rXiv Papers Database

Introduction

The "rXiv Papers Database Builder" is a Python script designed to efficiently process and organize academic paper data. It reads JSON data from a file (e.g., 'arXiv21.json') and populates a SQLite database ('rXivPapers.db') with information about academic papers, authors, citations, and submissions to arXiv. This script lays the foundation for further analysis and exploration of academic research data.

Part 1:

Data Import

1. “read_json_file.py”:
   - This script reads JSON data and imports it into 'rXivPapers.db'.
   - To access the database and its tables, use the 'Table-Data' folder provided.
   - Import the tables with their data by opening 'rXivPapers.db' via SQLite, then import TSV files.

2. “categoriesextract.py”:
   - This script is used in task 6 (3.6.sql) to create a unique categories view from the categories list.
   - This view, named "UniqueCategoriesView," is referenced in SQL script '3.6.sql'.

Database Import:

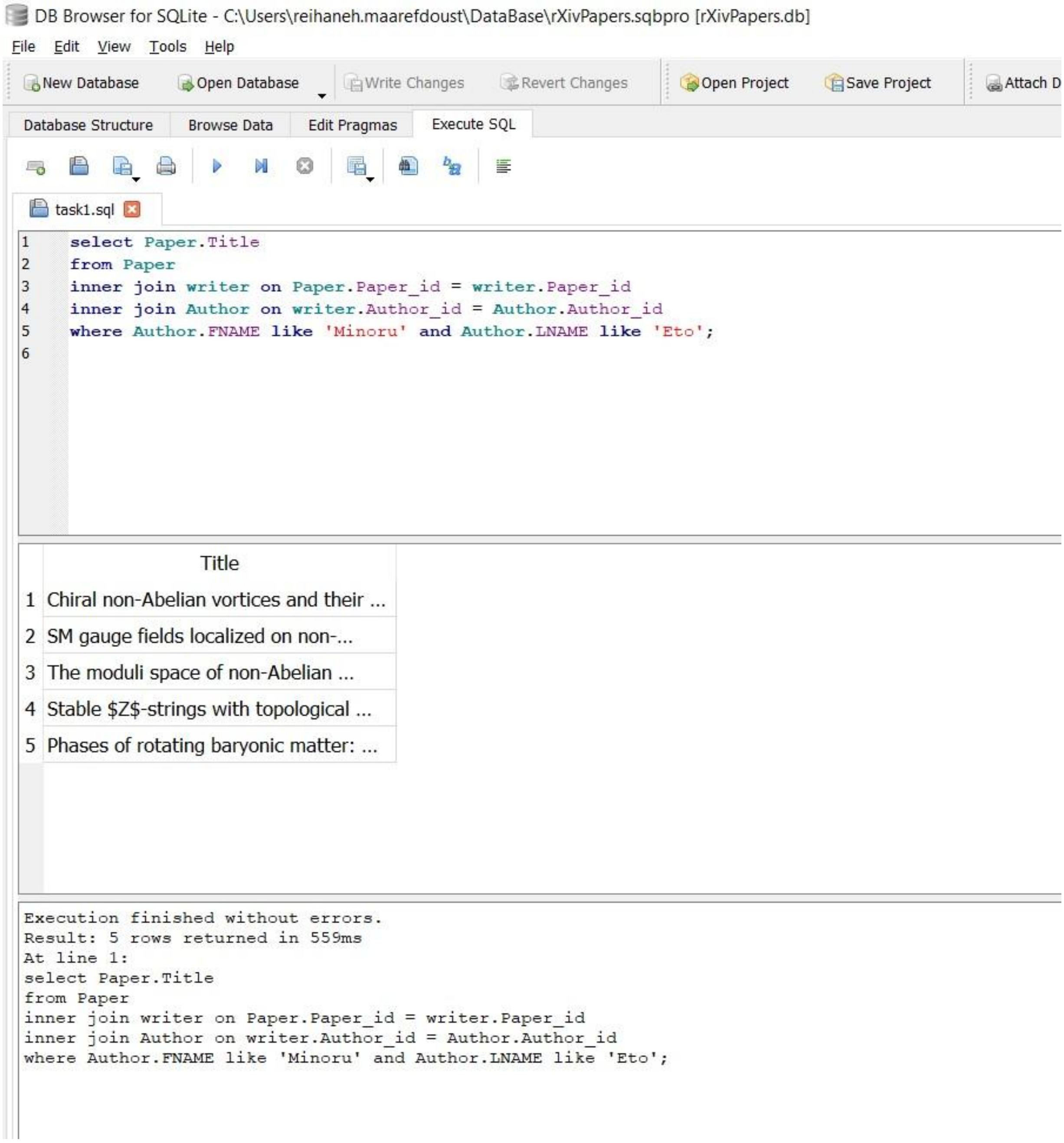
Begin by opening 'rXivPapers.db' (from the “DataBase” folder) using SQLite. Next, import the data from tables and views located in the "Table&Views-DataFiles" folder.

SQL Scripts

All SQL commands are provided as separate SQL files:
- '3.1.sql'
- '3.2.sql'
- '3.3.sql'(3.3-1.sql, 3.3-2.sql,3.3-3.sql)
- '3.4.sql'
- '3.5.sql'(3.5-1.sql, 3.5-2.sql)
- '3.6.sql'
- '3.7.sql'
- '3.8.sql'
- '3.9.sql'
- '3.10.sql'(3.10-1.sql, 3.10-2.sql)

These SQL scripts allow you to perform various queries and operations on the 'rXivPapers.db' database. Use these scripts to retrieve, manipulate, and analyze academic paper data based on your specific research needs.

3.1.sql

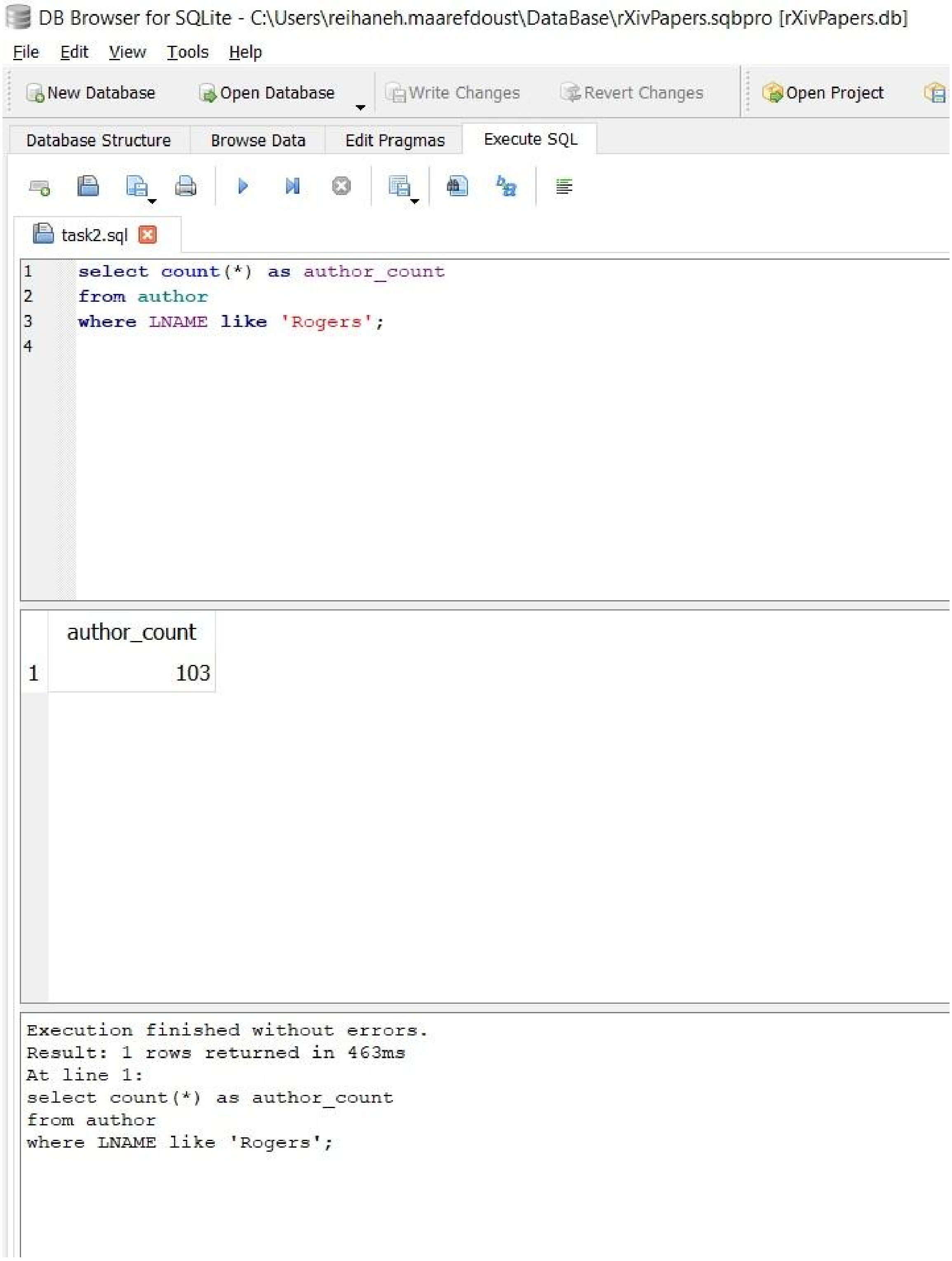


```sql
select Paper.Title
from Paper
inner join writer on Paper.Paper_id = writer.Paper_id
inner join Author on writer.Author_id = Author.Author_id
where Author.FNAME like 'Minoru' and Author.LNAME like 'Eto';
```

