e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:01/January -2021 Impact Factor- 5.354 www.irjmets.com

TEXT CLASSIFICATION FROM PDF DOCUMENTS

Hindu Kamakshi Ayinala [*1], Suresh Grandhi [*2]

[*1]M.Tech. Student, Department of CSE , Grandhi Varalakshmi Venkata Rao Institute of Technology, Bhimavaram, Andhra Pradesh, India

[*2]Assoc. Professor, Department of CSE , Grandhi Varalakshmi Venkata Rao Institute of Technology, Bhimavaram, Andhra Pradesh, India.

ABSTRACT

Government Agencies and Departments are using the Internet to disseminate information on policies and development requests. Most of the information is shared using a Portable Document Format (PDF). Heavy use of Portable PDF files has promoted research in analyzing the file layout for text extraction purposes. One of the PDF document's difficulties is that using the smart mobiles, users are extensively scanning the documents in PDF format using the phone camera. These documents are in image format. Optical Character Recognition (OCR) techniques need to be employed to get these images into text format.

The domain we considered for the project is to look at the documents from Government Agencies or departments and classify the documents and the paragraphs within the documents using Text Classifiers with Machine Learning Algorithms and other techniques. Classification techniques are to be developed to identify text patterns and classify them. The approach uses available text classification techniques and designing a method to split the paragraphs and also identifying the individual attributes or details within a paragraph.

We have used a dataset with 5000+ documents from the county or city planning departments. A sample set of text patterns is identified from the documents. Methods are to be implemented to automatically get these text patterns and use Machine Learning Algorithms to categorize the documents and the corresponding paragraphs into the identified categories. Algorithms developed can be applied to various applications that require text extraction from PDF documents.

Keywords: Text Classification, PDF to Text, OCR to Text, Text Parsing, Context Mapping, Entity Training, Knowledge Base Creation, Named Entity Recognition (NER).

I. INTRODUCTION

In 1991, Adobe co-founder Dr John Warnock launched the paper-to-digital revolution with an idea he called, The Camelot Project. The goal was to enable anyone to capture documents from any application, send electronic versions of these documents anywhere and view and print them on any machine. By 1992, Camelot had developed into PDF. Today, it is the format trusted by businesses around the world. The PDF is now an open standard, maintained by the International Organization for Standardization (ISO). PDF documents can contain links and buttons, form fields, audio, video and business logic. They can be signed electronically and you can easily view PDF files on Windows or Mac OS using the free Acrobat Reader DC software.

PDF leads the public sector electronic documentation. Documents in the form of laws, memos, reports, publications, contracts, policies, regulations, forms and so on are the infrastructure of government. The process of authoring, editing, reviewing, signing, managing, collecting, disseminating or destroying these documents is itself the work of the government, all day, every day. Reduced to its most basic elements, government is a collection of massive document systems, all interacting (and colliding) in various useful and useless ways. When it does work, it's "good government." When poorly considered or just plain dumb, we call government documentation requirements "red tape" to signify the bureaucratic obstacles these paper piles present.

Today long-standing calls for smaller government are producing a real evolution in document development and implementation. Adobe Acrobat and PDF are helping local, state and federal government offices and agencies around the globe do more with less paper and less inefficiency. In many ways, the U.S. federal government is leading the effort to transform the world of documents from paper to electronic, and Adobe's PDF has emerged as a common standard for the same three reasons that underlie the success of PDF everywhere:

PDF delivers an exact representation of an original document

PDF delivers a consistent experience on all computing platforms and printer

PDF is free to view and print with Adobe's free and ubiquitous Reader software

"There's not a federal agency that does not use PDF," says Greg Pisocky, Business Development Manager for Adobe Systems. "Acrobat software and Adobe PDF are key technologies in some capacity at all branches and levels of government, the military and virtually every agency," PDF adoption is hardly a U.S.-only phenomenon. Internationally, India, Australia and numerous other countries have adopted PDF in a big way, using it for publishing everything from voter-registration lists to family-court forms.

In this project, we have considered the documents from specific government agencies at city, county and state levels. And the departments are in particular are from the planning and zoning divisions.

