NOSIBLE Save Search - Data Dictionary V2.1.0

As of: 2026-07-20

This document describes every input and output field of the NOSIBLE Save Search endpoint so that a technical reader can understand the API at a glance. Save Search (`POST https://www.nosible.ai/search/v2/save-search`) lets NOSIBLE bring the web to you: it saves a search configuration to the crawl feed, which then runs it for you on an ongoing basis. Provide a `search_id` to update an existing saved search, or omit it to create a new one.

This is a beta endpoint. Access requires an alpha access code, supplied by NOSIBLE, passed as `internal_use.crawl_feed_alpha`. Requests without it fail with HTTP 401. Old pipe-format API keys are not supported; get a new key at app.nosible.com.

Interactive API documentation: this endpoint is documented live in the NOSIBLE Swagger UI at `https://www.nosible.ai/search/v2/docs/` (OpenAPI spec: `https://www.nosible.ai/search/v2/openapi.json`). The version of this data dictionary tracks the API version published there: this document describes NOSIBLE V2.1.0.

Contents

Request Headers

Every request to this endpoint is an HTTP POST with a JSON body. Authentication is handled by a single header, documented below. You should also send `Content-Type: application/json`, and `Accept-Encoding: gzip` is recommended to cut transfer sizes.

Header	Details
`api-key`	**Definition:** Your NOSIBLE Search API key. To get an API key please visit app.nosible.com, where you can manage organizations, users, subscriptions, and API keys. **Type:** string **Length:** 42–50 characters **Required:** Yes - requests without a valid key fail with HTTP 422. **Example:** `"nos_sk_XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX"`

Input Fields

All input fields are passed in the JSON body of the POST request. Every field is optional: any field you omit falls back to the default shown in its row. In practice, `question` is the only field you must think about; everything else tunes recall, filtering, and snippet length. The search_id and extra_args fields at the top of the table are specific to saved searches.

Input Field	Details
`search_id`	**Definition:** The ID of the search. Provide it to update an existing saved search; omit it to create a new one. **Type:** string or null **Length:** 1–100 characters **Default:** `null` **Example:** `"b3f5c1e2-8a9d-4f60-9c2b-7d41e8a3f0aa"`
`extra_args`	**Definition:** Extra arguments controlling where and how your saved search results are published to you. **Type:** object **Default:** `{}` **Example:** `{}`
`question`	**Definition:** The question you would like your search results to answer. This is the primary query that drives retrieval. **Type:** string

Input Field	Details
	Length: 1–500 characters **Default:** `"Why did Argus Filch confiscate the Marauder's Map?"` **Example:** `"What conservation efforts exist for the Queen Alexandra's Birdwing butterfly?"`

`expansions`	**Definition:** Semantically equivalent but lexically diverse rephrasings of the question. Optional and recommended: including ten expansions materially improves recall.
	Type: array of strings
	Items: 0–10
	Default: `[]`
	Example:
	`[`
	`  "How is the Queen Alexandra's Birdwing protected in PNG?",`
	`  "What threats face the largest butterfly species?"`
	`]`

`instruction`	**Definition:** The instruction passed along to the embedding model. USE WITH CAUTION. **Type:** string **Length:** 1–500 characters **Default:** `"Retrieve semantically similar text."` **Example:** `"Retrieve semantically similar text."`
`sql_filter`	**Definition:** A SQL filter to narrow down your search results. **Type:** string **Default:** `"SELECT loc FROM engine"` **Columns:** `netloc` - include or exclude specific websites from your search results `published` - a YYYY-MM-DD string for filtering on publish dates `modified` - a YYYY-MM-DD string for filtering on modified dates `visited` - a YYYY-MM-DD string for filtering on visited/crawled dates `language` - e.g. en-en (english document, english snippet) `stot` - include or exclude documents based on total number of snippets `snum` - include or exclude snippets based on their index on the page `words` - include or exclude snippets based on their word count `chars` - include or exclude snippets based on their character count **Example:** `"SELECT loc FROM engine WHERE words>=30"`

