

APPLIED DATABASE TECHNOLOGIES

REPORT ON IMDB MOVIE REVIEWS

SECTION 1

DEPLOYMENT URL:

<https://movie-app-agent-adt-project.streamlit.app/>

GitHub URL:

<https://github.com/Athe-kunal/Movie-QA-Agent-ADT-Project/tree/main>

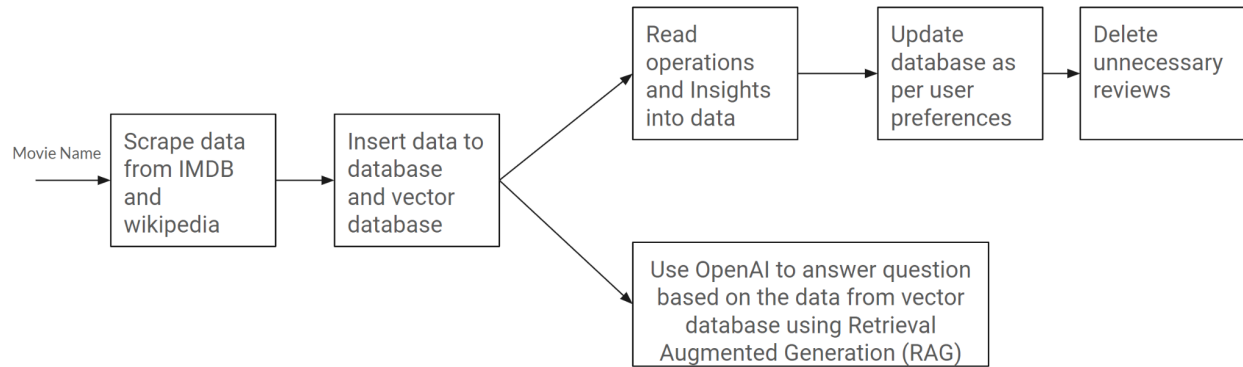
SECTION 2

PURPOSE

The purpose of our project is to create a system that enhances the understanding and accessibility of movies through an accessible and interactive platform for exploring IMDB movie reviews. Scraping data from IMDB and Wikipedia, the project aims to provide valuable insights into movies. The project also integrates a Large Language Model Chat Agent using OpenAI's API to answer questions about movie plot lines. This comprehensive approach allows for a deep, interactive exploration of movie reviews, enhancing users' experience and understanding of movie critiques and opinions.

HOW IT WAS BUILT

Diagram flow of our database



SCRAPING AND DATABASE CREATION

- First, we scrape the IMDB movie reviews website data using **Selenium** web scraper. It scrapes the date of review, review title, review comment, how many people find it helpful, how many people voted for it if the review contains spoilers, link to the review, rating and source
- Also, we have included the data from **Wikipedia**, where you can provide what data you want for the movie, like the plotline, cast etc. We just extracted the plotline. Then we pushed the data to **MongoDB Atlas using PyMongo**

CRUD OPERATIONS

- For read operations, we used PyMongo to read the database based on the user preferences on the attributes. Also, for a movie, we show insights like the rating distribution, number of reviews per year etc. The results were post-filtered to a limit of 10.
- For the update, we allowed users to vote for a review if they found it helpful. It will update the helpful and total votes counter in our database
- For delete, we are deleting generic and unnecessary reviews like the old reviews; that no one found helpful and the length is less than 20

words. It will reduce noise in our database, which will lead to better insights

LARGE LANGUAGE MODEL CHAT AGENT

- We used **OpenAI** chat functionality to answer questions on plot lines about a movie using a **Retrieval Augmented Generation (RAG)** pipeline. For this, we used the vector database functionality of MongoDB. Here we populated our database with the embeddings using a **Sentence Transformer**, which transforms natural language sentences into vectorized tokens. Below is the configuration file for **MongoDB Atlas**

```
{
  "mappings": {
    "dynamic": true,
    "fields": {
      "MovieName": {"type": "token"},
      "source": {"type": "token"},
      "embedding": {
        "type": "knnVector",
        "dimensions": 768,
        "similarity": "euclidean",
      },
    },
  },
}
```

- Hence, we are retrieving the most similar documents (reviews) based on the user question from our database based on **Euclidean** distance and giving context to our model to answer questions better. The vector embedding dimension is 768. We used **Langchain** to talk to **OpenAI LLMs** (gpt-3.5). To conserve the token limit of OpenAI LLMs,

which is around 4K, we split the reviews into chunk sizes of 512 and then populated the database.

