Building a Knowledge Graph Enriched With Large Language Models

Oğuzhan Güngör 24411006

Goal of the Presentation

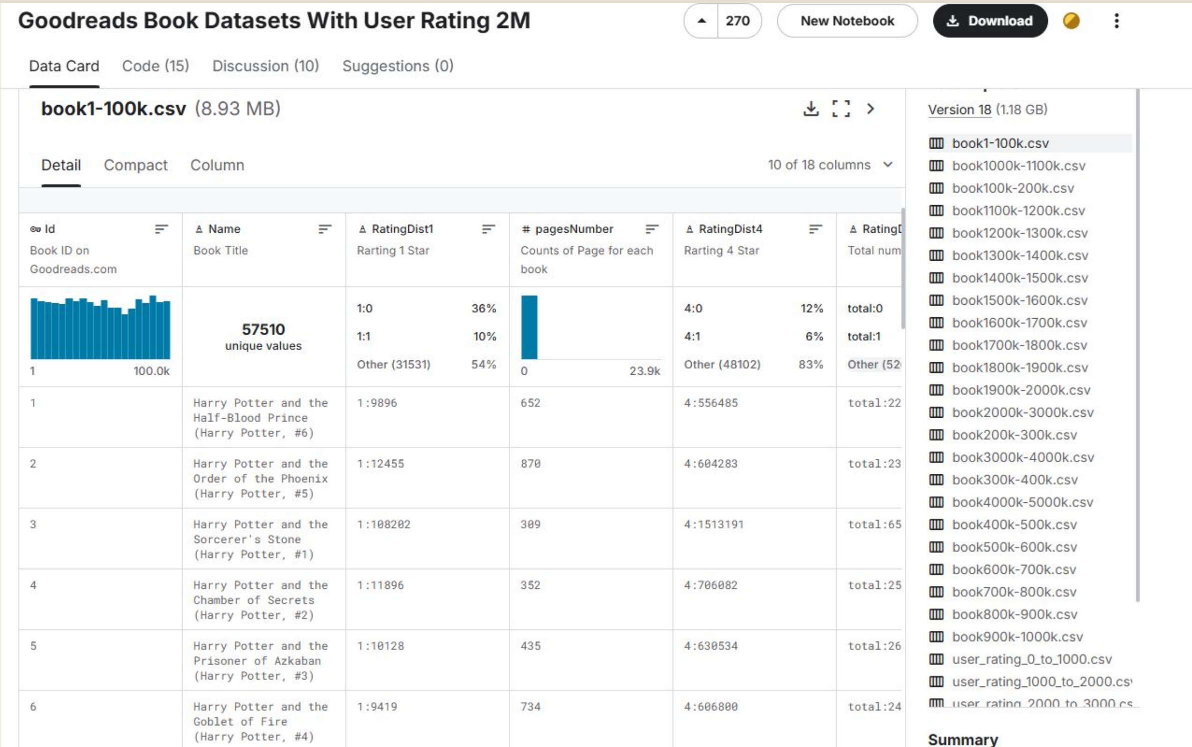
To explain the current state of the project and present the deliverables.

The project is creating a knowledge graph and enhance the Large Language Model (LLM).

Using the knowledge graph to enrich the LLM and link it to DBpedia for enhanced semantic connections and information retrieval.

Data Preparation

In this project, Goodreads Book Dataset has been used for generating the initial database.



Goodreads Book Datasets With User Rating 2M

book1-100k.csv (8.93 MB)

Id	Name	RatingDist1	pagesNumber	RatingDist4	RatingI
Book ID on Goodreads.com	Book Title	Rarting 1 Star	Counts of Page for each book	Rarting 4 Star	Total num
1 100.0k	57510 unique values	1:0 36% 1:1 10% Other (31531) 54%	0 23.9k	4:0 12% 4:1 6% Other (48102) 83%	total:0 total:1 Other (52
1	Harry Potter and the Half-Blood Prince (Harry Potter, #6)	1:9896	652	4:556485	total:22
2	Harry Potter and the Order of the Phoenix (Harry Potter, #5)	1:12455	870	4:604283	total:23
3	Harry Potter and the Sorcerer's Stone (Harry Potter, #1)	1:108202	309	4:1513191	total:65
4	Harry Potter and the Chamber of Secrets (Harry Potter, #2)	1:11896	352	4:706082	total:25
5	Harry Potter and the Prisoner of Azkaban (Harry Potter, #3)	1:10128	435	4:630534	total:26
6	Harry Potter and the Goblet of Fire (Harry Potter, #4)	1:9419	734	4:606800	total:24

Version 18 (1.18 GB)

- book1-100k.csv
- book1000k-1100k.csv
- book100k-200k.csv
- book1100k-1200k.csv
- book1200k-1300k.csv
- book1300k-1400k.csv
- book1400k-1500k.csv
- book1500k-1600k.csv
- book1600k-1700k.csv
- book1700k-1800k.csv
- book1800k-1900k.csv
- book1900k-2000k.csv
- book2000k-3000k.csv
- book200k-300k.csv
- book3000k-4000k.csv
- book300k-400k.csv
- book4000k-5000k.csv
- book400k-500k.csv
- book500k-600k.csv
- book600k-700k.csv
- book700k-800k.csv
- book800k-900k.csv
- book900k-1000k.csv
- user_rating_0_to_1000.csv
- user_rating_1000_to_2000.cs
- user_rating_2000_to_3000.cs

Summary

Data Pipeline Methods

There were 3 approaches when storing data from csv file to Neo4j graph database:

