Automating PDF data extractions.

MTEC uses Adobe PDF Extract API to improve speed and accuracy of automatic text extraction from financial data PDF files.

MTEC
MUFG 三菱UFJトラスト投資工学研究所

Established
1988
Employees: 45
Marunouchi, Chiyoda-ku, Tokyo
www.mtec-institute.co.jp/en/

Speeds up analysis and verification cycles

Products:
Adobe Acrobat Services
Adobe PDF Services API

Objectives

Automatically recognize awkward link breaks in sentences

Use data extraction that recognizes text styles and images in PDFs

Support further business growth

Decrease time to conduct PDF text extractions

Results

Enables **extraction of highly accurate sentences** rather than words with pre-OCR

Maintains document structure to enable analysis that captures meaning of sentences

Conducts more accurate surveys and expands scope of business

Accelerates analysis and verification cycle

Established in 1988, Mitsubishi UFJ Trust Investment Technology Institute (MTEC) provides asset management, risk management, data analytics, and data analysis consulting services primarily to its parent company, Mitsubishi UFJ Trust and Banking Corporation, and its group companies. In 2022, the company began utilizing the knowledge and expertise it had accumulated over many years to provide investment advisory services to users beyond the group's framework.

The scope of data science analysis is expanding to include unstructured data such as natural language. In this context, MTEC, which integrates mathematical science and information science to solve problems in financial operations, has focused on integrated reports of listed companies as one of its new targets for analysis. The company's efforts to adopt **Adobe PDF Extract API** as a tool for extracting text data from PDF files have contributed to improving the speed of the analysis and validation cycle, by completing the extraction of text data from integrated reports at high speed.

"When we talked to Microsoft about our need to convert documents to PDF quickly and accurately, Microsoft recommended Adobe PDF Services API."

Mr. Yusuke Naritomi
Financial Engineer, Development Group 2, Research Department, MTEC

Continued evolution of financial data science: analysis of natural language

MTEC focuses on analyzing market data, financial data, and other numerical data that affect the movement of stock market prices. It provides its parent company, Mitsubishi UFJ Trust and Banking Corporation, and its group companies with the mathematical models needed to make investment and financing decisions. The institute has been an early adopter in using text in financial statements and other documents in data analysis.

"It is becoming increasingly difficult to conduct more accurate analysis using conventional methods that only follow numerical data. There is a strong demand in the field of financial engineering for analysis that includes market psychology, in addition to numerical data," explains Yusuke Naritomi, a Financial Engineer in Development Group 2 of the Research Department. "Text extraction that preserves sentence structure is of great significance in the world of financial data science."

Under these circumstances, a new issue emerged. How can the text data of timely disclosure information and a variety of reports be extracted from PDF files with high accuracy and efficiency?

"In the past, we used free software to extract text information from timely disclosure information in PDF files. However, problems such as the incorrect interpretation of character strings at line breaks made it difficult to maintain the sentence structure of the extracted text," explains Mr. Naritomi. "That is fine if you simply want to pick out words contained in the text and quantify how frequently they occur. However, that method of analysis does not include the meaning of the text. In terms of improving our service quality, maintaining the sentence structure of text data extracted from PDF files has become important."

Using Adobe PDF Extract API to maintain sentence structure

Amid the challenges of extracting text data from PDF files, MTEC launched a new project to evaluate Environment, Social, and Governance (ESG) issues. ESG is attracting attention as it relates to a company's long-term growth. To objectively evaluate ESG, it is important to understand a company's initiatives, rather than just numerical figures. There is a focus on integrated reports, which add non-financial information such as corporate governance, corporate social responsibility (CSR), and intellectual property to a company's financial information. However, in order to fully understand the content of dozens of pages in an integrated report, maintaining the sentence structure of extracted text is essential. While researching ways to address this, Mr. Naritomi came across the **Adobe PDF Extract API**.

"I discovered a blog post in English and became interested after seeing the accuracy of the extracted sentence presented there as an example," says Mr. Naritomi. "There are several tools that recognize text in PDF files, but this was the only one that claimed to maintain the sentence structure. After consulting with Adobe, we decided to conduct an enterprise trial to verify the performance of the Adobe PDF Extract API."

"We found that Adobe Acrobat DC OCR and Adobe PDF Extract API resulted in the highest quality work."

Mr. Yusuke Naritomi
Financial Engineer, Development Group 2, Research Department, MTEC

Improving speed of the analysis and verification cycle

An integrated report issued by a company listed on the former First Section of the Tokyo Stock Exchange was used for this trial.

Masahiro Shimizu, Senior Financial Engineer in Development Group 1 of the Research Department, explains the aim of the project. "Information that is disseminated by ESG-related companies tends to be filled with similar words and phrases. Therefore, it is important for the text to be extracted in a manner that preserves the sentence structure and preserves the meaning of sentences and paragraphs to highlight the differences in the initiatives pursued by each company. Therefore, the most important point was to evaluate the ability of this tool to maintain sentence structure while extracting the text information from dozens of pages in an integrated report."

The Adobe PDF Extract API uses Adobe Sensei, the proprietary AI and machine learning engine from Adobe.

"Consistent and high-quality PDF conversion is very important to scaling text extraction," says Mr. Naritomi. "We found that Adobe Acrobat DC OCR and Adobe PDF Extract API resulted in the highest quality work. We took the PDF files in Amazon S3 that we wanted to convert and placed them in a separate folder, then we used Acrobat DC to convert the text in the PDF. Then we use the Adobe PDF Extract API to extract the text and output it as a JSON file. The text extraction process was fast, including the time required for OCR. This contributed to speeding up the analysis and verification cycle."

The data that is output as a JSON file requires a minimal amount of organization by Mr. Naritomi before being passed on to a researcher.

"In the past, we had to separate some sentences and splice others together to decipher the meaning of the text data extracted from PDF files," says Mr. Shimizu. "The Adobe PDF Extract API not only eliminates the need for such work, but also offers unprecedented features such as the ability to identify headings and body text. Another use is to determine the extent to which the 17 SDGs are present based on the text of an integrated report, and to determine from that where each company is focused related to the SDGs, or corporate materiality. By comparing these results with other companies in the same industry, it is possible to measure a company's level of focus on any particular issue."

"The Adobe PDF Extract API...offers unprecedented features such as the ability to identify headings and body text."

Mr. Masahiro Shimizu
Senior Financial Engineer, Development Group 1, Research Department, MTEC

Extracting text from a variety of PDF files, not just integrated reports

MTEC is currently working on automating the text extraction process with the Adobe PDF Extract API. It has also started using the tool for PDF files other than integrated reports.

"In addition to integrated reports, we are currently attempting to extract text from the Timely Disclosure network (TDnet) service provided by the Japan Exchange Group (JPX)," says Mr. Naritomi. "We are also looking into text extraction for a variety of reports, such as CSR reports, from approximately 4,000 companies listed on the JPX."

The company is also considering expanding these services to its parent company, Mitsubishi UFJ Trust and Banking Corporation, and other group companies.

"In the case of trust banks, many operations have documents that have not been converted to data," says Mr. Shimizu. "We believe that the conversion of such text into data is also meaningful from the perspective of improving operational efficiency."

Adobe Acrobat Services licenses are available for the Adobe PDF Extract API, as well as for the Document Generation API and the PDF Services API. MTEC is considering the use of Adobe Acrobat Services in the future for automatic document generation.

** This information is current as of September 2022.*