Use of OCR Technology for Data Extraction Using Amazon Textract

Sumit Muddalkar[1], Kiran Kolte[2], Ayush Batra[3], Ayush Naphade[4], Neha Lokhande[5]

Project Guide, Department of Information Technology[1]
Project Group Leader, Department of Information Technology[2]
Project Group Member, Department of Information Technology[3, 4, 5]
Shri Sant Gajanan Maharaj College of Engineering, Shegaon, Maharashtra, India

Abstract: *In the digital era of twenty first century, everything is becoming automated, and information is stored and transfer in digital forms. But there are many situations where data is not stored in digital form and it is essential to extract text from those hardcopies to store in digitized form. The latest technology such as Text recognition software has completely changed the process of text extraction using Optical Character Recognition. Therefore, this paper introduces the concept of OCR technology, explains the process of extraction using Amazon Textract tool and current research in the area. Detailed information and working methodology of Amazon Textract. Its comparison with other OCR tools and its scope. This paper will help other researchers in the field to get an overview of the technology.*

Keywords: Amazon Textract, Optical Character Recognition, Machine Learning, Google AI

I. Introduction

Optical Character Recognition (OCR) is the machine learning tool use for conversion of text or making a digital copy of the text. It takes resources through handwritten documents, printed text, or from natural images. Optical character recognition is a science that enables us to translate various types of documents or images into analysable, editable and searchable data. The objective of this research paper is to study the use of Optical Character Recognition with Amazon Textract tool and its comparison over other OCR tools available in the market.

In this Research paper, we collected and analyzed research articles on the topic of OCR technology and closely related topics which were published between year 2015 to 2021. Articles were searched using keywords, forward reference searching and backward reference searching in order to search all the articles related to the topic. OCR is a field that has applications in pretty much every other field like in Health care division, Data Extraction, Data Storage, Banking segment, etc.

II. Literature Review

For what reason to involve Amazon Artificial Intelligence and Machine Learning administrations for report robotization?

By utilizing Amazon Web Services Artificial intelligence and Machine Learning administrations to control report handling helps associations and reduce work to each shape type: Amazon Textract gets it and understands records and structures without requiring any broad pre-work to get the structure's design. All things considered, the AI-based approach gets the substance in view of the actual design, in any event, separating the information held in tables or structures and planning that into machine meaningful constructions to demonstrate what has been written in each piece of a structure by planning those qualities to their individual information fields.

- Increase and down depending on the situation: Business tasks are frequently tested by overseeing tops sought after, for instance, during application cut-off times or during occasions like the COVID-19 pandemic. Adaptability and current serverless cloud models are critical, which help rapidly increase to handle enormous volume of reports and afterward proportional down, limiting the continuous expenses right away.
- Join human and AI mastery to affirm or address information passage all the more effectively and rapidly: Tightly incorporated expanded AI banners to a human analyst the parts of structures which the AI couldn't peruse without

DOI: 10.48175/568

hesitation. The blend of AI and a human cooperating conveys an exceptionally hearty way to deal with productively mechanizing an archive work process.

- Perceive maintainability benefits: Organizations can lessen the carbon and energy consumed in actually moving huge loads of actual paper reports among locales, and afterward putting away something very similar in actual documents. Moving to electronic archive handling, with advanced sorting rooms ingesting and filtering the media, frees the labor force away from being truly co-situated with the records. A more extensive shift for the labor force that AI brings, is the capacity to depend on the AI for the commonplace errands and permit the human labor force to zero in on more worth adding undertakings that require interestingly human abilities.
- Remove additional worth from information to further develop processes: Using ML methods likewise increases current standards on how much worth can be extricated from records. Amazon Rekognition is utilized to recognize and separate pictures or charts installed inside archives, saving time and manual exertion by distinguishing and editing out pictures. The text inside reports is handled through administrations Amazon Translate, making it conceivable to help 55 dialects and variations from Afrikaans to Vietnamese, without expecting in-house interpreters. Amazon Comprehend utilizes normal language handling strategies to assist with getting a report. This is much of the time used to emergency inbound correspondence by getting the idea of the solicitation, and guiding the undertaking to the best work line. These can be taken care of straightforwardly into mechanical cycle mechanization driven work processes to some extent or completely attempt work that would require human groups.
- Assemble information bits of knowledge to further develop administrations: Extracted information can be driven into a diagram data set, like Amazon Neptune, for ensuing organization examination. This approach distinguishes application misrepresentation where organizations of partners, locations, and organizations are distinguished from the diagram that may be generally extremely difficult to perceive.

