
PDF Search Engine!:
An answer to hallucination

Issue:

Search Issues

Vector-based search still has inconsistent effectiveness when it comes to searches.

Question	Haystack Dense Passage Retriever (Whole PSSCOC)	Haystack Correct Answer (BM25 Generated)
What should the Contractor do if they believe there is a delay in the progress or completion of the Works due to certain events?	['The Superintending Officer shall take into account the effect, or extent, of any work omitted under the Contract and shall also take into account whether the event in question is one which will delay completion of the Works.']	['he shall forthwith notify the Superintending Officer in writing of such event and shall in any case do so within 60 days of the occurrence of such event.']

Did not answer "60 days"

Issue (Repeats):

Question	Haystack Dense Passage Retriever (Whole PSSCOC)	Correct Answer
"What items should be included in the Notice by the Contractor to the Superintending Officer under PSSCOC?"	['The Superintending Officer shall take into account the effect, or extent, of any work omitted under the Contract and shall also take into account whether the event in question is one which will delay completion of the Works. The Superintending Officer shall also take into account any delays which may operate concurrently with the delay due to the event or events in question and which are due to acts or default on the part of the Contractor.']	The notice given by the Contractor should include the reasons for the delay, the length of the delay, the extension of time required, and the impact of the event on the accepted program under Clause 9. The notice should also contain the appropriate Contract references.
Can the Superintending Officer request additional information from the Contractor regarding a delay application?	['The Superintending Officer shall take into account the effect, or extent, of any work omitted under the Contract and shall also take into account whether the event in question is one which will delay completion of the Works. The Superintending Officer shall also take into account any delays which may operate concurrently with the delay due to the event or events in question and which are due to acts or default on the part of the Contractor.']	Yes, if the Superintending Officer deems the notice and accompanying information insufficient, they have the authority to request the Contractor to provide further particulars within a specified period. The additional details may include information about the event, the circumstances of the delay, measures planned or taken to mitigate the delay, and any other information reasonably required by the Superintending Officer.

Cause of Issue:

Semantic Search issues

The system is unable to detect/find the relevant passage to "Copy+Paste" into the large language model to generate the "fancy" response.

Example :
(Story of little red riding hood):

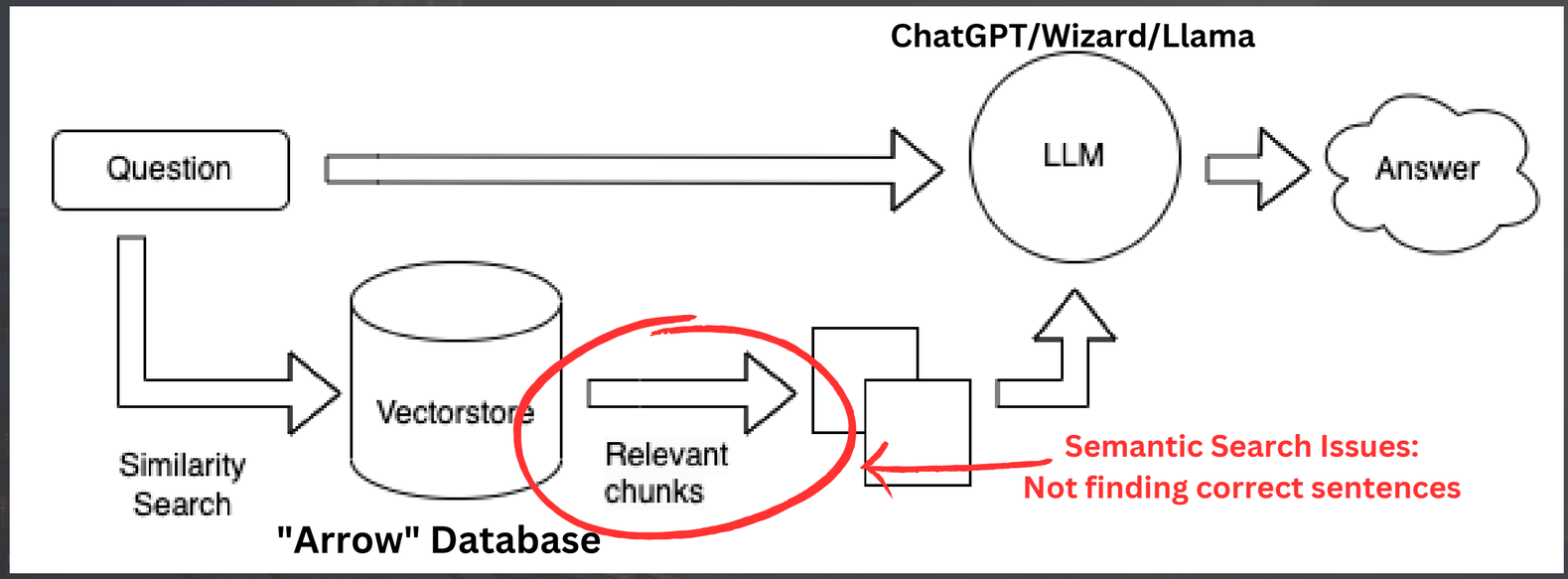
If the search is correct, the LLM will be given:

"The wolf dressed up as little red riding hood's grandmother."

But instead it was given:

"Justin Bieber's song is fantastic."

It will return a response to the user about Justin Bieber because that was all it was given as context from the search.

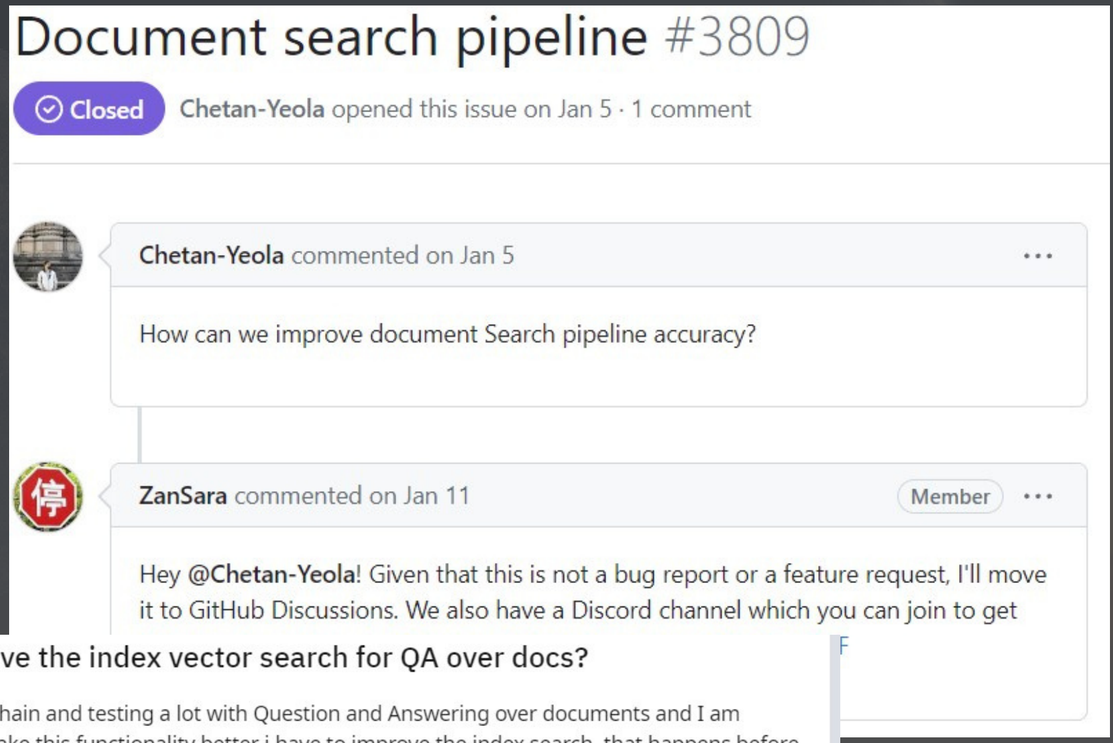


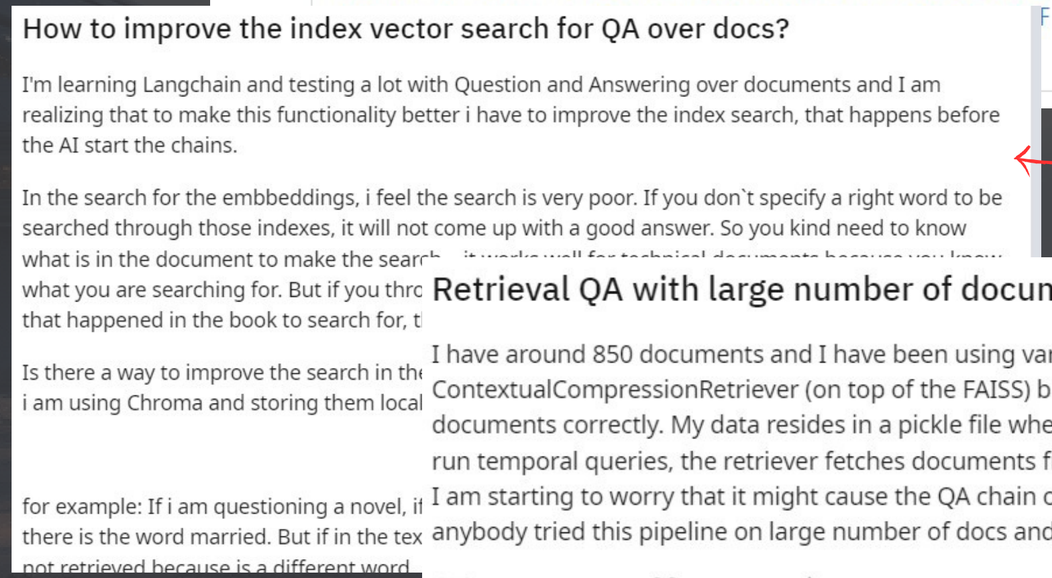
Issue not isolated:

Main issue plaguing vector store databases

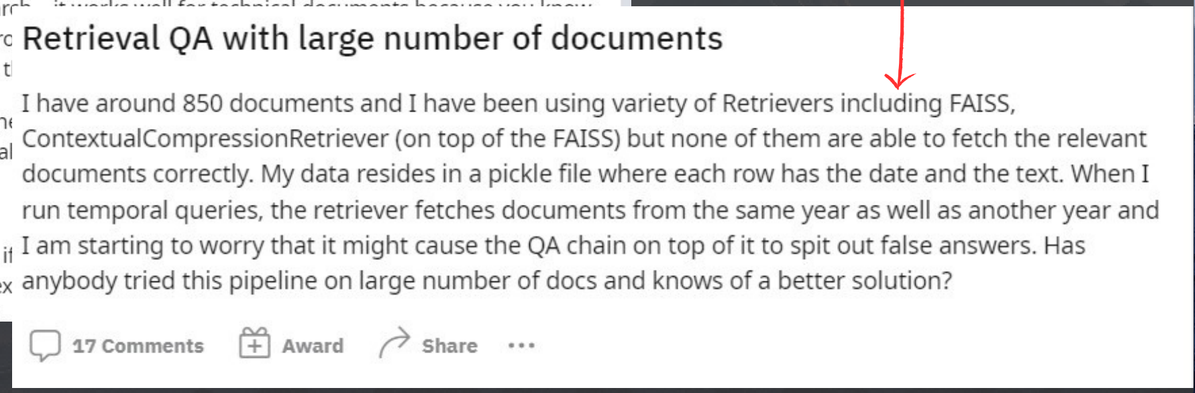
The issue isn't isolated to Langchain/Haystacks, but it is simply the nature of vector searches that are available in the open-source market now.

There are varying degree of solutions, but ultimately it revolves either heavy investment in pre-processing or just cutting down tremendously on the documents fed so there are less items to search.





Semantic Search Issues:
Too many pages = Poor results



Issue not isolated (2):

Semantic Search Issues:
Too many pages = Poor results

r/LangChain · Posted by u/CoffeeAndKnives 13 days ago

How to get better than semantic search?

i've been working with Langchain for a few weeks now. I've managed to upload PDFs to a vector database and get what seems to be an improvement over a typical semantic search results. Curious if anyone has feedback on how to get this to the next level where the LLM understands the uploaded data at a higher level. Maybe there's a way to preprocess the pdf for better information connectivity like a knowledge graph? Or a scheme with multiple agents that perform different tasks?

Hackerjurassicpark · 13 days ago

What's the issue that you face?

2 Reply Share ···

CoffeeAndKnives OP · 13 days ago

im working with a massive mechanical design pdf that contains almost 2000 pages plus tons of charts and tables. First, I'd love to get the tables and charts to be searchable/reference-able and I think that would require significant preprocessing to extrapolate that data into tabulated / CSV type format. Then I want to get an LLM to really absorb the document and be able to answer referential questions about the information. For now, it seems pretty good about semantic searching where it can locate the information I'm asking about along with some local insights but I want more im depth understanding.