	Title
1	Chiral non-Abelian vortices and their ...
2	SM gauge fields localized on non-...
3	The moduli space of non-Abelian ...
4	Stable Z-strings with topological ...
5	Phases of rotating baryonic matter: ...

```
Execution finished without errors.
Result: 5 rows returned in 559ms
At line 1:
select Paper.Title
from Paper
inner join writer on Paper.Paper_id = writer.Paper_id
inner join Author on writer.Author_id = Author.Author_id
where Author.FNAME like 'Minoru' and Author.LNAME like 'Eto';
```

3.2.sql

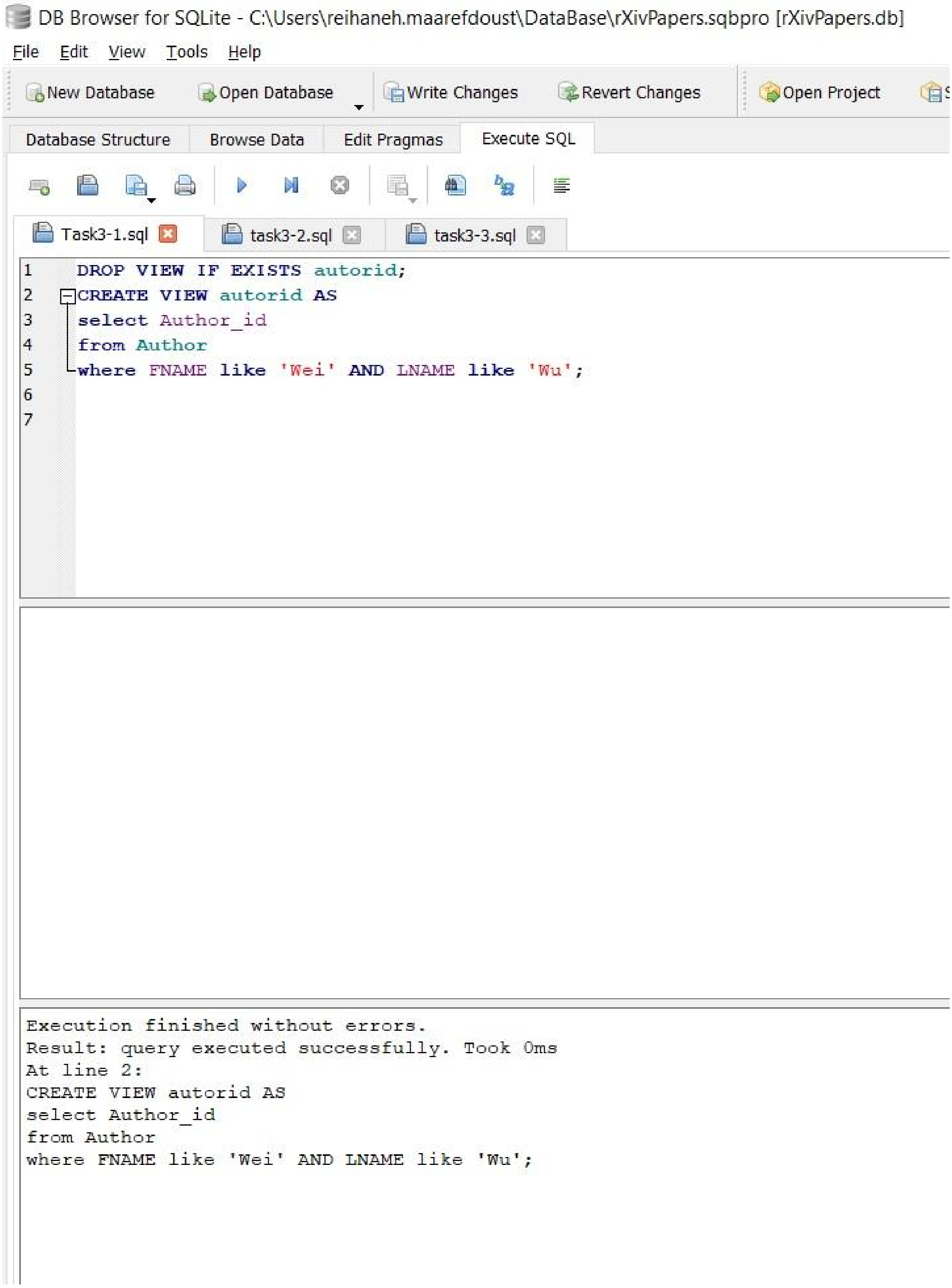


```sql
select count(*) as author_count
from author
where LNAME like 'Rogers';
```

	author_count
1	103

```
Execution finished without errors.
Result: 1 rows returned in 463ms
At line 1:
select count(*) as author_count
from author
where LNAME like 'Rogers';
```

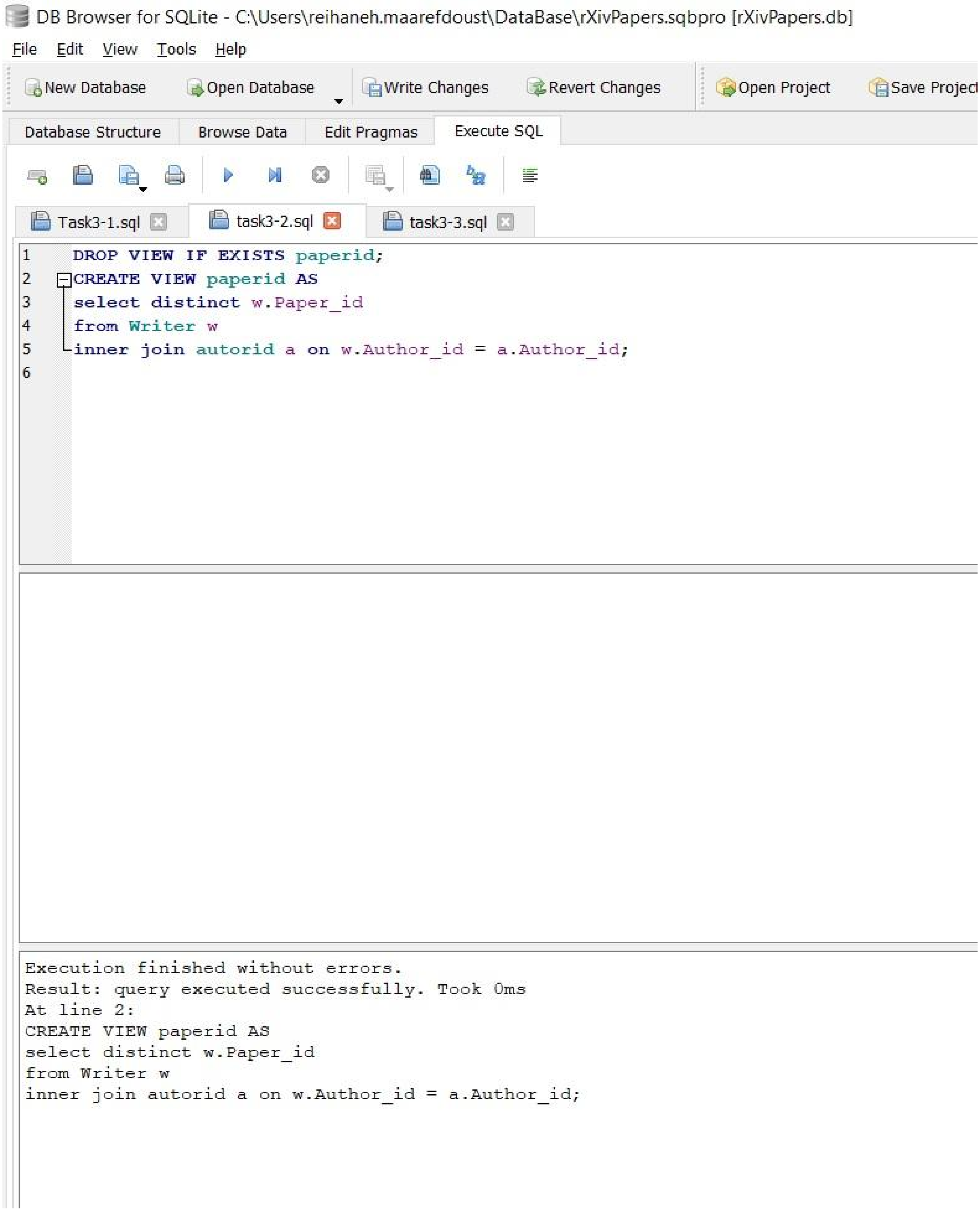
3.3-1.sql



```
DROP VIEW IF EXISTS autorid;
CREATE VIEW autorid AS
select Author_id
from Author
where FNAME like 'Wei' AND LNAME like 'Wu';
```

```
Execution finished without errors.
Result: query executed successfully. Took 0ms
At line 2:
CREATE VIEW autorid AS
select Author_id
from Author
where FNAME like 'Wei' AND LNAME like 'Wu';
```