AWS Textract is rethinking the way that organizations cycle records in a Digital World

Contemplate the last time you opened a financial balance, applied for protection, or renegotiated your home. It was most likely done on paper. The quantity of reports in a home loan bundle alone is more than 100 pages in length. How would you manage all that paper? For some organizations across an assortment of businesses, including monetary administrations, medical services, and assembling, it is careful to handle these archives. It's manual, slow, costly, and blunder inclined, and information is in many cases spread across unique sources. Subsequently, making and dealing with a record handling pipeline stays a test for some organizations. As indicated by Ritu Jyoti from IDC, "Supporting archive handling requires an AI-local stage that further develops precision, execution, dexterity and adaptability while supporting an expansive arrangement of record types. Fake Processing of Scanned Documents utilizing AWS Textract Dept of IT, SSGMCE, Shegaon Page 8 Intelligence (AI), can assist with smoothing out archive computerization giving better business results, further developed ROI, and decrease manual efforts."[AWS has sent off an answer for assist associations with separating bits of knowledge and robotize handling records of various arrangements (PDF, Word, crude message) and designs (shots, records) utilizing Amazon Comprehend. This new send-off joins the force of regular language handling (NLP) and Optical Character Recognition (OCR) to assist with decreasing how much pre-handling or post-handling expected to deal with records. You can now utilize specially named substance acknowledgment (NER) on greater archive types without expecting to change your records over to crude text. AWS has been advancing in the wise report handling (IDP) space for a really long time to change over information in records into usable data for archive driven processes. AWS sent off AI administrations like Amazon Textract, Amazon Comprehend, and others to assist with the computerization of separating experiences from reports. Since the send off of those administrations, upgrades in precision and speed have been ten times. These administrations offer new APIs like particular help for solicitations and receipts, penmanship and language support, in addition to enhancements in inertness.

AWS Textract is reclassifying the way that organizations interaction archives in a Digital World

Multi-modular transportation is probably the greatest advancement in the coordinated factors industry. There has been an effective coordinated effort across various transportation accomplices in inventory network cargo sending for a long time. Yet, there's as yet an impressive upward of desk work handling for every leg of the outing. A huge number of archives are handled in sea cargo sending alone. Utilizing physical work to deal with these archives (buy orders, solicitations, bills of filling, conveyance receipts, and that's only the tip of the iceberg) is both costly and blunder inclined. We really want to robotize the record handling in the operations business.

Supporting residents through COVID-19: Arizona State University Cloud Innovation Center (CIC)

The Arizona State University Cloud Innovation Center (CIC) assembled an open-source resource for refine the archive handling innovation of Amazon Textract for service bill and driver's permit information extraction. This arrangement was as of late utilized by Wildfire, a state relationship for Community Action Agencies, and Prefix Health Technologies (Prefix), an AWS Partner Network (APN) Technology Partner, to assist with giving help to residents during the COVID-19 pandemic. The Arizona benefits gateway permits COVID-19 affected families to prescreen and apply for help with lease, contract, gas, electric, and water. Candidates can join record pictures to the advantage applications utilizing their cell phone camera. Amazon Textract catches the information from the pictures and populates or confirms the information entered which takes out the requirement for manual confirmation and paces up the handling time. Generally speaking, not entirely settled at the place of passage and assets are credited to the client's record with practically zero deferral. For extra subtleties on the arrangement created read, "A smoothed out, portable first way to deal with administration conveyance for districts and states" where the arrangement they produced for Arizona came about in 49% of the applications to be consequently supported, in this way lessening the time expected to confirm and appropriate assets.

III. Comparative Study

In the section of comparative study, we have done comparison between Amazon Textract and other tools those are providing the same service as Amazon Textract is giving. For this we have considered other tools like Google's Document AI and Azure's Form Recognizer.

