Adobe PDF Extract APIにより PDFをテキスト自動抽出するスピードと精度が向上

MTEC
MUFG 三菱UFJトラスト投資工学研究所

「文章構造を維持したテキスト抽出の実現は、金融データサイエンスの世界でも大きな意義があると考えています」

研究部 開発第2グループ フィナンシャルエンジニア 成富 佑輔氏

導入製品

- Adobe PDF Extract API

導入メリット

文章の泣き別れを自動認識

事前のOCRで、単語ではなく
精度の高い文章としての抽出を実現

PDF内の文章スタイルや画像も判別

文書構造を維持したデータ抽出により
文意を含めた分析が可能に

事業のさらなる成長に寄与

より精度の高い調査が可能となり
業務の幅が拡大する可能性も

PDFのテキスト抽出を短期間で実現

分析・検証サイクルの
スピード向上に貢献

三菱UFJトラスト投資工学研究所

1988年設立の三菱UFJトラスト投資工学研究所（MTEC）は、主に親会社である三菱UFJ信託銀行やそのグループ企業に資産運用、リスク管理、データアナリティクス、データ分析コンサルティングなどのサービスを提供。2022年からはこれまで培ってきた知見やノウハウを生かし、グループの枠組みを越えたユーザーへの投資助言業務を開始している。

研究部 開発第2グループ
フィナンシャルエンジニア
成富 佑輔氏

データサイエンスの分析対象は、自然言語などの非構造化データに広がっている。こうした中、数理科学・情報科学の融合により、金融業務における課題解決に取り組む株式会社三菱UFJトラスト投資工学研究所は、新たな分析対象の一つに上場企業の統合報告書に注目。PDFのテキストデータ抽出ツールとしてAdobe PDF Extract APIを採用した同社の取り組みは、統合報告書のテキストデータ抽出を高速に終え、分析・検証サイクルのスピード向上に貢献している。

■ 導入の経緯

自然言語までを分析対象に進化を続ける金融データサイエンス

株式市場の値動きに影響のある市場データや財務データ等の数値データ分析に軸足を置き、親会社である三菱UFJ信託銀行やそのグループ企業に投資・融資判断をする際に必要となる数理モデルを提供してきた株式会社三菱UFJトラスト投資工学研究所（以下、MTEC）は、決算短信などに含まれる文章をデータ分析に活用する取り組みを早くから開始している。その背景を、研究部 開発第2グループ フィナンシャルエンジニアの成富 佑輔氏はこう説明する。

「やはり数値データだけを追う従来の分析方法だけでは、より精度が高い分析は難しくなっているため、金融工学の分野でも数値データ以外の市場心理まで含めた分析が強く求められるようになっています」

こうした中新たに浮上したのが、PDF 形式で配布される適時開示情報や各種報告書のテキストデータを、どうすれば高精度かつ効率的に読み取れるかという課題だった。

「以前はフリーソフトを使って適時開示情報のPDFファイルに含まれるテキスト情報を抽出していたのですが、改行時に文字列を正しく読み込まないなどの問題があり、文章構造を維持したままテキストを抽出することは困難でした。単純にテキストに含まれる単語を拾い出し、その出現頻度を数値化するような分析手法であればそれでも構わないのですが、これでは文意まで含めた分析は行えません。PDFファイルからの文章構造を維持したテキストデータ抽出は、私たちのサービス品質の向上という観点からも重要な課題になっていました」（成富氏）

研究部 開発第1グループ
主任フィナンシャルエンジニア
清水 正大氏

USER PROFILE

三菱UFJトラスト投資工学研究所

https://www.mtec-institute.co.jp/

所在地：東京都千代田区丸の内1-4-5 三菱UFJ信託銀行本店ビル22階

設 立：1988年1月14日

資本金：4億8,000万円

従業員数：45名（2022年4月1日現在）

事業内容：資産運用、リスク管理、データアナリティクス、データ分析コンサルティング、投資助言

詳細情報

adobe.com/go/dcsdk_home

Adobe

アドビ株式会社
〒141-0032 東京都品川区大崎1-11-2
ゲートシティ大崎イーストタワー
www.adobe.com/jp/

Adobe
345 Park Avenue
San Jose, CA 95110-2704
USA
www.adobe.com

■ 選択のポイント

文章構造の維持を唯一謳うAdobe PDF Extract APIのトライアルを実施

PDFファイルからのテキストデータ抽出に課題を抱える中、MTECにおいて新たに立ち上がったのが、Environment（環境）、Social（社会）、Governance（ガバナンス）の頭文字をとったESG銘柄の評価に関するプロジェクトだった。企業の長期的な成長という観点からも注目されるESGだが、その客観的評価には、数値だけに留まらない、企業の取り組みを知ることに大きな意味を持つ。そこで注目したのは、企業の財務情報に企業統治や社会的責任（CSR）、知的財産などの非財務情報を加えた統合報告書だった。だが、数十ページに及ぶ統合報告書の内容を正しく把握するには、文章構造を維持したテキスト抽出が避けて通れない課題だった。成富氏が情報収集する中で出会ったのは、リリースして間もないAdobe PDF Extract APIだった。