2 Reply Share ···

[Issue/Question] Too common keywords pollute the search in the vectorized database, any smart solutions ?

Hello guys, I have the following issue. I'm doing a chatbot with data from a website pages of an organisation. But when I ask question of the type "Who is the president of the [organisation]?" Because [organisation] is on every pages of the website, the page relative to who is the actual president does not show up in the documents found by the vectorsearch. When I ask "Who is the president" the result is correct (And the relevant page is actually in the context), so I'm sure it's a problem with [organisation] being everywhere

by another arbitrary word and his solution is not really good

abase : /ebsite-using-langchain/)

r/LangChain · Posted by u/nikhil_no_1 1 month ago

Understand why my LangChain-based Semantic PDF Search is working poorly compared to ChatGPT completion?

Hello, I followed this https://twitter.com/MisbahSy/status/1653105400403890176 to build a LangChain-based Semantic PDF Search app. However, the chatbot is unable to answer even simple questions. OTTH, when I use the code given in "ChatGPT Prompt Engineering for Developers - https://learn.deeplearning.ai/chatgpt-prompt-eng/lesson/8/chatbot", it works perfectly.

I am giving it my cover letter and then asking questions about myself.

For e.g.

LangChain-based Semantic PDF Search App:

Q) How many years of experience does Nikhil have?

A) Nikhil Utane has a lot of professional experience, as evidenced by the long list of skills he has listed.

ChatGPT Completion:

Q) How many years of experience does Nikhil have?

A) Nikhil has 22+ years of experience in diverse product development.

Could someone provide an explanation for this different behavior? Thanks.

3 Comments Award Share ···

nikhil_no_1 OP · 1 mo. ago

I found my mistake. I tried with different PDFs but used the same index. After deleting the index and trying with the single document, I am getting the results that I was looking for.

3 Reply Share ···

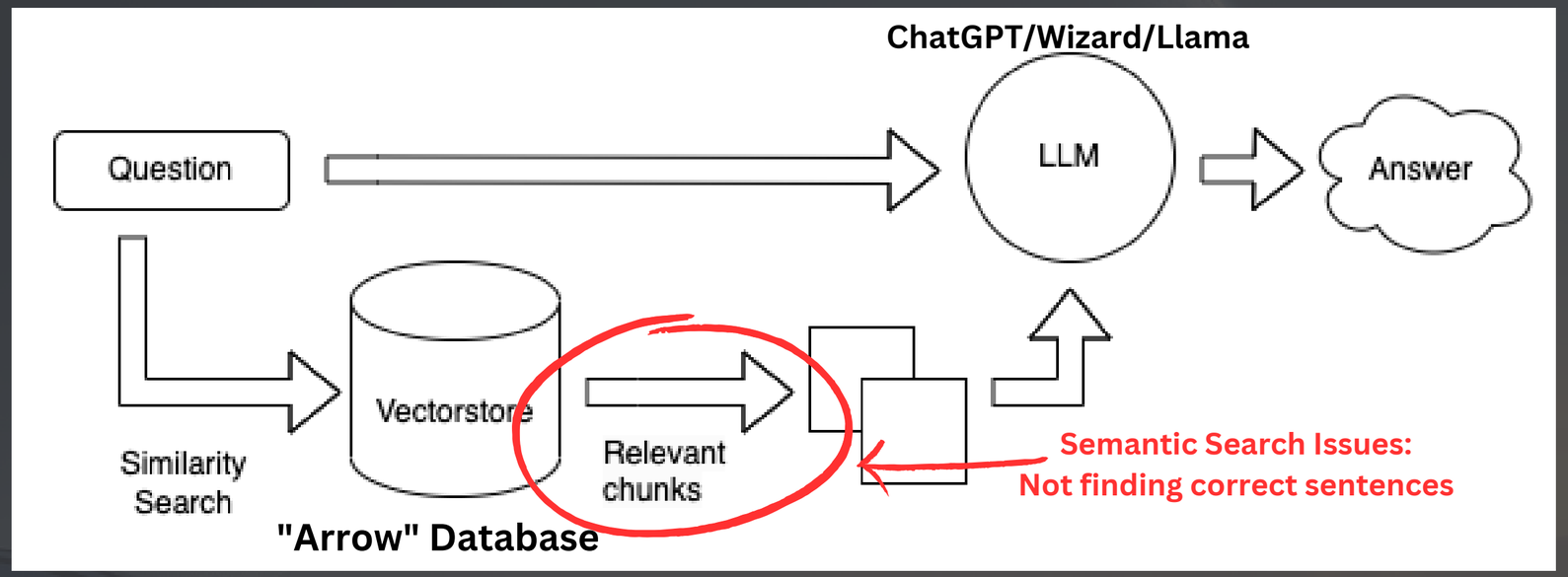
Semantic Search Issues:
Had to reduce pages/document size to get good results

