SimSage Search Syntax and AI

Introduction

This document explains the advanced query syntax used by SimSage to find items across documents and their metadata. SimSage has a sophisticated semantic search engine that automatically uses relationships, probabilities and distances between concepts and words to find the best results. SimSage can also perform searches within searches, and search for complex entities like credit-card numbers, social security numbers, names of people, countries, and many other entities.

SimSage searches in the body of documents by default, using concepts or words found in document titles (where possible) to emphasize the search score / importance of a document.

Beyond SimSage Search there is SimSage AI. A powerful integration of SimSage with Google or OpenAI Large Language Models. This will save you a lot of time and effort digesting your information.

Search operators

Double Quote Operator

This operator indicates an exact search, i.e. words without relationships, and close together. Exact searching in SimSage is much more complex than a normal keyword search engine, owing to its highly semantic data processing.

SimSage has mitigations to remove relationships from exact searches (behaving more like a keyword search engine). However there are some trade-offs. Most of the time a SimSage exact search will give you exactly what you want. There are cases however, where phrases with a lot of “noise words” (e.g. “the”, “a”, “an”, etc.) can fool SimSage in returning false positives.

Good Example

> “market value“

Examples that will trick SimSage

> “market the value“ vs “market a value”

both “the” and “a” are noise words in this example, and SimSage does not know what word is between “market” and “value”, only that there is a gap and it was taken by a noise word.

> “market the the the value“ vs “market a a a value”

Similarly, when SimSage sees a sequence of noise words, it will only mark a gaps between the instances “market” and “value”. Beware of these highly unusual cases that may not provide exactly what you’re looking for. Similarly, starting or ending an exact phrase with a noise word has no effect. In this case the noise words are stripped from the start or end of your exact search.

However, in both these cases, an exact search for “market value” is not likely to return false search results.

Title Searches

Documents usually have titles. SimSage will set the title of a document to its filename (without any path) if it cannot find a title. Where a document has a title, SimSage can use the “intitle:” directive to look for matching words or concepts.

Example

> intitle: baseball game

Document-type / File-type / File-extension Searches

Documents have types. SimSage uses file-extensions to identify document-types. For instance, a word document is either a “doc” (older Microsoft types) or a “docx” file. You can search for more than one document-type at a time. You can search using either “ext:” or “filetype:” as a keyword. Document type searches are filters and have to be used in addition with additional keywords to work.

Examples

the first two examples are equivalent, and search for all files containing HTML, JPG, and DOCX extensions.

> market forces ext: html, jpg, docx
>
> test filetype: html, jpg, docx

The next example searches for all HTML type files containing the word “test”

> test filetype: html

Archive files are not searchable, but the content of archive files is (where SimSage understands the archive format). So if, for instance, a zip file contains text files, search for “filetype: txt”.

Entity Searches

Entities are known objects and semantics in SimSage such as people’s names, credit-card numbers, and many more (see table below). Entity searches enables searching for more general terms. add entity: <entity-name> to your query. Where entity-name can be one of:

entity name	description
nin	UK national insurance number
ssn	US social security number
credit-card	A valid credit-card (Lunn verified) number
ip-address	an IPv4 or IPv6 address
mac-address	a network Media Access Control address
url	An HTTP or HTTPS based web address
person	A person (e.g. Jimmy Smith)
email	an email address

entity name	description
city	biggest cities in the world
country	all countries of the world
company	a small set of known companies
brand	a small set of known brands
continent	all continents of the world
capital	all capitals of the world
state	all states of the US
phone	valid phone numbers for the US and UK
zip	US zip codes
postcode	UK post codes
hashtag	hashtags, (e.g. #test, #market)
secret	A large set of secrets as defined by the truffleHog regex set (https://github.com/dxa4481/truffleHogRegexes)
vat	a UK tax number (VAT) starting with GB / GBHA / GBGD (includes EU format)
hip	SimSage home insurance policy numbers
pip	SimSage personal insurance policy numbers
cip	SimSage car insurance policy numbers
policy	SimSage policy numbers

You must use the exact “entity name” when using an “entity:” search. Your search will fail if you do not pass a known entity name after the “entity:” keyword.

Example

> entity: credit-card
>
> entity: mac-address

Document Language

SimSage can translate foreign languages to English (if enabled on your platform) as well as recognize languages (always enabled). You can search for documents that have been recognized as using a particular language. Use the language keyword to do so.

Example

> language: French
>
> language: dutch

Here is a list of the recognized languages:

Dutch, German, French, Spanish, Italian, Turkish, Greek, Portuguese, Arabic, Korean, Japanese, Russian, Hindi, Afrikaans, Chinese / Mandarin / Cantonese, Danish, Norwegian, Swedish, Finnish, Polish.

