Web Scraping, Database Storage, and Full-Text Search for Manufacturing Articles

Introduction

This project aims to scrape manufacturing-related articles from The Economic Times, store them in a SQLite database, and implement full-text search (FTS) and embedding-based search functionalities. The main objectives are:

1. **Web Scraping**: Extract article titles, links, and content from the webpage.
2. **Database Management**: Store the articles in a structured SQLite database.
3. **Full-Text Search**: Enable efficient retrieval of articles using SQLite's FTS5 functionality.
4. **Embedding-Based Search**: Use Sentence Transformers to generate embeddings for articles and retrieve similar articles based on query embeddings.
5. **Language Model Response Generation**: Use OpenAI's GPT-4 for summarizing and responding to user queries in a chatbot interface built using Streamlit.

Steps Followed

1.Web Scraping Using BeautifulSoup

1.1 **Getting the Webpage**: It uses the `requests` library to download the webpage's content.

1.2 **Reading the HTM**L: The `BeautifulSoup` library processes the downloaded webpage, allowing easy access to different parts of the HTML.

1.3 **Finding Article Titles and Links:** The code searches the webpage for article titles and their links, then saves them in a list for later use.

2.Database Management with SQLite

2.1 **Creating Tables**: Two tables are set up—one called "articles" to store the basic information and another called "articles_fts" for full-text searching.

2.2 **Inserting Data**: The scraped article details are then added into the database.

3. Fetching and Storing Article Content

`get_article_content` function retrieves the full text of each article by visiting its URL. It then stores the article's content in the SQLite database.

4.Full-Text Search with FTS5

4.1 **Creating a Virtual FTS5 Table**: A special table, called `articles_fts`, is created to make full-text searches faster and more efficient.

4.2 Searching Articles: Users can search for articles using keywords, and the code will return matching article titles along with brief excerpts from their content.

5 Embedding-Based Search Using Sentence-BERT

The project uses an embeddings-based search to find articles that are related to the user's query in a meaningful way. It uses Sentence-BERT to create embeddings (numeric representations) for both the articles and the user's query, and then compares them using cosine similarity to find the best matches.

1. **Creating Embeddings**: Sentence-BERT generates vector representations of the article content.
2. **Storing Embeddings**: These embeddings are saved in the SQLite database.
3. **Similarity Search**: The system compares the query's embeddings with the stored ones and retrieves the most relevant articles based on similarity scores.

6 Integration with GPT and Streamlit Interface

The project features a chatbot interface created with Streamlit, allowing users to ask questions about manufacturing and supply chains. The chatbot searches for relevant articles using either full-text search (FTS5) or embeddings, then uses GPT-4 to provide a detailed answer.

1. **RAG (Retrieval-Augmented Generation) Pipeline**: The chatbot first finds relevant articles in the database, then combines that information into a prompt for GPT-4, which generates a well-informed response.

Key Features

- **Web Scraping**: Extracts titles, links, and content from The Economic Times.
- **SQLite Storage**: Efficiently stores and retrieves articles using both traditional and FTS5 tables.
- **Full-Text Search**: Enables keyword-based retrieval of articles.
- **Embedding-Based Search**: Uses semantic search to find articles similar to a user's query.
- **Chatbot Interface**: Powered by GPT and integrated with Streamlit for interactive queries.

Libraries used

`requests`: Fetches web pages from the internet.

`BeautifulSoup`: Parses and extracts data from HTML content.

`sqlite3`: Manages the SQLite database for storing and retrieving article data.

`SentenceTransformer`: Converts article content and user queries into vector embeddings for semantic search.

`numpy`: Handles numerical operations, especially with embeddings.

