THE TEXT MINING SYSTEM FOR FORMLESS DATA TO STRUCTURED TABLE DATA

[1]Lavanya P R, [2]Dr. Shankaragowda B B

[1] Lavanya P R, MCA Student, Department of Master of Computer Application, BIET, Davanagere

[2] Associate Professor, Department of Master of Computer Application, BIET, Davanagere

---

ABSTRACT: Data extraction is concerned with using natural language processing to naturally remove the fundamental nuances from text records. The inherent reliance of current methodologies on the application space and the objective language is a tremendous burden. A few ML strategies have been used to work with the data extraction frameworks' portability. This task depicts an overall strategy for developing an in-line extraction framework using regular expression in conjunction with the technology. Working interaction in this task can be disseminated by the administrator and the client for a better presentation or to appropriately lead the undertaking. This text extraction method aided in the extraction of the file in table format. In this project, I am designing the application with two Python scripts for improved performance.

KEYWORDS: M L, Text, Algorithm, Extraction.

I. INTRODUCTION

Text extraction, also known as keyword extraction, uses machine learning to examine text and extract relevant or important words and expressions from unstructured data, such as news stories, reviews, and customer service protests. ML algorithms are used to aid in the content extraction and improvement strategies. Finally, the extracted text from the image is moved to the specified application or record type. There are various types of text extraction calculations and methods that are used for various purposes. To separate the content in our application, we use regular expressions and the tesseract technique.

The primary goal of our paper is to develop an application that will be useful in the college admissions process. Because this college management system project includes a student's admission procedure from the time he or she enrols in college in the first year until the time he or she completes his or her course. It takes time to separate student information into separate records. All of these documents must be reviewed and updated. The process of maintaining student records is becoming more difficult as the number of students grows and management is responsible for all of the students' records. The best way to make this process easier is to use machine learning to mine admission records.

Text mining admission records with machine learning simplifies and automates every aspect of the admission record. It is a list of every student who has been accepted into a college. According to department guidelines, the admittance register must be kept in the college indefinitely. As a result, it must be specially bonded and kept in secure possession. Because this register is occasionally requested by higher-ups, it must be error-free.

It aids in the correct and timely collection of information from people.

1. The student's serial number and name.
2. The name of his father.
3. Date of birth of the student
4. The date on which you were accepted into college.
5. Course
6. Entrance Fees

In this application, the module's operation is divided into two parts: admin and users, both of which have a login interface. After completing the registration process, the user can upload the images formatted admission form by using the upload file option in the menu. After submitting the file, the file will be extracted and the output will be in a tabled format. The view data menu allows users to view the data.

II. RELATED WORK

Y. Zhan et al. [1] proposed a robust method for segmenting text from colour images. To locate candidate text lines, the

proposed algorithm employs multiscale wavelet features and structural information. Then, from the candidate text lines, an SVM classifier was used to identify true text. This method was divided into four stages. Text blocks were enhanced in the preprocessing step by using cubic interpolation to rescale the input text blocks and a Gaussian filter to smooth the text blocks and remove noises. A component filtering procedure was used to remove non-text connected components from these image blocks after they were split into connected components. The left connected components were merged into several text layers using the K-means clustering algorithm, and a set of appropriate constraints were applied to find the true text layer. Finally, a post-processing step was used to refine the text layer.

Thai et al. [2] described a method for extracting text from graphical document images. The Morphological Component Analysis (MCA) algorithm was used, which is a sparse representation framework advancement with two appropriately chosen discriminative over complete dictionaries. Undecimated wavelet transform and curvelet transform were used to create two discriminative dictionaries. This method solved the problem of text and graphics touching and was also insensitive to different font styles, sizes, and orientations.

S.Audithan et al. [3] developed a method for extracting text regions from documents that is both efficient and computationally fast. To detect the edges of candidate text regions, they proposed the Haar discrete wavelet transform. The thresholding technique was used to remove non-text edges. The isolated candidate text edge was connected using the morphological dilation operator, and a line feature vector graph was generated based on the edge map. To detect text pixels, this method used an improved canny edge detector. The spatial distribution of edge pixels was used to extract stroke information. Finally, line features were used to generate and filter text regions.

Grover et al. [4] described a method for detecting text in documents where the text was embedded in complex coloured document images. To accomplish this task, they proposed a simple edge-based feature. By forming a weighted sum of the R, G, and B components, the image was converted to grey scale. The grayscale image was then edge detected by convolving it with Sobel masks, one for each horizontal and vertical edge. Convolution was then followed by the removal of non-maxima and the thresholding of weak edges. The edge image was then divided into small non-overlapping blocks of m x m pixels, where m is determined by the image resolution. They used a pre-defined threshold to distinguish the text from the image during block classification.