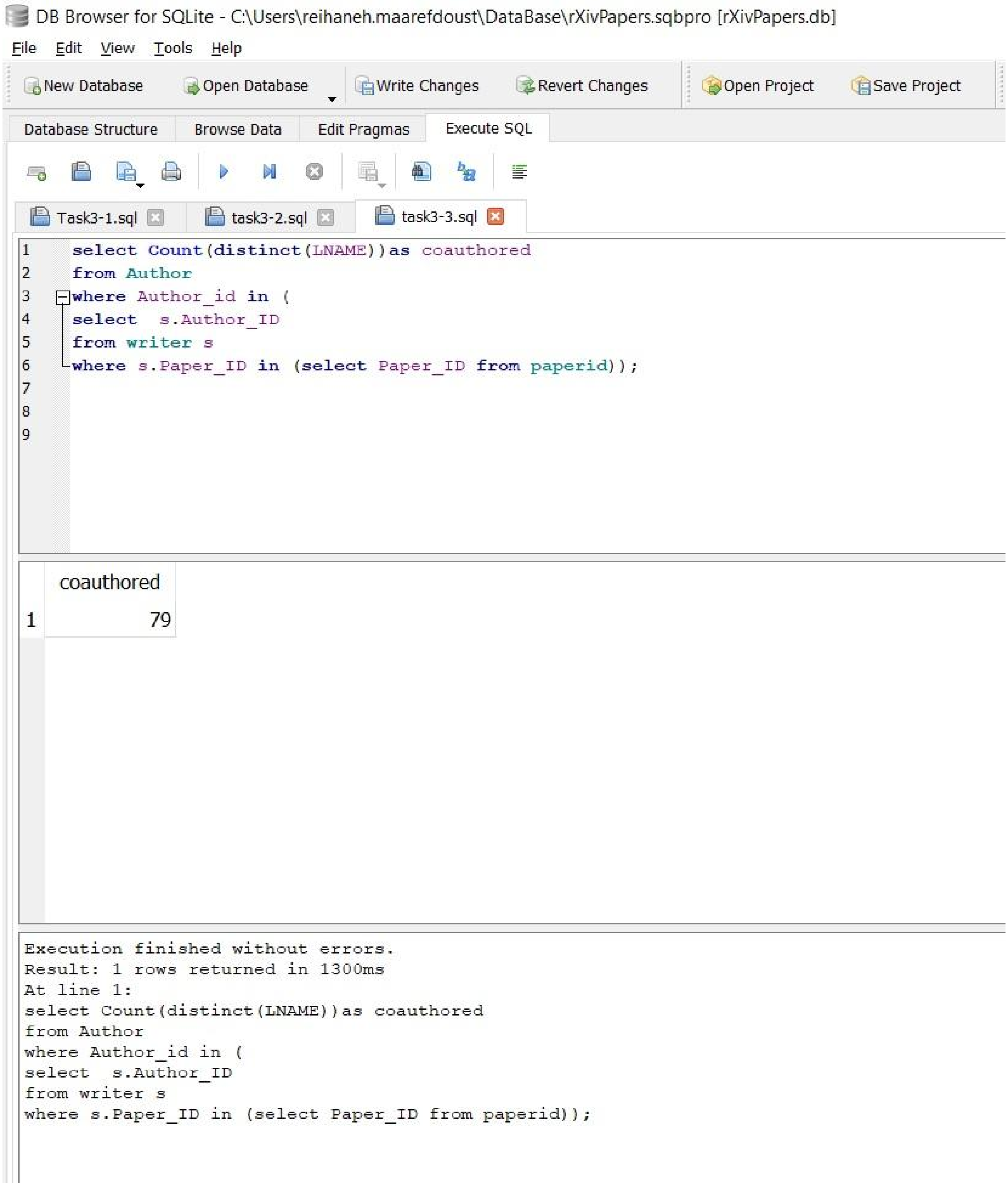
3.3-2.sql



```sql
DROP VIEW IF EXISTS paperid;
CREATE VIEW paperid AS
select distinct w.Paper_id
from Writer w
inner join autorid a on w.Author_id = a.Author_id;
```

```
Execution finished without errors.
Result: query executed successfully. Took 0ms
At line 2:
CREATE VIEW paperid AS
select distinct w.Paper_id
from Writer w
inner join autorid a on w.Author_id = a.Author_id;
```

3.3-3.sql

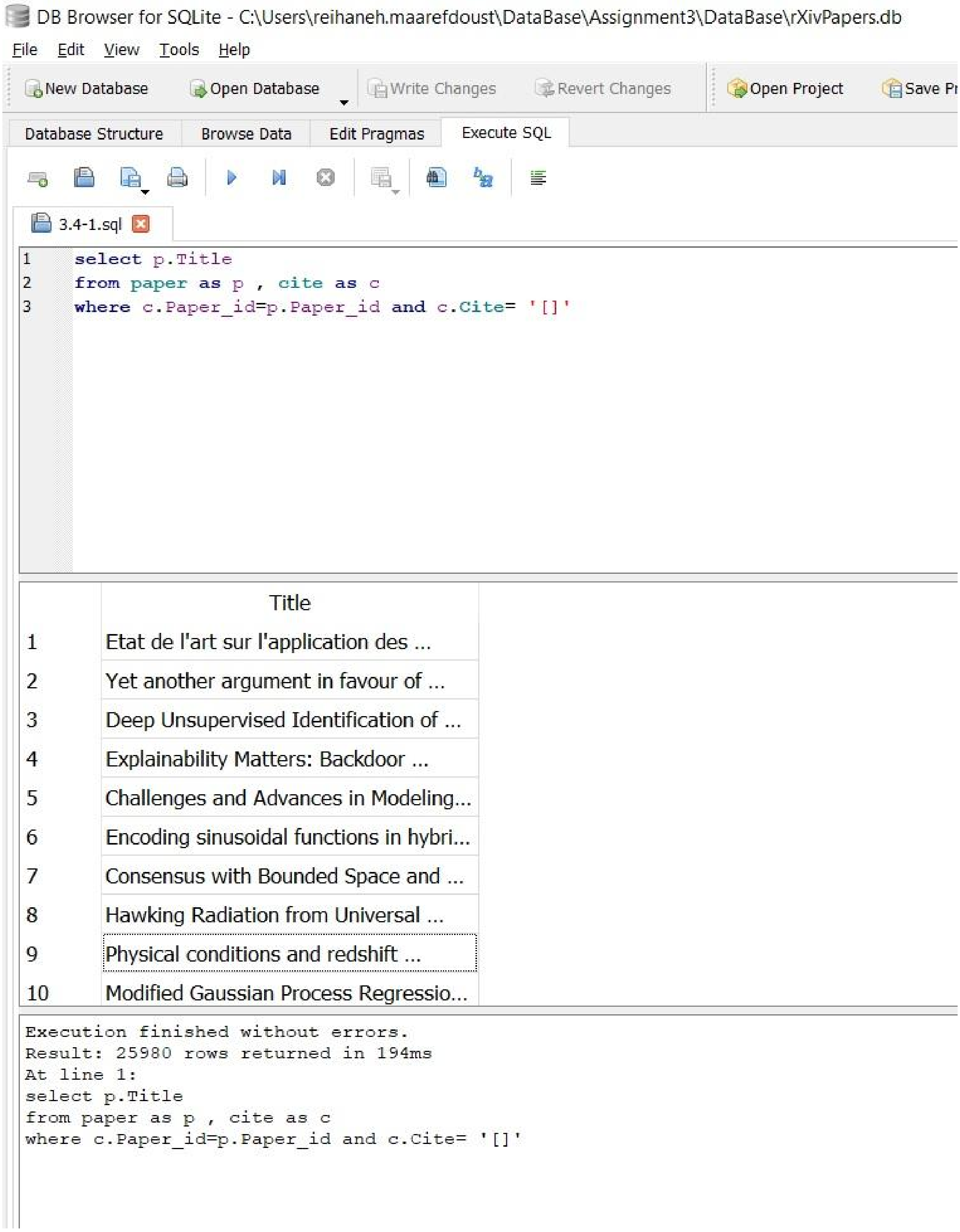


```sql
select Count(distinct(LNAME))as coauthored
from Author
where Author_id in (
select  s.Author_ID
from writer s
where s.Paper_ID in (select Paper_ID from paperid));
```

```
  coauthored
1         79
```

```
Execution finished without errors.
Result: 1 rows returned in 1300ms
At line 1:
select Count(distinct(LNAME))as coauthored
from Author
where Author_id in (
select  s.Author_ID
from writer s
where s.Paper_ID in (select Paper_ID from paperid));
```