URL/Path Searches

URLs in SimSage are the primary keys of whatever data type you're searching for. In some cases these aren't actually URLs. Websites, and most web-based systems do use URLs. The *inurl:* <many words> can be used to look for a series of words occurring inside the URLs / primary keys of your data.

Example

> inurl: research jobs
>
> inurl: research facilities in Japan

Some path searches can include unusual characters like "%", "~", "&" etc. Use double quotes in your URL / path search for such characters.

Example

> inurl: "/path~%20/&file1.html"

In text / main content of your data searching

This filter is provided to enable a user to switch back to the default search of searching inside document body text. This is SimSage's default for searching and does not need to be specified ordinarily. However, it can be explicitly used in combination with other filters to switch between them. The syntax for this search is: *intext:* <many words>

Example

> intext: the effects of radiation

User defined Semantics

In addition to the above entity types, users can define entity types in SimSage. This is done by creating or importing semantics in the admin UI. An example might be that of a *semantic* named “staff”. Suppose our staff consists of “Joan Bobbat”, “Bruce Willis”, and “John Elderberries”. We can then define Joan Bobbat: staff, Bruce Willis: staff, John Elderberries: staff. Once any existing documents have been re-processed, or new content is added, any of these people can then be found search for “entity: staff” in SimSage search.

Another example might be a medical research category like “grade”. Grade can be one of mild, moderate, severe, life-threatening, and death. By defining mild: grade, moderate: grade, severe: grade, life-threatening: grade, and death: grade in the admin UI, you can simplify the search across documents for “grades” using “entity: grade”, which will include any references found in documents to mild, moderate, severe, life-threatening, and death.

Example

> entity: staff

Metadata searches

Arbitrary name values can be searched too. If your data provides, for instance, a “status” field in its metadata you can search for a metadata-name equals value. How metadata is mapped, and what metadata is available and what it is set too all depend on your sources. In our example we assume there is a metadata field called “status” and it has values like “closed”, “released”, “open” etc.

Example

> status: closed

Source Filters

In many cases your SimSage system can have many “sources”. A source is where your information comes from / external integration points. These sources will have been given names by your administrator and can be referenced as part of a filter using the source: keyword. The usage is *source:* <unique name of a source>. Your search will be rejected with an error if the source name is incorrect or does not exist. Sources are filters, like document-types. This means you need to add other keywords in addition to the source keyword for your searches to work.

Example

> market forces source: second floor server
>
> test source: google drive one

Exclude Filters

You can chose to exclude a single word / concept by prefixing the word with a hyphen (-) as shown in the example below.

Example

> -second

This operator applies to inurl / allinurl / intitle / allintitle items too. The operator only applies to the exact word, not its relationships if applicable.

Time Based Searching

We will group these into one category. Time based searches are modeled after the Google time based searches. *Before* and *After* can only be used *with other keywords* (i.e., they need to be used as part of a search). The two time-based searches must have this exact syntax:

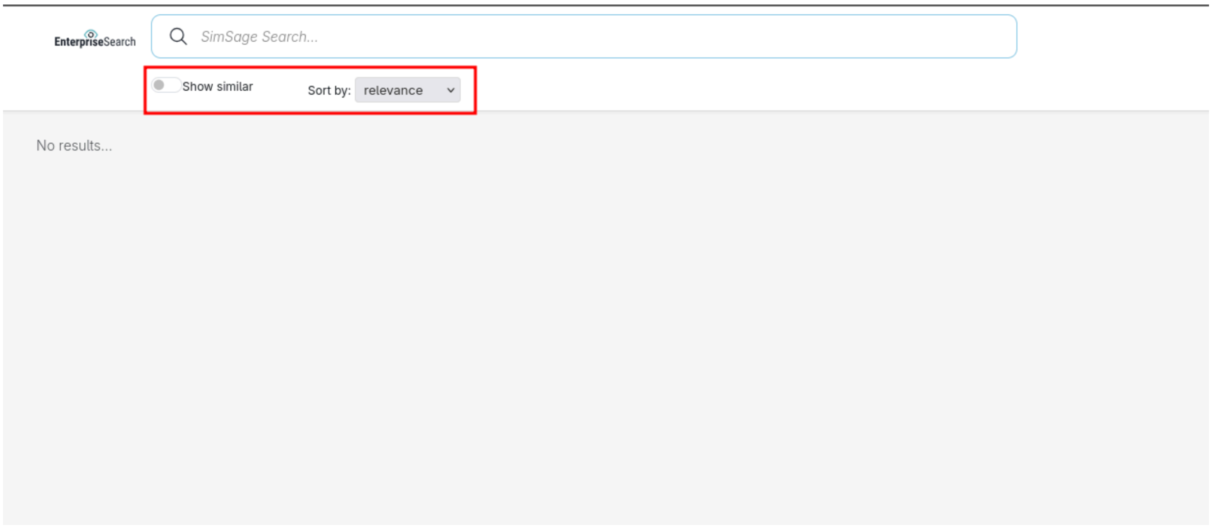
- before: yyyy-mm-dd
- after: yyyy-mm-dd

Example

> test instrumentation before: 2020-05-02

Dates for before and after must be between 1970-01-01 and 2200-12-31. Any other value will result in an "invalid date following before: or after:" error.