P. Nagabhushan et al. [5] proposed an innovative method for extracting text from complex background colour document images. To detect edges, the proposed method employed a canny edge detector. When the dilation operation was applied to the edge image, it resulted in holes in the majority of the connected components that correspond to character strings. Connected components that did not have a hole(s) were removed. Other non-text components were eliminated by computing and analysing each connected component's standard deviation. To perform foreground segmentation in detected text regions, an unsupervised local thresholding method was devised. Finally, the noisy text regions were found and reprocessed to improve the quality of the retrieved foreground.

Davodet.al[6] proposed a robust and efficient wavelet-based algorithm for automatic text extraction from coloured book and journal cover sheets. To detect edges from detail wavelet coefficients, a dynamic threshold was used. By blurring approximate coefficients with alternative heuristic thresholding, more effective edges were obtained. Finally, the Region of Interest (ROI) technique was used to extract text. They tested the algorithm's performance on 80 images gathered from the internet.

Karin et al. [7] proposed yet another algorithm for Automatic text location and identification on coloured book and journal covers. A clustering algorithm was used to reduce the number of colours. A top-down analysis based on successive horizontal and vertical splitting was used to locate text candidates. A bottom-up analysis used a region growing method to detect homogeneous regions; a grouping step was used to find subsets of regions. Finally, text and non-text regions were separated.

Zhixin Shi et al. [8] proposed an extraction algorithm for a complex handwritten historical document based on connectivity features. This paper described an algorithm that employs the adaptive local connectivity map (ALCM) technique. When the grey scale image is thresholded, it reveals distinct text-line patterns as connected components. The connected components were grouped into location masks for each text line using a grouping algorithm. The text line components were extracted by mapping the location masks back onto the binary image. The splitting algorithm solved the issue of components touching multiple lines. This method dealt with skewed or fluctuating text lines and was used for other types of images such as binary images, machine printed images, and even mixed script.

Syed Saqibet et al. [9] described a method for extracting curled textline information from grayscale camera-captured document images. Multi-oriented multi-scale anisotropic Gaussian smoothing was used to improve the grayscale textline. Ridges were used to detect the central lines of curled textlines. This method was based on differential geometry, and it measured curvature using the local direction of gradients and second derivatives. The Hessian matrix was used to determine the direction of gradients and derivatives. Ridges were detected using this information by locating the zero-crossing of the appropriate directional derivatives of the smoothed image. For estimating x-line and baseline pairs from detected textlines, a modified coupled snakes model was used[10,11]. Their method is resistant to high degrees of curl and requires no post-processing[12].

III. EXISTING SYSTEM

Currently, many methods are implemented to convert different file formats into normal text format, and many algorithms and machine learning techniques are implemented to convert unstructured or different format files into desired format files. Somehow, using the applications or the concept is very tedious for the common people who require those applications for their purposes.

IV. PROPOSED SYSTEM

Using a regular expression, we will extract the image formatted file into the desired tabled format file in this proposed system. In this section, we will describe the application using two modules: users and admin. Users who want the desired output can use this application by entering their authenticated username and password, making the application safe and secure from unauthorised users.

Screenshots Regarding to the system:

User Upload File page:

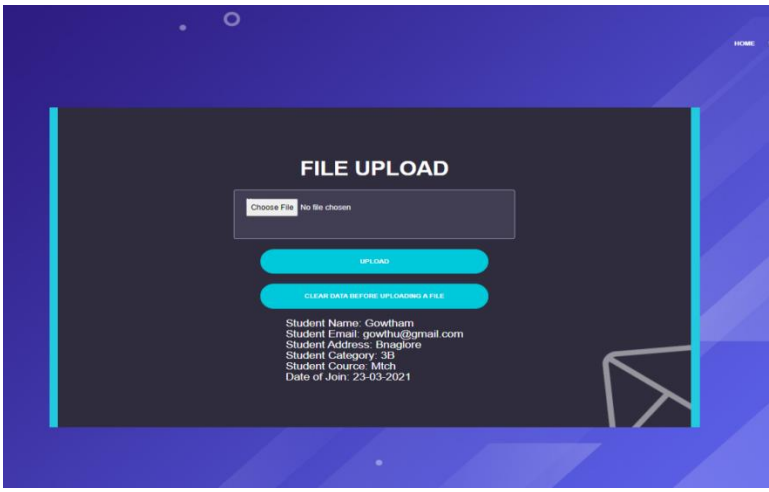

Fig: User Upload File page

The above figure is the user upload file page where user can upload the student information.