3.4-1.sql (papers that have no references)

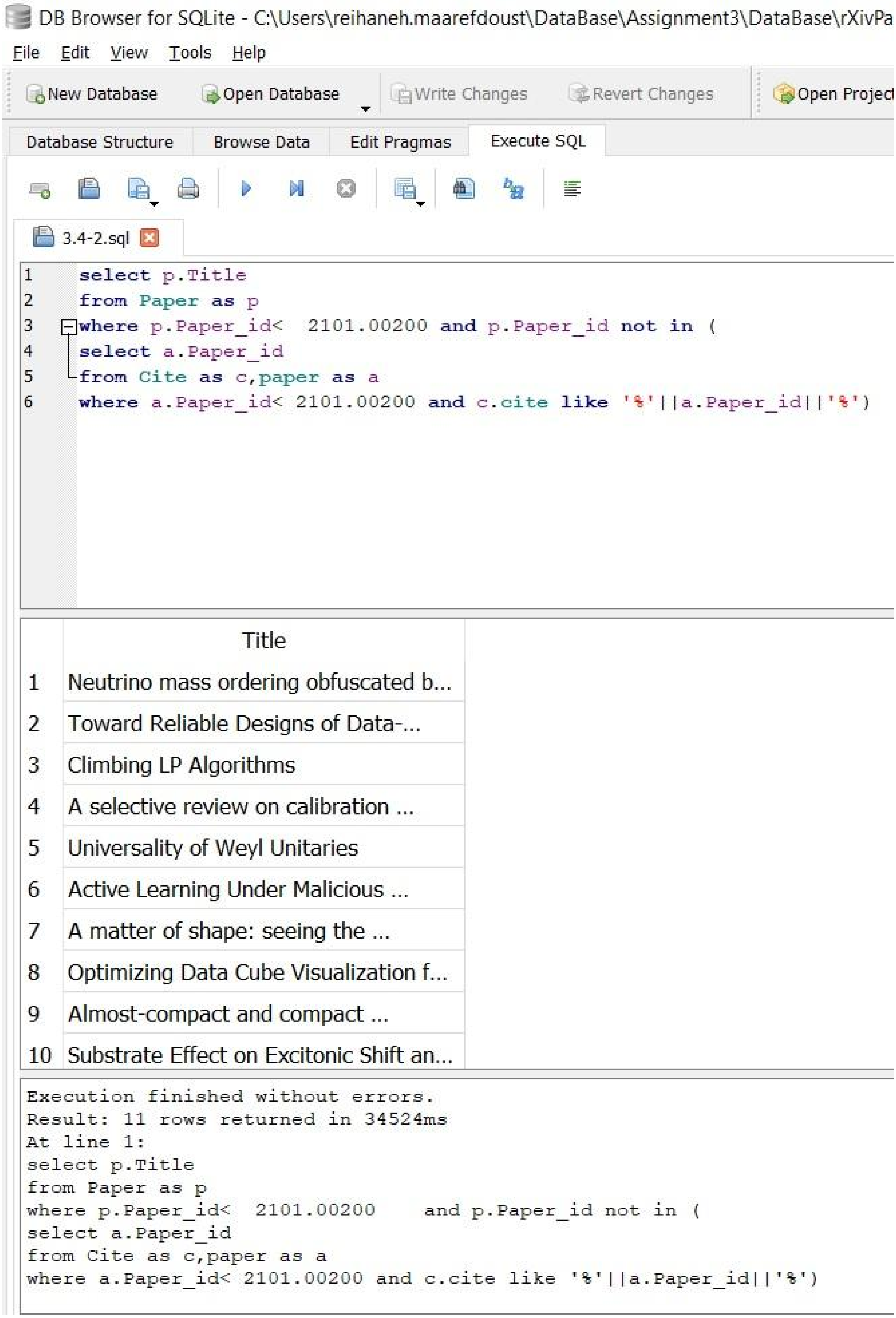


```sql
select p.Title
from paper as p , cite as c
where c.Paper_id=p.Paper_id and c.Cite= '[]'
```

	Title
1	Etat de l'art sur l'application des ...
2	Yet another argument in favour of ...
3	Deep Unsupervised Identification of ...
4	Explainability Matters: Backdoor ...
5	Challenges and Advances in Modeling...
6	Encoding sinusoidal functions in hybri...
7	Consensus with Bounded Space and ...
8	Hawking Radiation from Universal ...
9	Physical conditions and redshift ...
10	Modified Gaussian Process Regressio...

```
Execution finished without errors.
Result: 25980 rows returned in 194ms
At line 1:
select p.Title
from paper as p , cite as c
where c.Paper_id=p.Paper_id and c.Cite= '[]'
```

3.4-2.sql (papers not use as references in any paper)

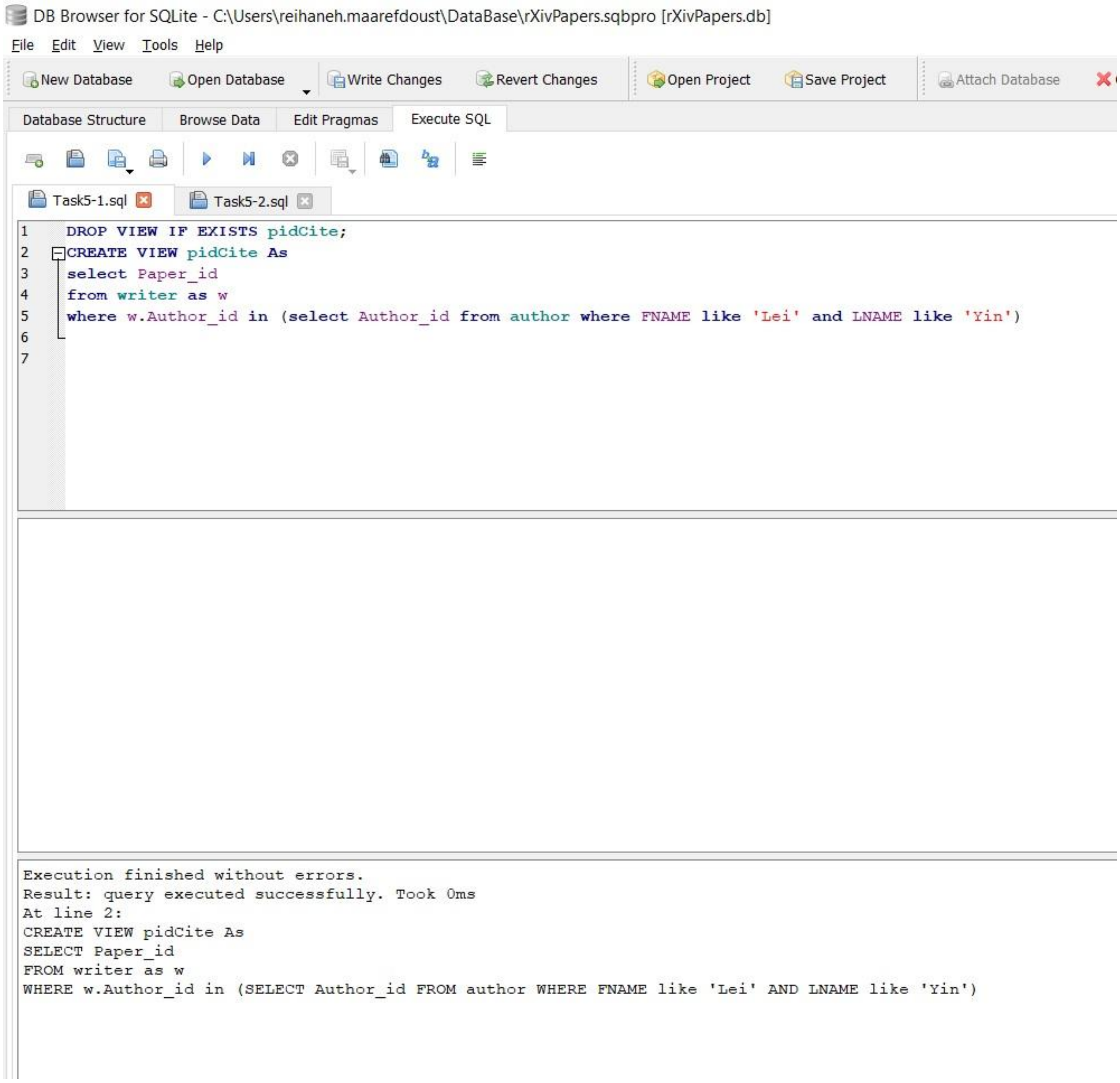


```sql
select p.Title
from Paper as p
where p.Paper_id<  2101.00200 and p.Paper_id not in (
select a.Paper_id
from Cite as c,paper as a
where a.Paper_id< 2101.00200 and c.cite like '%'||a.Paper_id||'%')
```