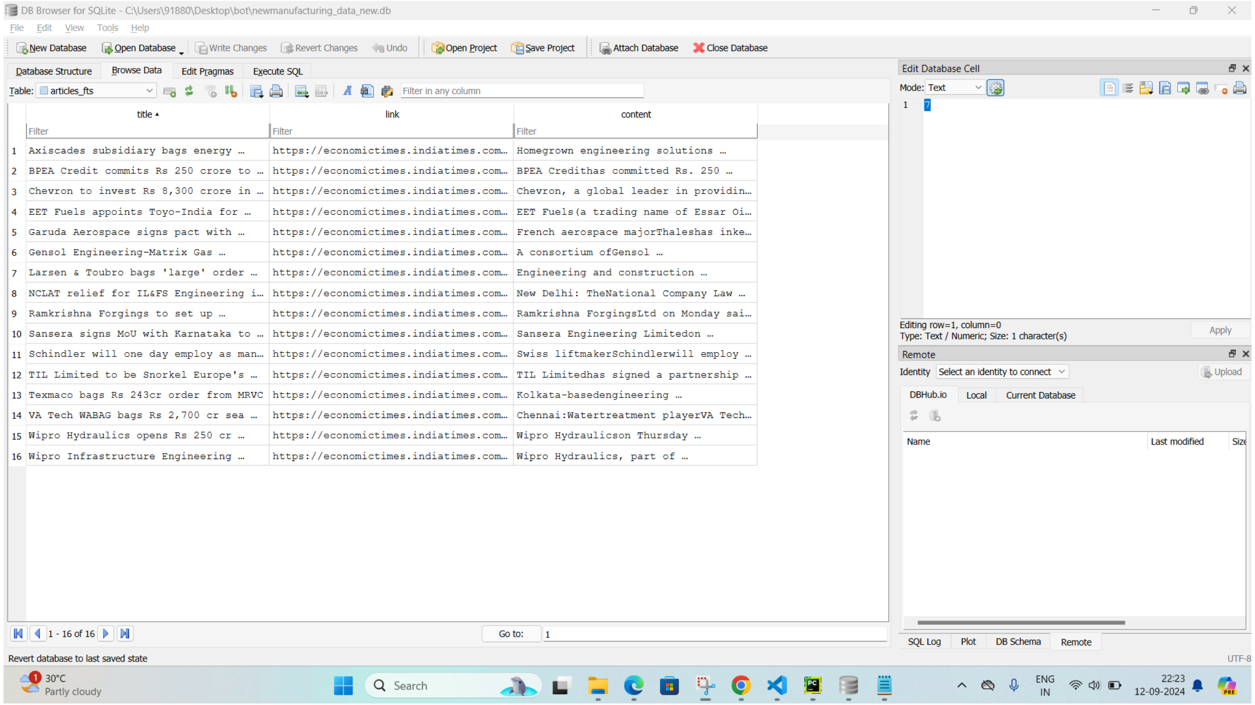
`pickle`: Serializes and deserializes data (like embeddings) for storage in the database.

`scipy.spatial.distance`: Calculates cosine similarity to compare embeddings.

`openai`: Accesses GPT to generate detailed responses based on article data.

`streamlit`: Creates a web interface for the chatbot, allowing user interaction.

Databases



DB Browser for SQLite - C:\Users\91880\Desktop\bot\newmanufacturing_data_new.db

Table: articles_fts

	title ▴	link	content
1	Axiscades subsidiary bags energy …	https://economictimes.indiatimes.com…	Homegrown engineering solutions …
2	BPEA Credit commits Rs 250 crore to …	https://economictimes.indiatimes.com…	BPEA Credithas committed Rs. 250 …
3	Chevron to invest Rs 8,300 crore in …	https://economictimes.indiatimes.com…	Chevron, a global leader in providin…
4	EET Fuels appoints Toyo-India for …	https://economictimes.indiatimes.com…	EET Fuels(a trading name of Essar Oi…
5	Garuda Aerospace signs pact with …	https://economictimes.indiatimes.com…	French aerospace majorThaleshas inke…
6	Gensol Engineering-Matrix Gas …	https://economictimes.indiatimes.com…	A consortium ofGensol …
7	Larsen & Toubro bags 'large' order …	https://economictimes.indiatimes.com…	Engineering and construction …
8	NCLAT relief for IL&FS Engineering i…	https://economictimes.indiatimes.com…	New Delhi: TheNational Company Law …
9	Ramkrishna Forgings to set up …	https://economictimes.indiatimes.com…	Ramkrishna ForgingsLtd on Monday sai…
10	Sansera signs MoU with Karnataka to …	https://economictimes.indiatimes.com…	Sansera Engineering Limitedon …
11	Schindler will one day employ as man…	https://economictimes.indiatimes.com…	Swiss liftmakerSchindlerwill employ …
12	TIL Limited to be Snorkel Europe's …	https://economictimes.indiatimes.com…	TIL Limitedhas signed a partnership …
13	Texmaco bags Rs 243cr order from MRVC	https://economictimes.indiatimes.com…	Kolkata-basedengineering …
14	VA Tech WABAG bags Rs 2,700 cr sea …	https://economictimes.indiatimes.com…	Chennai:Watertreatment playerVA Tech…
15	Wipro Hydraulics opens Rs 250 cr …	https://economictimes.indiatimes.com…	Wipro Hydraulicson Thursday …
16	Wipro Infrastructure Engineering …	https://economictimes.indiatimes.com…	Wipro Hydraulics, part of …