1) Utilizing Concurrency in Python

2) Utilizing Multiprocessing in Python

3) Utilizing Concurrency at Database Level

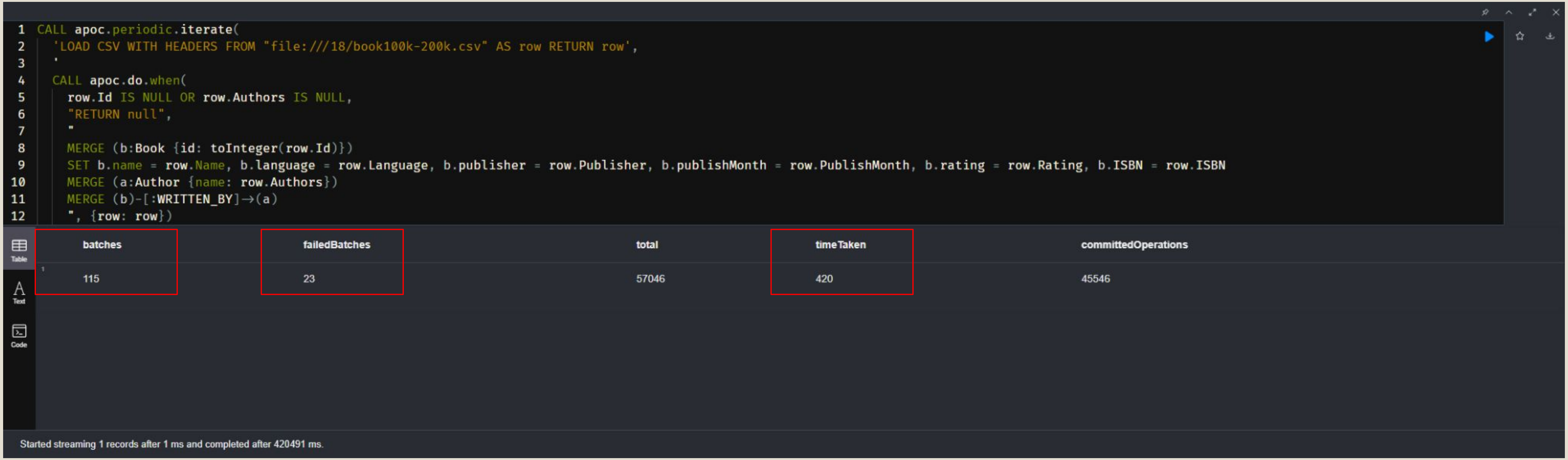
Data Pipeline Methods Pros and Cons

Concurrency in Python: It is fast. Comes with race conditions, managing connection pools is hard. Lacks data integrity.

Multiprocessing in Python: Easier solution. It is slow. Don’t need to manage connection pools. Data integrity is full.

Concurrency in Neo4j using APOC: It is fastest. Comes with race conditions therefore comes with failed insertions. Lacks data integrity. This method was picked due to its accuracy.

Data Pipeline Methods- APOC



```
CALL apoc.periodic.iterate(
  'LOAD CSV WITH HEADERS FROM "file:///18/book100k-200k.csv" AS row RETURN row',
  '
  CALL apoc.do.when(
    row.Id IS NULL OR row.Authors IS NULL,
    "RETURN null",
    "
    MERGE (b:Book {id: toInteger(row.Id)})
    SET b.name = row.Name, b.language = row.Language, b.publisher = row.Publisher, b.publishMonth = row.PublishMonth, b.rating = row.Rating, b.ISBN = row.ISBN
    MERGE (a:Author {name: row.Authors})
    MERGE (b)-[:WRITTEN_BY]->(a)
    ", {row: row})
```

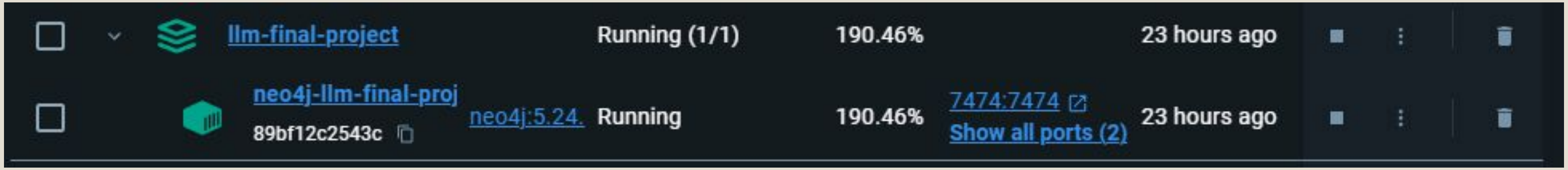
	batches	failedBatches	total	timeTaken	committedOperations
1	115	23	57046	420	45546

Started streaming 1 records after 1 ms and completed after 420491 ms.

Environment Generation

For the development environment, Docker container has been utilized for easier deployment of Neo4j. For Python scripts, Python 3.12 has been picked due to speed benefits of Python 3.12.

For version control, Git and Github has been utilized. Since the project is a solo project with time constraints, no branch controls have been implemented.



□	llm-final-project		Running (1/1)	190.46%		23 hours ago
□	neo4j-llm-final-proj 89bf12c2543c	neo4j:5.24.	Running	190.46%	7474:7474 Show all ports (2)	23 hours ago

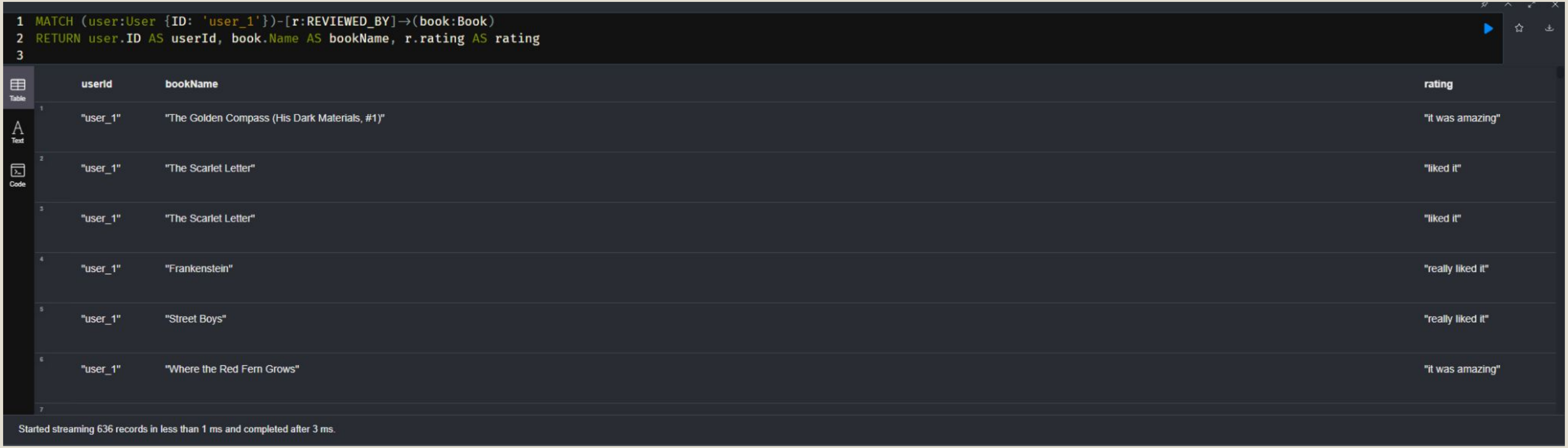
Graph Version 1:
Generated With Only The Dataset