Grouping and sorting



The left top of the screen has a toggles and a drop down list. These are labelled Show similar and Sort by. Show similar shows similar documents with each search result. This has to be enabled for individual sources. Once enabled, a source has to be re-processed (if this wasn't set up before). An a knowledge-base schedule for similarity calculations needs to be set.

Sort by, has three options, relevance, newest first and oldest first. Selecting relevance, sorts documents by highest relevance in relation to your search terms. Selecting newest first, sorts documents by newest first (modified most recently). Selecting oldest first, sorts documents by the oldest last modified documents first.

Searches within Searches

You can search again within the results of a search. SimSage uses the "sub" keyword to add searches within searches. The second search isn't strictly tied to the search that came before it but will try and find the closest result to the previous search. A document is rejected if it does not match all requirements of each and every sub-search. A sub search is an "infix" operator, meaning that it appears between two search directives as defined by the rest of this document.

Syntax:

some search terms **sub** *some other search terms*

You can create up to nine sub-searches in a row. However, each search counts as a new additional search and requires additional processing power of the SimSage platform (equivalent to another search each time).

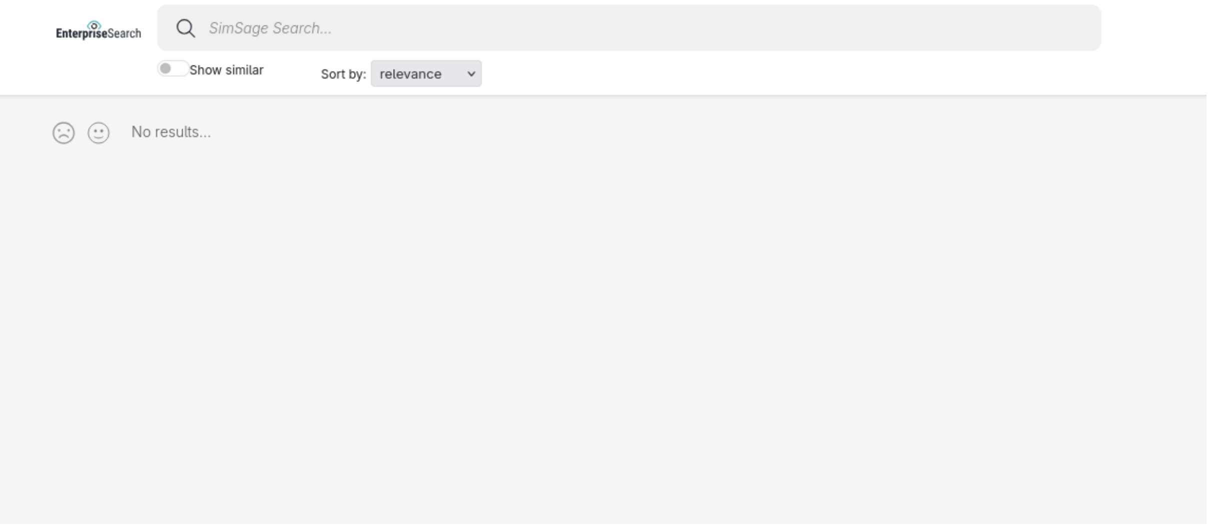
Example

- Suppose you wanted to find an email address for “John Smith”. You can first search for “John Smith” and then instruct SimSage to find the closest email address to any of those terms found by using the “sub” keyword and the “entity: email” selector. Search for:
    - *John Smith sub entity: email*
        - SimSage will return no results if it cannot find any email addresses or John Smith in any combination together
        - SimSage will return the closest email address it can find to John Smith. This might not be John Smith’s email address depending on the data.
- Suppose you want to see if there are any sensitive MAC addresses in your router documentation. You know all your router documentation has the word “router” in the title of each document, so you can then search for
    - *intitle: router with entity: mac-address*
- Suppose you want to find documents where people, email addresses but also social security numbers are mentioned. You can do a triple “sub” search by entering:
    - *entity: person entity: email entity: ssn*
        - SimSage effectively executes three searches and needs to do 3x the amount of work
        - The “sub” keyword act as an AND across the content of the document. Any document not containing all three will not be shown.
        - SimSage will show only the first “person” it finds in each document followed by the closest email and social security number relative to that person.
        - “entity:” type searches implicitly use “sub” relationships when used.

