



NextPick

Students: Shelby Larrucea, Justin Kurz, Alex Cardoso, Imran Imran,
Sebastian Perez Fematt

Instructor/Faculty: Masoud Sadjadi, Florida International University



FLORIDA
INTERNATIONAL
UNIVERSITY



Purpose

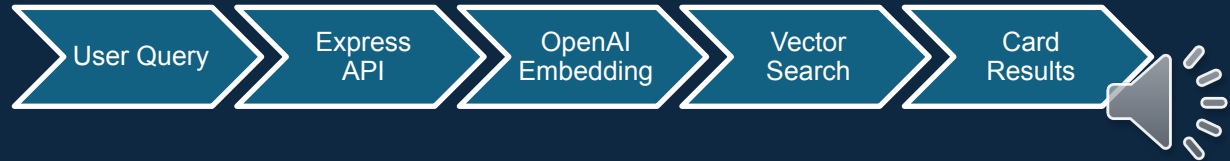
Avid reader reach a point where they have read so many books that they aren't sure what to read next. Not because of lack of but due to the difficulty of selecting a book that aligns with their preference. Traditional book sites recommend books based on genres, authors and keywords. Often times these recommendations do not align with the deeper feelings a book portrays. Our project aims at addressing this gap by creating a system that focuses on conceptual similarity rather than matching genres and authors.





System Architecture

Our capstone project is an AI driven website designed to change the way readers discover books. A user inputs their favorite or last read book and our sites takes care of the rest. The site uses a semantic search over a vectorized database of books which is powered by OpenAI embedding and Pinecone, to then return a ranked list of recommendations to the user. Our app also includes two features to alter results, excluding certain books by title/ISBN and an exclude series choice. Upon user input our site displays recommendation cards with book cover, genre, ISBN, and an AI optimized description of each result. We aim to prove how modern retrieval and AI can be combined to bring readers their next best pick.



FLORIDA
INTERNATIONAL
UNIVERSITY



Technology Stack/ Implementation

The frontend was built with HTML, CSS, and JavaScript. The backend runs on Node.js and Express, connecting the UI to the text model and Pinecone via the API. Python was also used for dataset cleaning, vectorizing, and uploading. Pinecone is what stores our vectorized database, while we use OpenAI text embedding model. While Vercel takes care of the hosting/deployment.



FLORIDA
INTERNATIONAL
UNIVERSITY



OpenAI



Vercel



Pinecone



Book Dataset

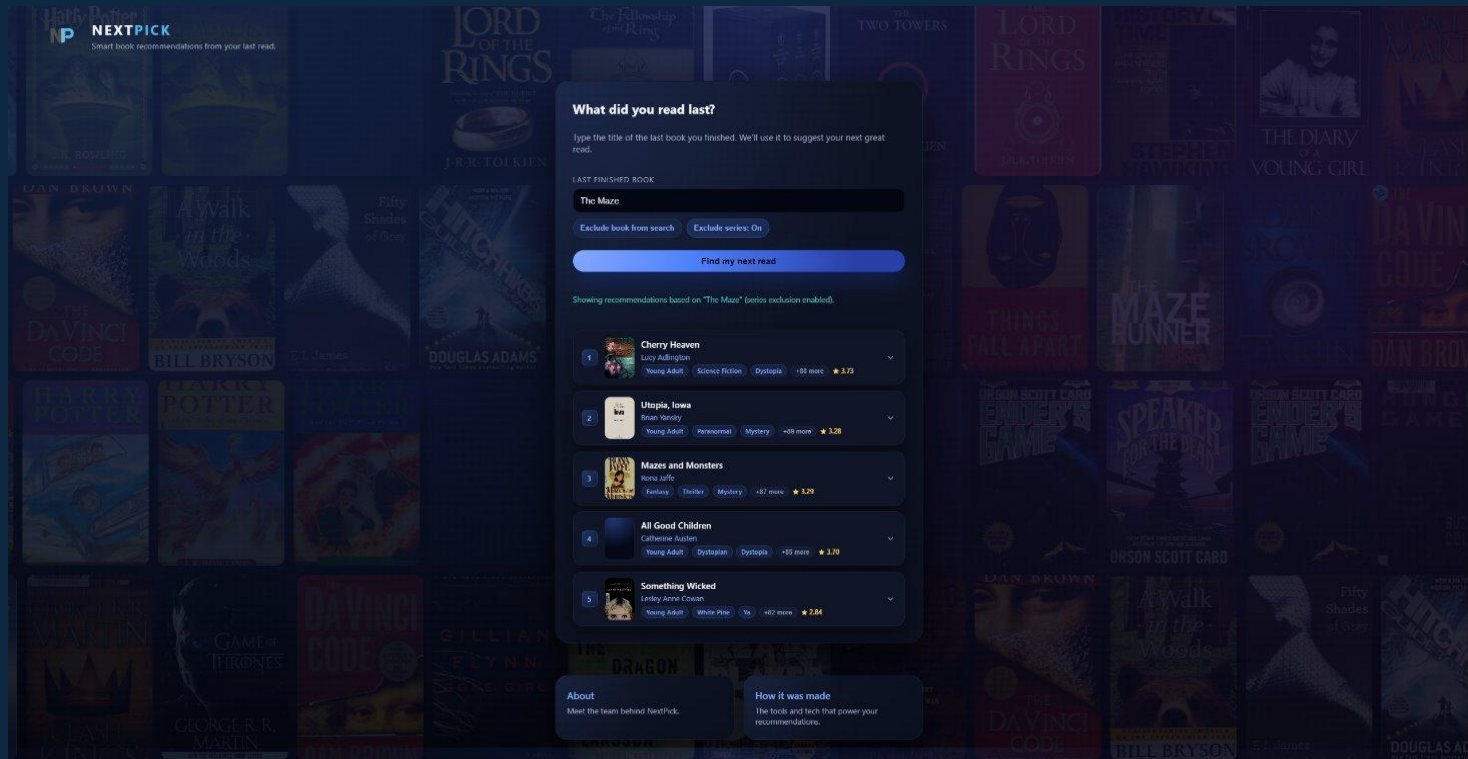
- Book data was gathered from publicly available sources
 - Goodreads
 - Kaggle
 - Universities
 - Forums
- Focused on books with appropriate data for recommendation algorithm
 - English text
 - Title and genre tags
 - User ratings
 - ISBN-13
- Data set required additional cleanup







