



データシート・キャプチャ・WEB サイトの情報源

情報はすべて Web サイトにあるのに、 誰が書き写すのですか？ Web サイトが、エージェントの知識に なります。

Web サイトの情報源は、お客様の Web サイトの公開 HTTPS ページを、スケジュールに従って読み込みます。各ページは、選択したナレッジベースの内部記事になり、エージェントは電話と Web サイトのチャットでそれをもとに回答します。

概要

情報源ごとに最大 2000 ページ、リンクの深さは最大 6、更新は 6 時間ごとから毎月までの間で設定できます。変更されたページは書き直され、消えたページは二回の実行後に取り下げられます。

01・概要

同じ答えが二か所にある。 これからは Web サイトが基準になります。

営業時間、製品ページ、ヘルプの文章は、ずっと前から Web サイトにあります。エージェント用にそれを書き写すのは時間がかかり、Web サイトを変更するたびにコピーが古くなります。Web サイトの情報源はページそのものを読み、次の実行で変更を反映します。

入力

https の開始 URL。任意でサイトマップと、同じホスト上の追加の開始 URL

結果

ページごとに一つの内部記事。ページ URL を出典とし、検出された言語を付与

スケジュール

6 時間ごと、12 時間ごと、毎日、毎週、14 日ごと、毎月

読者

電話と Web サイトのチャットの会話エージェント、Agent AI のチャットエージェント



一つのホスト、多くのページ

開始 URL から、www の有無にかかわらず、同じホスト上のリンクをたどります。robots.txt のサイトマップや、独自のサイトマップで一覧を広げられます。

範囲を設定

情報源ごとに、ページ数の上限とリンクの深さを設定できます。さらに、パスの接頭辞、除外する glob パターン、クエリ文字列付きの URL、/tag のような一覧パスを指定できます。

JavaScript にも対応

Fast は HTML を読み、Full はブラウザで描画し、Auto はページが空に見える場合に自動で切り替えます。クリックのセレクターで、例えば FAQ のアコーディオンを開けます。

すっきりした記事

CSS の除外設定で、ナビゲーションやフッターなどの領域の記事から外します。ご希望であれば、ページから短い要約の記事に取り込めます。

Web サイトのルール

robots.txt と noindex を尊重します。インデックスされたページが後からブロックされた場合、クローラーはその記事を取り下げます。

全面的な書き直しではなく、変更分だけ

クローラーは内容によって変更されたページを見つけ、変更のないページには手を付けません。消えたページは、二回の実行でヒットしなかった後に取り下げられます。

02・仕組み

情報源を追加し、確認し、エージェントに割り当てる。

Web サイトの情報源は、数項目を入力するだけで作成できます。プレビューには、最初の実行の前にクローラーが見つかるものが表示され、その後、カタログには各 URL の結果が表示されます。

1 情報源を追加

Knowledge Base AI › Webサイトソース で **ソースを追加** を使い、名前、開始 URL、対象のナレッジベース、更新間隔を入力します。

2 プレビューを確認

プレビュー には、何もインデックスせずに robots.txt、サイトマップ、候補が表示されます。そのあと、範囲とクローラーを設定してください。

3 実行して監視

最初の実行は、保存の数分後に始まります。**クロール開始** で次回の実行を予約できます。**インデックスされたページ** には、URL ごとのステータスと HTTP コードが表示されます。

4 エージェントに割り当て

Conversation AI › 会話エージェント で、**ナレッジ** タブからナレッジベースを選択します。その後、エージェントは Web サイトの記事から回答します。

技術データ

機能	詳細
入力	一つのホストの公開 HTTPS ページ、HTML のテキスト
情報源ごとの範囲	最大 2000 ページ（既定値 200）、リンクの深さは最大 6（既定値 3）
アカウントごとの情報源	最大 25 件
クローラー	Fast、Full、Auto。クリックと待機のセレクトター付き
出力	ページごとの内部記事。ページ URL、検出された言語、タグ website と host 付き
概要	インデックスされたページ : ステータス、描画方式、HTTP、最終取得。情報源ごとの 同期履歴
アクセスと料金	Knowledge Base AI › Webサイトソース の下の Web アプリ。利用量に応じて課金されます

03・実際の活用

Web サイトの知識が 電話とチャットに届く場所。

Web サイトのチャットのハウツー

あるソフトウェア会社は、ハウツーを /docs の下で管理しています。パスの接頭辞 /docs を指定した Web サイトの情報源が毎日それを読み込み、Web サイトのチャットは回答の中で該当ページを示します。

ホットライン用の製品ページ

ある小売業者は、製品ページを毎週読み込ませています。電話の会話エージェントは、サイズや素材について、お客様がオンラインで見ると同じ文章をもとに回答します。

Web サイトと自社の記事

休業日など、Web サイトにないことは、同じナレッジベースに自社の記事として追加します。エージェントは両方をまとめて読みます。

よくある質問

Web サイトの情報源はどのページを読みますか？

一つのホストの公開 HTTPS ページで、開始 URL から、必要に応じてサイトマップや追加の開始 URL を経由して読み込みます。HTML ページのテキストを読み、ログイン、Cookie、認証情報は使いません。

Web サイトの記事に自分で加えた編集はどうなりますか？

ページが変更されると、次の実行で記事が書き直されます。したがって、追加内容は同じナレッジベース内の独自の記事に書いてください。

エージェントの知識はどのくらい新しいですか？

直近の実行と同じ程度に新しいです。6 時間ごとから毎月までの間でスケジュールを選択でき、クローリング開始で追加の実行を予約できます。手動の実行を二回続けて行う場合、その間隔は 15 分です。

Web サイトのページはカスタマーポータルに表示されますか？

Web サイトの記事は社内向けです。チームと、そのナレッジベースを読むエージェントが参照し、カスタマーポータルには表示されません。

はじめに

Web サイトを最初の情報源にしましょう。

Web アプリで **Knowledge Base AI** > **Webサイトソース** を開き、**ソースを追加** を使って開始 URL を入力し、プレビューを確認して保存してください。最初の実行は数分後に始まります。

paperoffice.ai/ja/ai-web-sources-website-crawler-knowledge



モジュールの詳細

paperoffice.ai/ja/ai-web-sources-website-crawler-knowledge

料金

paperoffice.ai/en/pricing

お問い合わせ

paperoffice.ai/ja/otoiawase