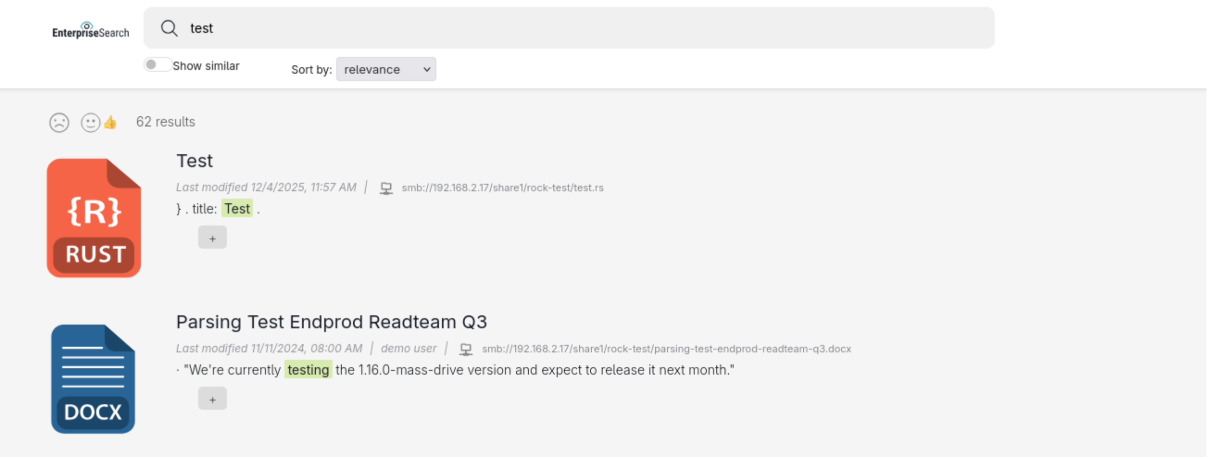
> John Smith sub entity: email
>
> intitle: router sub entity: mac-address
>
> entity: person entity: email entity: ssn
>
> entity: person sub entity: email sub entity: ssn   *(identical to the previous line)*

User feedback

Search users can send feedback to their SimSage system using the scowl and smiley faces. These faces will only show if this feature has been enabled on your platform and after a search.



Clicking the smiley face provides a positive feedback, and responds immediately with a thumbs up.



Test

Last modified 12/4/2025, 11:57 AM | smb://192.168.2.17/share1/rock-test/test.rs

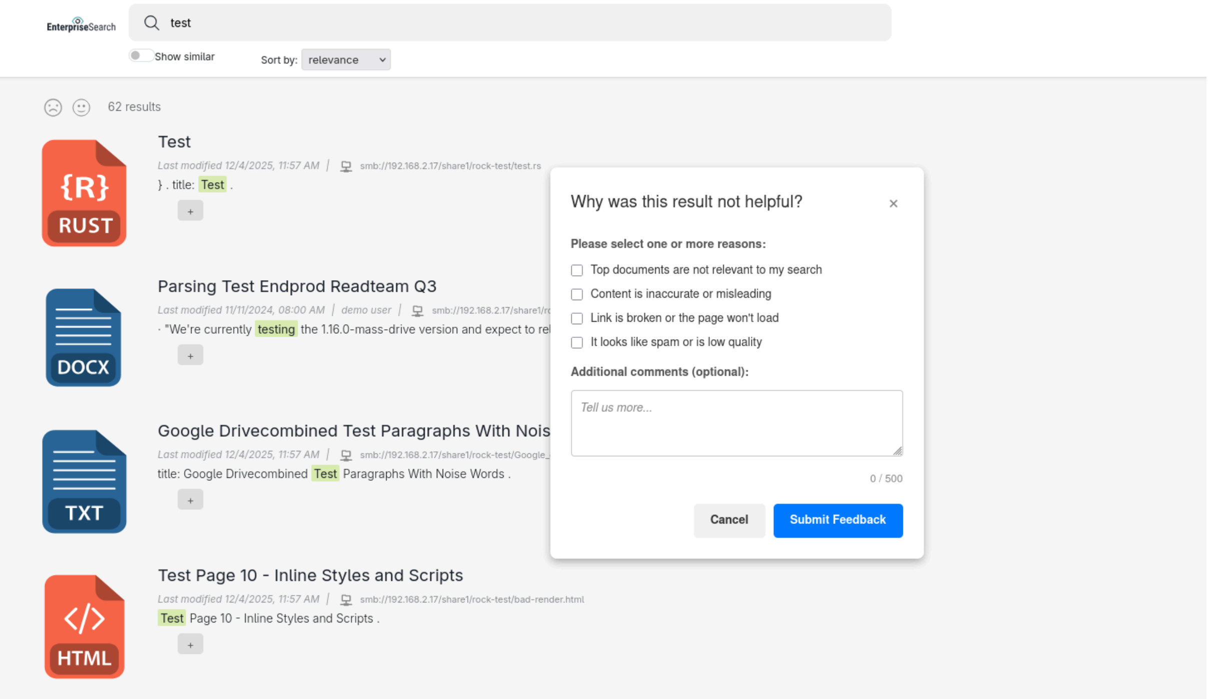
} . title: Test .

