Book Recommendations Using RAG

Rohan Arora

DATS 4001: Data Science Capstone

Abstract

This project introduces a content-based book recommendation system that uses large language models to suggest titles based on book descriptions and emotional tone, rather than user history. The system integrates zero-shot genre classification, sentiment analysis, and vector similarity search, and is built using Python tools such as LangChain, Hugging Face, and Gradio. Drawing on data from Goodreads and Open Library, it recommends books through a simple web interface.

Video Presentation with Demonstration of Application

I. Introduction

With millions of new books published each year, readers often struggle to find novels that match their interests. Recommendation systems on sites like Amazon or Goodreads aim to solve this, but these systems rely on collaborative filtering, which depends on user ratings, reviews, and the individuals' purchasing data. These systems fail when data is sparse or when books lack user feedback. A content-based approach offers an alternative by focusing on the content of the books themselves, rather than user behavior. This project follows this approach using large language models (LLMs). Using natural language processing techniques, the system analyzes book descriptions and emotional tone to match users with titles that match their interests. The recommendation engine combines vector-based similarity searches, zero-shot classification, and sentiment analysis to provide meaningful suggestions without requiring user history.

This system is built in Python, using packages like LangChain for document handling, Hugging Face Transformers for classification, and Gradio to create an interactive web interface. Users begin by entering a short book description, selecting a genre, and choosing a desired emotional tone. The system uses this input to filter and rank books from a pre-processed dataset.

| 🖻 Book Recommender L | nmender Using RAG = book Category Emotional Tone Find Recommendations All - Happy - | | | |
|---------------------------------|---|----------|----------------|----------------------|
| Enter a description of the book | | Category | Emotional Tone | Find Recommendations |
| A book on philosophy | © | | Нарру | |
| Recommendations | | | | |

It first narrows down the dataset using zero-shot genre classification, then applies emotion-based filtering using sentiment analysis. Finally, it runs a vector similarity search between the user's description and the remaining candidates to find the most semantically similar titles. Unlike systems based on popularity and user history, the model evaluates both what the book is about

and how it feels. Emotion classification helps match with moods such as sadness, hope, or excitement. The system then returns the top eight most relevant books, each accompanied by key metadata, such as title, author, and description. This setup enables a more nuanced book discovery process that goes beyond surface-level features to capture a book's content and quality.

II. Data

A. Sources

This project draws on two sources to create a dataset of books: a Kaggle dataset scraped from the Goodreads API and the Open Library API. Together, these sources provide a blend of structured and unstructured text, which supports both the machine learning pipeline and the user interface. The Kaggle dataset includes 6,810 books containing fields such as ISBN-13, title, subtitle, authors, categories, description, published year, average rating, and number of pages. These features provide a solid foundation for evaluating basic content and bibliographic details. Still, the descriptions in this dataset can be brief or missing, which limits their use for content-based modeling. To add to this existing dataset, the Open Library API was used to retrieve 26,266 books along with richer metadata. This API also provides access to edition counts, ebook availability, and preview links, adding depth to the dataset.

B. Data Cleaning and EDA

Before building the recommendation system, both datasets required significant cleaning to ensure consistency and quality in the suggestions. An initial inspection of the Kaggle dataset revealed that 262 books did not have descriptions, which are necessary for the recommendation system, so these entries were removed. After analyzing the distribution of description lengths, most entries were short, with a length under 250 words and a mean of 66 words. However, descriptions under 20 words were often vague or lacked meaningful detail. Filtering out these

2

descriptions left 5,595 books. Lastly, the dataset splits titles into two different fields: title and subtitle. To improve clarity, the fields were aggregated into a single string. This step preserved important context; for example, combining "I Am That" with its subtitle, "Talks with Sri Nisargadatta Maharaj", helps the reader and model understand the subject matter better.



A similar process was applied to the Open Library data, starting with removing books with descriptions, which was a sizable chunk of the data - 12,719 books. Compared to the Kaggle dataset, the descriptions were much longer and more descriptive, with a mean length of 114 words. Like the Kaggle dataset, books with fewer than 20 words were excluded due to the lack of detail. This filtering left 8,751 books for the merged datasets, which were merged on the unique identifier, ISBN-13. By combining two datasets, this provided the model with more information that can be used to acquire accurate book recommendations.