In an era of increased need for coordination, cooperation, and regionalization between local governments, the role of county planning is much more important. Public involvement is a major tool for three successful county planning programs. A planning commission is not set up to be doing staff work, It is an advisory and a policy making body. There are some things that might accomplish a little bit: scheduling combined planning commission meetings, to keep a light shining on planning issues. But until resources can be put behind those efforts not much can be expected. One of the powers of any planning commission is its ability to create committees – including committees whose membership are not from the planning commission. This gives a planning commission's committee the ability to be set up many different ways. Also, members of such committees can consist of experts, in a topic without having to worry about residency and other such issues.

II. METHODOLOGY

The process of converting the PDF documents which have unstructured text into a structure text involves several phases. The following methodology is followed to convert unstructured text into structure text from the PDF documents.

PDF to Text Conversion

Considered Input documents are PDF documents. These documents are first converted into text format for further analysis and splitting the data into records. PDF documents are of two types. One having text data that can be easily converted using the pdf to text tools. We have used available Python libraries to convert PDF documents into text documents. Most of the PDFs are now are in image format that are scanned using several scanning devices available today including Scanners, Tabs, Mobiles etc. The challenging part is to convert text in image format. We have used OCR (Optical Character Recognition) techniques to convert image data into text.

Text to Records

Once the PDF document text is extracted, this text is in unstructured format. We proposed a process to extract records from the text paragraphs using a record separator mechanism. Each Document type is associated with the document configuration that contains the record separators (record beginning) and record end markers. Record end markers are used to reduce the noise which is unwanted text along with the record text.

The following diagram explains the record separator concept. In this format, ") Case" is considered as a record separator. The record parser identifies the record separator pattern and then splits each record.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:01/January -2021 Impact Factor- 5.354 www.irjmets.com

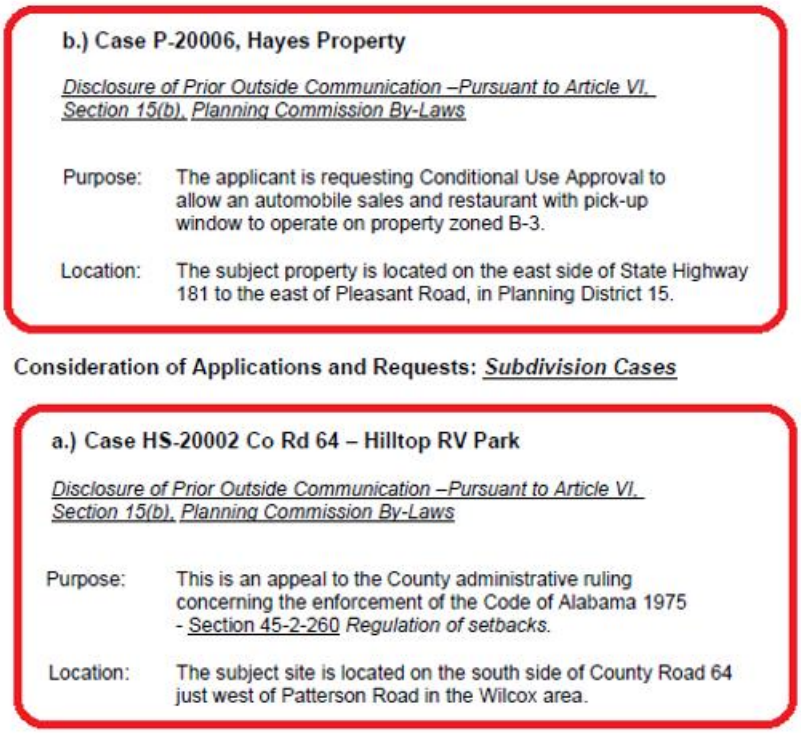

Fig.-1: Record Separation

Records to Data Fields

Each record extracted in the previous step is now processed to split it into data fields which can be inserted into an RDBMS for easy querying and processing of the data. A template driven approach is designed for handling different record types within a document. Each template is further supported by field tokens for identifying the data fields in the text.

{TT0}Purpose : → Project Purpose data field
{TT1}zoned → Zoning data field
{TT2}Location: → Site Location data field

Here, the record mentioned int the figure 1, contains 3 different fields. The first one is the purpose of the record, second is zoning information and the third information is address location of the record. These data fields are identified by the token fields and the field parser works on the text record to split them to the corresponding data fields using the token information.

Dummy tokens can be introduced wherever require to reduce the noise which is the text that is not part of the data field.