Parsing Test Endprod Readteam Q3

Last modified 11/11/2024, 08:00 AM | *demo user* | smb://192.168.2.17/share1/rock-test/parsing-test-endprod-readteam-q3.docx

· "We're currently testing the 1.16.0-mass-drive version and expect to release it next month."

Clicking the scowl face brings up a feedback dialog where you can provide some more information to your own internal team with what is not working well for you.



Test

Last modified 12/4/2025, 11:57 AM | smb://192.168.2.17/share1/rock-test/test.rs

} . title: Test .

Parsing Test Endprod Readteam Q3

Last modified 11/11/2024, 08:00 AM | demo user | smb://192.168.2.17/share1/r

· "We're currently testing the 1.16.0-mass-drive version and expect to rel

Google Drivecombined Test Paragraphs With Nois

Last modified 12/4/2025, 11:57 AM | smb://192.168.2.17/share1/rock-test/Google_

title: Google Drivecombined Test Paragraphs With Noise Words .

Test Page 10 - Inline Styles and Scripts

Last modified 12/4/2025, 11:57 AM | smb://192.168.2.17/share1/rock-test/bad-render.html

Test Page 10 - Inline Styles and Scripts .

Why was this result not helpful? ×

Please select one or more reasons:

- □ Top documents are not relevant to my search
- □ Content is inaccurate or misleading
- □ Link is broken or the page won't load
- □ It looks like spam or is low quality

Additional comments (optional):

Tell us more...

0 / 500

Cancel | Submit Feedback

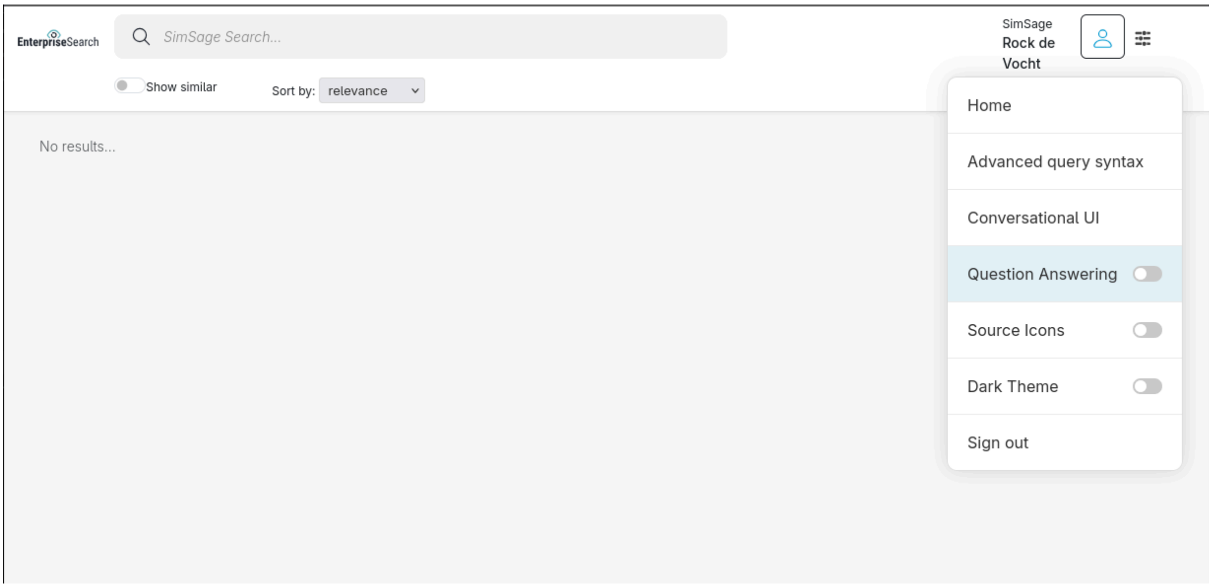
Once you describe your issue and click Submit Feedback, you'll see the same thumbs up icon shown above.

SimSage AI

You have additional menu items available to you if AI has been enabled for your SimSage platform. SimSage can use its search engine result output with language models (Gen AI) such as Google's Gemini, OpenAI, or other LLMs. This process uses Retrieval Augmented Generation (RAG) and passes snippets of your search results to Large Language Models (LLMs) along with your search query.