This version made with only using the contents of the data. In this version, graph database have meaningful connections but lacks the depth.

Nodes:
Books, Authors, Users

Edges:
WRITTEN_BY: Book -> Author
REVIEWED_BY: Book -> User

Graph Version 1:
Generated With Only The Dataset

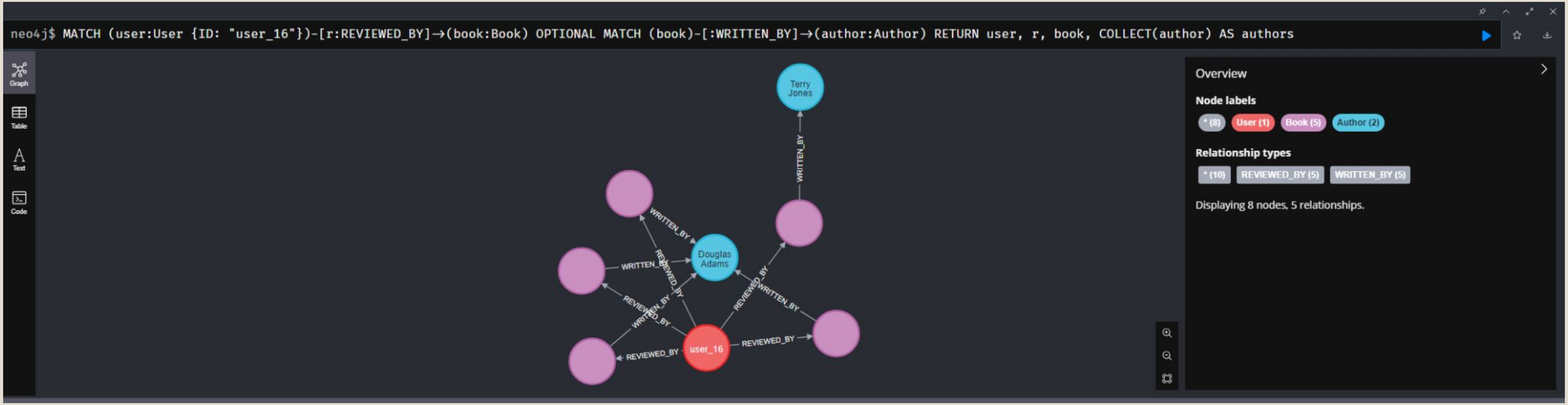


```
MATCH (user:User {ID: 'user_1'})-[r:REVIEWED_BY]->(book:Book)
RETURN user.ID AS userId, book.Name AS bookName, r.rating AS rating
```

	userId	bookName	rating
1	"user_1"	"The Golden Compass (His Dark Materials, #1)"	"it was amazing"
2	"user_1"	"The Scarlet Letter"	"liked it"
3	"user_1"	"The Scarlet Letter"	"liked it"
4	"user_1"	"Frankenstein"	"really liked it"
5	"user_1"	"Street Boys"	"really liked it"
6	"user_1"	"Where the Red Fern Grows"	"it was amazing"
7			

Started streaming 636 records in less than 1 ms and completed after 3 ms.

Graph Version 1:
Generated With Only The Dataset



```
neo4j$ MATCH (user:User {ID: "user_16"})-[r:REVIEWED_BY]→(book:Book) OPTIONAL MATCH (book)-[:WRITTEN_BY]→(author:Author) RETURN user, r, book, COLLECT(author) AS authors
```

Overview

Node labels

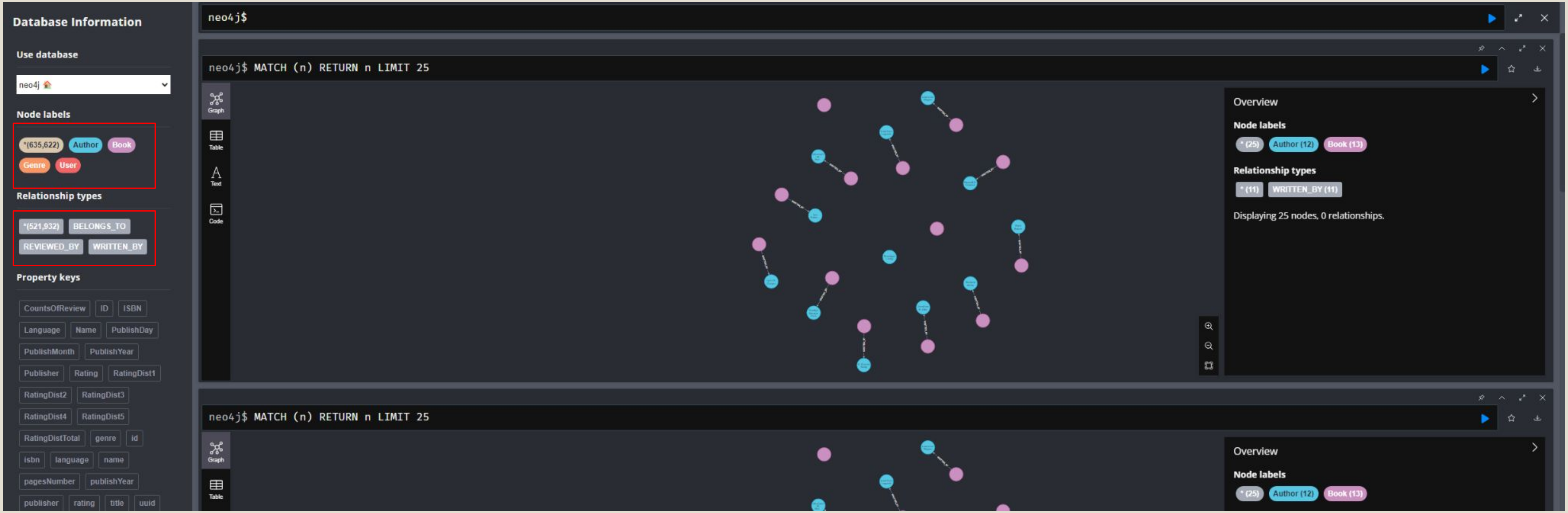
* (8) User (1) Book (5) Author (2)

