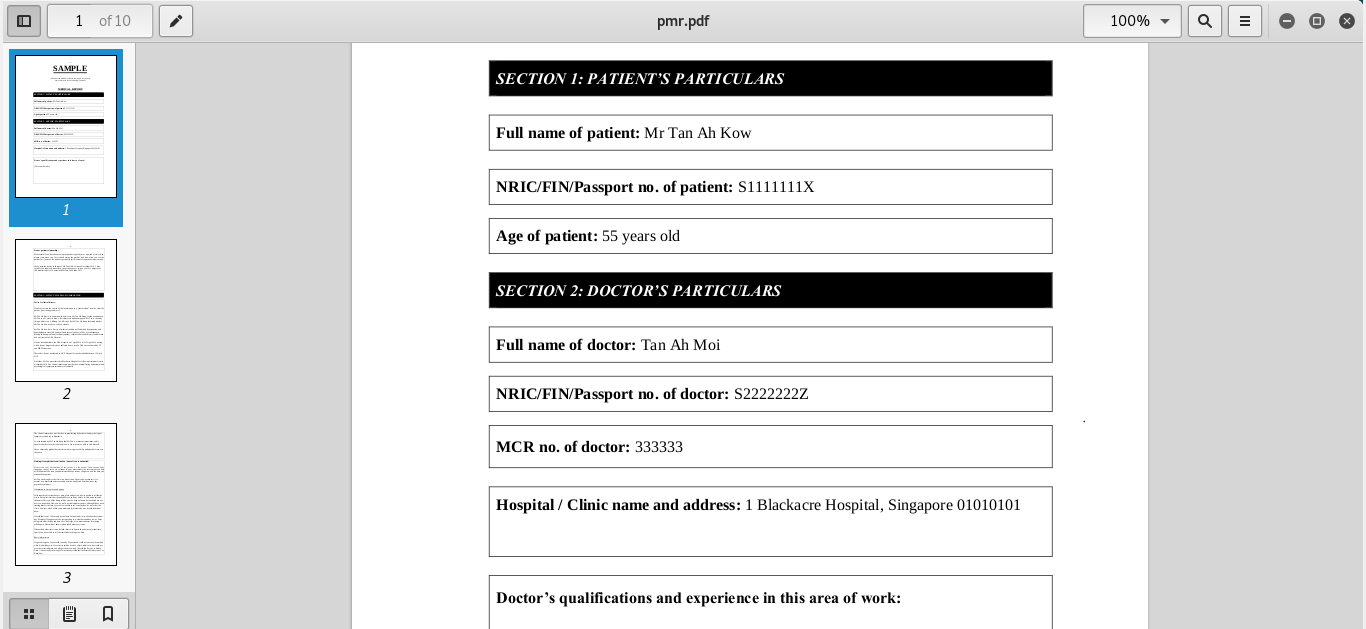
Modules available in this project are:-

* PDF file as input
* Character Recognition using OCR
* Pdf to text file conversion
* Heading Extraction from the Text file
* Heading Comparison with Database Columns for Similarity
* Data Entry in Respective Database Columns