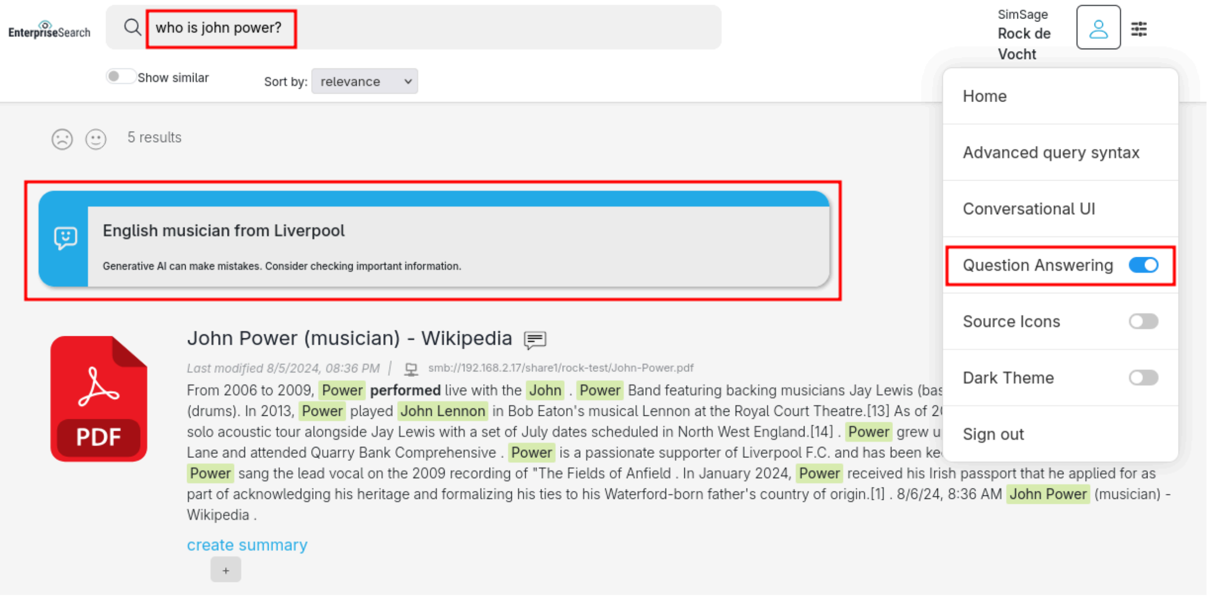
In addition SimSage can use these LLMs to go into a Question and Answer mode for any of your documents where you can interrogate the content of a document found in your searches.

NB. the "Question Answering" menu will only show if AI has been enabled on your platform.



No results...

Once the AI is enabled (which can be either Google Gemini or OpenAI depending on your platform configuration) SimSage will use their LLMs to pass small snippets of text of your search results and try and answer the user's query.



who is john power?

5 results

English musician from Liverpool

Generative AI can make mistakes. Consider checking important information.

John Power (musician) - Wikipedia

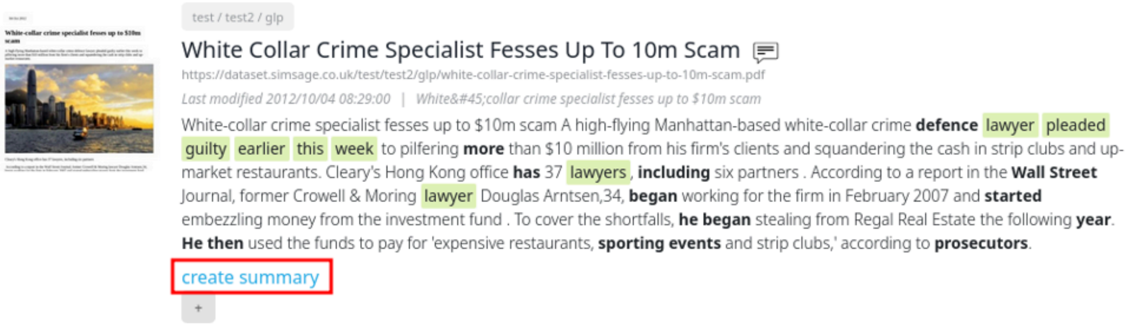
Last modified 8/5/2024, 08:36 PM | smb://192.168.2.17/share1/rock-test/John-Power.pdf

From 2006 to 2009, Power **performed** live with the John . Power Band featuring backing musicians Jay Lewis (bas
(drums). In 2013, Power played John Lennon in Bob Eaton's musical Lennon at the Royal Court Theatre.[13] As of 2
solo acoustic tour alongside Jay Lewis with a set of July dates scheduled in North West England.[14] . Power grew u
Lane and attended Quarry Bank Comprehensive . Power is a passionate supporter of Liverpool F.C. and has been ke
Power sang the lead vocal on the 2009 recording of "The Fields of Anfield . In January 2024, Power received his Irish passport that he applied for as part of acknowledging his heritage and formalizing his ties to his Waterford-born father's country of origin.[1] . 8/6/24, 8:36 AM John Power (musician) - Wikipedia .

