# AIツールとAI人材の育成

AI Singaporeは、Adobe PDF Extract APIを使用して深層学習モデルを加速しています。

AI SINGAPORE

創業
**2017年**
従業員数　90名
Singapore
https://aisingapore.org

**40%**
機械学習（ML）モデルを提供するために必要なスプリントの削減

**導入製品**

Adobe Acrobat Services ›
Adobe PDF Extract API ›

## 課題

自然言語のパイプライン処理によって取り込むデータの品質向上

企業の持続可能性報告書（サステナビリティレポート）に必要なコンテンツの提供

10スプリントかつ7か月以内に、最小限の実用的な機械学習モデル開発

## 成果

異なる500のドキュメントのコンセプトを証明するために、**高精度**でPDFからデータを抽出

PDF Extract APIを**わずか2週間**で実装

より優れたモデル結果を得るために、**より良いコンテキストと構造を持つデータ**を提供

計画よりも**40%早く**展開可能なモデルを提供

シンガポールでは、ビジネスの課題解決のために人工知能（AI）を活用したいと考える企業は、AI Singaporeに相談をします。AI を活用して将来のシンガポールのデジタル経済を強化するために立ち上げられたこの国家プログラムは、シンガポールに拠点を置くすべての研究機関と、AI製品を開発するスタートアップや企業のエコシステムを結集し、AIへの取り組みを強化するために、実利用を想定した研究を行い、知識を深め、ツールを開発し、AI取り組みを支える人材を育成しています。

Siavash Sakhaviは、シンガポールのAIイノベーションを促進するために設立された「100 Experimentsプログラム」の副代表です。この組織は、既成のAIソリューションが存在しない領域において問題を提起し、シンガポールの研究者のエコシステムとAISGのエンジニアリングチームによって、9〜18ヶ月以内に解決することを目指しています。

Sakhaviのチームは、大規模な多国籍金融サービスクライアントから持続可能性報告書プロジェクトに関する依頼を受けました。クライアントは、さまざまな情報源からPDF形式で提供される多様なレポートやパンフレットからテキストを抽出するのに苦労していました。プロジェクトチームは複数のAI・データ・プラットフォームエンジニア、およびAISGの実習生からなり、抽出された情報を自然言語のパイプライン処理に入力したいと考えていましたが、当時使用していたPDF抽出ツールによって大量の非構造化の意味不明なテキストが返されており、自然言語のパイプライン処理が期待通りに機能していないことに気付きました。

通常のプロジェクトでは、チームは1〜2か月でモデルを開発することができます。しかし、このときはすでに10スプリントのうちの6スプリント目に入っており、結果を出すためのプレッシャーが高まっていました。幸いなことに、彼らは当時ベータ版から一般提供に移行していたAdobe PDF Extract APIについて知りました。アドビの新しいWebサービスであるPDF Extract APIは、ネイティブおよびスキャンされたPDFファイルからデータとコンテキストを解析し、構造化されたJSONファイル内のテキスト、表、および画像要素を抽出できます。

---

**「Adobe PDF Extract API はまさに救世主でした。
これがなければ、さまざまなデータを使った自然言語処理ソリューションを
計画通りに構築することは難しかったでしょう。」**

Siavash Sakhavi
Assistant Head, 100E, AI Singapore

---

## より良い構造とコンテンツ抽出の実現

「1回目のデモで、私たちはそれまで使っていたPDF抽出ツールからはAdobe PDF Extract APIに完全に切り替えることを決めました。」とSakhaviは述べています。スプリントの終わりまでに、チームは大きな進歩を遂げていました。自然言語のパイプライン処理への迅速なデータ取り込みを実現し、最終的にはプロジェクトのスポンサーに約束された作業を予定よりも早く提供することができました。

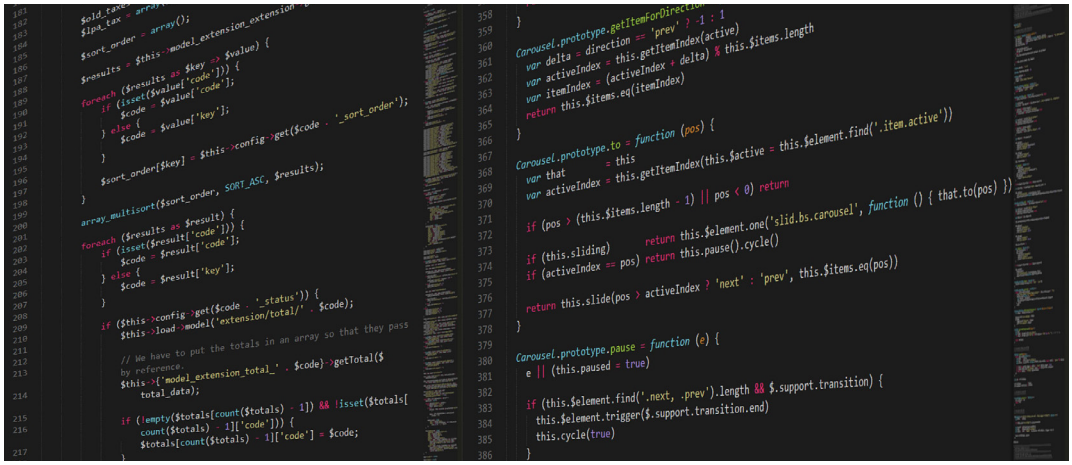
「Adobe PDF Extract APIはまさに救世主でした。これがなければ、さまざまなデータを使った自然言語処理ソリューションを計画通りに構築することは難しかったでしょう。」とSakhaviは述べています。「それまで、機械学習モデルがどのように機能すべきかを構築し、改良してきました。しかし、PDF抽出ツールでの抽出結果がボトルネックになっていたのです。」