III. Model Setup

The model pipeline integrates three core components: zero-shot genre classification, emotion-based sentiment analysis, and vector-based similarity search. Each stage filters or ranks books to help return recommendations that match the user's description, preferred genre, and desired emotional tone.



A. Genre Classification

The initial step narrowed down the large variety of category labels, totaling 9,099 distinct categories, using a combination of rule-based mapping and zero-shot classification. First, the most common genres, such as biography, autobiography, and history, were manually categorized into fiction, nonfiction, children's fiction, and children's nonfiction. For books with multiple categories or ultra-specific genres, like children of physicians, a zero-shot classification model (facebook/bart-large-mnli) was used to assign them to the four categories listed above. This model evaluates the likelihood of a category belonging to a given set of categories without

requiring retraining. It correctly predicted the genre 89% of the time when tested against the existing labeled examples.

B. Sentiment/Emotion Classification

To further refine results by tone, the system uses the emotion-english-distilroberta-base model to analyze the emotional content in each book description. After testing this model against the entire book description, it yielded inaccurate results, as the model would often focus on one or two sentences in the description for its analysis. To address this, descriptions were split into sentences, truncated based on token length, and passed through the classifier to identify seven emotions: anger, disgust, fear, joy, sadness, surprise, and neutral. The model output scores for each emotion, and the highest score across all sentences were retained for each label.

C. Vector Embedding and Similarity Search

The system then generated vector embeddings from the book descriptions to support content matching. Vector embeddings are a numerical representation of text in a high-dimensional space, where similar texts are positioned closer to each other. To create these embeddings, a simplified text file was created containing only the ISBN-13 and description variables. This smaller file was passed into a LangChain-based embedding pipeline using the OpenAI Embeddings API. LangChain handles this through a structured pipeline that splits the documents into manageable chunks, processes them via a



transformer-based language model, and generates dense 1536-dimensional vectors. Each vector captures features of the text, such as thematic content, genre-related vocabulary, and emotional undertones, allowing the system to capture nuanced similarities even when explicit keywords differ. These vectors were stored in a vector database, enabling fast similarity searches across all of the embedded documents. This workflow follows the retriever and embedding concepts defined in LangChain's documentation: embeddings act as a projection of meaning into a vector space, while the receiver uses similarity metrics to return the most relevant documents given a query vector.

When a user inputs a custom query or mentions a specific book, that input is embedded using the same process. The system then computes the cosine similarity between the user's input vector and the precomputed book vectors to rank books based on semantic closeness. Cosine similarity compares the angle between vectors, making it ideal for detecting shared meaning. For

$$\cos(heta) = rac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = rac{\sum\limits_{i=1}^n A_i B_i}{\sqrt{\sum\limits_{i=1}^n A_i^2} \sqrt{\sum\limits_{i=1}^n B_i^2}}$$

example, if a user searches for a "gritty detective thriller," the system may give a book described as a "dark mystery" if the underlying semantic vectors are close. The top matches are then reconnected to the full dataset using ISBN-13 as the key, and the genre and emotion filters are applied. Ensuring that only books that match the user's selected category and tone are included in the recommendations.

IV. Results

The final system offers an intuitive dashboard for book recommendations. Users enter a sort description of the type of book they are looking for, select a genre, and choose an emotional tone. In return, the system displays eight recommended books that align with the user's inputs across all three dimensions: content, category, and tone. The recommendations are displayed with relevant metadata, the title, author(s), and descriptions.