Relationship types

* (10) REVIEWED_BY (5) WRITTEN_BY (5)

Displaying 8 nodes, 5 relationships.

Graph Version 1:
Generated With Only The Dataset



Database Information

Use database

neo4j

Node labels

*(635,622) Author Book Genre User

Relationship types

*(521,932) BELONGS_TO REVIEWED_BY WRITTEN_BY

Property keys

CountsOfReview ID ISBN Language Name PublishDay PublishMonth PublishYear Publisher Rating RatingDist1 RatingDist2 RatingDist3 RatingDist4 RatingDist5 RatingDistTotal genre id isbn language name pagesNumber publishYear publisher rating title uuid

```
neo4j$
```

```
neo4j$ MATCH (n) RETURN n LIMIT 25
```

Overview

Node labels

* (25) Author (12) Book (13)

Relationship types

* (11) WRITTEN_BY (11)

Displaying 25 nodes, 0 relationships.

```
neo4j$ MATCH (n) RETURN n LIMIT 25
```

Overview

Node labels

* (25) Author (12) Book (13)

DBpedia Integration

Since the data for books are lacking Genre of the books, the awards that author or the book got using DBpedia to obtain genres of the books was the solution for improving the graph database.

Graph Version II:
Improved Version With DBpedia

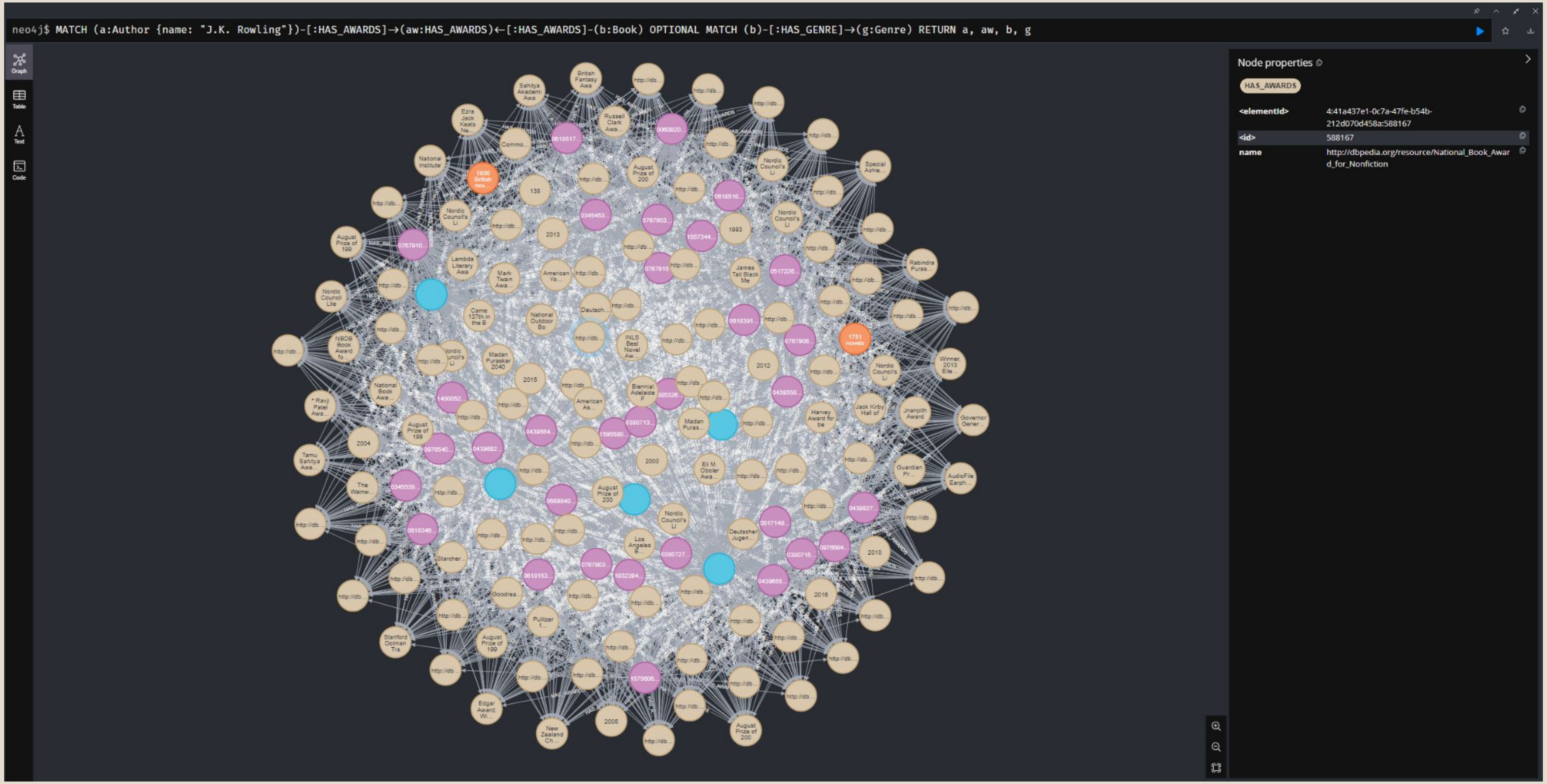
This version made with addition to DBpedia data. In this version, graph database have meaningful connections but still lacks the depth.

