

# GPT APIで作る！AIクローラー＆AI-OCR入門

情報収集と文字認識の自動化で業務効率アップ



## 現場のよくある困りごと

- 大量の紙資料の手入力
- 同じWebサイトからの繰り返しコピー
- 単純作業による時間のロス

## 前 手動作業の現状

- コピー＆ペーストの繰り返し
- 紙資料からの手入力
- 単調な作業でミスが発生



## 後 GPT APIで自動化

- Web情報を自動収集・整形
- 紙資料をAI-OCRでデジタル化
- 作業時間半減＆ミス削減

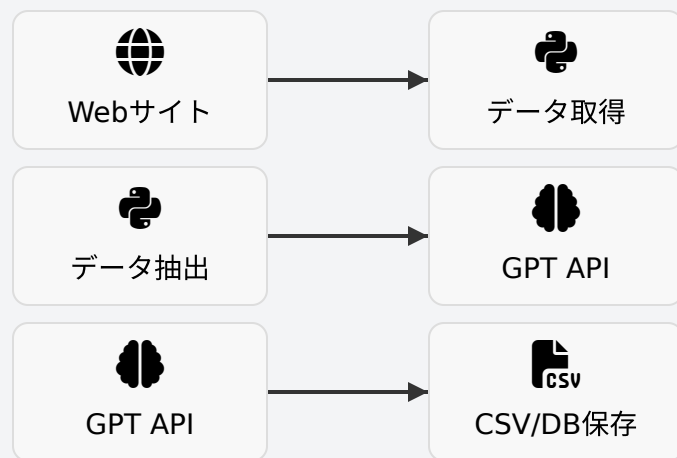
## 安心ポイント

- ✓ 専門知識がなくても始められる
- ✓ サンプルコード付きでわかりやすい
- ✓ 段階的に改善していける

# Webクローラー概要（GPT API活用）

Webクローラーとは指定したWebサイトから自動的に情報を収集する仕組みです。GPT APIと組み合わせれば、要約や抽出、整形まですべて自動化が可能！

## 処理フロー



## 現場課題例

- 毎日の価格調査・競合サイトチェック
- 求人情報の定期収集
- ニュース記事の要約と整理

## ✂ 構成例

`</>` **Python:** Requests/BeautifulSoup

+ **API連携:** OpenAI API (GPT-4/GPT-3.5)

**出力形式:** CSV/JSON/データベース

## 💡 活用例

ニュース収集

競合価格チェック

求人情報抽出

SNS分析

## ✔ メリット

**時間短縮:** 手動作業時間の80%削減

**自動整形:** 不要部分の除去と文章整理

**分析強化:** データの要約・構造化が可能

# Google Colabでの開発環境構築

Google Colaboratory（Colab）は、ブラウザからPythonを実行できる無料サービスです。環境構築不要で、機械学習用ライブラリも標準搭載されています。

## 1 環境準備

- ➡ Googleアカウントでログイン
- 🔗 colab.research.google.com  
へアクセス
- ➕ 「ノートブックを新規作成」を  
クリック



## 2 コードセルの操作

- ⌂ セル内にPythonコードを入力
- ▶ 実行ボタンまたはShift+Enterで  
実行
- ⬇ 実行結果がセル下に表示される



## 3 テキストセルの使用

- A テキストセルボタンをクリック
- 📝 マークダウン記法でメモを作成
- ☑ セル外クリックで表示確定

## 🖥️ Google Colab画面例

📄 **untitled.ipynb**

ファイル 編集 表示 挿入 ランタイム

```
[ ] import numpy as np
import pandas as pd

# データの作成
data = np.random.randn(100, 4)
df = pd.DataFrame(data, columns=['A', 'B', 'C', 'D'])
df.head()
```

|   | A      | B      | C      | D      |
|---|--------|--------|--------|--------|
| 0 | 0.324  | -1.043 | 0.762  | -0.132 |
| 1 | -0.845 | 0.523  | -0.421 | 1.346  |

```
[ ] # 次のコードセル（空の状態）
```