| Construction Provide Harden Har | | | | | | |
|--|---|------------------------------|---|--|---|--|
| <form> Image of a longer of a longer Image of a longer Image of a longer Image of a longer Image of a longer Image of a lon</form> | S Book Recommender Using | g RAG | | | | |
| <form> in the control Control <</form> | Enter a description of the book | Category | | Emotional Tone | Find Recommendations | |
| Procursmediations Procursmediations <t< td=""><td></td><th>Fiction</th><th></th><th></th><th></th><td></td></t<> | | Fiction | | | | |
| The region of the loss of the region of t | Pecommendations | | | | | |
| The vacuum of th | Johnny Got His Gun by Delton Trumbo | | | | | |
| He aag way uut. It is should up voent, tarrifying hontids, uucompromising brug, emouwees and guesome - but to is war - Putableu: He aag way uut. It is should up voent, tarrifying hontids, uucompromising brug, emouwees and guesome - but to is war - Putableu: He was not should up voent, tarrifying hontids, uucompromising brug, emouwees and guesome - but to is war - Putableu: A Que to the Western Franct by frich Maria Amangue A Que to the Western Franct by frich Maria Amangue A Que to the Western Franct by frich Maria Amangue A Que to the Western Franct by frich Maria Amangue A Que to the Western Franct by frich Maria Amangue A Que to the Western Franct by frich Maria Amangue A Que to the Western Franct by frich Maria Amangue Maria and the constance of the shade and the constance of the shade and the constance of the shade and the shade and the constance of the constance of the constance of the shade and the constance of the constance | This was no ordinary war. This was a war to make the w | orid safe for democracy. An | d if democracy was made safe, then nothing | ise matterednot the millions of dead bodies, nor the thousan | ds of ruined lives This is no ordinary novel. This is a novel that never takes | |
| The task inform \$2. Percentaining The task inform \$2. Percentaining A between Proof by probably the meet-news a notes itsustain eminazity. If the there a denotes of Europers underground-set in motion is concatenation of world abadring. Ustimutely find exerts. Ad Quere on the weetern Froot by probably the meet from probably the meet from and an one leave entern. The story is tool by a young underone solitar in the mentohes of Biodenic aduring the First. World Weet. Through The specifies and denoticities which iterate in motion is advected by a story of under from a networks of the relation is advected by a story of under from a networks of advected by a story of under from a networks of advected by a story of under from a networks of advected by a story of under from a networks of advected by a story of under from a networks of advected by a story of under from a networks of advected by a story of under from a networks of advected by a story of under from a networks of advected by a story of under from a networks of advected by a story of the freeling of pound betraped and a deceptively simple indictionent of war- of advected by a story of the freeling of pound betraped and a deceptively simple indictionent of war- of advected by a story of the best of the story of the freeling of pound betraped and a deceptively simple indictionent of war- of advected by a story of the best of the story of the best of | the easy way out: it is shocking, violent, terrifying, horrib | ole, uncompromising, brute | , remorseless and gruesome but so is war. | Publisher. | | |
| be take from 54. Beterstung to yoor failer: Act betree Voord Wir 1 wor men- one is note it austain emissary, the other is denoteen of Europe's underground-set in motion is concatenation of world abadeing ultimately fatel events. A Color on the Weetern Front to globability the motil famous and incern note is event states to do you you you known states if in the entrope's and abadeing uberstual diverse and its event states and the entrope's and denote you world you known states if in the entrope's and denote you world you known states if in the entrope's and denote you world you known states if in the entrope's and denote you world you have and the entrope's and denote you world you have and abadeing uberstual diverse and its event and a denote you have and you | | | | | | |
| In a bandwords is detailed by the owner and observed the other a detailed on the studies and detailed by the owner and water of Europees underground-set in motion a concentration of world abailing. Litimately field events. A Cache on the Weetern Front by <i>Dich Alada</i> Remarks, the other a detailed on the studies and detailed by a more and event water. The assay is under y survey water motion is an observe of Europees and detailed and an event of Europees and detailed and an event of the field event. A Cache on the Weetern Front by <i>Dich Alada</i> Remarks, the other a detailed of the detail to depend as a detailed of the detailed and detailed and detailed and an event of the field of youth betrayed and a detailed and an event of the field of youth betrayed and a detailed and an event to the integend and a detailed are the world gene med. Robert Respective of the field of youth betrayed and a detailed are the world gene med. Robert Respective are integend and an event to the integend of the data and an event of the field of the data and an event of the field of the data and an event of the field of the data and an event of the data and event of the data a | | | | | | |
| Al Que on the Western Front by <i>Drich Maria America</i> in the part of the way and a second product with the second of the term on base, and honoray and a second of the part of the term of the second of the term of | The Man from St. Petersburg by Ken Posetr | emicrany the other a depi | an of Europe's underground uset in motion a | representation of world staking ultimately fatal sweets | | |