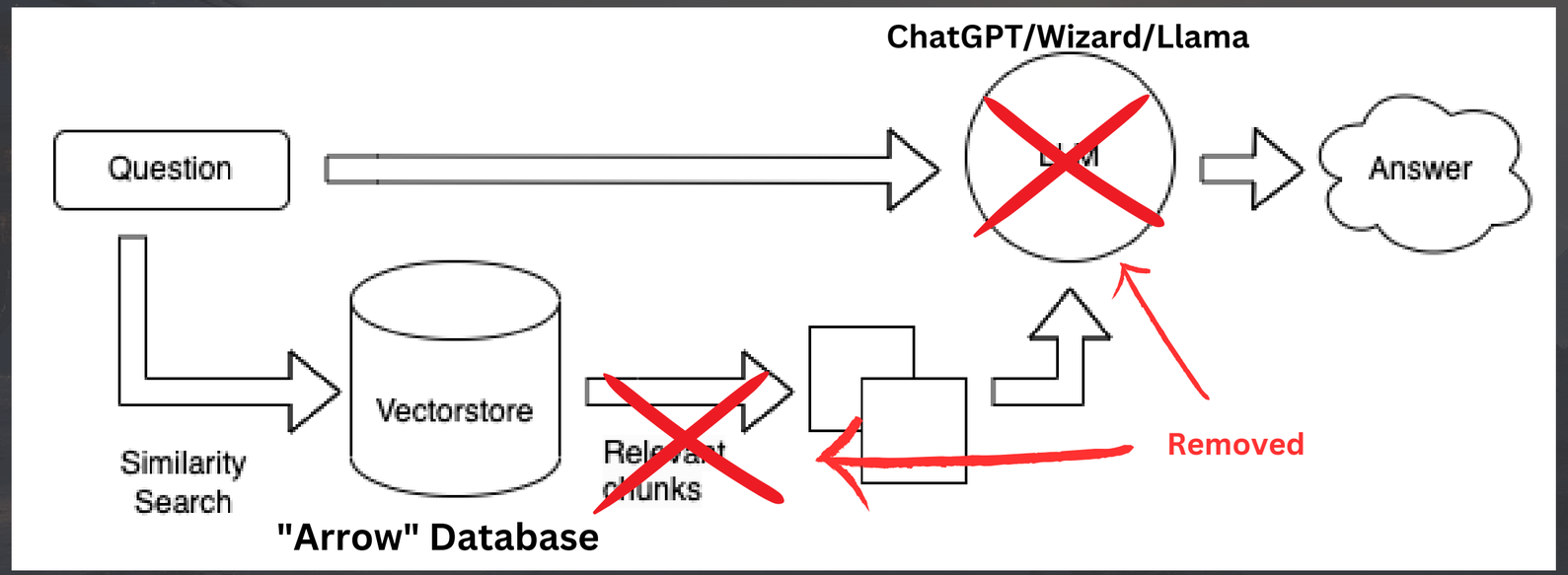
Solution?

Stop the Algo from searching

By removing the part that is unable to find the relevant chunk, and just allow users to query the vectorstore/database directly.