3.1 Amazon Textract

Amazon Textract is a Machine Learning (ML) administration that consequently removes text, penmanship, and information from filtered records. It goes past basic optical person acknowledgment (OCR) to distinguish, comprehend, and extricate information from structures and tables. Today, many organizations physically remove information from checked records like PDFs, pictures, tables, and structures, or through basic OCR programming that requires manual arrangement. To beat these manual and costly cycles, Textract utilizes Machine Learning to peruse and handle any sort of record, precisely separating text, penmanship, tables, and different information with no manual exertion. You can rapidly control report handling and follow up on the data removed, whether you're computerizing credits handling or extricating data from solicitations and receipts. Textract can remove the information in only couple of moments rather than hours or days.

Amazon Textract makes it simple to add report text discovery and examination to your applications. Utilizing Amazon Textract clients can:

- Detect typed and handwritten text in a variety of documents, including financial reports, medical records, and tax forms.
- Extract text, forms, and tables from documents, using the Amazon Textract Document Analysis API.
- Process invoices and receipts with the Analyze Expense API.
- Process ID documents such as drivers' licenses and passports issued by U.S. government, using the Analyzes ID API.

Amazon Textract depends on a similar demonstrated, exceptionally versatile, profound learning innovation that was created by Amazon's PC vision researchers to dissect billions of pictures and recordings every day. You needn't bother with any AI skill to utilize it. Amazon Textract incorporates straightforward, simple to-utilize APIs that can dissect picture records and PDF documents. Amazon Textract is continuously gaining from new information, and Amazon is consistently adding new elements to the help.

3.2 Azure Form Recognizer

Sky blue structure recognizer is a cloud-based Azure applied AI administration that utilizations AI models to remove key-esteem matches, text, and tables from your archives. Structure recognizer dissects your structures and archives, extricates text and information, maps field connections as key-esteem matches, and returns an organized JSON yield. You can rapidly come by exact outcomes that are custom-made to your particular substance without unnecessary manual mediation or broad information science mastery. Use structure recognizer to robotize your information handling in applications and work processes, improve information driven systems, and advance record search abilities.

DOI: 10.48175/568

3.3 Google document AI

Google AI mechanizes information handling of archives at scale. It is worked from the times of AI examination of Google, and thusly gives itemized data with respect to a specific report beyond anything that can be described. Google Document AI uses computer vision and optical character recognition (OCR), along with natural language processing (NLP), to create pretrained models for extracting information from the documents. Google's Document AI provides a variety of parsers across industries. Google's Lending Document AI and Procurement Document AI can help organizations process high volumes of documents and optimize the processing time. Document AI also has generic parsers like OCR and form parsers that can be used to provide some structure to the data and easily extract values. These parsers reside in a unified dashboard from where they can be tested by uploading a document directly in the console. Other than giving a nonexclusive report investigation and recovery, Google Document AI likewise upholds explicit arrangements, for example, receipts, solicitations, pay slips, and explicit structures that are much of the time handled in enormous bunches by associations.

3.4 Proposed Work

The main motive of this research paper is to study Amazon Textract and its working. For that we had studied Amazon Textract in detail. We have started it with the question, what is Amazon Textract? So let get started with its answer.

3.5 What is Amazon Textract?

Amazon Textract mechanizes information handling of reports at scale. It is worked with the assistance of AI and consequently gives point by point data in regards to a specific record beyond anything that can be put into words. Other than giving a conventional report examination and recovery, Amazon Textract likewise upholds explicit arrangements, for example, receipts, solicitations, pay slips, and explicit structures that are many times handled in enormous clusters by associations. One might make a beeline for Amazon Textract and try out one of their reports or one of your own to see the nature of extraction. The result will be as JSON, PDF, and word and so on design that could be downloaded and examined.

3.6 What type of Data is Supported by Amazon Textract?

The fundamental objective of Amazon Textract is to separate the text inside pdf/images. This would back off the method involved with looking over structures that require huge human exertion. Other than the message, Amazon Textract additionally figures out where there are line breaks and sentence breaks. This permits further personalization and handling by clients on the ideal result in the wake of recovering the significant data from Amazon Textract. For instance, contingent upon the business/reason for using the administrations, one might perform further information examination or give reactions to structures subsequent to extricating the essential data from a PDF archive. The text could likewise be composed or transcribed, giving greater adaptability which Amazon Textract can deal with. Amazon Textract is also able to scan and analyse handwritten documents. Their machine learning features make it capable of understanding handwritings of people and extract information from it.