| Al Quere on the Western Front. Is probably the most demonstration of lease index in the story is dod by a purp yurknown kiddler in the teachers during the Frist World West. Through the was wester all the makines of devolution of the story of the feeling of purp benerged and a developing time president of the sensities of the story on teace, and in hospitals and developing time in the story of the feeling of purp benerged and a developing time per video dev | | | an or Europea underground-set in motion a | An a series of the second strategy and the system over the | | |
| A Quet on the Weetern Floret to plotably the most flores and varies on root a low most and with the The tary is to dd by s young vinces on solder in the menches of indexes during the Tits. World with Through his eyes we see all the needless of warr, under fine, on passes, which and the tar is no sense of advecture inter, only the feeling of youth beersyed and a deception of any under fine, on passes, which and the tar is no sense of advecture inter, only the feeling of youth beersyed and a deception of advectore inter, only the feeling of youth beersyed and a deception of advectore inter, only the feeling of youth beersyed and a deception of advectore inter, only the feeling of youth beersyed and a deception of advectore inter, only the feeling of youth beersyed and a deception of advectore inter of war - of advectore inter on the advectore inter, only the feeling of youth beersyed and a deception of advectore inter of war - of advectore inter on the advectore inter, only the feeling of youth beersyed and a deception of advectore inter of war - of advectore inter on the advectore inter, only the feeling of youth beersyed and a deception of advectore inter, only the feeling of youth beersyed and a deception of advectore inter, only the feeling of youth beersyed and a deception of advectore inter, only the feeling of youth beersyed and a deception of advectore inter, only the feeling of youth beersyed and a deception of advectore inter, only the feeling of youth beersyed and a deception of advectore inter, only the feeling of youth beersyed and a deception of advectore inter, only the feeling of youth beersyed and a deception of advectore inter, only the feeling of youth beersyed and eater advectore inter, of the site of advectore inter on the site of advectore inter on the advectore inter, only the feeling of youth beersysted and eater advectore inter, only the feeling of youth beersysted and eater advectore inter, only the feeling of youth beersysted and eater advectore inter, only the feeling of youth beersyst | | | | | | |
| All operation that weatern fronts generally in propulsity the most atmonue and weatern. The story is toold by a young variescent solidary in the matches of flanders guing the Fitst. World Web. Through the space weater at weate | All Quiet on the Western Front by Erich Maria Remarge | | | | | |
| In the final and wave, and on indexes and undexes and undexplaced and undexpla | All Quiet on the Western Front is probably the most fam | nous anti-war novel ever w | itten. The story is told by a young 'unknown s | ldier' in the trenches of Flanders during the First World War. T | hrough his eyes we see all the realities of war; under fire, on patrol, waiting | |
| The Want by Findely Robert Disk, a sandbar inforcement year disk Caladian officer, went to war- the Warts Din duit Yors: Na found Inmart for the night make world of tranch warters of muil and enosis, of choiring gas and noting corpans. In this world yors mult, hoart hous partormed a disk term by Findely Findely Faster by Funders act to decide the commitment to the in the midter of dealth. The Warts Ball do simply one of the bars novels out the First World War. Faster by Funders Burdis: The question of accidential were is examined in this novel about American planes which By past the point of reals to dip nuclear borrist on Moscow. A Persian requirement a novel by Simin Duinchear Chronicies the life of Zari, the wile of a fundal landiency as the attempts to preserve the finally emidiat the turbulence of the Bitish occupation of southern Persa during World War II With the Bits by Lawson Classop Result of the Simon covel based on the subtra experiences as one of the TRats of Totoruk in the Middle East during WWII. First published in 1944. Part of the Turbutalian World Classor's series introduced by Sir Edward Tweey Cuntop. | war - told for a | aressing stations. Although | there are vividly described incidents which h | main in mino, there is no sense or adventure here, only the ree | ing of youth becayed and a deceptively simple indictment of war - of any | |
| The Yan's by Throshy Findsly Rebert Risk, a sensitive intensensyses did Canadian officer, went to war - the War is Did All Wars. He found himself in the nightmare world of trench warder, of nuclear did did noties, of choine gas and rotting corposes. In this world gone mad, Robert Risks, performed a ind sequence act to decide the commitment to life in the model of didt. The Wars is quite sembly one of the best tooles ever written about the Field World War. Fed setts for gapene Burdick: The question of accidential war is examined in this novel about American planes which fly gass the point of nuclear bornts on Moscow. | | | | | | |
| The Watch Test Journal of Fondage The Watch Test Journal of Fondage The Watch Test Journal of Test Journal International Charakterian officers, works to war - the Watch Test Journal of Journal In the ingetterman works of drawnth warker, of invul and smoke, of chlorine gas and rothing corpaes. In this works joore mad, Babert Bass performed a last espenses act to declare his commitment to life in the middle of datath. The Wats is quite simply one of the best novels ever written about the First World Wat. The question of accubertal works of statements and the first World Wat. The question of accubertal works of statements and the first World Wat. The question of accubertal works of statements and the first World Wat. The question of accubertal work is searmined in this novel about American planes which fly past the point of recall to drop nuclear bornts on Mozcow. | | | | | | |