- We also provide the links to the review that the LLM is using to answer the question, which is the metadata for the IMDB reviews.

DATA

- The provided dataset represents a MongoDB document capturing information about a movie review for "Parasite_2019," likely scraped from IMDb and Wikipedia. The unique identifier (oid) distinguishes this particular entry, and the date field records when the review is given by the user, offering a temporal context. The title succinctly characterizes the movie, portraying it as an engaging drama with an intelligent social aspect, providing a quick overview of its thematic content.
- The comment field is a noteworthy component, containing an extensive review of the movie. It offers insights into the reviewer's perspective on the plot, characters, and the film's broader social commentary. This detailed commentary allows for a deeper understanding of the reviewer's thoughts and contributes to a more comprehensive analysis of the movie's reception.
- The rating field, with a value of 0, reflects the reviewer's subjective evaluation of the movie on a scale ranging from 0 to 10. The rating provides a numerical representation of the reviewer's opinion, allowing for a quantitative assessment of the movie's perceived quality.
- The helpful field indicates the review's impact, revealing that it garnered 20 helpful votes out of a total of 24 votes. This indicates the review's perceived value within the IMDb community. The if_spoiler field, set to false, informs that the review doesn't contain spoilers,

providing potential readers with information about the review's content.

- Finally, the source field specifies that the data originates from IMDb, offering transparency about the platform from which the movie information was extracted. In summary, this dataset encapsulates a comprehensive snapshot of a movie review, including contextual details, reviewer sentiment, and community feedback, making it a valuable resource for understanding the reception of "Parasite_2019."

```
{
  "_id": {
    "$oid": "656e911aa61bcd32ec46ed50"
  },
  "MovieName": "Parasite_2019",
  "date": {
    "$date": "2020-08-16T00:00:00.000Z"
  },
  "title": "Engaging as a drama, with an intelligent social aspect to it",
  "comment": "A poor family see an opportunity whenever their son starts tutoring English for a wealthy family if they can engineer it they can each get one of the jobs within the household This is the basis for a film that starts out as a sort of con story seeing the rich family as the marks As it plays out though it keeps this assumption in the background eating at the viewer as an idea before then making it very relevant in the closing aspects of the plot Between the start and then the focus is on the various twists and turns of the drama itself In this the film engages and I found it easy to engage with it on the basis of what was happening The later develops do work better though when viewed in the context of the social aspect There are lots of clever critics that can talk to you about the meaning and hidden depths of commentary from the film for me the key one was the falseness of the con itself Although the family mock the gullibility of the rich family it is not like they are stealing money from them or somehow dislodging them from their position in life no they are just providing labour to them in exchange of comparatively low wages They are doing this at the expense of other working people just trying to keep a job and the rich family could probably not care less about the truth as long as their needs continue to be met This aspect is important for the direction of the later stages of the film and adds sense to what happens and why but it is interesting in and of itself Technically the film looks great and the director builds mood and tone well Performances are strong across all the cast but the turn from Song Kang Ho probably was my favourite as he was the most subtle and
```

```
had the most space to shift across the running time There is a lot of talk for
Oscar recognition but it is a handsome clever film and the timing in the year is
right I don t see it being the first foreign language film to win best picture but
this is more to do with the system than with the film Regardless of awards or not
it is an engaging drama with unusual developments and built on top of an
intelligent social aspect which links well to the direction of the narrative",
  "rating": 0,
  "link": "https://www.imdb.com/review/rw6001308/?ref_=tt_urv",
  "helpful": 20,
  "total_votes": 24,
  "if_spoiler": false,
  "source": "imdb"
}
```

FUNCTIONALITIES

LLM Chat Functionality

The LLM chat works by taking a user question and calling the MongoDB vector database to get semantically similar split reviews. These semantically similar reviews are fed into the prompt to get the final grounded answer. Below is a screenshot of the application using OpenAI LLMs

Movie QA



Hi, how can I help you?



What was the significance of the rock in the movie?



The rock in the movie symbolized hope and wealth for the Kims. Despite their difficult financial situation, the Kims were able to find and hold onto a symbol of their dreams of wealth and success. The rock was also used as a weapon against Ki woo, symbolizing how his hope had been betrayed. Ultimately, the rock was a tool to help the family achieve their dreams of wealth and success. By placing the rock back in its original location, it symbolized the completion of their fantasy.