チームがテストしていたオープンソースのPDF抽出ツールは、段落を正確に識別することができませんでした。大半のテキストは文章の途中で途切れているなど活用が出来ず、また図表の数字やラベルが本文として誤って抽出されるなど、文章の構造情報が完全に欠落していました。

「これらのテキストから必要な情報を手動で抽出し、最終的にそれらを分類するのは簡単な仕事ではありません。自動化と効率性を提供するより良いPDF抽出ツールを見つける必要がありました。幸いなことに、私たちは自然言

語処理量を急速に増やし始めるタイミングでPDF Extract APIを組み込むことができました。」Sakhaviは述べています。

PDF Extract APIは、単なる文や断片ではなく、段落に基づいて出力できます。「段落をグループ化するコンテキストデータと機能は非常に貴重でした。これにより、自然言語のパイプライン処理が大幅に強化され、機械学習アルゴリズムでもより良い結果が得られるようになりました。」とSakhaviは述べています。



## 上位のモデル結果のための関連性向上

Sakhaviは次のように述べています。「私たちの結果は驚くべきものでした。PDF Extract APIは非常にうまく組み込まれました。私たちは、必要なものだけを抽出するためのアダプタを作成しました。APIは驚くほど正確であり、プロジェクトスケジュールを加速させました。」

AISGチームは、Bidirectional Encoder Representations from Transformers（BERT）の深層学習モデルを開発することを計画していました。このモデルにデータを供給するために、プロジェクトのスポンサーは、特定の企業ESG（環境・社会・ガバナンス）イニシアチブに重要な要因を示すキーワード用語集の定義を提供しました。AISGチームの目標は、これらの定義と分析対象ドキュメントの間で類似性マッチングプロセスを実行し、テキストのどの部分がプロジェクトに関連するかを特定することでした。

「これらは環境持続可能性に固有のコンテキストだったので、用語集の情報と一致する関連テキストを抽出するまでモデルをトレーニングすることができませんでしたが、PDF Extract APIは環境持続可能性に関連するトピックに関する段落の文脈で重要な文レベルの情報を正しく識別することができ、高品質なデータの取り込みを実現しました。」とSakhaviは述べています。

---

**「私たちの結果は驚くべきものでした。**
**PDF Extract API は非常にうまく組み込まれました。API は驚くほど正確であり、**
**プロジェクトスケジュールを加速させました。」**

Siavash Sakhavi
Assistant Head, 100E, AI Singapore

---

企業のスポンサーは、様々なソースからの500もの多様な文書を分析したがっており、Sakhaviのチームに最初の10文書を処理し、コンセプトを証明した後、残りの490文書を処理できる展開可能なモデルを提供するよう求めていました。「関係するレポートやパンフレットは、どれも簡単ではありませんでした。テキスト要素の他に、ページ

の各所に画像がたくさんあり、サイズや内容も様々でした」と彼は述べています。Sakhaviと彼のチームは、たった2週間でPDF Extract APIを実装し、データ品質の大幅な向上を実現しました。「学習曲線は非常に短かったです」と彼は述べています。チームはすぐに、データのリストをパイプラインに組み込み、パス属性に基づいてすべてを解析して、望ましい結果を得る方法を学びました。

「スプリント6の終わりには、私たちは有望な結果をクライアントに提示することができ、チームのさらなる作業によって本番環境のソリューションを実現することができました。」と、Sakhaviは述べています。

## Innovative Singaporean AI ecosystemの育成

プロジェクトの範囲はその後、新しい領域の研究調査に拡張され、AISGのAIエンジニアリングチームと連携して作業しているシンガポールの大学に引き継がれました。企業のスポンサーにとっての良いニュースは、AISGプロジェクトチームで働いていた同じ人々が、引き続きこのイニシアチブを実現させるために大学研究チームに参加し、最終的にはラベル付きデータを使用して新しいモデルをトレーニングし、優れた結果を出していることです。

Sakhaviは次のように述べています。「AISGにはPDF解析が必要なプロジェクトが多数あります。私からすべてのチームへの推奨事項は、PDF Extract APIを活用することであり、スポンサー企業にも継続的な使用を提案することです。」