1 - 16 of 16

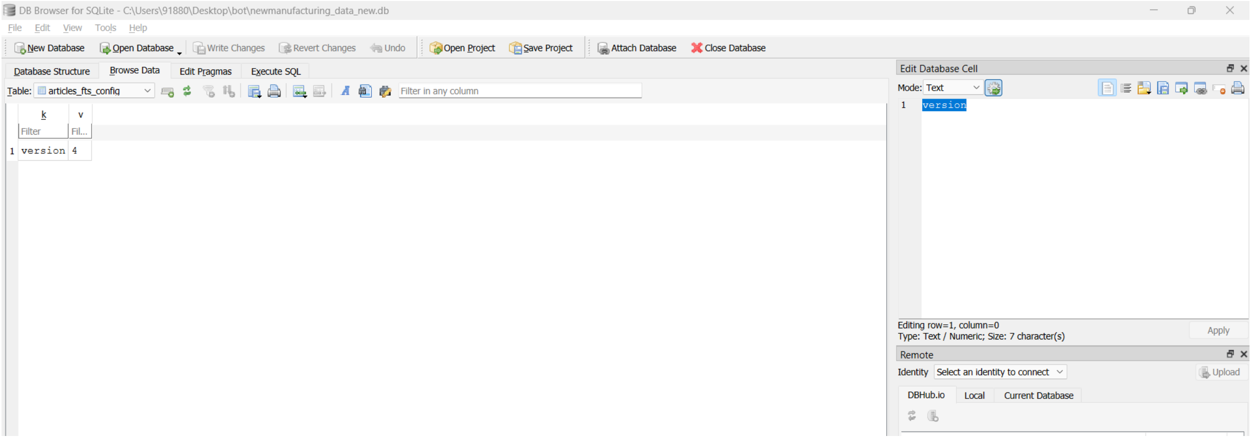
Mode: Text

Editing row=1, column=0
Type: Text / Numeric; Size: 1 character(s)

Apply

Identity: Select an identity to connect

Revert database to last saved state



DB Browser for SQLite - C:\Users\91880\Desktop\bot\newmanufacturing_data_new.db

Table: articles_fts_config

	k	v
1	version	4

Mode: Text

1 version

Editing row=1, column=0
Type: Text / Numeric; Size: 7 character(s)