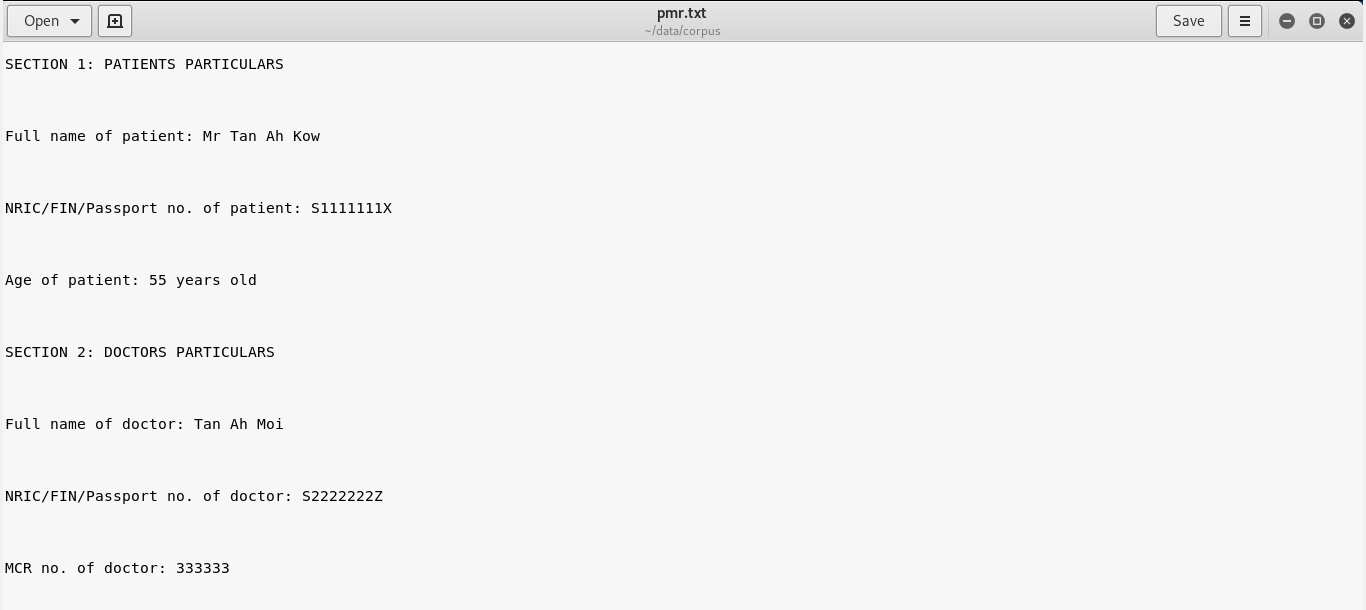
**RESULT**

**The results we got while developing each module is given below**

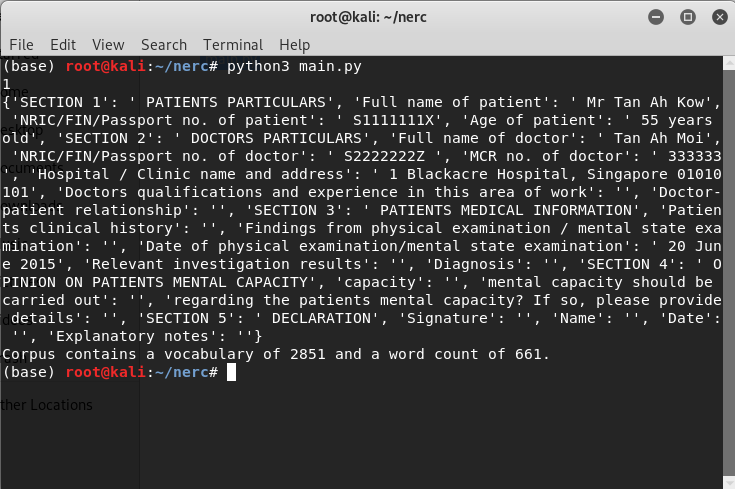
* PDF File As Input



* OCR Character Recognition And Conversion To Text File



* Heading Extraction From Text File



* Database with Entries

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| S.NO | Date of visit | Name of patient |  |  |  |  | Doctors Recomendation |
|  |  |  |  |  |  |  |  |

**DIFFICULTIES FACED**

During the engineering of this product, we faced some problems. They are listed below

1. We have used Natural Language Processing but since we have not yet studied anything in it, it is taking a lot of time.
2. Search for the best database to store the data.

**FINAL DELIVERABLES**

A software which will be capable of extracting all the text from any pdf document and storing the desired data in a database in a structured manner.

**REFERENCES**

Entity Extraction:

<https://medium.com/district-data-labs/named-entity-recognition-and-classification-for-entity-extraction-6f23342aa7c5>

NLP Tutorial:

<https://www.tutorialspoint.com/natural_language_processing/index.htm>