	Title
1	Neutrino mass ordering obfuscated b...
2	Toward Reliable Designs of Data-...
3	Climbing LP Algorithms
4	A selective review on calibration ...
5	Universality of Weyl Unitaries
6	Active Learning Under Malicious ...
7	A matter of shape: seeing the ...
8	Optimizing Data Cube Visualization f...
9	Almost-compact and compact ...
10	Substrate Effect on Excitonic Shift an...

```
Execution finished without errors.
Result: 11 rows returned in 34524ms
At line 1:
select p.Title
from Paper as p
where p.Paper_id<  2101.00200    and p.Paper_id not in (
select a.Paper_id
from Cite as c,paper as a
where a.Paper_id< 2101.00200 and c.cite like '%'||a.Paper_id||'%')
```

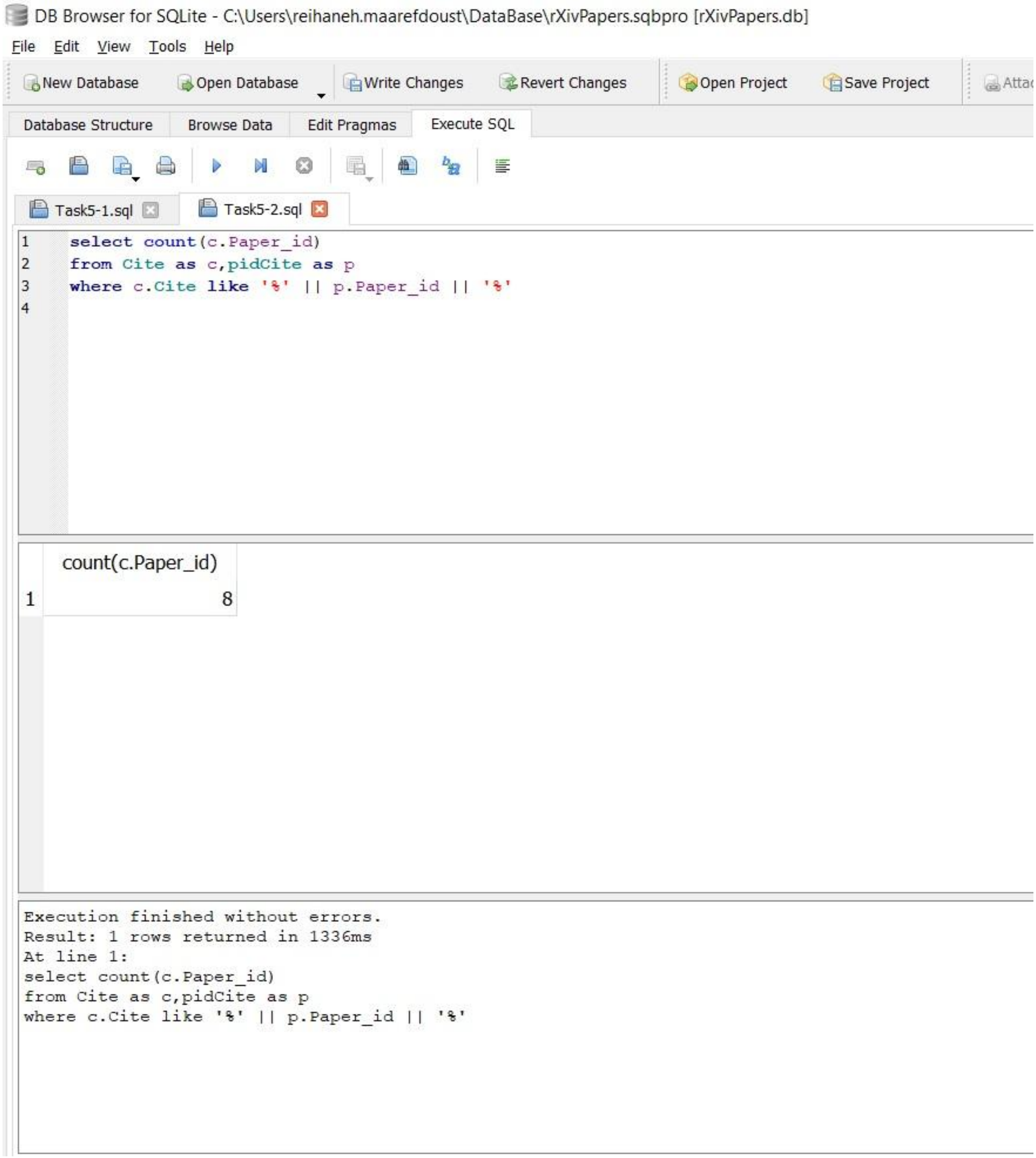
3.5-1.sql



```
DROP VIEW IF EXISTS pidCite;
CREATE VIEW pidCite As
select Paper_id
from writer as w
where w.Author_id in (select Author_id from author where FNAME like 'Lei' and LNAME like 'Yin')
```

```
Execution finished without errors.
Result: query executed successfully. Took 0ms
At line 2:
CREATE VIEW pidCite As
SELECT Paper_id
FROM writer as w
WHERE w.Author_id in (SELECT Author_id FROM author WHERE FNAME like 'Lei' AND LNAME like 'Yin')
```

3.5-2.sql

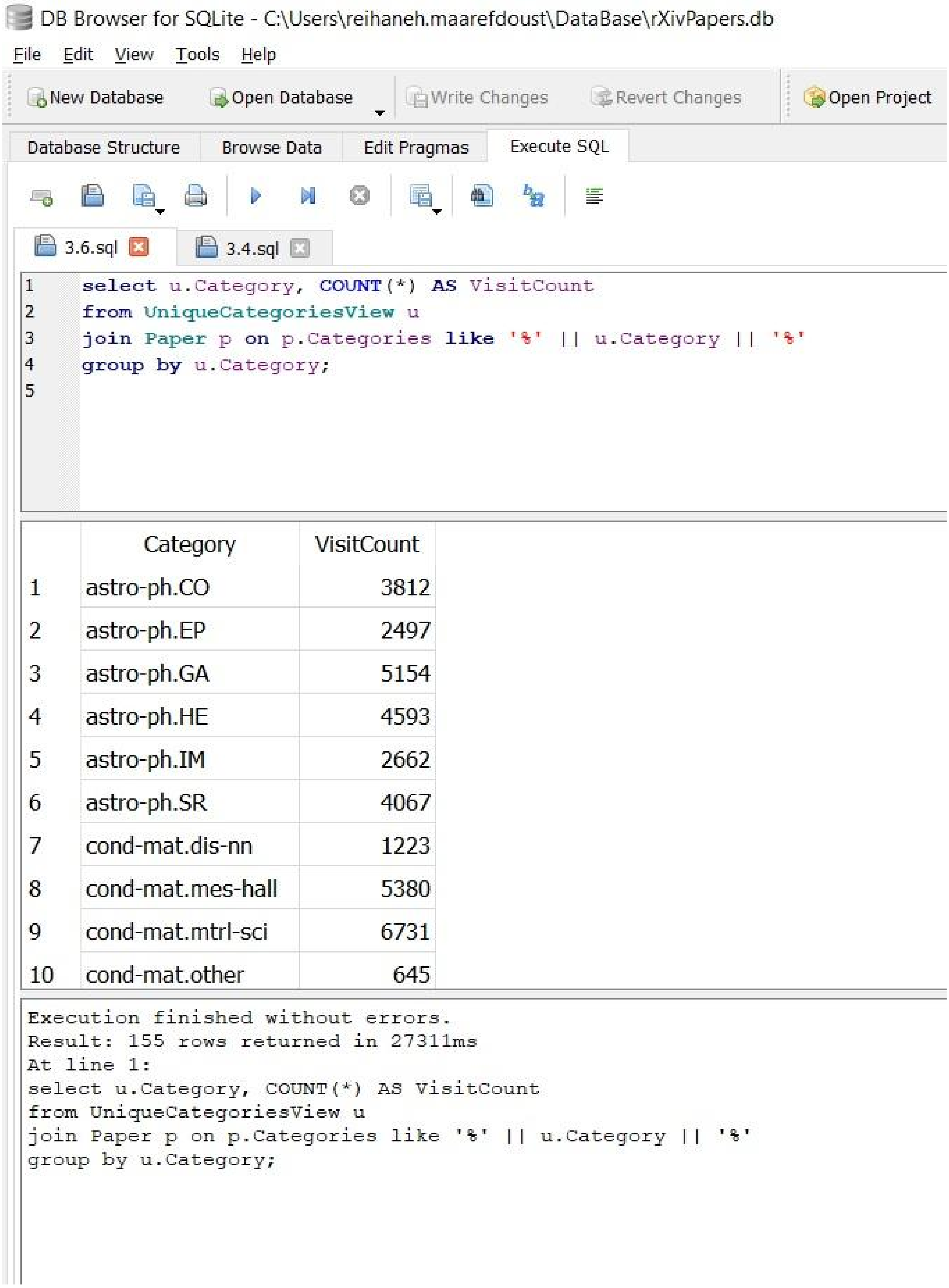


```sql
select count(c.Paper_id)
from Cite as c,pidCite as p
where c.Cite like '%' || p.Paper_id || '%'
```

```
   count(c.Paper_id)
1                  8
```

```
Execution finished without errors.
Result: 1 rows returned in 1336ms
At line 1:
select count(c.Paper_id)
from Cite as c,pidCite as p
where c.Cite like '%' || p.Paper_id || '%'
```