See relevant IMDB movie review links

[Source Link 0](#)

[Source Link 1](#)

[Source Link 2](#)

[Source Link 3](#)

[Source Link 4](#)

Also, we provide the citations in the form of movie review links provided by IMDB, so that users can verify the answers generated by OpenAI LLMs.

Insights

The Insights give us a whole lot of information about a movie selected by the user such as the distribution of ratings using a pie chart for that particular movie, word cloud for all the reviews given for that movie, the top 10 recent reviews that are given by the users based on the date, the bar chart which gives us information about how many reviews are received by that movie

in each year, the top 10 reviews that are helpful to other users, and the top 10 reviews that explain about the aspects the user has given as input.

Movie Information

Select Movie Name:

Parasite_2019

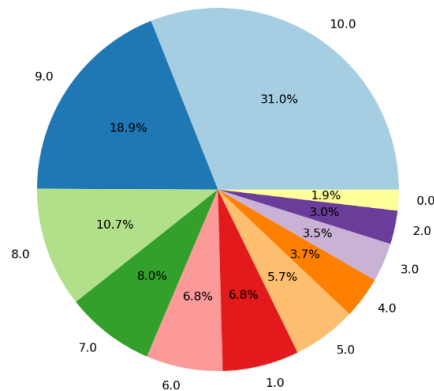
Enter Comma-Separated Words:

music,hero

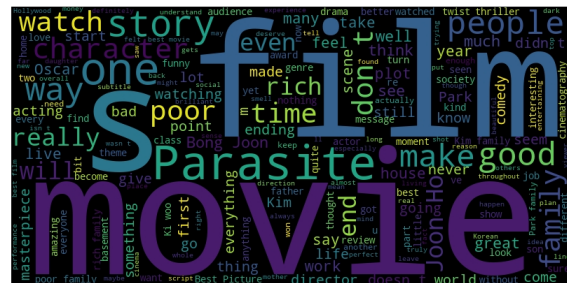
Show Results

Reset Input

Rating Distribution



Word Cloud based on Reviews

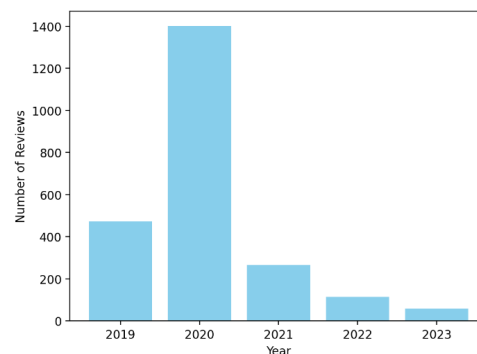


Word Cloud for Parasite_2019

Top 10 Recent Reviews

	date	comment
1,224	2023-11-24 00:00:00	Almost four years ago South Korea s PARASITE made history by becoming the first mo
1,650	2023-11-07 00:00:00	It starts off somewhat light hearted It then shows how people act inauthentically to f
638	2023-10-13 00:00:00	Parasite 7 10 Positive The premise is incredibly exciting and builds up a great game w
1,695	2023-10-06 00:00:00	Considered a cinematic masterpiece Parasite 2019 stands out for its unparalleled allu
1,984	2023-10-05 00:00:00	On one side is the flooded cottage and on the other side is the birthday Party On one
2,238	2023-10-05 00:00:00	The series is all about a family showing through the differences in between the classe
1,985	2023-09-26 00:00:00	This film has one of the most compelling premises and buildups in of any film in rece
1,027	2023-09-25 00:00:00	I wish I could hop in a DeLorean Phone Booth and travel back in time Not all that long
1,482	2023-09-23 00:00:00	The director seems to have a bone to pick with poor folks He s like Hey look poor peo
1,628	2023-09-14 00:00:00	I think the film does a good job of exploring the theme of class The Kim family is a po