How much content it passes to Google Gemini, OpenAI, or other LLMs depends on

- how the SimSage platform is configured
- the limits of the language model used

The "create summary" link at the creates a summary of the document's content.

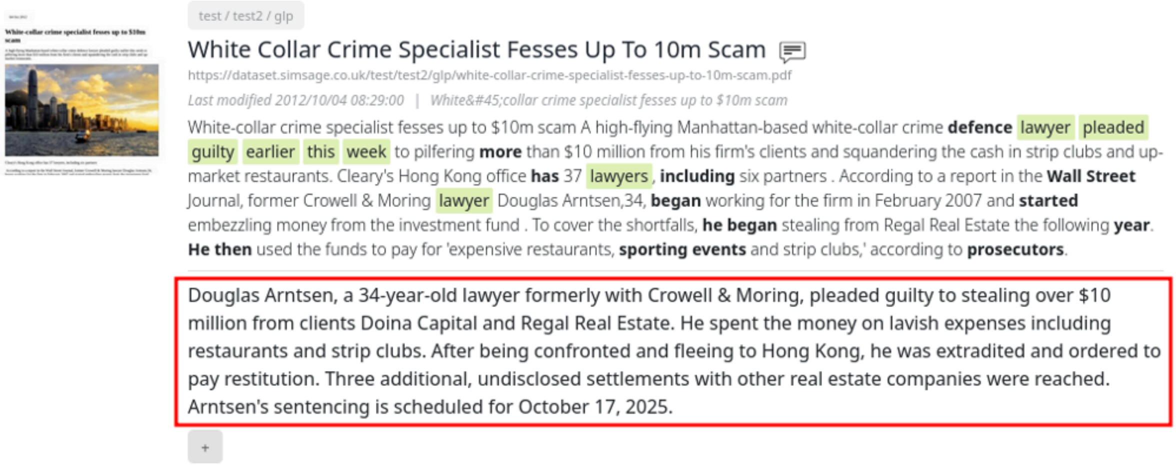


White Collar Crime Specialist Fesses Up To 10m Scam

https://dataset.simsage.co.uk/test/test2/glp/white-collar-crime-specialist-fesses-up-to-10m-scam.pdf

Last modified 2012/10/04 08:29:00 | White&#45;collar crime specialist fesses up to $10m scam

White-collar crime specialist fesses up to $10m scam A high-flying Manhattan-based white-collar crime **defence** lawyer pleaded guilty earlier this week to pilfering **more** than $10 million from his firm's clients and squandering the cash in strip clubs and up-market restaurants. Cleary's Hong Kong office **has** 37 lawyers , **including** six partners . According to a report in the **Wall Street** Journal, former Crowell & Moring lawyer Douglas Arntsen,34, **began** working for the firm in February 2007 and **started** embezzling money from the investment fund . To cover the shortfalls, **he began** stealing from Regal Real Estate the following **year**. **He then** used the funds to pay for 'expensive restaurants, **sporting events** and strip clubs,' according to **prosecutors**.