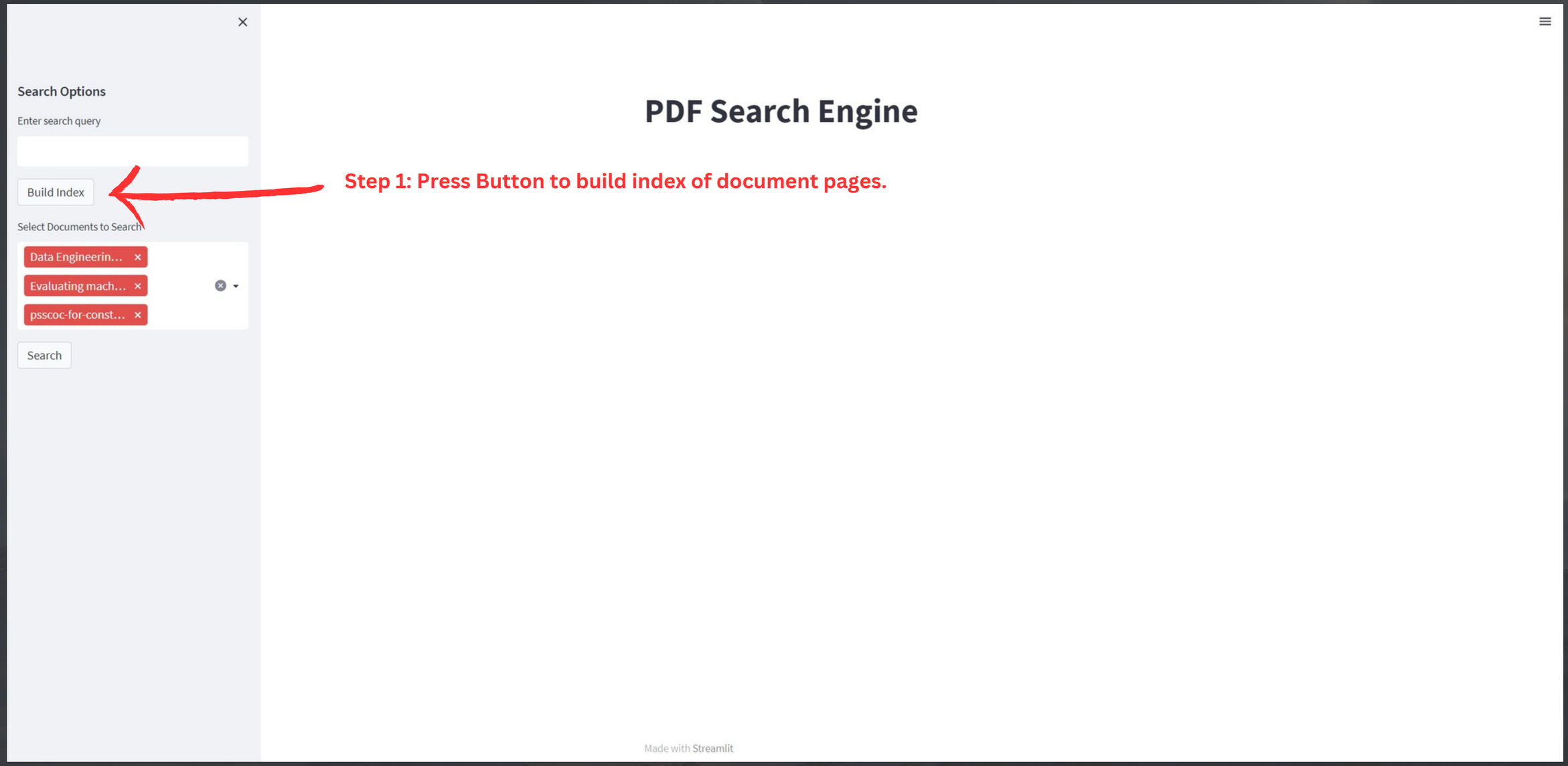
What this simply results in is essentially a search engine.