Number of Reviews per Year



Top 10 Reviews based on Helpful Votes

		comment	helpful
229	17 00:00:00	Is the fact that people who voted for it to win the Cannes film festival award and those	396.000000
9	19 00:00:00	Well written and performed also technically shines cinematography bgm are too good	379.000000
7	12 00:00:00	As a film about a family imposing on another and keeping dark secrets hidden beneath	366.000000
369	15 00:00:00	Do you want to see a movie that will have your stomach in knots anxious during man	306.000000
1,416	17 00:00:00	I would like to start out by saying that I am a Korean American college student studyi	295.000000
136	18 00:00:00	After reading all the glowing reviews especially about how this film is one of the best	279.000000
38	15 00:00:00	Huge Oscar Winner What a JOKE There's no way or simply not enough to convince me	278.000000
78	12 00:00:00	A family of incompetents who struggle to fold a pizza carton miraculously transform t	272.000000
26	10 00:00:00	How has this film has won the best film Oscar What a travesty It's an ine rebarbative e	256.000000
85	10 00:00:00	This had the makings of a fun farce As I watched the first hour or so I thought it could	243.000000

Top 10 Reviews Containing Words: music, hero

		comment	total_votes
1,416		I would like to start out by saying that I am a Korean American college student studyi	705.000000
85		This had the makings of a fun farce As I watched the first hour or so I thought it could	483.000000
65		Following the step of Get Out it is a powerful film that comprises of metaphors and al	315.000000
387		This movie without a doubt is the best movie I've seen in years 2019 has been a slow	109.000000
1,368		Parasite by Bong Joon ho is like a Hollywood almost shallow version of Burning by Le	98.000000
738		Richly appointed a glitterati champagne Cannes launch to an inebriated fanfare of str	84.000000
499		This movie is clever on so many levels the acting is impeccable the story line unexpect	80.000000
97		PROS EVERY SINGLE THING The acting is great The pacing is amazing it always feels li	54.000000
260		At the beginning of this movie we watch the Kim family struggle to survive in our moc	53.000000
30		Parasite is the best movie of the year This Korean movie won the renowned Palme d'Or	50.000000

Read Functionality

Using the read functionality the user will be able to get information about a movie by applying different filters. The user has to select the movie name, the reviews that are received after a particular date, the rating range, helpful votes range, total votes range, if_spoiler or not (True/False), and the source from which the review is extracted. Based on all these filters top 10 reviews will be displayed.

Movie Reviews Explorer

Select Movie Name:

Parasite_2019



Select Start Date:

2020/01/01

Select Rating Range:



Select Helpful Votes Range:



Select Total Votes Range:



Select Spoiler:

False



Select Source:

imdb



Show Results

Reset Input

Top 10 Reviews for Parasite_2019 based on Helpful Votes

	MovieName	date	title
2	Parasite_2019	2020-02-08 00:00:00	My God People, Really?? Get Your heads out of the sand
0	Parasite_2019	2020-01-12 00:00:00	An original dark comedy about class struggles
5	Parasite_2019	2020-02-27 00:00:00	Not that deep, severely overrated
3	Parasite_2019	2020-02-18 00:00:00	Overhyped, not that interesting and long
1	Parasite_2019	2020-02-04 00:00:00	Thy shall not ridicule fellas with body odor. Definitely deserves the best
6	Parasite_2019	2020-01-11 00:00:00	Best movie
4	Parasite_2019	2020-01-08 00:00:00	Learn to think for yourself.
8	Parasite_2019	2020-01-04 00:00:00	It told A story, nothing about people. That's all about the movie!
7	Parasite_2019	2020-03-15 00:00:00	good movie but i couldn't understand how it won the best picture

Update Functionality

Update works on updating attributes like “helpful” and “total votes”. Users will select the comments they found helpful and would like to give votes to them. And after submitting the update the counter will be increased.