| In the species are to decide by Committee to Shin in the most of dealth the Was is quite simply one of the best novels and the final action of a shink of the best novel give at the be | The Wars by Timothy Findley | | | | | |
| Pak sets by Jisgons Burdick. The question of accidential war is examined in this nowel about American planes which fly past the point of neal to drop nuclear bornts on Moscow. A Persian requirement a nowel by Simin Distributor Crements the life of Zeri, the write of a floadal landtore, as the attempts to preserve her family artistict the turbulence of the British occupation of southern Pensa during World War I Were the Bits of Lesson Classop Ressult of the Simola novel based on the subort experiences as one of the Tates of Tobruic in the Middle East during WWN. First published in 1944 Part of the Turbutanian wer Classop: anisot and the subort experiences as one of the Tates of Tobruic in the Middle East during WWN. First published in 1944 Part of the Turbutanian wer Classop: | last desperate act to declare his commitment to life in t | he midst of death.The War | s is quite simply one of the best novels ever w | tten about the First World War. | as and rocong corpses. In this world gone mad, Robert Rois performed a | |
| Parkane by Eugene Burdick The question of accidental we're aanined in this novel about American planes which fly past the point of recall to drop nuclear bombs on Moscou. A Parkan requirem: a novel by Simin Duinshear Choncies: the life of Zari, the write of a fould landford, as the attempts to preserve her family amidst the turbulence of the British occupation of southern Persia during World Wer II We Were the Bats by Lanson: Classop Revent the Bats by Lanson: Classop Concisient the Bats on novel bases on the suchors experiences as one of the TRats of Tooruk's in the Middle East during WWIL Piss published in 1964. Part of the Taluarsilian Wer Classor's arries introduced by Sir Edward TWeary Europe. | | | | | | |
| Patisation by (signer allow) The question of accidental war is examined in this novel about American planes which fly past the point of recall to drop nuclear bombs on Moscow. A Persian requirem a novel by Simin Duhinhear Chronices the life of Zivi, the write of a found landord, as the attempts to presence her family ansids the turbulence of the British occupation of southern Press during World War II We Wave the Rets by Lesson Classop Reasone of the Simon David based on the authors experiences as one of the TBase of Toorisk' in the Middle East during Wwit, Fred published in 1944. Part of the TAustralian War Classor'sames introduced by Sir Edward TWeary Duriop. | | | | | | |
| A Pursian request to account a were searning on the new account which a particular with part to the and up notice to the British occupation of southern Persa Auring World War I | Fail-safe by Eugene Burdick | | | | | |
| A Persian requires a novel by Simih Dánkhvar Chronicies the life of Zari, the write of a fundal landtord, as the attempts to preserve her family ansids the turbulence of the lifetish occupation of southern Pensia during World War II We Were the Bes by Lenson Classop Ressue of the Samod new Ibased on the authors experiences as one of the Tats of Tobruk in the Middle East during WWN. First published in 1944 Part of the TAustralian War Classocraines introduced by Sir Edward TWeary Dunios. | The question of accidental war is examined in this nove | rabout schencari planes w | nich hy past the point of recail to drop hocies | bombs on Moscow. | | |
| A Pardian regularis a novel by Shinh Dúnishvar Chronicies the life of Zar, the wife of a fould landbord, as the attempts to preserve her family antidis the Lubulence of the British occupation of locativem Persa during World Wer I | | | | | | |
| Chronicles the life of Zari, the wife of a fuedal landlord, as the attempts to preserve her family antidist the turbulence of the British occupation of southern Pensia during World War II We Were the Rats by Lawoon Classop Resource of the Brances novel based on the authors experiences as one of the Thats of Tobruk's in the Middle East during WWII. Prist published in 1944, Part of the Tututralian War Classocr series introduced by Sir Edward TWeary Duritop. | A Persian requiem: a novel by Simin Dänishvar | | | | | |
| We Were the Bat's by Lawson Classop Resourd of the famous novel based on the author's experiences as one of the TBats of Tobrus' in the Middle East during WWIL Fast published in 1944. Part of the TAudoslan War Classics' series introduced by Sir Edward TWeary During. | Chronicles the life of Zari, the wife of a feudal landlord, a | as she attempts to preserve | her family amidst the turbulence of the Britis | occupation of southern Persia during World War II | | |
| We Were the Rats by Lesson Glassop Reason of the Bandus novel based on the author's experiences as one of the TRats of Tokruk's in the Middle East during WWI. First published in 1944. Part of the TAustralian War Classor's areas introduced by Sir Edward TWeary Durings. | | | | | | |
| We wave the Bate by Laward Castop Resource of the Bamous novel based on the author's experiences as one of the TBate of Tobruk'in the Middle East during WWIL First published in 1544. Part of the TAustralian War Classics' same introduced by Sir Edward TWeary Curriop. | | | | | | |
| Resour of the famous novel based on the author's experiences as one of the Teals of obruik in the Middle East during WWIL Frid published in 1944. Part of the Tulustralian War Classics' series introduced by Sit Edward TWeary During. | We Were the Rats by Lawson Glassop | | | | | |
| | Reissue of the famous novel based on the author's expe | eriences as one of the TRats | I OF TODRUK IN THE MIDDLE East during WWII. F | st published in 1944. Part of the TAustralian War Classics' serie: | s introduced by Sir Edward TWeary' Dunlop. | |
| | | | | | | |