White Collar Crime Specialist Fesses Up To 10m Scam

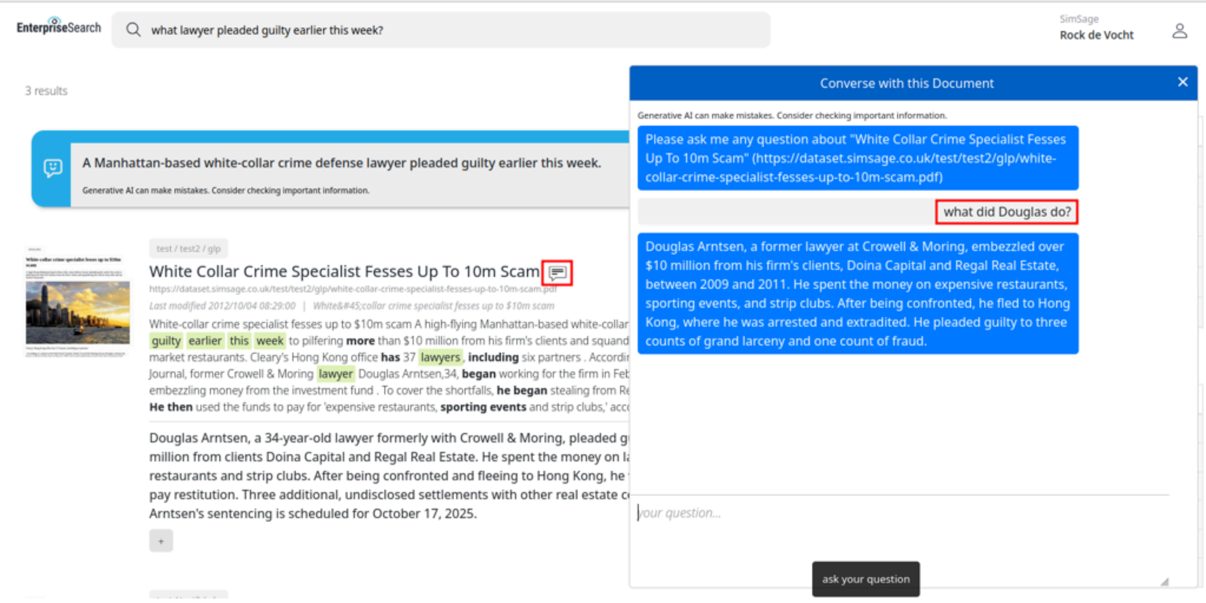
https://dataset.simsage.co.uk/test/test2/glp/white-collar-crime-specialist-fesses-up-to-10m-scam.pdf

Last modified 2012/10/04 08:29:00 | White&#45;collar crime specialist fesses up to $10m scam

White-collar crime specialist fesses up to $10m scam A high-flying Manhattan-based white-collar crime **defence** lawyer pleaded guilty earlier this week to pilfering **more** than $10 million from his firm's clients and squandering the cash in strip clubs and up-market restaurants. Cleary's Hong Kong office **has** 37 lawyers , **including** six partners . According to a report in the **Wall Street** Journal, former Crowell & Moring lawyer Douglas Arntsen,34, **began** working for the firm in February 2007 and **started** embezzling money from the investment fund . To cover the shortfalls, **he began** stealing from Regal Real Estate the following **year**. **He then** used the funds to pay for 'expensive restaurants, **sporting events** and strip clubs,' according to **prosecutors**.

Douglas Arntsen, a 34-year-old lawyer formerly with Crowell & Moring, pleaded guilty to stealing over $10 million from clients Doina Capital and Regal Real Estate. He spent the money on lavish expenses including restaurants and strip clubs. After being confronted and fleeing to Hong Kong, he was extradited and ordered to pay restitution. Three additional, undisclosed settlements with other real estate companies were reached. Arntsen's sentencing is scheduled for October 17, 2025.

The “dialog” icon shown to the right of the title can be clicked to go into document Question and Answer mode.



what lawyer pleaded guilty earlier this week?

A Manhattan-based white-collar crime defense lawyer pleaded guilty earlier this week.

Generative AI can make mistakes. Consider checking important information.

White Collar Crime Specialist Fesses Up To 10m Scam

https://dataset.simsage.co.uk/test/test2/glp/white-collar-crime-specialist-fesses-up-to-10m-scam.pdf

Last modified 2012/10/04 08:29:00 | White&#45;collar crime specialist fesses up to $10m scam

White-collar crime specialist fesses up to $10m scam A high-flying Manhattan-based white-collar guilty earlier this week to pilfering **more** than $10 million from his firm's clients and squand market restaurants. Cleary's Hong Kong office **has** 37 lawyers , **including** six partners . Accordir Journal, former Crowell & Moring lawyer Douglas Arntsen,34, **began** working for the firm in Feb embezzling money from the investment fund . To cover the shortfalls, **he began** stealing from Re **He then** used the funds to pay for 'expensive restaurants, **sporting events** and strip clubs,' acc

Douglas Arntsen, a 34-year-old lawyer formerly with Crowell & Moring, pleaded g million from clients Doina Capital and Regal Real Estate. He spent the money on la restaurants and strip clubs. After being confronted and fleeing to Hong Kong, he pay restitution. Three additional, undisclosed settlements with other real estate c Arntsen's sentencing is scheduled for October 17, 2025.

Generative AI can make mistakes. Consider checking important information.

Please ask me any question about "White Collar Crime Specialist Fesses Up To 10m Scam" (https://dataset.simsage.co.uk/test/test2/glp/white-collar-crime-specialist-fesses-up-to-10m-scam.pdf)

what did Douglas do?

Douglas Arntsen, a former lawyer at Crowell & Moring, embezzled over $10 million from his firm's clients, Doina Capital and Regal Real Estate, between 2009 and 2011. He spent the money on expensive restaurants, sporting events, and strip clubs. After being confronted, he fled to Hong Kong, where he was arrested and extradited. He pleaded guilty to three counts of grand larceny and one count of fraud.

your question...