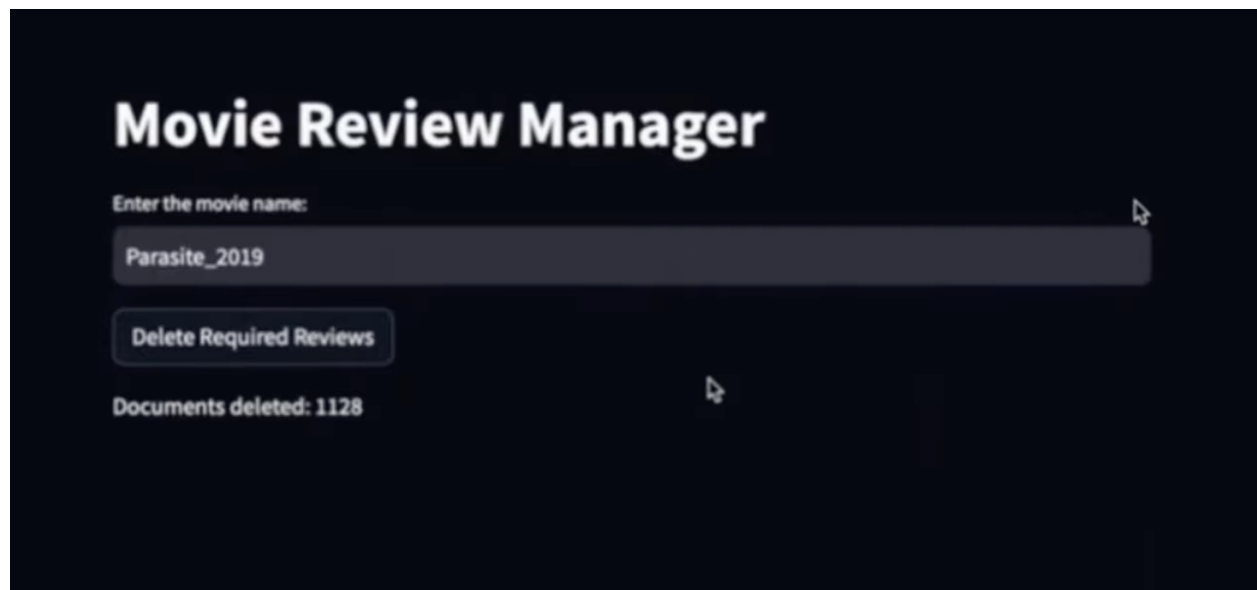
Update

comment	rating	helpful/votes	found helpful?
Is the fact that people who voted for it to win the Cannes film festival award and those who ga	5	396/760	<input type="checkbox"/>
Well written and performed also technically shines cinematography bgm are too good and the	10	379/590	<input type="checkbox"/>
As a film about a family imposing on another and keeping dark secrets hidden beneath the su	9	366/570	<input type="checkbox"/>
Do you want to see a movie that will have your stomach in knots anxious during many scenes	10	306/481	<input type="checkbox"/>
I would like to start out by saying that I am a Korean American college student studying at an /	4	295/705	<input type="checkbox"/>
After reading all the glowing reviews especially about how this film is one of the best of the de	6	279/533	<input type="checkbox"/>
Huge Oscar Winner What a JOKE There s no way or simply not enough to convince me that thi	0	278/575	<input type="checkbox"/>
A family of incompetents who struggle to fold a pizza carton miraculously transform to ingeni	1	272/524	<input type="checkbox"/>
How has this film has won the best film Oscar What a travesty It s inane rebarbative empty chi	0	256/486	<input type="checkbox"/>
This had the makings of a fun farce As I watched the first hour or so I thought it could be a ma	1	243/483	<input type="checkbox"/>

Submit Update

Delete Functionality

This function is designed to remove specific documents (entries) from a MongoDB collection based on given criteria. For our application, the criteria for the delete function is to remove all the documents people found unhelpful, and comments which have less than 20 words in them. Therefore, we deleted attributes like “helpful” is 0 and “comment” is less than 20 words.



SECTION 3

ISSUES

We faced a few issues while designing the JSON schema for our database. We followed official MongoDB guidelines and took help from the MongoDB community to resolve errors. We also faced issues while connecting the client to a server at a specific time. This issue SSL certification error. After updating the certificate using the bash command, the error was resolved.

For the LLM Application, we faced the issue of chunking the reviews. The OpenAI context length is around 4K, so we had to chunk the documents in such a way that we could provide continuity in context for LLMs to answer. So we decided to chunk the document into a length of 512 words and a chunk overlap of 35. So the current chunk has 35 from the previous chunk so it provides some continuity in context.

Also, we faced an issue with the 512 MB MongoDB Atlas size. Hence, we are not able to test our database with more than two movies. Probably with higher tiers we can test our database. But for now, we have a limited number of movies.

CONTRIBUTIONS

Name(IU email id)	Contributions
Astarag Mohapatra (astmohap@iu.edu)	Scrapping IMDB reviews website and Wikipedia data; Inserting to the database and vector database; Large Language Model Application using Langchain and OpenAI. Implementation of the Update application for the database.
Diksha Adke (dadke@iu.edu)	Working on creating databases, Creating JSON Schema, Implementation of updates and deleting parts of the application.
Chandini Sai Kumar Dukka (cdukka@iu.edu)	Written queries to extract the insights.Implemented the Insights and read part of the application.