III. MODELING AND ANALYSIS

The system architecture is depicted in the below diagram. PDF documents are collected from the department or agency websites. These documents are converted into text format and then are parsed into records and further processed into data fields.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:01/January -2021 Impact Factor- 5.354 www.irjmets.com

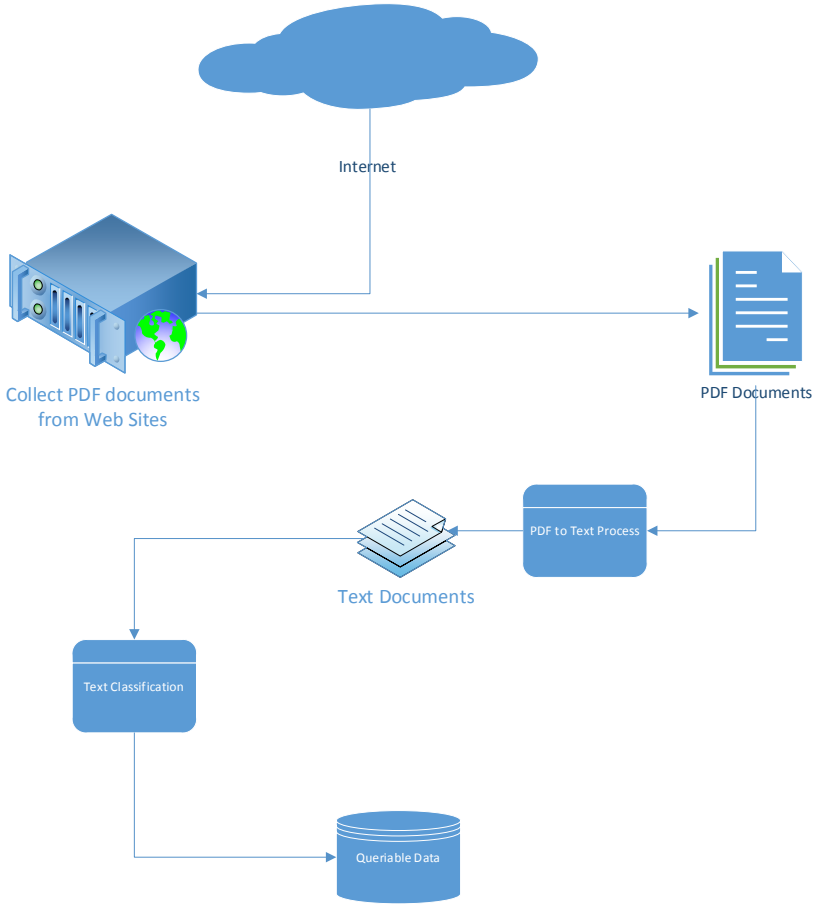

Fig.-2: System Architecture.

The following are the steps to process the input PDF files.

1. Identify the Web Sites where the documents are available
2. Identify the Frequency when the documents are posted
3. Create a schedule using the frequency to gather posted documents
4. Download the Files when they are available
5. Convert the PDF files into text files by a cron job
6. Use PDFToText to convert PDF to Text
7. Handle exceptions and OCR files
    a. Some PDF files may need to be printed to extract only few pages out of huge documents
    b. May need to re-print PDF to clear any special characters.
8. Identify the Record Separators
9. Run the process to split records from the text files
10. Parse the records into data fields and insert them into RDBMS

IV. RESULTS AND DISCUSSION

Once the records are gathered in the database, a separate geo service is used to get the geo-locations of the addresses. Geo-location helps us to present the data in map visualization for easier understanding of the area where the applications are located. The following screenshot shows the map visualization of the applications collected and parsed.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:03/Issue:01/January -2021 Impact Factor- 5.354 www.irjmets.com

Fig.-3: Application Locations.

The following visualization shows the cities with the projects by area. The heat map shown below gives the cities where the developments are happening with the quantity in the units for example, the picture shows the units sqft.