Apply

Identity: Select an identity to connect

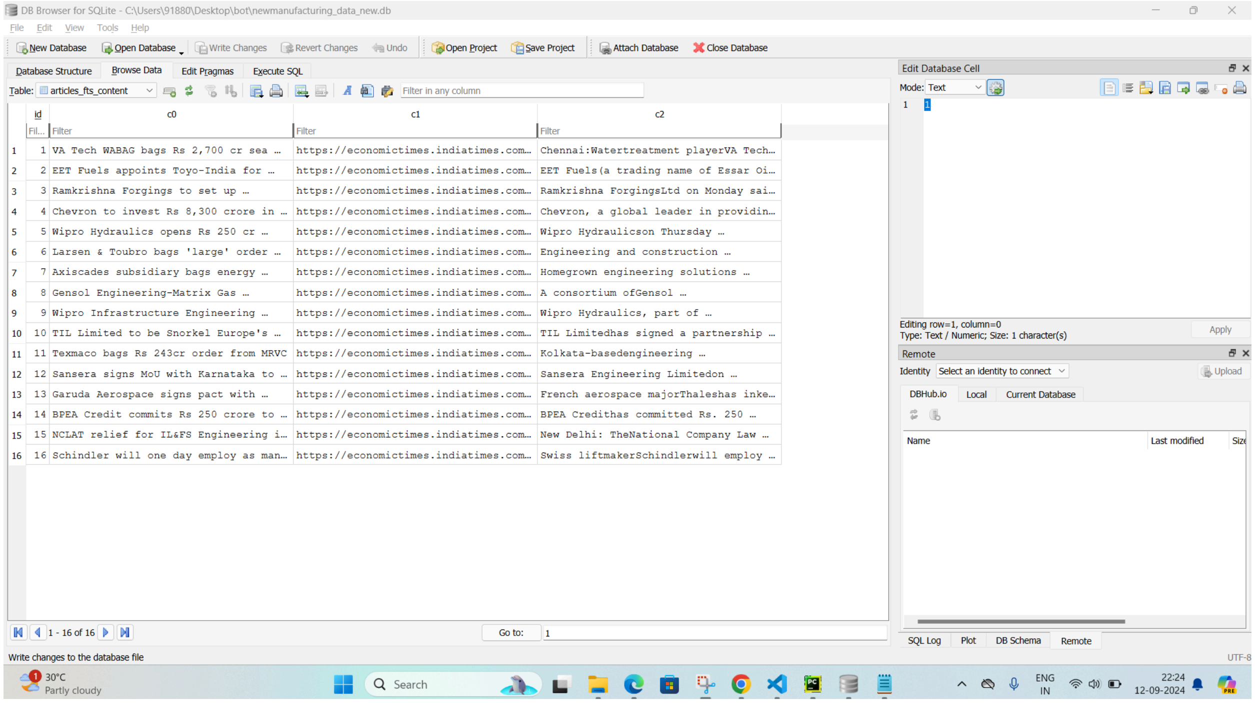


Table: articles_fts_content

Filter in any column

	id	c0	c1	c2
1	1	VA Tech WABAG bags Rs 2,700 cr sea …	https://economictimes.indiatimes.com…	Chennai:Watertreatment playerVA Tech…
2	2	EET Fuels appoints Toyo-India for …	https://economictimes.indiatimes.com…	EET Fuels(a trading name of Essar Oi…
3	3	Ramkrishna Forgings to set up …	https://economictimes.indiatimes.com…	Ramkrishna ForgingsLtd on Monday sai…
4	4	Chevron to invest Rs 8,300 crore in …	https://economictimes.indiatimes.com…	Chevron, a global leader in providin…
5	5	Wipro Hydraulics opens Rs 250 cr …	https://economictimes.indiatimes.com…	Wipro Hydraulicson Thursday …
6	6	Larsen & Toubro bags 'large' order …	https://economictimes.indiatimes.com…	Engineering and construction …
7	7	Axiscades subsidiary bags energy …	https://economictimes.indiatimes.com…	Homegrown engineering solutions …
8	8	Gensol Engineering-Matrix Gas …	https://economictimes.indiatimes.com…	A consortium ofGensol …
9	9	Wipro Infrastructure Engineering …	https://economictimes.indiatimes.com…	Wipro Hydraulics, part of …
10	10	TIL Limited to be Snorkel Europe's …	https://economictimes.indiatimes.com…	TIL Limitedhas signed a partnership …
11	11	Texmaco bags Rs 243cr order from MRVC	https://economictimes.indiatimes.com…	Kolkata-basedengineering …
12	12	Sansera signs MoU with Karnataka to …	https://economictimes.indiatimes.com…	Sansera Engineering Limitedon …
13	13	Garuda Aerospace signs pact with …	https://economictimes.indiatimes.com…	French aerospace majorThaleshas inke…
14	14	BPEA Credit commits Rs 250 crore to …	https://economictimes.indiatimes.com…	BPEA Credithas committed Rs. 250 …
15	15	NCLAT relief for IL&FS Engineering i…	https://economictimes.indiatimes.com…	New Delhi: TheNational Company Law …
16	16	Schindler will one day employ as man…	https://economictimes.indiatimes.com…	Swiss liftmakerSchindlerwill employ …

1 - 16 of 16

Go to: 1

Mode: Text

Editing row=1, column=0
Type: Text / Numeric; Size: 1 character(s)