3.6.sql

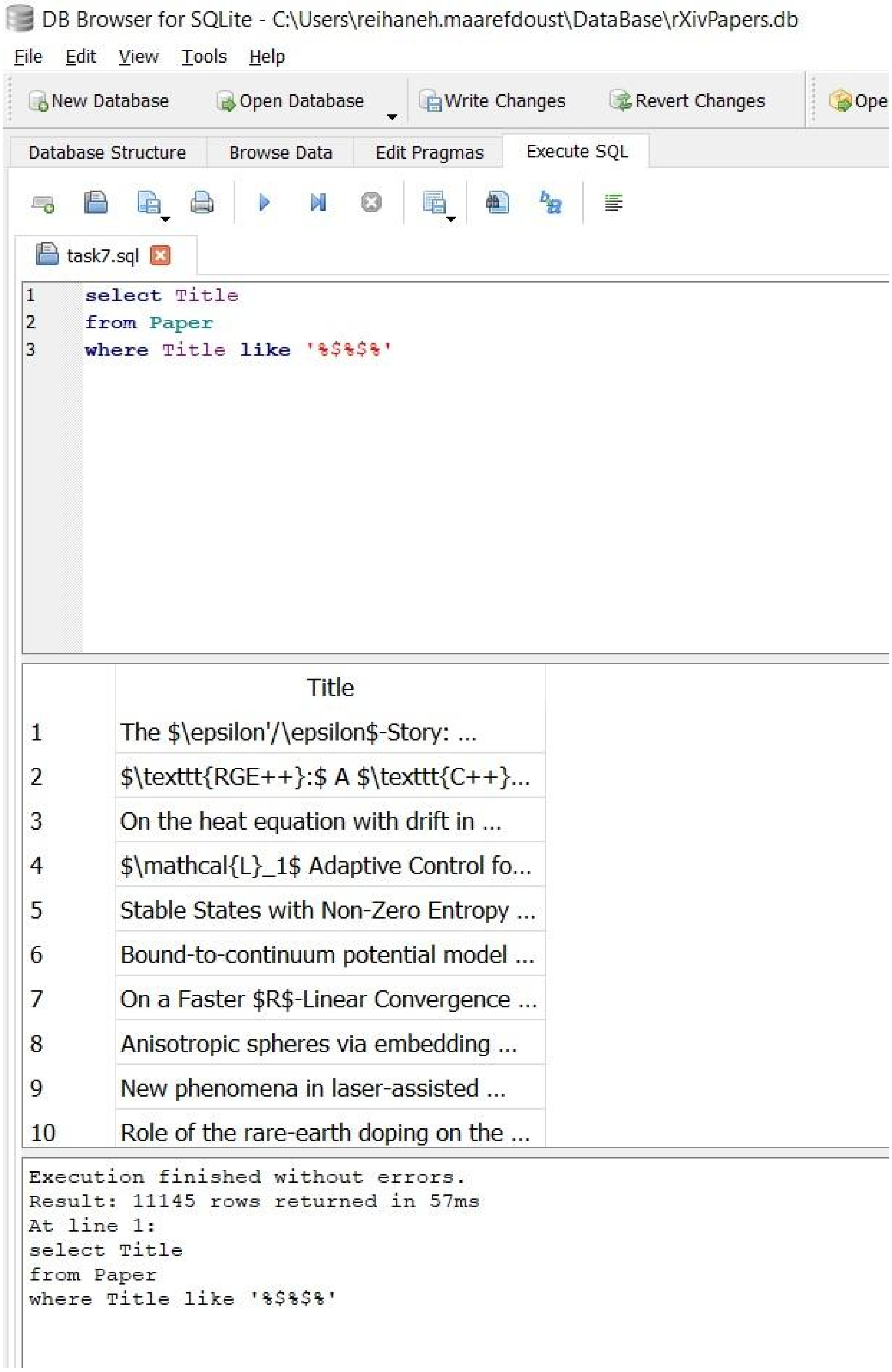


```sql
select u.Category, COUNT(*) AS VisitCount
from UniqueCategoriesView u
join Paper p on p.Categories like '%' || u.Category || '%'
group by u.Category;
```

	Category	VisitCount
1	astro-ph.CO	3812
2	astro-ph.EP	2497
3	astro-ph.GA	5154
4	astro-ph.HE	4593
5	astro-ph.IM	2662
6	astro-ph.SR	4067
7	cond-mat.dis-nn	1223
8	cond-mat.mes-hall	5380
9	cond-mat.mtrl-sci	6731
10	cond-mat.other	645

```
Execution finished without errors.
Result: 155 rows returned in 27311ms
At line 1:
select u.Category, COUNT(*) AS VisitCount
from UniqueCategoriesView u
join Paper p on p.Categories like '%' || u.Category || '%'
group by u.Category;
```

3.7.sql

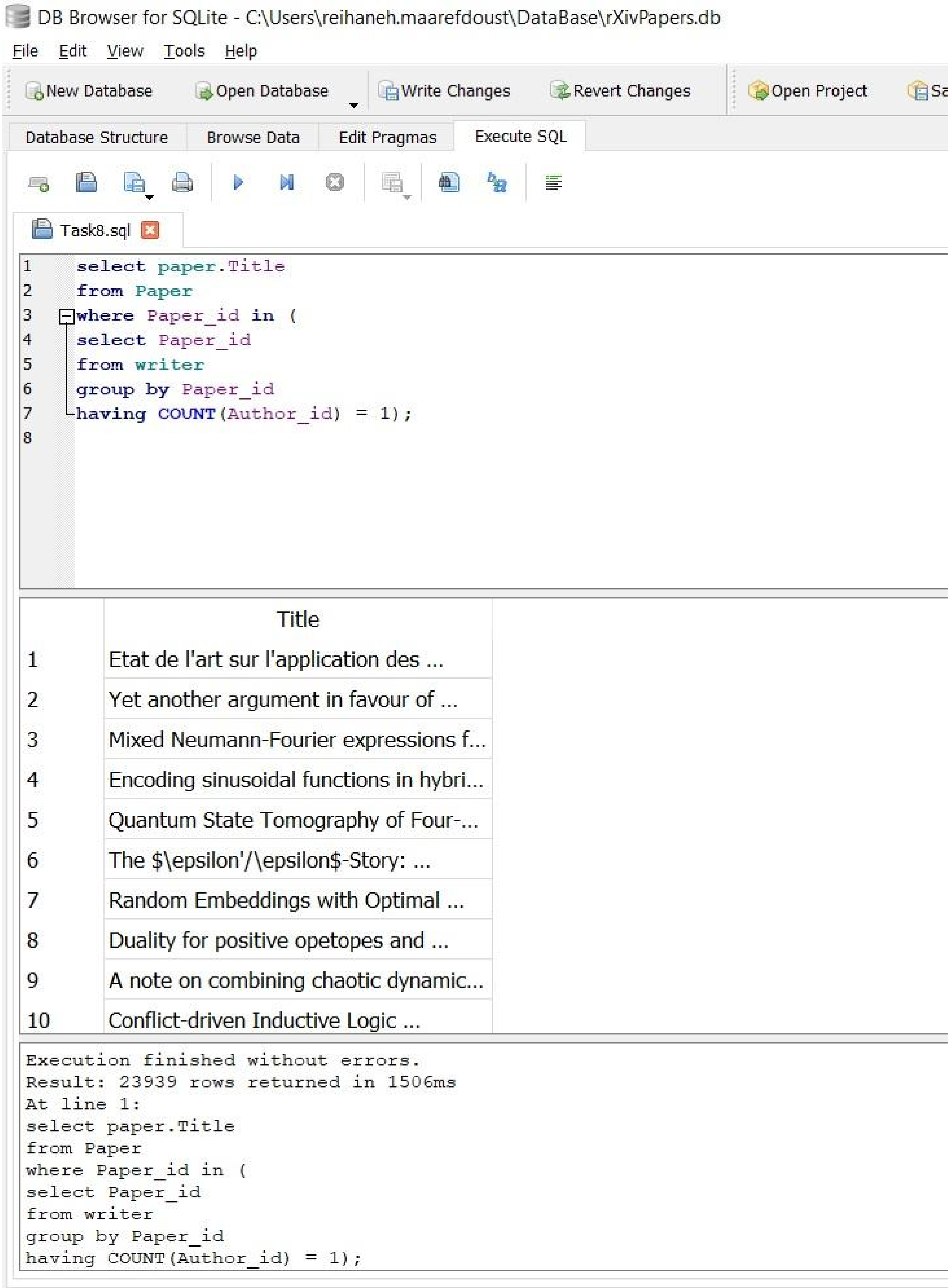


```
select Title
from Paper
where Title like '%$%$%'
```