Improvements:
Added Genre, Award, Descriptions

Nodes:
Books, Authors, Users, Genre, Award

Edges:
WRITTEN_BY: Book -> Author
REVIEWED_BY: Book -> User
HAS_GENRE: Book -> Genre
HAS_AWARDS: Book -> Award

Graph Version II:
Improved Version With DBpedia



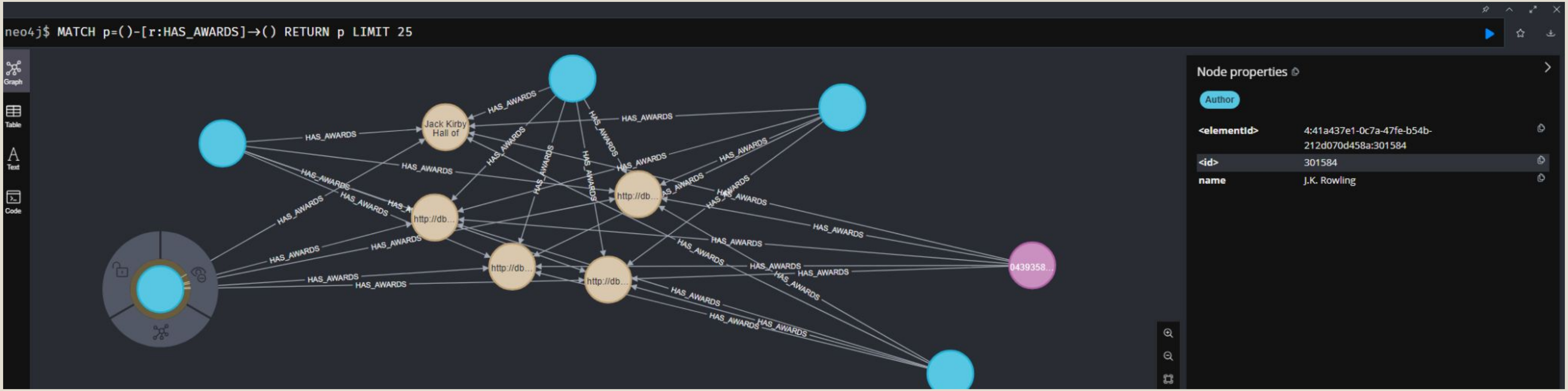
```
neo4j$ MATCH (a:Author {name: "J.K. Rowling"})-[:HAS_AWARDS]→(aw:HAS_AWARDS)←[:HAS_AWARDS]-(b:Book) OPTIONAL MATCH (b)-[:HAS_GENRE]→(g:Genre) RETURN a, aw, b, g
```

Node properties

HAS_AWARDS

<elementId>	4:41a437e1-0c7a-47fe-b54b-212d070d458a:588167
<id>	588167
name	http://dbpedia.org/resource/National_Book_Award_for_Nonfiction

Graph Version II:
Improved Version With DBpedia



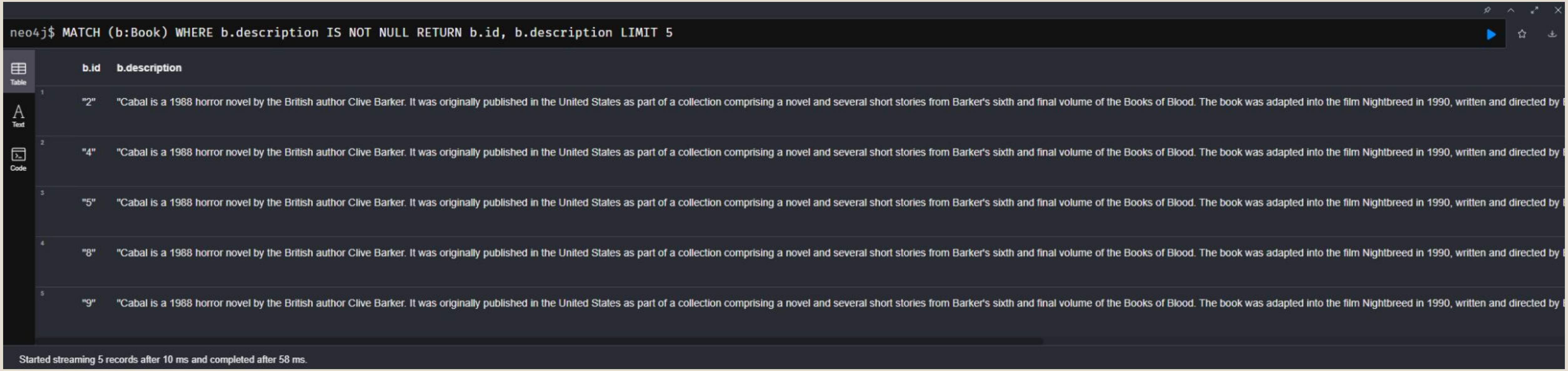
```
neo4j$ MATCH p=()-[r:HAS_AWARDS]→() RETURN p LIMIT 25
```

Node properties

Author

<elementId>	4:41a437e1-0c7a-47fe-b54b-212d070d458a:301584
<id>	301584
name	J.K. Rowling

Graph Version II:
Improved Version With DBpedia
(Bugged Version)