## 🔧 GPUの設定方法

➤ 「ランタイム」→「ランタイムのタイプを変更」

ハードウェアアクセラレータ

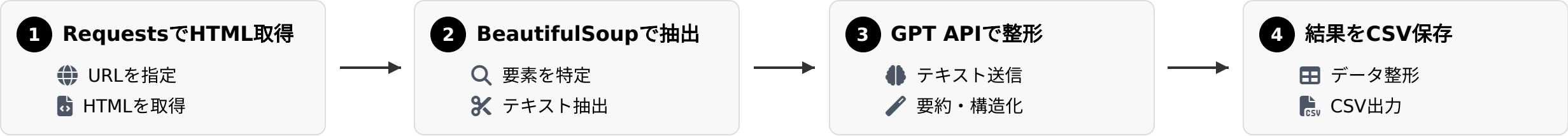
- ☒ None
- ☐ GPU（機械学習に推奨）
- ☐ TPU

### Google Colabのメリット

- 環境構築が不要で、すぐに開発開始可能
- NumPy、pandas、Tensorflowなど主要ライブラリが標準搭載
- 基本無料でGPUが使える
- コードとメモを一体管理できる
- Googleドライブと連携が簡単

# Python+GPT APIによるクローラー実装例

実際にPythonでWebページを取得し、抽出したテキストをGPT APIに渡して、CSVとして保存するまでのサンプルフロー。



## </> サンプルコード

```
import requests
from bs4 import BeautifulSoup
import csv
from openai import OpenAI

# 1. RequestsでHTML取得
url = "https://example.com/news"
response = requests.get(url)
html = response.text
```

## 📄 出力結果 (CSV)

| タイトル       | カテゴリ | 要約            |
|------------|------|---------------|
| AIの進化が加速   | 技術   | 新たなAI技術が業界に変革 |
| 市場予測2025年版 | 経済   | デジタル分野の成長続く   |
| データ活用の新手法  | 技術   | 効率性が30%向上     |

✔️ 自動生成完了

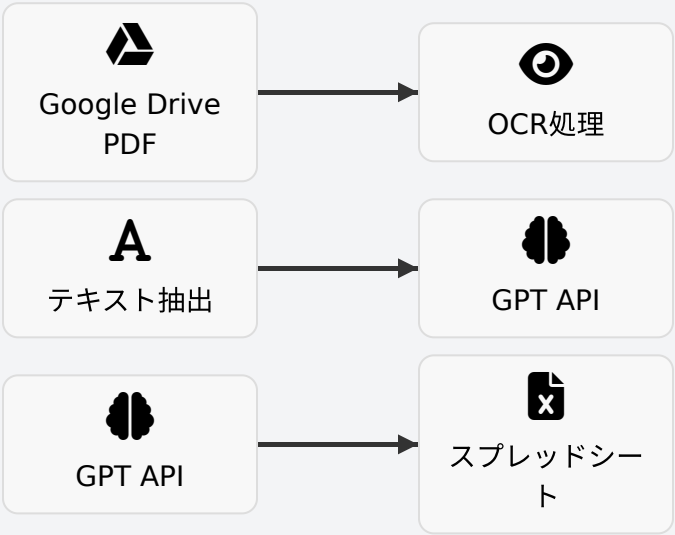
### ポイント

- GPT APIへの適切な指示が重要
- 定期実行でデータを自動収集可能
- 出力形式はJSON/DBにも変更可能

# Google Drive PDF + GPT API → スプレッドシート転記

Google Driveに保存したPDFファイルの内容をGPT APIで解析し、自動でスプレッドシートに転記するワークフローを構築できます。複数のPDFも一括処理が可能です。

## 処理フロー



## 現場課題例

- 請求書PDFから項目ごとの金額を手入力
- 契約書PDFから重要条項を抜き出す作業
- 複数のPDFレポートからデータを集計

## API構成

- Google Drive API:** PDFファイルアクセス
- Google Cloud Vision API:** PDF文字認識
- OpenAI API:** テキスト解析・構造化