Input Field	Details
`algorithm`	**Definition:** The algorithm to use when scoring search results. Saved searches currently support only `hybrid-3`. **Type:** string **Options:** `hybrid-3` **Default:** `"hybrid-3"` **Example:** `"hybrid-3"`
`min_similarity`	**Definition:** The minimum similarity that a result must have to be returned. **Type:** float **Range:** 0.0–1.0 **Default:** `0.0` **Example:** `0.65`

`must_include`	**Definition:** Any of the following strings must be included in each result. Be careful when using this: aggressive lists can filter out everything.
	Type: array of strings
	Items: 0–100
	Default: `[]`
	Example:
	`[`
	`  "Ornithoptera alexandrae"`
	`]`

`must_exclude`	**Definition:** All of the following strings must be excluded from each result. Be careful when using this: aggressive lists can filter out everything.
	Type: array of strings
	Items: 0–100
	Default: `[]`
	Example:
	`[`
	`  "casino",`
	`  "betting"`
	`]`

`brand_safety`	**Definition:** The brand safety level to apply to the source websites. **Type:** string or null **Options:** `Safe`, `Sensitive`, `Unsafe` **Default:** `null` **Example:** `"Safe"`

Input Field	Details
`language`	**Definition:** The language you would like your search results to be refined to. **Type:** string or null **Options:** one of 94 ISO 639-1 two-letter codes, e.g. `en`, `fr`, `de`, `es`, `pt`, `zh`, `ja`, `ar` **Default:** `null` **Example:** `"en"`
`continent`	**Definition:** The continent you would like your search results to be refined to. **Type:** string or null **Options:** `Africa`, `Asia`, `Europe`, `North America`, `Oceania`, `South America`, `Worldwide` **Default:** `null` **Example:** `"Oceania"`
`region`	**Definition:** The geographic region you would like your search results to be refined to. **Type:** string or null **Options:** one of 31 named world regions, e.g. `Western Europe`, `Middle East`, `Southeast Asia`, `Sub-Saharan Africa`, `The Balkans`, `The Caucasus` **Default:** `null` **Example:** `"Southeast Asia"`
`country`	**Definition:** The country you would like your search results to be refined to. **Type:** string or null **Options:** one of 203 full English country names (including `Worldwide`), e.g. `United States`, `Germany`, `South Africa` **Default:** `null` **Example:** `"Papua New Guinea"`
`sector`	**Definition:** The GICS sector you would like your search results to be refined to. **Type:** string or null **Options:** `Other`, `Communication Services`, `Consumer Discretionary`, `Consumer Staples`, `Energy`, `Financials`, `Health Care`, `Industrials`, `Information Technology`,

Input Field	Details
	`Materials`, `Real Estate`, `Utilities` **Default:** `null` **Example:** `"Materials"`
`industry_group`	**Definition:** The GICS industry group you would like your search results to be refined to. **Type:** string or null **Options:** one of 26 GICS industry groups, e.g. `Banks`, `Capital Goods`, `Media & Entertainment`, `Software & Services` **Default:** `null` **Example:** `"Food, Beverage & Tobacco"`
`industry`	**Definition:** The GICS industry you would like your search results to be refined to. **Type:** string or null **Options:** one of 75 GICS industries, e.g. `Biotechnology`, `Aerospace & Defense`, `Automobiles` **Default:** `null` **Example:** `"Biotechnology"`
`sub_industry`	**Definition:** The GICS sub-industry you would like your search results to be refined to. **Type:** string or null **Options:** one of 164 GICS sub-industries, e.g. `Advertising`, `Apparel Retail`, `Aluminum` **Default:** `null` **Example:** `"Environmental & Facilities Services"`
`iab_tier_1`	**Definition:** The tier 1 IAB content category you would like your search results to be refined to. **Type:** string or null **Options:** one of 39 tier 1 IAB categories, e.g. `Automotive`, `Business and Finance`, `Technology & Computing` **Default:** `null` **Example:** `"Science"`
`iab_tier_2`	**Definition:** The tier 2 IAB content category you would like your search results to be refined to. **Type:** string or null **Options:** one of 388 tier 2 IAB categories