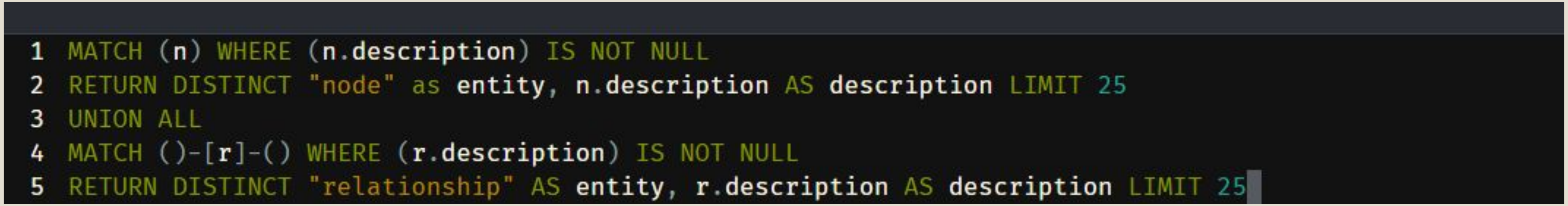


```
neo4j$ MATCH (b:Book) WHERE b.description IS NOT NULL RETURN b.id, b.description LIMIT 5
```

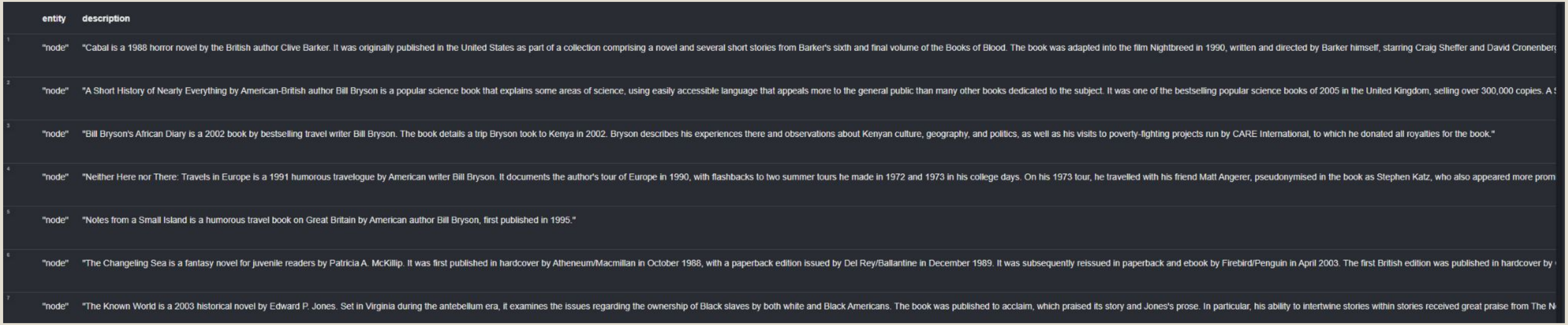
	b.id	b.description
1	"2"	"Cabal is a 1988 horror novel by the British author Clive Barker. It was originally published in the United States as part of a collection comprising a novel and several short stories from Barker's sixth and final volume of the Books of Blood. The book was adapted into the film Nightbreed in 1990, written and directed by
2	"4"	"Cabal is a 1988 horror novel by the British author Clive Barker. It was originally published in the United States as part of a collection comprising a novel and several short stories from Barker's sixth and final volume of the Books of Blood. The book was adapted into the film Nightbreed in 1990, written and directed by
3	"5"	"Cabal is a 1988 horror novel by the British author Clive Barker. It was originally published in the United States as part of a collection comprising a novel and several short stories from Barker's sixth and final volume of the Books of Blood. The book was adapted into the film Nightbreed in 1990, written and directed by
4	"8"	"Cabal is a 1988 horror novel by the British author Clive Barker. It was originally published in the United States as part of a collection comprising a novel and several short stories from Barker's sixth and final volume of the Books of Blood. The book was adapted into the film Nightbreed in 1990, written and directed by
5	"9"	"Cabal is a 1988 horror novel by the British author Clive Barker. It was originally published in the United States as part of a collection comprising a novel and several short stories from Barker's sixth and final volume of the Books of Blood. The book was adapted into the film Nightbreed in 1990, written and directed by

Started streaming 5 records after 10 ms and completed after 58 ms.

Graph Version 11:
Improved Version With DBpedia
(Fixed Version)

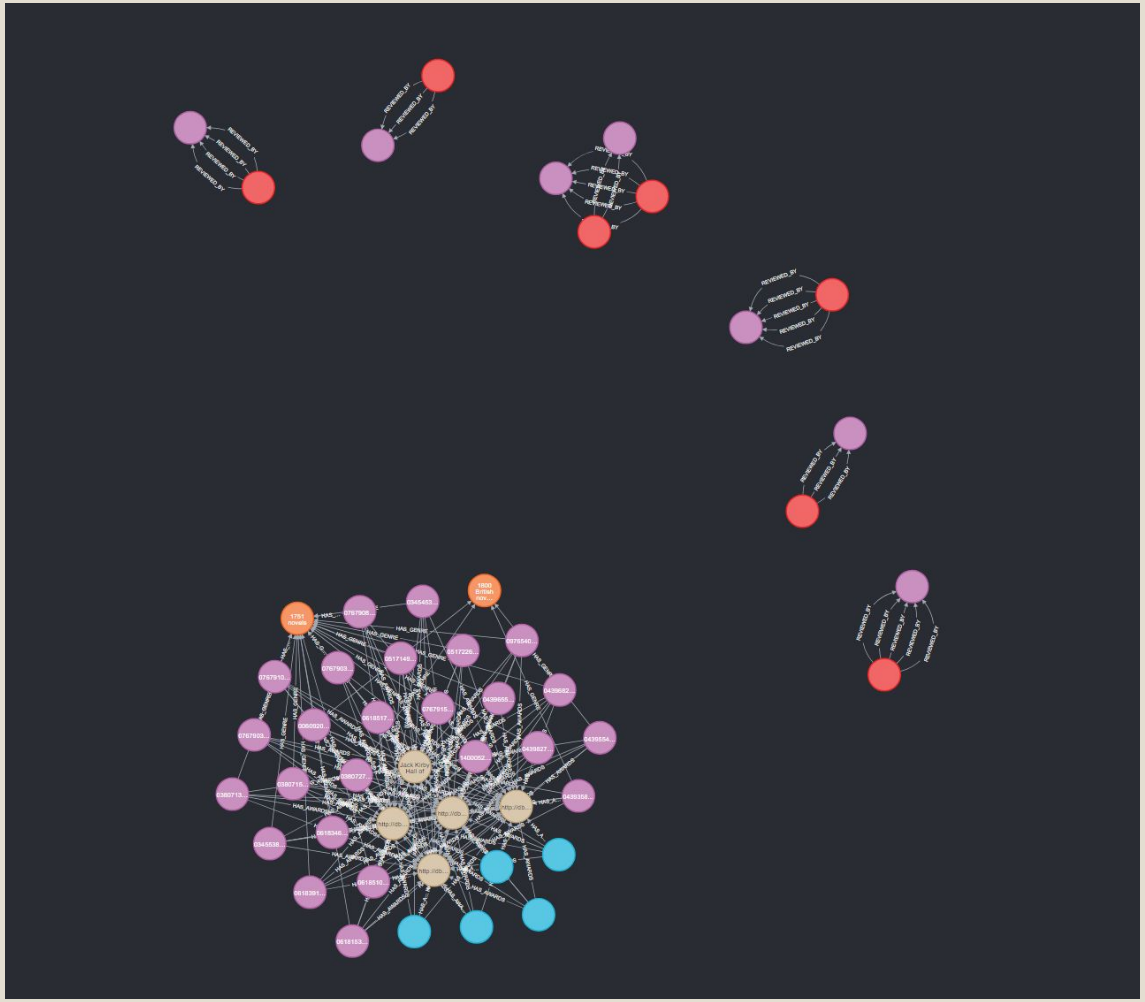


```
MATCH (n) WHERE (n.description) IS NOT NULL
RETURN DISTINCT "node" as entity, n.description AS description LIMIT 25
UNION ALL
MATCH ()-[r]-() WHERE (r.description) IS NOT NULL
RETURN DISTINCT "relationship" AS entity, r.description AS description LIMIT 25
```