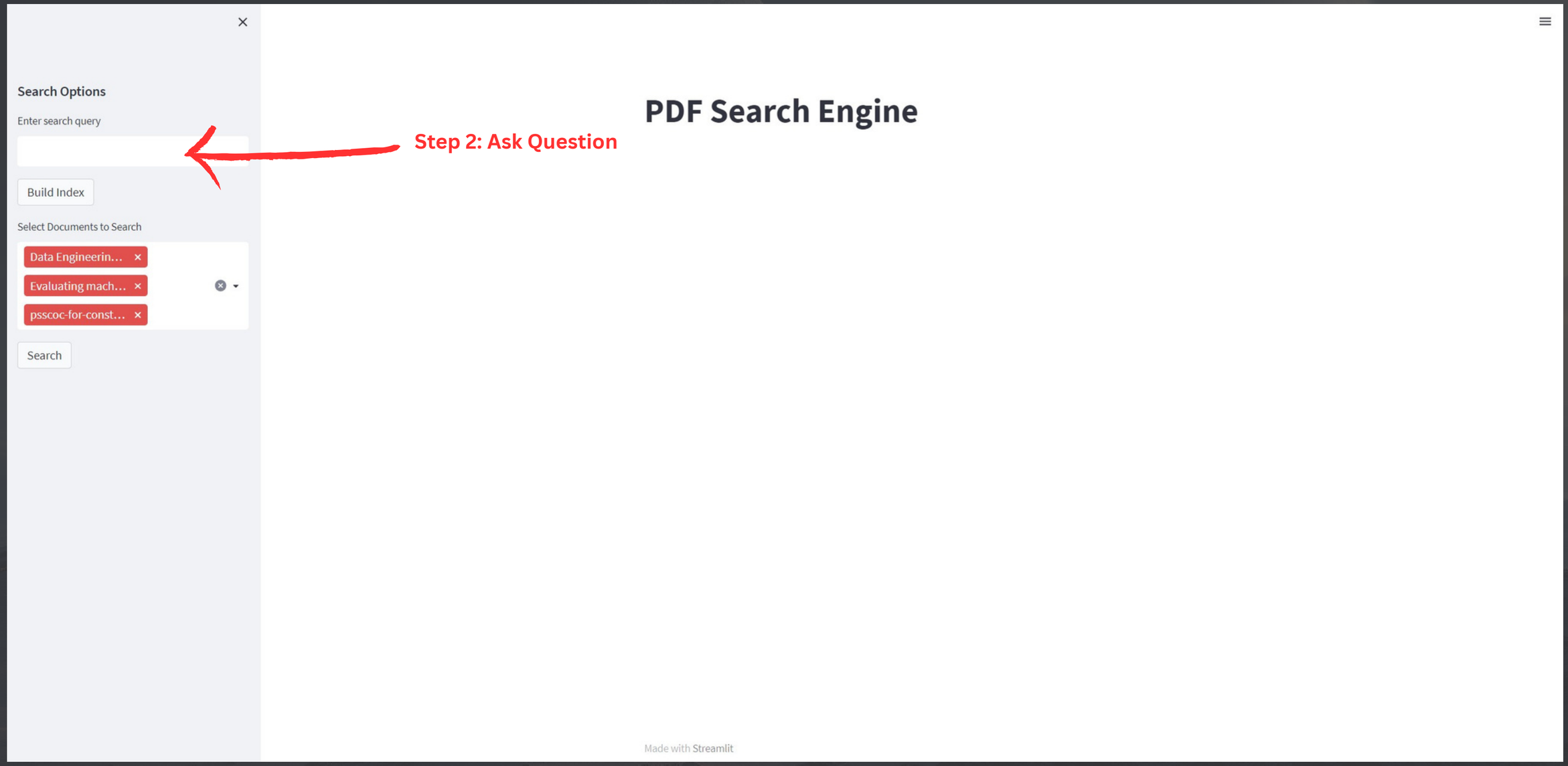


Result: PDF Search Engine

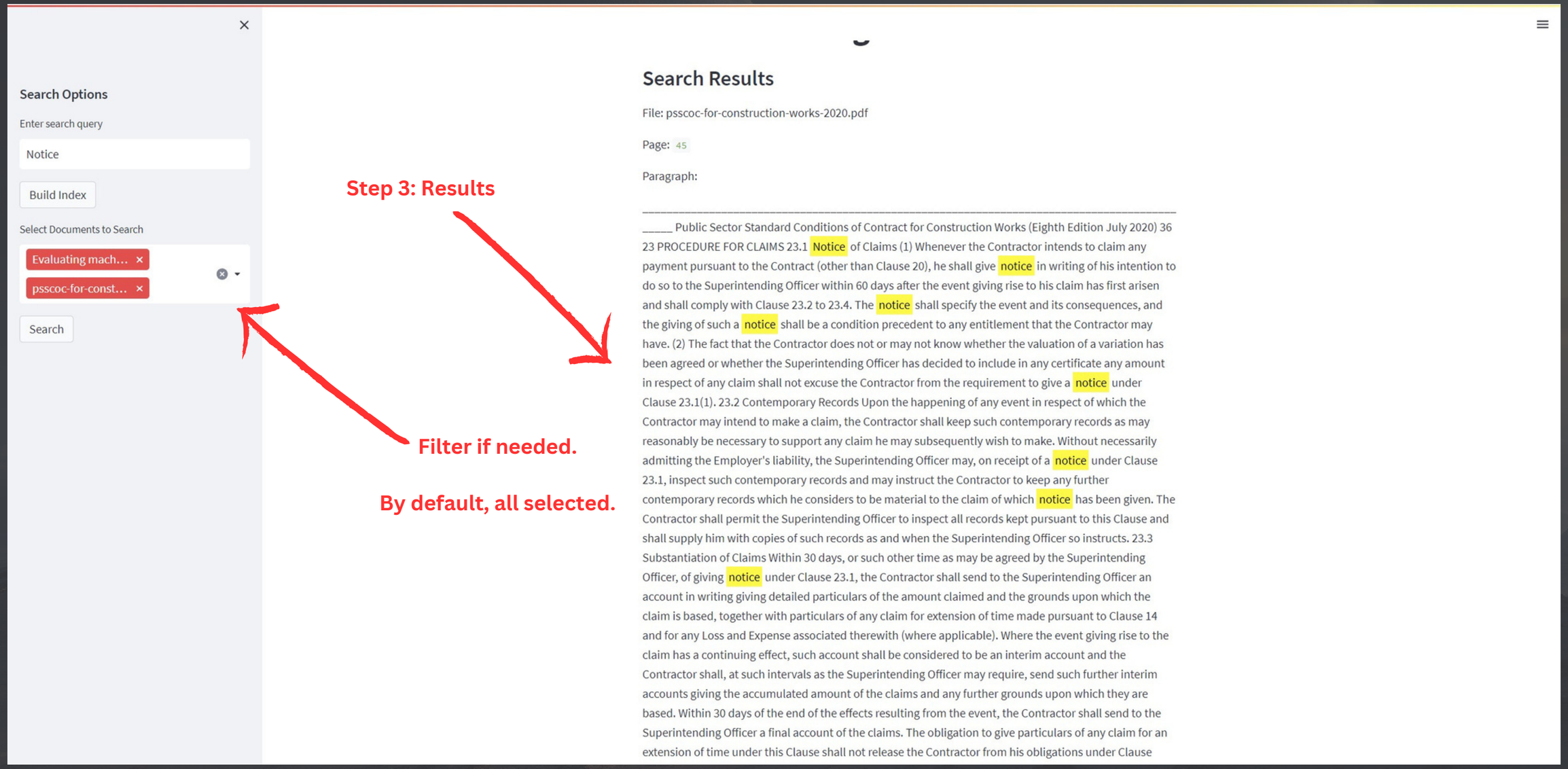
Result: PDF Search Engine



Result: PDF Search Engine



Result: PDF Search Engine



Result: PDF Search Engine

Search Options

Enter search query

Notice

Build Index

Select Documents to Search

Evaluating mach… ×

psscoc-for-const… ×

Search

Page: 16

Paragraph:

_____ Public Sector Standard Conditions of Contract for Construction Works (Eighth Edition July 2020) 7 3.4 Need for Further Drawings etc. The Contractor shall give adequate notice in writing to the Superintending Officer: (a) of any further drawing, specification or other information which the Superintending Officer is required to provide under the Contract; (b) of any drawing, specification, instruction or other information which is required by any specific time, whenever the planning or execution of the Works is likely to be delayed or disrupted by its lack, and whether or not the need for it is shown on any programme accepted by the Superintending Officer under Clause 9. The notice shall also state the consequences in terms of delay to the progress or completion of the Works or any part of the Works and any financial consequences should the Superintending Officer not comply with any of the requirements of the notice. The Superintending Officer shall on receipt of the notice comply with its requirements, provided that it is given in sufficient time for the Superintending Officer reasonably to prepare and issue the information required. 3.5 Further Supplementary Drawings etc. and Instructions The Superintending Officer shall issue to the Contractor, from time to time, such further or revised drawings, specifications or instructions as may in his opinion be necessary for the purposes of the execution and completion of the Works. The Contractor shall carry out and be bound by the same. 3.6 Delay and Time If: (a) the Contractor shall have duly given notice pursuant to Clause 3.4 and if the Superintending Officer shall not have complied with any of its requirements; or (b) the Superintending Officer shall not have issued any further or revised drawing specification or instruction as required by Clause 3.5, and if thereby the progress or completion of the Works or any part of the Works has been materially affected then, subject to compliance by the Contractor with Clauses 14, 23 and 32, the Superintending Officer may grant an extension of time pursuant to Clause 14 and may certify pursuant to Clause 32 such sum as may be reasonable in respect of any Loss and Expense incurred by the Contractor.

Download Page 16

Download the page you want!

File: psscoc-for-construction-works-2020.pdf

Page: 40

Paragraph:

THANK YOU!