Input Field	Details
	Default: `null` **Example:** `"Environment"`
`iab_tier_3`	**Definition:** The tier 3 IAB content category you would like your search results to be refined to. **Type:** string or null **Options:** one of 644 tier 3 IAB categories **Default:** `null` **Example:** `"Environment"`
`iab_tier_4`	**Definition:** The tier 4 IAB content category you would like your search results to be refined to. **Type:** string or null **Options:** one of 704 tier 4 IAB categories **Default:** `null` **Example:** `"Nature"`

`companies`	**Definition:** A list of company names. This will be used to refine the search results.
	Type: array of strings
	Items: 0–3
	Default: `[]`
	Example:
	`[`
	`  "Tesla"`
	`]`

`collection`	**Definition:** The collection to search over. Use `this-week` to search only recently added data. **Type:** string **Options:** `everything`, `this-week` **Default:** `"everything"` **Example:** `"this-week"`
`deduplicate`	**Definition:** Whether to deduplicate search results based on duplicate headlines. **Type:** Boolean **Default:** `false` **Example:** `true`
`n_results`	**Definition:** The number of search results you would like per feed run.

Input Field	Details
	Type: integer **Range:** 10–100 **Default:** `100` **Example:** `100`
`n_probes`	**Definition:** The number of shards (collections) to probe. **Type:** integer **Range:** 5–1,000 **Default:** `500` **Example:** `500`
`n_contextify`	**Definition:** The amount of context you would like per search result, in words. **Type:** integer **Range:** 64–1024 **Default:** `256` **Example:** `256`
`internal_use`	**Definition:** Internal-only. Used to enable hidden features and new algorithms from time to time. Hidden from the public API docs. **Type:** object **Default:** `{}` **Example:** `{}`

Output Fields

A successful save returns HTTP 200 with the ID of the saved search. A failed update or invalid configuration returns HTTP 400.

Output Field	Details
`message`	**Definition:** A human-readable acknowledgement. **Type:** string **Example:** `"Search saved successfully."`
`response`	**Definition:** The wrapper of the result payload. **Type:** object **Fields:** `search_id` (str) - the ID of the saved search (a UUID for new searches,

Output Field	Details

	or the ID you supplied when updating)
	Example:
	`{`
	`  "search_id": "b3f5c1e2-8a9d-4f60-9c2b-7d41e8a3f0aa"`
	`}`

Streaming Results via the Firehose

Saving a search is only half of the story: the crawl feed runs your saved searches against freshly crawled data on an ongoing basis, and streams every new match to you over the `/search/v2/firehose` websocket. Connect with `firehose=private` to receive the live stream of results from your own saved searches. The rows below describe the connection parameters and the message format.

Parameter	Details
`endpoint`	**Definition:** The firehose is a websocket, not an HTTP endpoint. Connect to it to receive your saved-search results as a continuous stream, pushed to you as the crawl feed finds fresh matches. **Type:** websocket **Example:** `wss://www.nosible.ai/search/v2/firehose?firehose=private`
`api-key`	**Definition:** Your NOSIBLE Search API key, sent as a connection header. Only new-format keys (`nos_sk_...`) are accepted. Invalid or missing keys close the connection with code 4401. **Type:** string (header) **Required:** Yes **Example:** `"nos_sk_XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX"`
`internal_use`	**Definition:** A JSON string, sent as a connection header, carrying your alpha access code as `crawl_feed_alpha`. Missing or invalid codes close the connection with code 4401. **Type:** string (header) **Required:** Yes **Example:** `"{\"crawl_feed_alpha\": \"YOUR_ALPHA_ACCESS_CODE\"}"`
`firehose`	**Definition:** Query parameter selecting the stream to read. `private`