	Title
1	The ϵ'/ϵ-Story: ...
2	$\texttt{RGE++}:$ A $\texttt{C++}...
3	On the heat equation with drift in ...
4	$\mathcal{L}_1$ Adaptive Control fo...
5	Stable States with Non-Zero Entropy ...
6	Bound-to-continuum potential model ...
7	On a Faster R-Linear Convergence ...
8	Anisotropic spheres via embedding ...
9	New phenomena in laser-assisted ...
10	Role of the rare-earth doping on the ...

```
Execution finished without errors.
Result: 11145 rows returned in 57ms
At line 1:
select Title
from Paper
where Title like '%$%$%'
```

3.8.sql

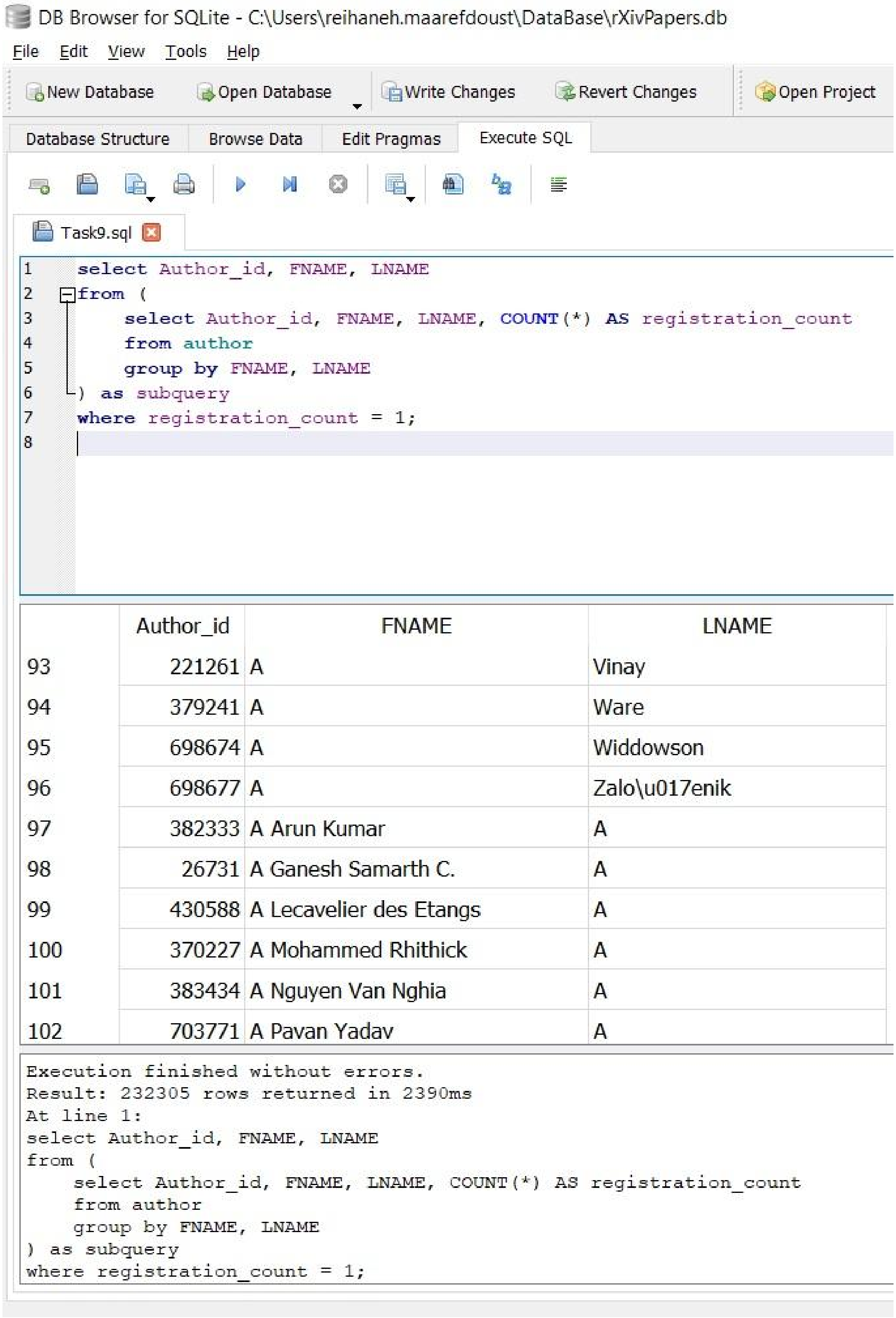


```sql
select paper.Title
from Paper
where Paper_id in (
select Paper_id
from writer
group by Paper_id
having COUNT(Author_id) = 1);
```

	Title
1	Etat de l'art sur l'application des ...
2	Yet another argument in favour of ...
3	Mixed Neumann-Fourier expressions f...
4	Encoding sinusoidal functions in hybri...
5	Quantum State Tomography of Four-...
6	The ϵ'/ϵ-Story: ...
7	Random Embeddings with Optimal ...
8	Duality for positive opetopes and ...
9	A note on combining chaotic dynamic...
10	Conflict-driven Inductive Logic ...

```
Execution finished without errors.
Result: 23939 rows returned in 1506ms
At line 1:
select paper.Title
from Paper
where Paper_id in (
select Paper_id
from writer
group by Paper_id
having COUNT(Author_id) = 1);
```