Admin View Record page:

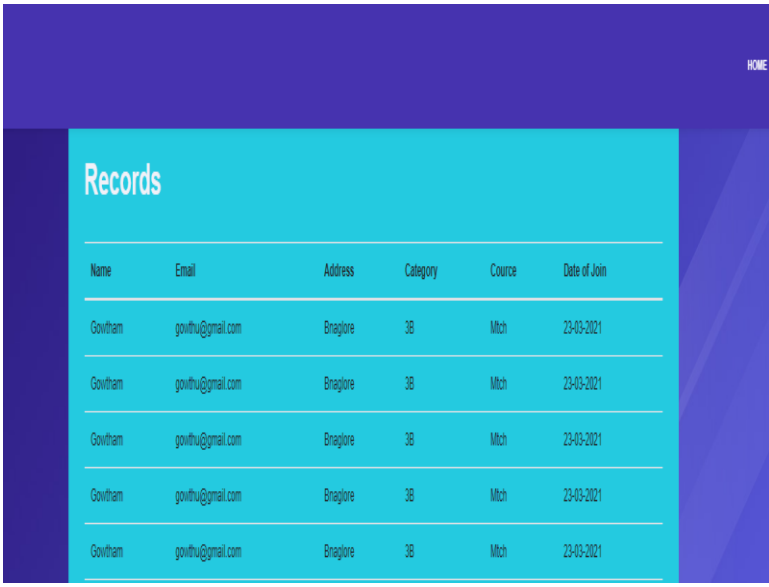

Fig: Page for the administrator to see records

V. CONCLUSION

In this manner, we will build an Admission management system, which will aid in the reduction of Clark's manual work, resulting in a reduction in manpower requirements. Students' records can be accessed in a matter of seconds. There should be more clarity in the account area. Our system is mostly concerned with data extraction. Transform the entire information on the Admission Receipt, which is unstructured data, to a structured format.

Students' databases can be retrieved quickly, allowing for proper record keeping. The Admission receipt will be entered into the system by the administrator so that it can be processed automatically and the data can be stored in table format without having to manually enter each field.

REFERENCES

[1] Y. Zhan, W. Wang, W. Gao (2006), "A Robust Split-And-Merge Text Segmentation Approach For Images", International Conference On Pattern Recognition,06(2):pp 1002-1005.

[2] Thai V. Hoang , S. Tabbone(2010),"Text Extraction From Graphical Document Images Using Sparse Representation"in Proc. Das, pp 143–150.

[3] Audithan,,R.M.Chandrasekaran (2009), "Document Text Extraction From Document Images Using Haar Discrete Wavelet Transform",European Journal Of Scientific Research, Vol.36 No.4 , pp.502-512.

[4] Sachin, Grover,Kushal Arora,,Suman K. Mitra(2009),"Text Extraction From Document Images Using Edge Information",IEEE India Council Conference.

[5] P. Nagabhushan, S. Nirmala(2009) ,"Text Extraction In Complex Color Document Images For Enhanced Readabi

[6] lity",Intelligent Information Management, pp: 120-133.

[7] Davod Zaravi, Habib Rostami, Alireza Malahzaheh, S.S Mortazavi(2011)," Journals Subheadlines Text Extraction Using Wavelet Thresholding And New Projection Profile", World Academy Of Science, Engineering And Technology .Issue 73.

[8] Karin Sobottka, Horst Bunke and Heino Kronenberg(2009), "Identification Of Text On Colored Book And Journal Covers", ICDAR.

[9] Zhixin Shi, Srirangaraj Setlur And Venu Govindaraju(2005), "Text Extraction From Gray Scale Historical Document Image Using Adaptive Local Connectivity Map", Proceeding Of The Eighth International Conference On Document Analysis And Recognition, Vol. 2, pp: 794–798.

[10] Syed Saqib Bukhari , Thomas M. Breuel,Faisal Shafait(2009), "Textline Information Extraction From Grayscale Camera-Captured Document Images ", ICIP Proceedings Of The 16th IEEE International Conference On Image Processing, pp: 2013 – 2016.

[11] Boussellaa , Aymen Bougacha, Abderrazak Zahour, Haikal El Abed, Adel Alimi(2009) ,"Enhanced Text Extraction From Arabic Degraded Document Images Using Em Algorithm", 10th International Conference On Document Analysis And Recognition.

[12] S. A. Angadi , M. M. Kodabagi(2009) , "A Texture Based Methodology For Text Region Extraction From Low Resolution Natural Scene Images ", International Journal Of Image Processing (Ijip) Volume(3), Issue(5).