Parameter	Details
	streams the results of your own saved searches; `public` (the default) streams the public firehose. **Type:** string (query parameter) **Options:** `private`, `public` **Default:** `public` **Example:** `?firehose=private`
`last_id`	**Definition:** Query parameter to resume the stream without gaps after a disconnect: pass the `id` of the last message you received. Defaults to `$`, meaning only new messages. The private stream retains roughly the 100 most recent results. **Type:** string (query parameter) **Default:** `$` **Example:** `?firehose=private&last_id=1784560567545-0`

`message`	**Definition:** Each server frame is a JSON object (sent as bytes). `data` carries one saved-search result. Results are deduplicated against the last 24 hours by URL, title, and best chunk before being streamed, so you only see fresh content.
	Type: object
	Fields:
	`id` (str) - the stream ID of the message; use it as last_id to resume
	`data.search_id` (str) - the saved search that produced this result
	`data.search_result` (dict) - one flat search result (the same format as Fast Search: url_hash, url, netloc, published, visited, language, author, title, description, best_chunk, content, semantics)
	`data.timestamp` (str) - the UTC time the result was published to the stream
	Example:
	`{`
	`  "id": "1784560567545-0",`
	`  "data": {`
	`  "search_id": "b3f5c1e2-8a9d-4f60-9c2b-7d41e8a3f0aa",`
	`  "search_result": {`
	`  "url_hash": "aB3dE_fG7hIj-KlMnOpQrStU",`
	`  "title": "New sanctuary offers hope for the world's largest butterfly",`
	`  "best_chunk": "The Queen Alexandra's Birdwing ...",`
	`  ...`
	`  },`
	`  "timestamp": "2026-07-20T14:12:05.421000+00:00"`

Parameter	Details
	`  }` `}`
`ping / pong`	**Definition:** Keep-alive: send the text frame `ping` and the server replies with `pong`. **Type:** text frame **Example:** `ws.send("ping")  ->  "pong"`

Example Request and Response

The examples below save a search end to end, then connect to the firehose to stream its results. All field values are illustrative.

Example Request:

```
import requests

response = requests.post(
    url="https://www.nosible.ai/search/v2/save-search",
    json={
        "question": "Conservation efforts for the Queen Alexandra's Birdwing?",
        "algorithm": "hybrid-3",
        "n_results": 100,
        "n_probes": 500,
        "n_contextify": 256,
        "sql_filter": "SELECT loc FROM engine",
        "internal_use": {
            "crawl_feed_alpha": "YOUR_ALPHA_ACCESS_CODE"
        }
    },
    headers={
        "Accept-Encoding": "gzip",
        "Content-Type": "application/json",
        "api-key": "YOUR_NOSIBLE_API_KEY"
    }
)

print(response.json())
```

Example Response:

```
{
    "message": "Search saved successfully.",
    "response": {
        "search_id": "b3f5c1e2-8a9d-4f60-9c2b-7d41e8a3f0aa"
    }
```

```
}
```

Example Firehose Client (streams your saved-search results; reconnects and resumes from the last received message without gaps):

```
import asyncio
import json

import websockets

async def stream_results():
    url = "wss://www.nosible.ai/search/v2/firehose?firehose=private"

    headers = {
        "api-key": "YOUR_NOSIBLE_API_KEY",
        "internal_use": json.dumps({
            "crawl_feed_alpha": "YOUR_ALPHA_ACCESS_CODE"
        })
    }

    last_id = "$"  # Or the id of the last message you received.

    while True:
        try:
            full_url = f"{url}&last_id={last_id}"
            async with websockets.connect(
                full_url, additional_headers=headers
            ) as ws:
                async for frame in ws:
                    message = json.loads(frame)
                    last_id = message["id"]
                    result = message["data"]["search_result"]
                    print(message["data"]["search_id"], "->", result["title"])

        except Exception as e:
            # Reconnect and resume from last_id without missing results.
            print(f"Reconnecting after: {e}")
            await asyncio.sleep(5)

asyncio.run(stream_results())
```