3.7 Final Takeaway

Amazon Textract helps businesses to be more efficient as it helps to manage the data without any hassle or errors. But Klearstack's solutions are much more efficient than Amazon Textract. While Textract stores data on the cloud directly, KlearStack provides an option to extract data in excel and therefore, provides flexibility to upload the data wherever you would like to or keep it in an excel file.

IV. Working Methodology

4.1 Key-Value Pairs and Table Extractions

Other than text and outline data about an archive, one most significant component that should be separated from reports is information. Manual information extraction is a monotonous cycle that could be overwhelming and mistake inclined, also the hardships when archives are checked as pictures and not text.

By and large, information isn't put away in passages or sentences, yet in even structures and key-esteem matches (KVPs), which are basically two connected information things, key and worth, where the key is utilized as a special identifier for the worth (i.e., Name: John or Age: 19). In particular, while managing records, for example, shapes, these information types exist more than frequently, and text extraction will just not be sufficient. Furthermore, not at all like tables, KVPs could frequently exist in obscure configurations and are frequently to some extent transcribed in structures. Indeed, even with best-in-class text extraction, it might in any case be challenging to decide KVPs with just text and not considering the elements on paper (e.g, bouncing boxes, lines).

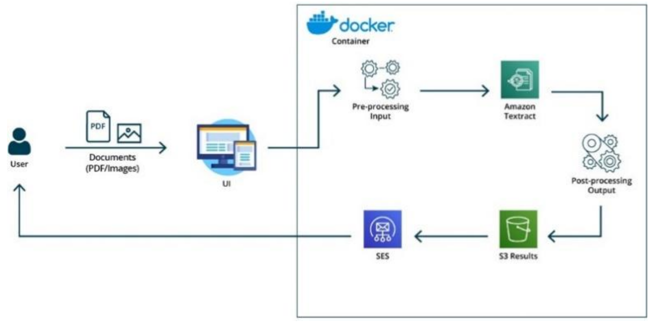

Figure 1.1: Working of Amazon Textract

Step 1: Scan the Document

The first step is to scan the document from which the data has to be extracted. Below is the list of some types of documents, but not limited to, from which data can be extracted:

Customary Invoices and Bills

- Monetary Documents
- Clinical Documents
- Manually written Documents and letters
- Pay slips or Employees Documents

Make sure paper is put in place properly before scanning the document. Amazon Textract may fail to recognize some part of the document if it is left out of the scanning area.

Step 2: Reading the Data

After the document is appropriately placed for scanning, Amazon Textract starts a virtual scan of the document. The tool basically reads the data. This helps to extract and map the data at the later stages. This process is almost instantaneous and happens quite quickly, with respect to the size of the document.

Step 3: Identifying Key Information

Once a thorough scan is done of the document, Amazon Textract automatically identifies key and vital information that has to be extracted and stored. Since it is based on a deep-learning technology, the identification of the information is very accurate.

Step 4: Matching & Data Integration

Using the JavaScript Object Notation (JSON) format, the data is then extracted and stored. JSON is a standard file and data exchange format that helps the human-readable text to be stored on web servers. Since Amazon Textract is a product of Amazon Web Services (AWS), data can be integrated with other AWS products such as Amazon Comprehend, Amazon DynamoDB and so on.

V. Result Analysis: Applications Of Amazon Textract

Utilizations of Amazon Textract are tremendous and exceptionally requested by enterprises and individual clients. Here we generally partition the applications into individual and business use and give a couple of models inside each field.

Individual:

While robotization of report perusing is generally utilized for enormous scope creation to diminish work expenses, quick and exact extraction of information and text can likewise be advantageous for working on private daily practice and association.