Identity: Select an identity to connect

Name	Last modified	Size

Write changes to the database file

UTF-8

30°C
Partly cloudy

ENG IN 22:24 12-09-2024

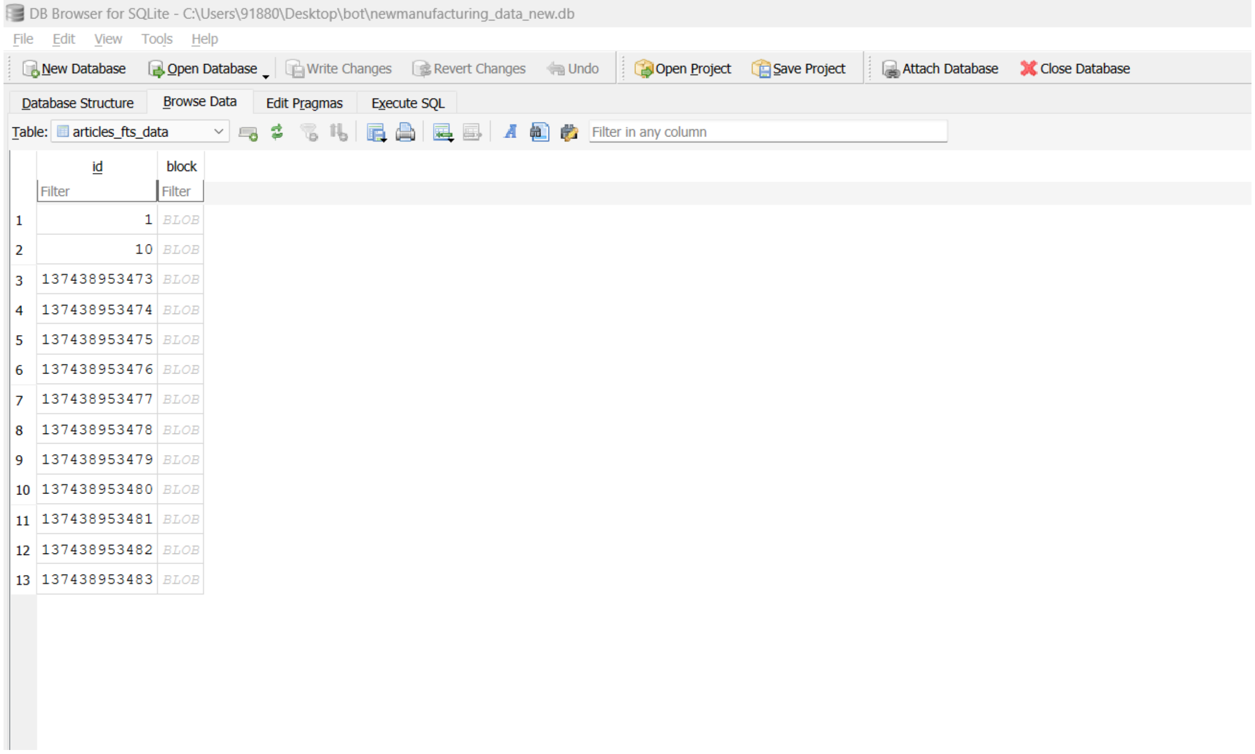


Table: articles_fts_data

Filter in any column

	id	block
1	1	BLOB
2	10	BLOB
3	137438953473	BLOB
4	137438953474	BLOB
5	137438953475	BLOB
6	137438953476	BLOB
7	137438953477	BLOB
8	137438953478	BLOB
9	137438953479	BLOB
10	137438953480	BLOB
11	137438953481	BLOB
12	137438953482	BLOB
13	137438953483	BLOB