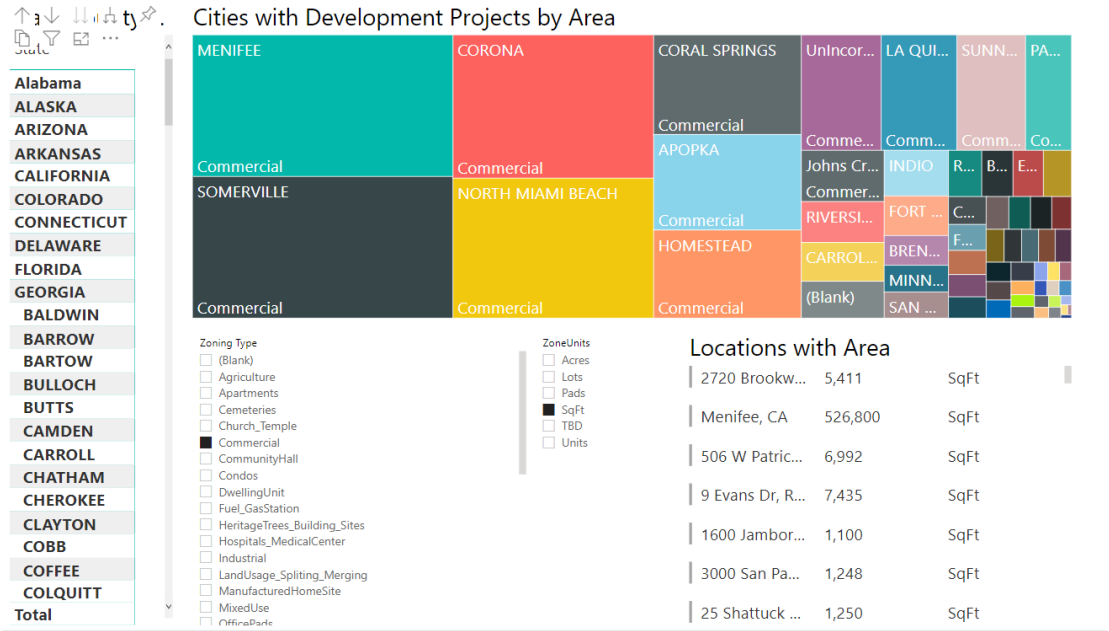

Fig.-4: Cities with Development Projects by Area

V. CONCLUSION

Text processing is an essential task as we have more digital content available on the internet today. We have attempted a problem of available textual data in public domain from the government agencies. The difficulty of locating and analyzing the textual information is the toughest task now a days. Our approach here is to propose a mechanism to convert unstructured text into a structure text in a database format so that the data can be easily queried. Machine learning techniques can be used to classify text when sufficient feature data is gathered.

During the execution tests, it is observed that the proposed scheme works with a greater number of records and PDF files. The system has used the Agenda Files which consists of one or two paragraphs for a case. The same system can be extended to Agenda Packets which contain more information about the proposed projects including the diagrams, more details, project dates etc.

VI. REFERENCES

[1] Holmes, M. (1961). Steel frames with brickwork and concrete infilling, Proceedings of the Institution of Civil Engineers, 473-478.

[2] Smith, B. S. and Carter, C. (1969). A method of analysis for infilled frames, Proceedings of the Institution of Civil Engineers, Vol.7218, 31-48.

[3] Mainstone, R. J. and Weeks, G. A. (1970). The influence of bounding frame on the racking stiffness and strength of brick walls, in Proc. 2nd International Brick Masonry Conference, Building Research Establishment, Watford, England, 165-171.

[4] ATC (1996). Seismic Evaluation and retrofit of Concrete buildings, Vol. 1, ATC-40 Report, Applied Technology Council, Redwood City, California.

[5] Federal Emergency Management Agency (1998). Evaluation of Earthquake Damaged Concrete and Masonry Wall Buildings: Basic Procedures Manual, FEMA-306, Applied Technology Council, Washington DC.

[6] FEMA-356 (2000). Prestandard and Commentary for the Seismic Rehabilitation of Buildings, Building Seismic Safety Council, Washington DC.

[7] E Umamaheswari Vasanthakumar and Francis Bond. A semantic multi-field clinical search for patient medical records. 2018.

[8] Chirag Patel, Atul Patel, and Dharmendra Patel. Optical character recogni- tion by open source ocr tool tesseract: A case study. International Journal of Computer Applications, 55(10), 2012.

[9] Y.Shinyama. Pdfminer: Python pdf parser and analyzer, 2010. http://www. unixuser.org/~euske/python/pdfminer/,.