- ID-Scans and Data Conversion: Personal IDs and travel papers are frequently put away in PDF records and checked across various sites. They contain different information, especially KVPs (e.g., given name to date of birth), which are frequently required for online applications, however we need to physically find and type in the indistinguishable data over and over. Appropriate information extraction from PDFs can permit us to change over information into machine-reasonable text rapidly. Processes like filling in structures will then, at that point, become inconsequential assignments for a considerable length of time and the main manual endeavors left would be speedy output throughs for twofold checking.
- Receipt Data Extraction: Budgeting is a urgent part of our regular routines. While the improvement of calculation sheets has worked on the assignments as of now, programmed extraction of information actually stays a cycle that, whenever engaged by machines, backs off a significant part of the planning system. Clients can rapidly perform examination in view of the consequences of Google Document AI and decide/observe buys that are unusual or are not reasonable.

5.1 Business

Business companies and huge associations manage great many desks work with comparative arrangements everyday - - Big banks get various indistinguishable applications, and research groups need to investigate heaps of structures to direct measurable examination. Accordingly, robotization of the underlying advance of separating information from reports fundamentally decreases the overt repetitiveness of HR and permits laborers to zero in on investigating information and checking on applications as opposed to entering in data.

- Instalment Reconciliation: Payment Reconciliation is the method involved with contrasting bank articulations against your bookkeeping with ensure the sums are matched accurately. For little firms where their clients and incomes are from less sources and banks, compromise might be genuinely clear. Nonetheless, as organization scale extends, and cash inflows and outpourings become more different, this cycle will before long become overwhelming and work escalated, dramatically expanding the likelihood of blunder. In this manner, various computerized techniques were proposed to ease the pipeline from human endeavours. The underlying phase of instalment compromise is information extraction from records, which can be a difficult issue for an organization with significant size and different areas.
- Measurable Analysis: Feedback from clients, residents, or even test members are expected by partnerships and associations to enhance their item/administration and arranging. To extensively assess input, measurable investigation is frequently required. Notwithstanding, the study information might exist in various configurations or might be concealed between text either composed or written by hand.

VI. Future Scope

Record Analysis and fortify and uphold research. Report gives valuable information that is the reason making record examination helpful and useful strategy for most exploration. (Why archive examination?)

Record investigation is a proficient and viable approach to get-together information since archives are reasonable and common-sense assets. (Benefit of Document Analysis?)

The worldwide record investigation market is supposed to develop from USD 438 million of every 2019 to USD 3,855 million by 2024 at CAGR of 54.5% during the conjecture time frame. (Future of Document Analysis?)

There are 16.3M US mortgage applications from 20161 and the nearly quarter of a billion W2 tax forms expected to be processed in the US in 2018. Same scenario in our country too and in various department and sector. Document Analyzer can be useful in following industries/sectors

1. Banking Sector
2. Public Sector (Insurance and Tax Department)
3. Healthcare and life Science
4. Manufacturing Industries
5. Education Sector
6. Wholesale and Retail Shops

6.1 Opportunities in the Market

- Factors, for example, mix of trend setting innovations with record examination arrangement and expanding need to further develop client experience, offering are the open doors in report investigation market.
- Developing drives to digitalize content across undertakings is one of the central point driving the development of the record examination market.

VII. Conclusion

Amazon Textract enables applications to integrate with SDK APIs so that the documents or images with textual data from various representations of text in form of raw text, forms, tables are easily extractable. Now with the expense analysis support, Textract goes a level ahead to consolidate the items and also extract key information from the invoice or receipts. Textract also provides the confidence level / percentage of the extracted text making it a choice for the integrating applications to either consider it or neglect it.

References

[1]. Jamshed, Memon Maira, Sami, Rizwan Ahmed Khan and Mueen Uddin, “Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR)”, School of Computing, Quest International University Perak, Ipoh 30250, Malaysia, Department of Computer Science, Shaheed Zulfiqar Ali Bhutto Institute of Science and Technology, Karachi 75600, Pakistan. Faculty of IT, Barrett Hodgson University, Karachi 74900, Pakistan. Department of Software Engineering, Faculty of Science and Technology, Ilma University, Karachi 75190, Pakistan.

[2]. Rishabh Mittal, “Text extraction using OCR: A systematic review”, Department of Computer Science and Engineering, Amity school for Engineering and Technology, Amity university Uttar Pradesh, Noida (UP), India.

[3]. Thomas Hegghammer, “OCR with Tesseract, Amazon Textract, and Google Document AI: a benchmarking experiment”