	entity	description
1	"node"	"Cabal is a 1988 horror novel by the British author Clive Barker. It was originally published in the United States as part of a collection comprising a novel and several short stories from Barker's sixth and final volume of the Books of Blood. The book was adapted into the film Nightbreed in 1990, written and directed by Barker himself, starring Craig Sheffer and David Cronenber
2	"node"	"A Short History of Nearly Everything by American-British author Bill Bryson is a popular science book that explains some areas of science, using easily accessible language that appeals more to the general public than many other books dedicated to the subject. It was one of the bestselling popular science books of 2005 in the United Kingdom, selling over 300,000 copies. A
3	"node"	"Bill Bryson's African Diary is a 2002 book by bestselling travel writer Bill Bryson. The book details a trip Bryson took to Kenya in 2002. Bryson describes his experiences there and observations about Kenyan culture, geography, and politics, as well as his visits to poverty-fighting projects run by CARE International, to which he donated all royalties for the book."
4	"node"	"Neither Here nor There: Travels in Europe is a 1991 humorous travelogue by American writer Bill Bryson. It documents the author's tour of Europe in 1990, with flashbacks to two summer tours he made in 1972 and 1973 in his college days. On his 1973 tour, he travelled with his friend Matt Angerer, pseudonymised in the book as Stephen Katz, who also appeared more prom
5	"node"	"Notes from a Small Island is a humorous travel book on Great Britain by American author Bill Bryson, first published in 1995."
6	"node"	"The Changeling Sea is a fantasy novel for juvenile readers by Patricia A. McKillip. It was first published in hardcover by Atheneum/Macmillan in October 1988, with a paperback edition issued by Del Rey/Ballantine in December 1989. It was subsequently reissued in paperback and ebook by Firebird/Penguin in April 2003. The first British edition was published in hardcover by
7	"node"	"The Known World is a 2003 historical novel by Edward P. Jones. Set in Virginia during the antebellum era, it examines the issues regarding the ownership of Black slaves by both white and Black Americans. The book was published to acclaim, which praised its story and Jones's prose. In particular, his ability to intertwine stories within stories received great praise from The N