However, this system does have limitations. Since it only uses the descriptions and not full book content, recommendations are based solely on how they are described, which can lead to surface-level matches when descriptions are vague or poorly written. Emotion classification also depends on sentence-level context and can misinterpret descriptions with shifting or ambiguous tone. These constraints highlight the importance of high-quality descriptions and suggest opportunities for future refinement, such as incorporating full-text analysis or additional metadata like user tags or themes.

V. Bibliography

Trivedi, Ayushi. (2024, October 4). *Top 6 Books on Retrieval Augmented Generation (RAG)*. Analytics Vidhya.

https://www.analyticsvidhya.com/blog/2024/10/books-on-rag/

Castillo, D. J. (n.d.). 7k Books with Metadata [Data set]. Kaggle.

https://www.kaggle.com/datasets/dylanjcastillo/7k-books-with-metadata

Dhapre, M. (2024, February 2). *Book Recommendation using Retrieval Augmented Generation*. Medium.

https://medium.com/@mrunmayee.dhapre/book-recommendation-using-retrieval-augmen ted-generation-52965b71ed16

Jonathandika. (n.d.). *llm-recommender-system*. GitHub.

https://github.com/Jonathandika/llm-recommender-system

Open Library. (n.d.). Developer Center / APIs. Open Library.

https://openlibrary.org/developers/api