「英語のブログ記事を発見し、そこで紹介されていた例文の抽出精度を見て興味を持ちました。PDFのテキスト認識ツールは複数ありますが、それらの中で唯一文章構造の維持を謳っていたこともあり、アドビさんに相談してAdobe PDF Extract APIのエンタープライズトライアルによる検証を行うことにしました」（成富氏）

■ 導入効果

分析・検証サイクルのスピード向上に貢献

トライアルの対象になったのは、旧東証一部上場企業が発行した統合報告書である。研究部 開発第1グループ 主任フィナンシャルエンジニアの清水 正大氏はその狙いをこう説明する。

「ESG関連企業の情報発信は同じような言葉が並びがちです。それだけに、各社の取り組みの差異を浮かび上がらせるためには、文章構造を維持した形でテキストを取り出し、センテンスやパラグラフの単位で文意を抽出することが重要となります。そのため、数十ページに及ぶ統合報告書のテキスト情報が文章構造を維持した状態で抽出できることが最大の評価ポイントになりました」

アドビ独自のAI・機械学習エンジンAdobe Senseiを利用したAdobe PDF Extract API は、2021年夏の英語版から提供を開始した新しいサービスだ。特に日本語版の場合、MTEC をはじめとするユーザーの利用に基づき、精度向上を図る段階にある。こうした中、MTECはトライアンドエラーを経て、PDFデータをOCR化するなどの独自の運用プロセスを構築し、高精度なテキスト抽出を実現した。

「当初の問題はPDF作成時の仕様に関連する文字化けでしたが、その多くは一度OCR化した上でAdobe PDF Extract APIに送ることで解決できています。当社ではAmazon S3にアップロードしたPDFをAcrobat ProでOCR化し、Adobe PDF Extract APIでテキストを抽出してJSONファイルで出力するという流れで運用しています。OCRは、用意した別フォルダにPDFを入れ、Acrobat Proでフォルダ指定を行うという手順で行っています。今回のテキスト抽出は、OCR化に要した時間を含め高速に処理し、分析・検証サイクルのスピード向上に貢献しています」（成富氏）

JSONファイルで出力されたデータは、成富氏が最小限の整理を行った上で研究員に渡される。

「従来は、PDFファイルから抽出されたテキストデータから文意を読み解くには、文章を区切ったり、つなぎ合わせたりする作業が必要でした。Adobe PDF Extract APIが抽出するテキストデータはこうした作業が不要になるだけでなく、見出しと本文が分かるなどこれまでにない特長を備えています。他にも、統合報告書にSDGsの17個のテーマがどの程度存在するかを文章ベースで判別し、そこから各企業がSDGsのテーマのどこに注力しているか、いわゆる企業のマテリアリティを判別することにも活用しています。さらに同業他社と比較することで、その企業のテーマに対する注力度を測ること等も可能です」（清水氏）

■ 今後の展望

統合報告書だけでなく、多様なPDFファイルからのテキスト抽出に活用

MTECが現在取り組んでいるのは、Adobe PDF Extract APIによるテキスト抽出プロセスの自動化だ。また、統合報告書以外のPDFファイルへの応用も既にスタートしている。

「統合報告書以外では、現在JPX（日本取引所グループ）が提供するTDnet（適時開示情報閲覧サービス）のテキスト抽出を試みています。また、JPXに上場する約4000社のCSRレポートなど各種報告書のテキスト抽出も現在検討しています」（成富氏）

さらに、親会社である三菱UFJ信託銀行をはじめとするグループ企業へのサービス展開も検討中だ。

「信託銀行の場合、多くの業務でデータ化されていないドキュメントが存在します。こうしたテキストのデータ化は、業務効率の向上という観点からも意義があると考えています」（清水氏）

Adobe Document Serviceのライセンスは、Adobe PDF Extract APIのほか、Document Generation APIやPDF Services APIの利用も可能だ。同社は今後、文書の自動生成などにもAdobe Document Serviceの活用を検討しているという。

※掲載された情報は、2022年9月取材現在のものです。