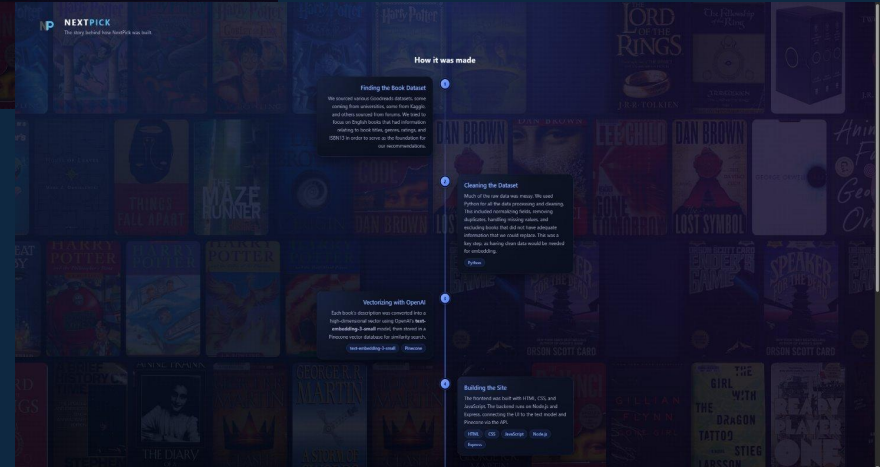
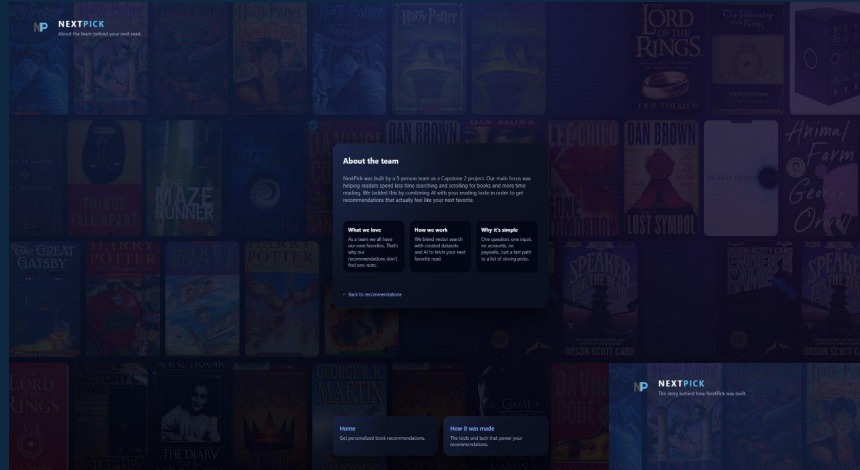
Demonstration



FLORIDA
INTERNATIONAL
UNIVERSITY



Demonstration



Challenges

- One of the biggest challenges faced is the size of book databases. Many large companies with book information, like Goodreads have enormous book datasets.
- The issue with this is the cost and time of cleaning, vectorizing, and uploading.
- With this also comes the cost of hosting as Pinecone and other DB sites charge depending on data size.





Lessons Learned

One lesson learned is the importance of checking dataset against each other. The first dataset we used was clean but small at around only 10k books. The second dataset we found was much larger around 300k+. This dataset was not clean and was missing vital information like ISBN. After cleaning this set and removing what we thought were all the null values we came to find that many ISBN were missing and had a different ID under ISBN which we had uploaded. We then had to go back and find all of these and try to run a python script with openbook api to try and find their real ISBN13.



FLORIDA
INTERNATIONAL
UNIVERSITY



Future Work

- Addition of user accounts
- Including wish lists, favorites, and perhaps an onsite voting system.
- Massive expansions to the book database.
- Search feature which brings up specific book information on page.





Conclusion

Our Ai based site has been able to demonstrate how semantic search and vector embedding are capable of solving choice paralysis, that many readers face. Going past generic recommendations off genres and authors, we have created a platform that captures the main point in recommendations. Our implementation proves that using a vectorized database and an OpenAi model allows for a more human recommendation logic. Essentially ensuring that every is just one query away from their next favorite read.