Graph Version II:
Improved Version With DBpedia



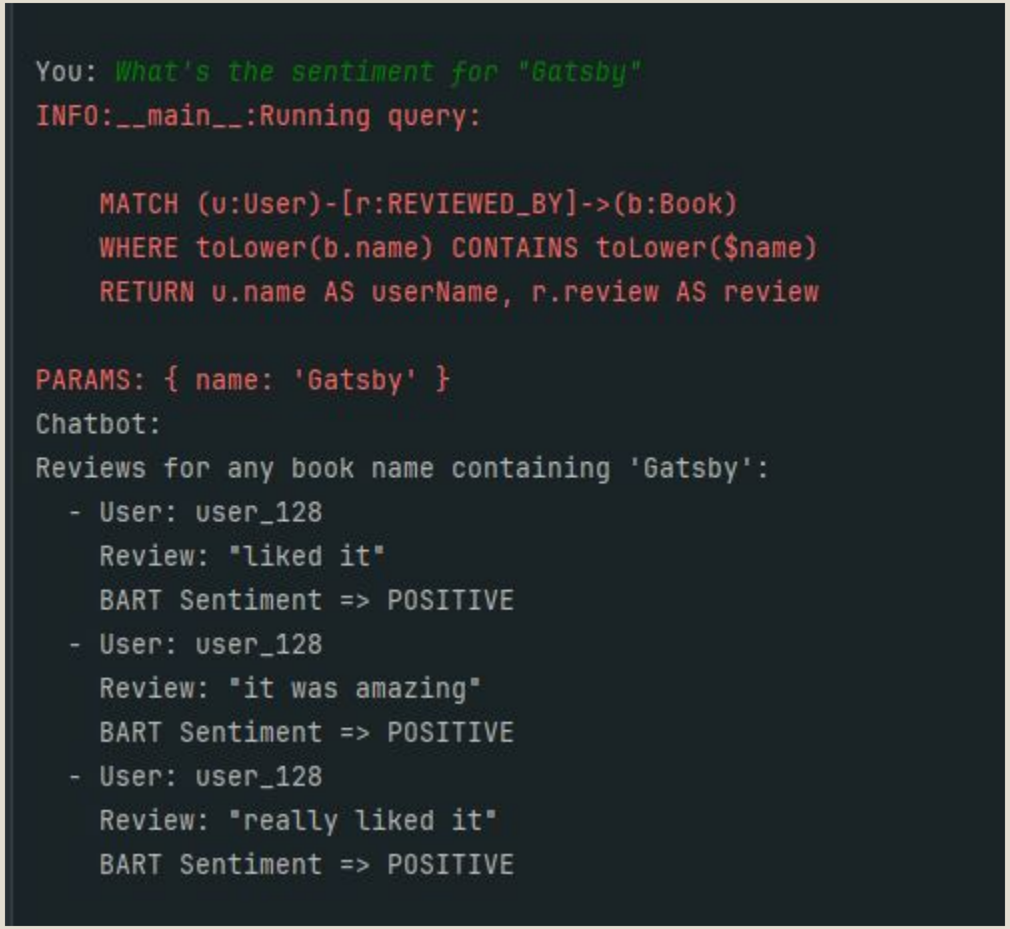
LLM Integration

Since the data for books are lacking semantic relationships with comments and other books, utilizing LLM to improve the graph database was crucial.

This step is work in progress at the moment. This step was missing the crucial connection between similar books.

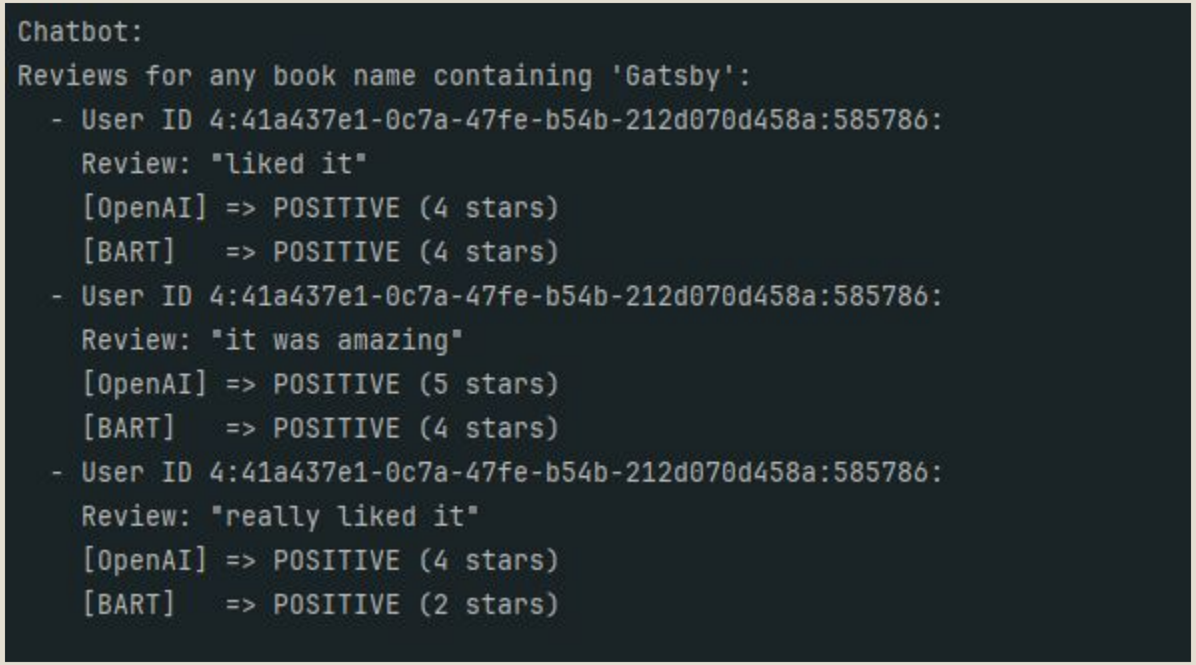
In this step, sentiment analysis was performed using multiple LLMs, mainly BART and OpenAI. In the sentiment analysis, partial matching is also supported.

Graph Version III:
Improved Version With LLM Integration
BART Only



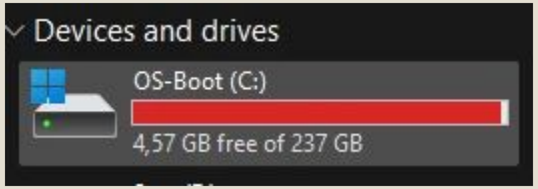
```
You: What's the sentiment for "Gatsby"
INFO:__main__:Running query:

    MATCH (u:User)-[r:REVIEWED_BY]->(b:Book)
    WHERE toLower(b.name) CONTAINS toLower($name)
    RETURN u.name AS userName, r.review AS review

PARAMS: { name: 'Gatsby' }
Chatbot:
Reviews for any book name containing 'Gatsby':
  - User: user_128
    Review: "liked it"
    BART Sentiment => POSITIVE
  - User: user_128
    Review: "it was amazing"
    BART Sentiment => POSITIVE
  - User: user_128
    Review: "really liked it"
    BART Sentiment => POSITIVE
```

Graph Version III:
Improved Version With LLM Integration
BART and OpenAI



```
Chatbot:
Reviews for any book name containing 'Gatsby':
  - User ID 4:41a437e1-0c7a-47fe-b54b-212d070d458a:585786:
    Review: "liked it"
    [OpenAI] => POSITIVE (4 stars)
    [BART]   => POSITIVE (4 stars)
  - User ID 4:41a437e1-0c7a-47fe-b54b-212d070d458a:585786:
    Review: "it was amazing"
    [OpenAI] => POSITIVE (5 stars)
    [BART]   => POSITIVE (4 stars)
  - User ID 4:41a437e1-0c7a-47fe-b54b-212d070d458a:585786:
    Review: "really liked it"
    [OpenAI] => POSITIVE (4 stars)
    [BART]   => POSITIVE (2 stars)
```

Storage



Devices and drives

OS-Boot (C:)

4,57 GB free of 237 GB

References

1) Goodreads Book Datasets With User Rating 2M
2)

Thanks for listening