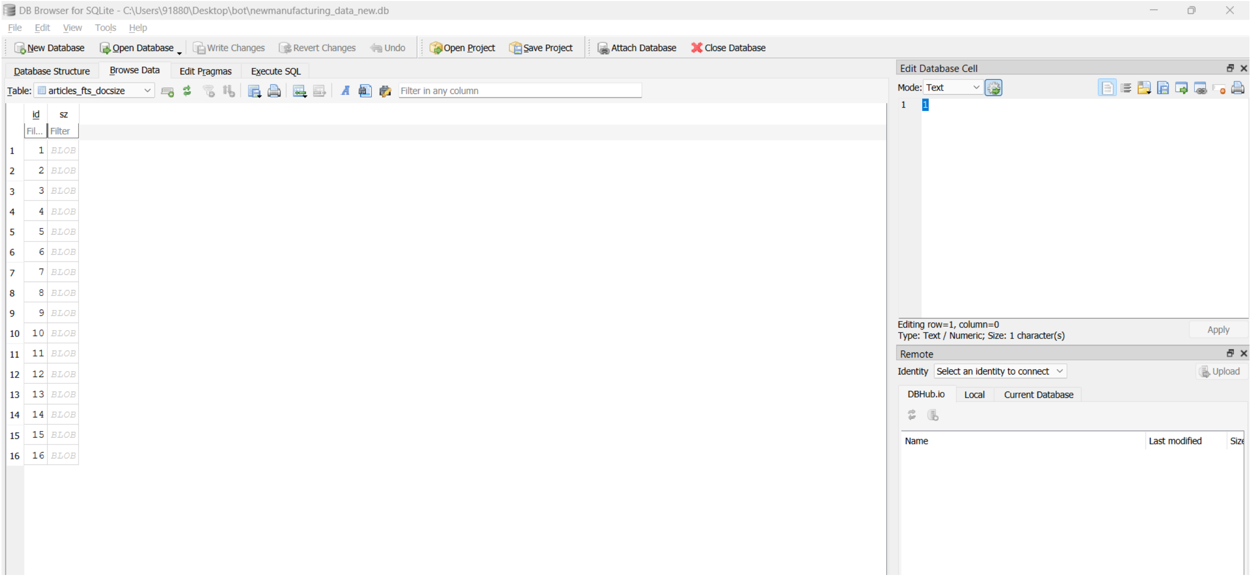


Table: articles_fts_docsize

Mode: Text

Editing row=1, column=0
Type: Text / Numeric; Size: 1 character(s)

Identity: Select an identity to connect

	id	sz
1	1	BLOB
2	2	BLOB
3	3	BLOB
4	4	BLOB
5	5	BLOB
6	6	BLOB
7	7	BLOB
8	8	BLOB
9	9	BLOB
10	10	BLOB
11	11	BLOB
12	12	BLOB
13	13	BLOB
14	14	BLOB
15	15	BLOB
16	16	BLOB

```
1  1
```