3.9.sql



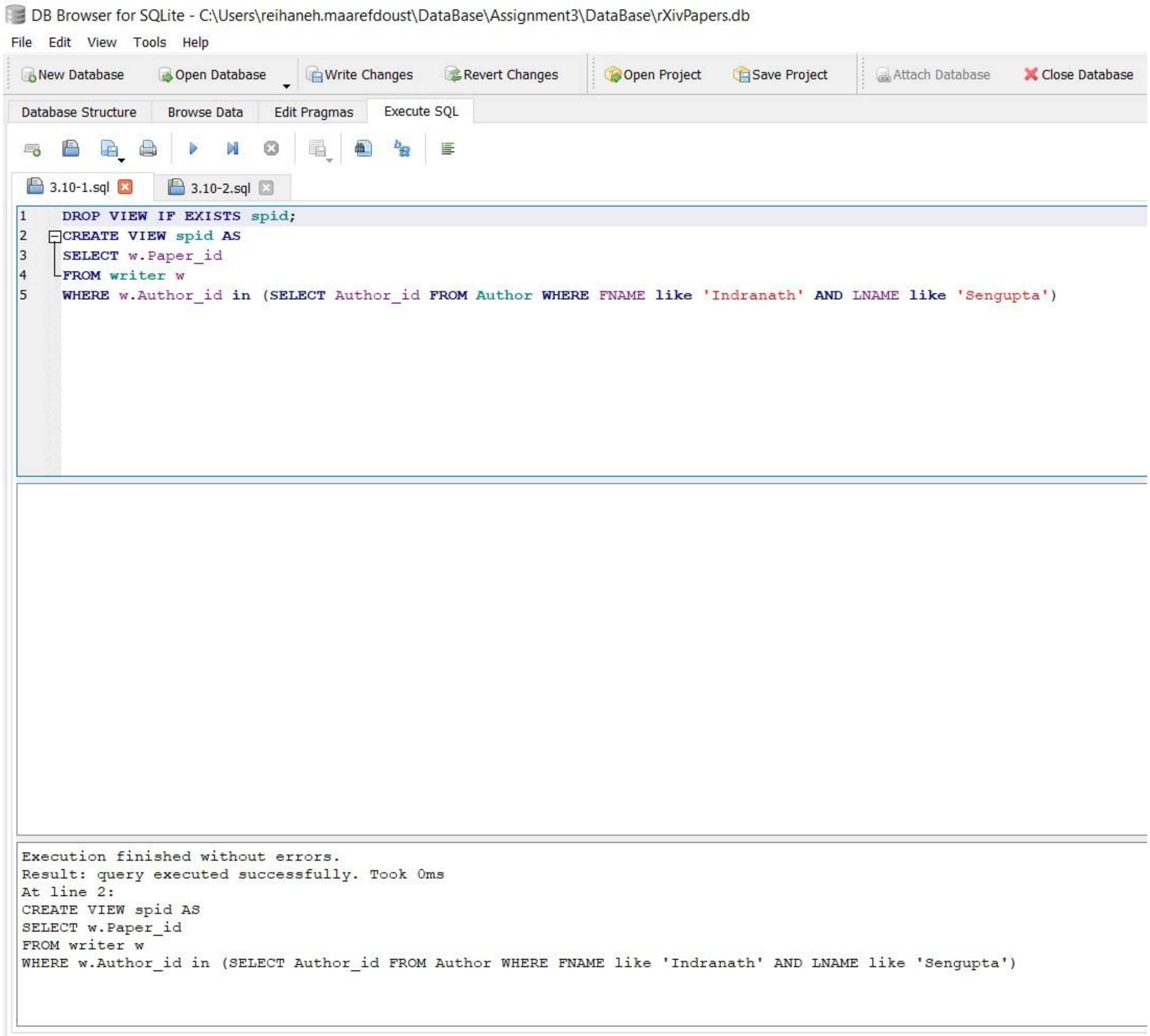
```sql
select Author_id, FNAME, LNAME
from (
    select Author_id, FNAME, LNAME, COUNT(*) AS registration_count
    from author
    group by FNAME, LNAME
) as subquery
where registration_count = 1;

```

	Author_id	FNAME	LNAME
93	221261	A	Vinay
94	379241	A	Ware
95	698674	A	Widdowson
96	698677	A	Zalo\u017enik
97	382333	A Arun Kumar	A
98	26731	A Ganesh Samarth C.	A
99	430588	A Lecavelier des Etangs	A
100	370227	A Mohammed Rhithick	A
101	383434	A Nguyen Van Nghia	A
102	703771	A Pavan Yadav	A

```
Execution finished without errors.
Result: 232305 rows returned in 2390ms
At line 1:
select Author_id, FNAME, LNAME
from (
    select Author_id, FNAME, LNAME, COUNT(*) AS registration_count
    from author
    group by FNAME, LNAME
) as subquery
where registration_count = 1;
```

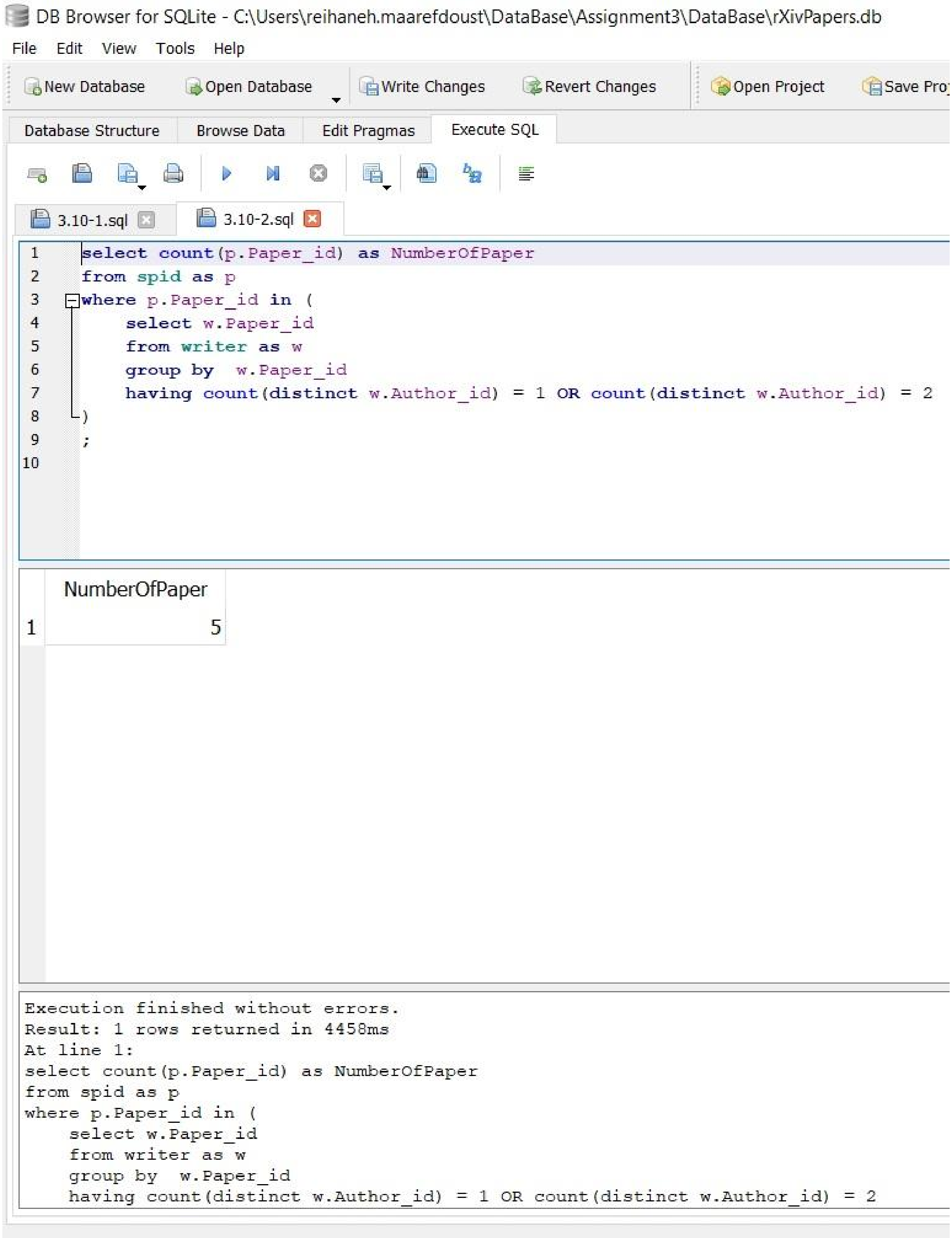
3.10-1.sql



```sql
DROP VIEW IF EXISTS spid;
CREATE VIEW spid AS
SELECT w.Paper_id
FROM writer w
WHERE w.Author_id in (SELECT Author_id FROM Author WHERE FNAME like 'Indranath' AND LNAME like 'Sengupta')
```

```
Execution finished without errors.
Result: query executed successfully. Took 0ms
At line 2:
CREATE VIEW spid AS
SELECT w.Paper_id
FROM writer w
WHERE w.Author_id in (SELECT Author_id FROM Author WHERE FNAME like 'Indranath' AND LNAME like 'Sengupta')
```

3.10-2.sql



```
select count(p.Paper_id) as NumberOfPaper
from spid as p
where p.Paper_id in (
    select w.Paper_id
    from writer as w
    group by  w.Paper_id
    having count(distinct w.Author_id) = 1 OR count(distinct w.Author_id) = 2
)
;
```

	NumberOfPaper
1	5

```
Execution finished without errors.
Result: 1 rows returned in 4458ms
At line 1:
select count(p.Paper_id) as NumberOfPaper
from spid as p
where p.Paper_id in (
    select w.Paper_id
    from writer as w
    group by  w.Paper_id
    having count(distinct w.Author_id) = 1 OR count(distinct w.Author_id) = 2
```

PART2

a.

$\pi_{\text{paper-id}}\ (\sigma_{\text{Cite = '2107.06267'}}\ (\text{Cite}))$

$\{\ \text{t.paper-id}\ |\ \exists\ t\ \ t \in \text{cite}\ (\text{t.cite} = \text{'2107.06267'})\ \}$

b.

$\pi_{\text{(Title, Submitter)}}(\sigma(\text{Category = 'cs.DC' AND Paper\_id NOT IN }(\pi_{\text{(Cite)}}(\text{Cite})),\ \text{Paper}))$

$\{\langle \text{p.Title, p.Submitter} \rangle\ |\ p \in \text{Paper} \land \text{p.Category = 'cs.DC'} \land \neg(\exists c \in \text{Cite}\ (\text{c.Cite} = \text{p.Paper\_id}))\}$