## 活用例

- 請求書処理
- 契約書管理
- レポートデータ集計
- 多言語PDF翻訳

## メリット

- 処理時間:** 手動転記比で最大95%削減
- 一元管理:** Drive上のPDFを自動集計
- 自動化:** 定期実行で最新状態を維持

# Google Drive PDF → GPT API → スプレッドシート連携

Google Driveに保存されたPDFファイルをPythonで読み取り、GPT APIを使って内容を構造化し、スプレッドシートに転記する自動化フローを実装します。

## 1 Google DriveからPDF読取

- Google Drive APIで認証
- PDFファイルをダウンロード
- PyPDF2でテキスト抽出

## 2 GPT APIで内容解析

- 抽出テキストを送信
- 内容の意味を理解・分類
- 構造化データに変換

## 3 スプレッドシート出力

- Google Sheets APIで認証
- データをセルに書き込み
- 自動フォーマット適用

### </> サンプルコード

```
import os
import io
from googleapiclient.discovery import build
from googleapiclient.http import MediaIoBaseDownload
from google.oauth2 import service_account
import PyPDF2
from openai import OpenAI
import gspread
import json
```

### スプレッドシート出力例

| Google スプレッドシート |            |                 |            |          |    |          |
|-----------------|------------|-----------------|------------|----------|----|----------|
| 請求書番号           | 請求日        | 取引先名            | 商品/サービス    | 単価       | 数量 | 金額       |
| INV-20250815    | 2025/08/15 | 株式会社 テックソリューション | AIコンサルティング | 50,000円  | 2  | 100,000円 |
|                 |            |                 | データ分析      | 30,000円  | 1  | 30,000円  |
|                 |            |                 | システム構築     | 120,000円 | 1  | 120,000円 |
| 合計金額（税込）        |            |                 |            |          |    | 275,000円 |

#### 活用のポイント

- 複数の請求書PDFを一括処理可能
- GPTプロンプトを調整して抽出項目をカスタマイズ
- スプレッドシートでそのまま集計・分析が可能
- 定期実行で月次請求書の自動処理も実現

# まとめ・応用例

## 今日のまとめ



### Webクローラー × GPT API

Webサイトから自動収集したデータをGPT APIで整理・要約し、有益な情報に変換



### AI-OCR × GPT API

紙資料や画像からOCRで抽出したテキストをGPT APIで構造化・校正

## 効率化の流れ



収集・スキャン



AI処理・変換



活用・保存

時間削減効果：最大80%の作業時間短縮

## 業務応用例



### 営業リスト自動作成

顧客情報サイトから潜在顧客データを自動収集し、営業優先度を付けて整理



### 契約書デジタル化

紙の契約書をOCRでデジタル化し、GPT APIで重要条項を抽出・整理



### 競合価格モニタリング

競合他社の価格情報を自動収集し、価格変動を分析してレポート化



### 問い合わせ自動処理

顧客からのメール・問い合わせを自動分類し、回答案や優先度を付与



### 発展的活用

- 複数のAIツールを組み合わせでワークフローを自動化
- 定期的なデータ更新で常に最新情報を維持
- 単純作業から解放され、創造的業務に集中

# ワーク：自分の業務に置き換えてみよう

今日学んだGPT API活用法を、自分の仕事のどの作業に応用できるかを考えてみましょう。小さな一歩が効率化の始まりです。

## 1 繰り返し作業の洗い出し

日々の業務で繰り返し行っている作業を3つ書き出してみましょう

例：製品価格の定期チェック

## 2 自動化の方針検討

左で選んだ作業から一つを選び、どのように自動化できるか考えましょう

☐ Webクローラー×GPT API    ☐ OCR×GPT API



自動化の実装の流れ、具体的なアクションプランを書いてみよう。

### ヒント

- 今日学んだ技術は小規模な実装からスタートできます
- 既存のAPI連携サービスを活用すると導入のハードルが下がります
- まずは単一業務の部分的な自動化から始めてみましょう