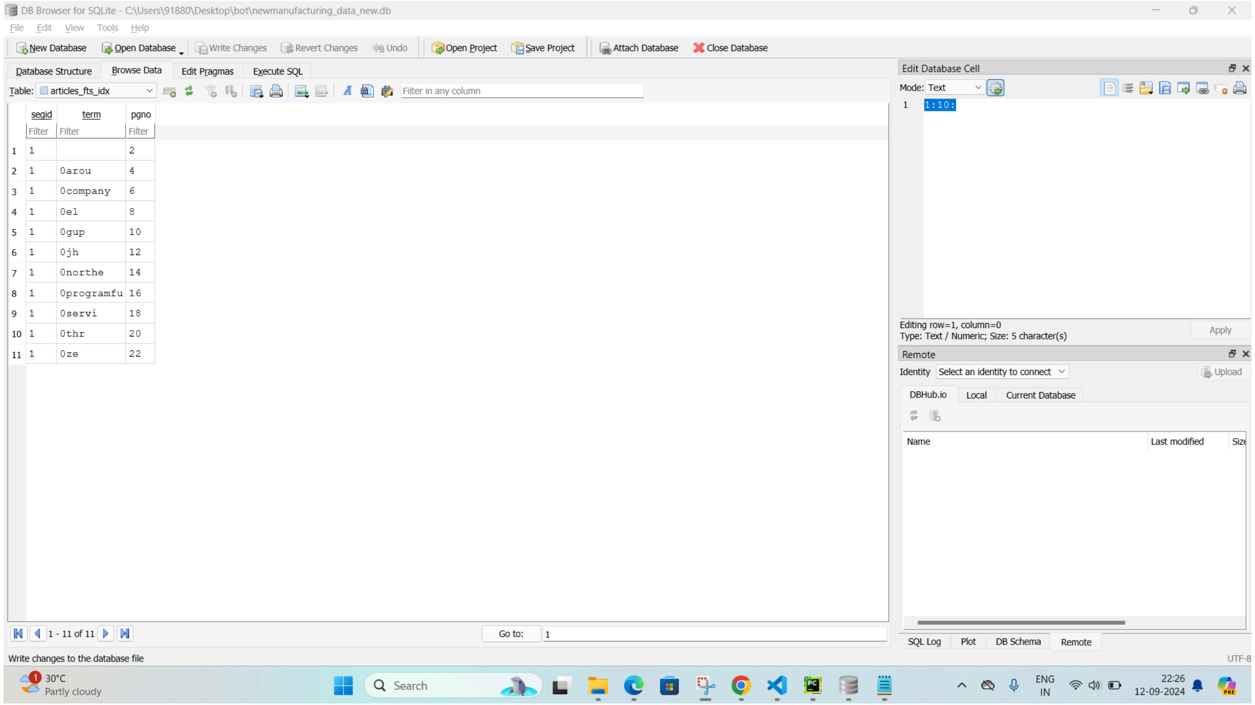


Table: articles_fts_idx

Mode: Text

Editing row=1, column=0
Type: Text / Numeric; Size: 5 character(s)

Identity: Select an identity to connect

	segid	term	pgno
1	1		2
2	1	0arou	4
3	1	0company	6
4	1	0el	8
5	1	0gup	10
6	1	0jh	12
7	1	0northe	14
8	1	0programfu	16
9	1	0servi	18
10	1	0thr	20
11	1	0ze	22

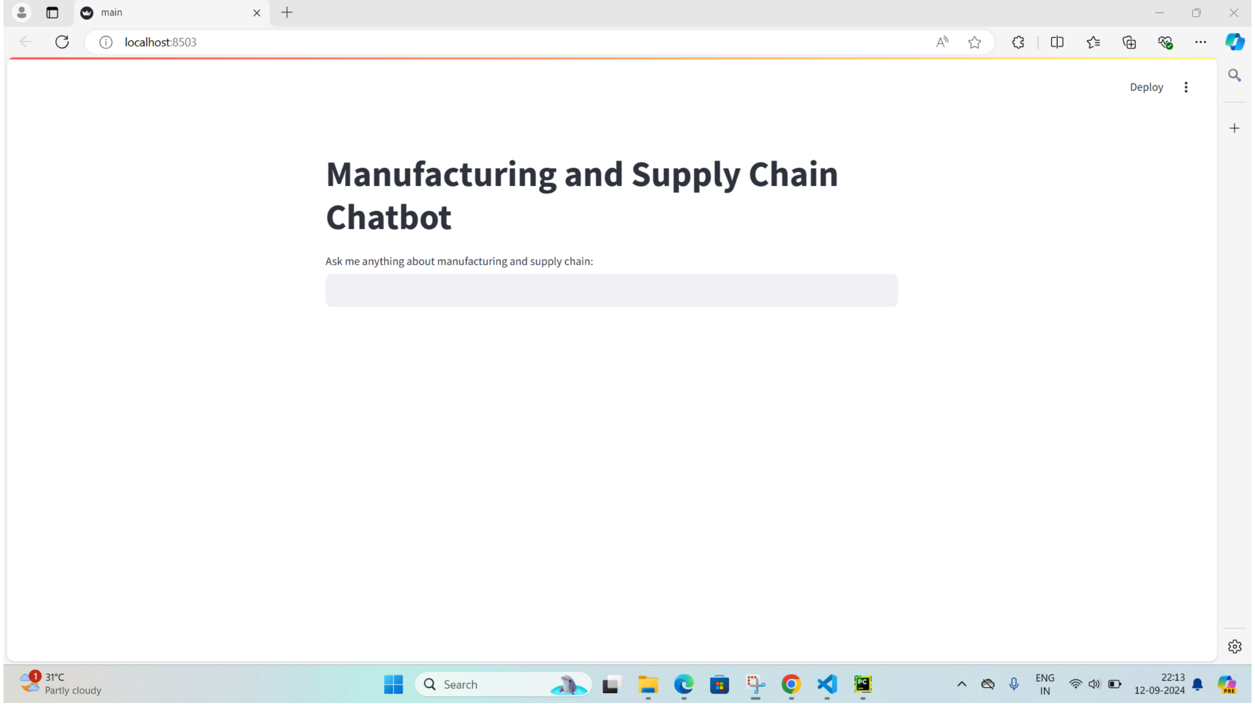
```
1  1:10:
```

1 - 11 of 11

Go to: 1

Write changes to the database file

Result



Manufacturing and Supply Chain Chatbot

Ask me anything about manufacturing and supply chain: