NLP Final Individual Report

Title: AI Book Recommendations

Completed by:

Caitlin Bailey

Due May 1, 2024

Introduction

In this project my groupmates and I developed three different NLP models to recommend books to users using the Amazon Books Reviews dataset [1]. The overall project idea was developed by Nina, and Nina identified the dataset for the project, as well. We each took the lead on developing one of the three recommendation models. I took the lead on developing the SVD model, while Liang developed the KNN model and Nina developed the DeepLake model [2]. I also led development of the streamlit demo application and drafting our initial group report. Liang took the lead on developing our streamlit presentation. We each contributed to the group report and the presentation. Overall, this project was a very productive and complimentary collaboration.

Description of My Individual Work

I developed the SVD model. I wrote all SVD model code and performed all experimental tests, including identifying the optimal number of components for the LSA and qualitatively analyzing the model output. In addition, I developed the preprocessing code, the streamlit application (including merging all of our models into one demo), and I drafted the initial version of our group project report. I also collaborated with my groupmates' to troubleshoot their models as issues arose.

SVD Model Development. I researched and drafted the background information for the SVD model as follows below.

The SVD model is a classical NLP model typically used for dimensionality reduction and latent semantic analysis. It was initially developed in the late 1960s to decompose matrices and extract meaningful patterns from high-dimensional data [3]. In the context of book reviews and recommendations, the SVD is a clear choice from the classical NLP model toolkit because it allows us to capture the latent semantic structure of the reviews and identify underlying topics or themes. This enables us to generate recommendations based on review similarities and identify related books.

The model was implemented using the sklearn package, which efficiently implements various machine-learning algorithms. The SVD model architecture I developed for this task takes in user input as a short (recommended: 3-5 sentence) book review. Using cosine similarity, the model outputs the top most similar reviews (deduplicated) and relevant linked data from the merged Amazon Books Reviews datasets (i.e., book title, author, book review summary, review rating).

Strengths of the SVD model include its ability to capture complex relationships in high-dimensional data, its interpretability, and its flexibility in handling different types of input data. Drawbacks of the model include the lack of CPU optimization for sklearn, which can lead to longer training times for large datasets. However, this limitation can often be mitigated by leveraging parallel processing or using optimized libraries for specific tasks.

The SVD equation can be written as:

$$A=U\Sigma V^T$$

Where A is the original matrix (e.g., document-term matrix), U is the left singular vectors matrix (e.g., a representation of the “concepts” or latent factors, capturing relationships between documents [i.e., rows]), Σ is the diagonal matrix of singular values (e.g., the representation of the significance of each latent factor), and V^T is the transpose of the right singular vectors matrix (e.g., a representation of the latent factors, capturing relationships between terms [i.e., columns]). SVD decomposes the original matrix A into orthogonal basis vectors (in U and V^T) and scaling factors (in Σ), collectively representing the data's latent structure. This decomposition enables dimensionality reduction and semantic analysis.

Experiment. My experimental setup for developing the SVD model is described next.

I designed the SVD model architecture to process natural language text as a short book review and output the top most similar book reviews (using combined title and summary review data) using cosine similarity. I used the NLTK package to preprocess the text, including tokenization, removal of special characters and stopwords, and lemmatization. The sklearn package performed the TF-IDF vectorization, SVD, and LSA pipeline. The number of components for the model was selected so that 90% of the variance was explained, striking a balance between maximizing information retention and preventing overtraining (see Figure 1) [4]. The final trained model was saved as a pickle file for later access.

Given the unsupervised nature of the SVD model, conventional assessment methodologies like train-test splitting, accuracy, and F1 scores were not applicable. Instead, I developed a novel, fit-for-purpose way to evaluate the model's efficacy. After using scree plots of cumulative explained variance and explained variance ratio to gauge the extent to which the model captured and retained essential information from the dataset, I manually assessed the model output to provide qualitative insights into the model's performance. My user-generated assessment ensured alignment of the model with the intended application (book recommendation). The assessment process I designed consisted of inputting eight book reviews of different genres (i.e., mystery, romance, science fiction, fantasy, thriller, historical fiction, young adult, and nonfiction) and manually assessing how many of the model’s recommended books were similar to the review prompt in terms of genre/topic. Book reviews for the model assessment were generated by ChatGPT (see Table 1) [5].

Table 1. Book reviews generated by ChatGPT for eight different genres.

Genre	Review
Mystery	With clever twists and turns, this mystery novel keeps readers on the edge of their seats as they follow the detective's relentless pursuit of the truth. Every clue unravels another layer of intrigue, leading to a shocking revelation that will leave you guessing until the very end. A gripping page-turner that masterfully blends suspense and deduction, making it a must-read for fans of the genre.
Romance	In this heartwarming tale of love and second chances, sparks fly when two unlikely souls cross paths. With its tender moments and passionate encounters, this romance novel sweeps readers off their feet into a whirlwind of emotions. The characters' chemistry is palpable, drawing you into their journey of self-discovery and redemption. A captivating story of love's transformative power that will leave you longing for more.
Science Fiction	Set in a distant future where technology reigns supreme, this science fiction epic explores humanity's quest for survival in the face of existential threats. With its futuristic landscapes and visionary ideas, the novel immerses readers in a world of awe-inspiring possibilities. From space exploration to artificial intelligence, it delves into the ethical dilemmas and moral quandaries of a technologically advanced society. A thought-provoking adventure that pushes the boundaries of imagination and intellect.
Fantasy	Journey to a realm of magic and wonder in this enchanting fantasy novel filled with mythical creatures and epic quests. With its richly imagined world and vibrant characters, the story transports readers to a place where anything is possible. From ancient prophecies to epic battles between good and evil, the narrative weaves a tapestry of adventure and intrigue. A spellbinding tale that captivates the imagination and leaves a lasting impression.
Thriller	Prepare to be on the edge of your seat with this pulse-pounding thriller that delivers non-stop action and suspense. From the opening scene to the heart-stopping climax, the tension builds with each twist and turn of the plot. With its complex characters and high stakes, the novel keeps you guessing until the very end. A gripping rollercoaster ride of thrills and chills that will leave you breathless.
Historical Fiction	Step back in time to an era of intrigue and upheaval in this meticulously researched historical fiction novel. Through vivid storytelling and evocative prose, the author brings the past to life, immersing readers in the sights, sounds, and struggles of bygone eras. From sweeping sagas of war and conquest to intimate portraits of ordinary lives, the novel paints a vivid tapestry of history. A captivating glimpse into the past that resonates with timeless themes of love, loss, and resilience.
Young Adult	Navigating the tumultuous waters of adolescence has never been more captivating than in this poignant young adult novel that explores the trials and triumphs of growing up. From first love to friendship struggles, the story delves into the complex emotions and experiences of teenage life. With its relatable characters and authentic voice, it resonates with readers of all ages, capturing the essence of youth with honesty and empathy. A coming-of-age tale that speaks to the heart and soul of every teenager.
Nonfiction	From riveting biographies to compelling exposés, this collection of nonfiction essays offers a fascinating glimpse into the diverse tapestry of human experience. With its thought-provoking insights and meticulously researched facts, each essay sheds light on a different aspect of the world around us. From science and history to politics and culture, the authors explore a wide range of topics with depth and clarity. A compelling anthology that challenges assumptions, sparks conversation, and broadens horizons.

Results. My experimental results are described below.

For the SVD model, 17 components were found to explain 90% of the variance (see Figure 1). I selected 90% cumulative variance as my threshold, *a priori*, to ensure that a significant portion of the original variance in the dataset was captured and preserved in the reduced-dimensional space. This can help maintain the richness and complexity of the data while still achieving benefits of dimensionality reduction.

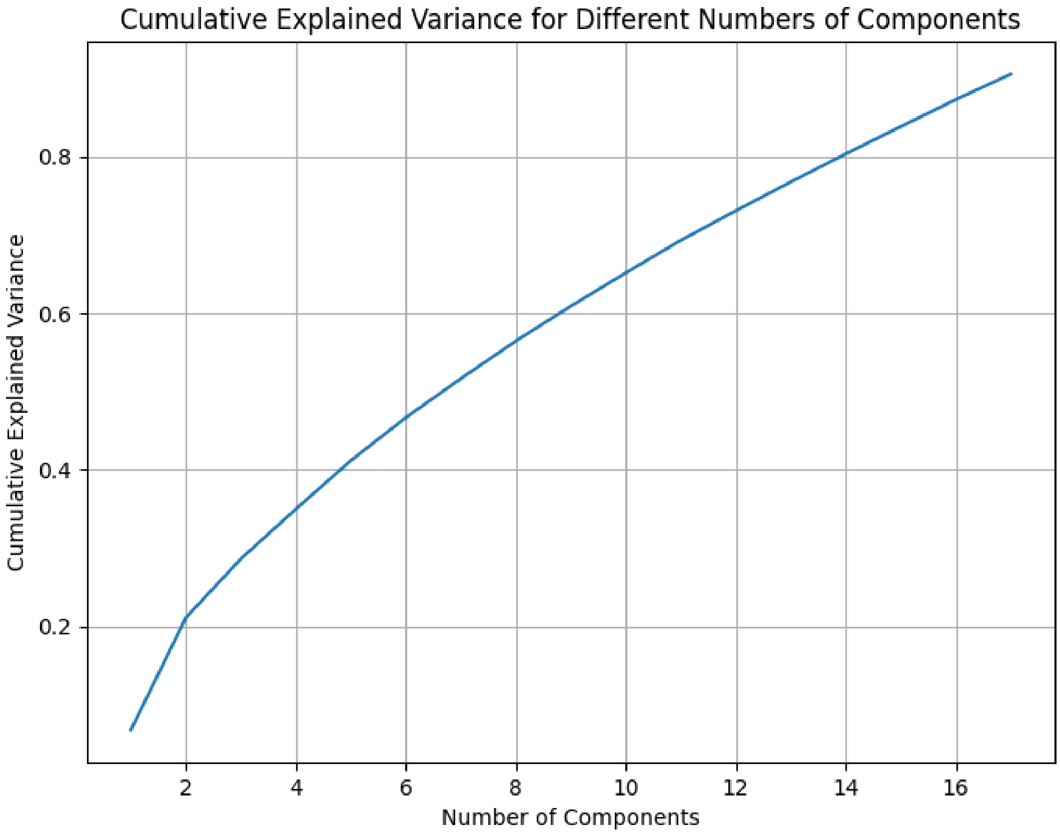

Figure 1. Scree plot of the cumulative explained variance per number of components. Seventeen components explained 90% of the cumulative variance for the SVD model.

I also examined a scree plot of explained variance ratio (see Figure 2). This plot indicates that the explained variance ratio peaks at component 2. However, the magnitude of the explained variance ratio is small across components. While the explained variance ratio plot shows diminishing returns after two components, I considered the model's overall performance in achieving its intended task prior to selecting the final number of components. I knew that retaining more components might lead to better performance in tasks such as similarity calculation and recommendation generation, where capturing subtle nuances in the data is crucial. To test this hypothesis, in my manual model output assessment, I tested three SVD models: an SVD model with 17 components (90% of cumulative variance explained), an SVD model with 2 components (peak of explained variance ratio), and a saturated SVD model with 20 components.

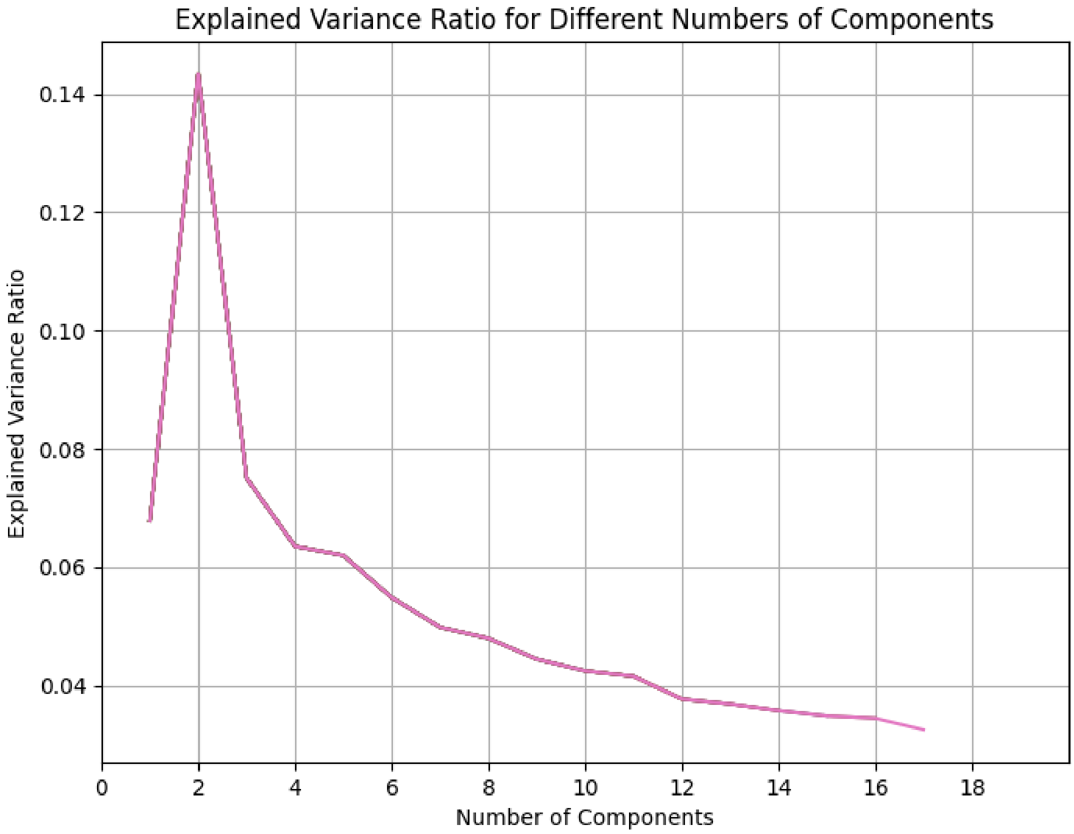

Figure 2. Scree plot of the explained variance ratio per component. Explained variance ratio peaks at component 2 for the SVD model.

Based on my assessment of model outputs, I found that the SVD model with 17 components was significantly better at correctly recommending books in the same genre or topic area as the review prompt compared with the 2-component model (see Table 2). Compared to the SVD model with 17 components, the SVD model with 20 components was slightly better for prediction in some genres/topics and slightly worse in others. Thus, for my final demonstration, the SVD model with 17 components was selected for implementation.

Table 2. Proportion of book recommendations similar to the review prompt in terms of genre/topic.

	SVD (2 components)	SVD (17 components)	SVD (20 components)
Mystery	0/10	8/10	8/10
Romance	1/10	9/10	7/10
Science Fiction	0/10	7/10	4/10
Fantasy	0/10	7/10	9/10
Thriller	1/10	8/10	6/10
Historical Fiction	1/10	7/10	8/10
Young Adult	2/10	6/10	6/10
Nonfiction	10/10	10/10	10/10

Summary and Conclusions

In summary, my project contribution was mainly focused on developing the SVD book recommendation model. In addition, I lead the development of the preprocessing code, as well as the drafting and development of both our streamlit application and our group report. I also assisted in troubleshooting groupmates' models at various points.

The SVD model had strengths and limitations. For example, the model was relatively straightforward to develop and implement (being a classical NLP model) and did not require GPU resources. This could be beneficial for a small company or an individual looking to implement a model without large resources. However, it is also a limitation of the model in that it ran on CPU and therefore took some time to initially reduce dimensionality and train with the 3GB Amazon Books Reviews dataset. It also required more attention to rule-based text preprocessing tasks and manual qualitative assessment of output, given that it was a classical model using unsupervised learning techniques.

Overall, the comparison between SVD, KNN, and DeepLake models was an interesting thought experiment. Comparing the application of more resource intensive models, such as transformers, to classical models, such as SVD, is an important experiment to ensure that a specific application does in fact benefit from utilizing more resource intensive models. Sometimes, simple may be effective enough, particularly for small focused tasks and/or small companies. I encourage others to build on our work in future applications of book (or other product) recommendation.

Code found on Internet: Approximately 40-50% of my code was found on the internet, taking into account modification and addition of my own code.

References

1. Bekheet, M. (2022, September 13). *Amazon Books Reviews*. Kaggle. https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews

2. DeepLake version 3.9.0. Deep Lake Docs. (2024). https://docs.activeloop.ai/?utm_source=github&utm_medium=github&utm_campaign=github_readme&utm_id=readme

3. Stewart, G. W. (1993). On the Early History of the Singular Value Decomposition. *SIAM Review*. 35(4):551-566. doi:10.1137/1035134.

4. Bushel, P.R. (2021). Principal Variance Component Analysis. National Institute of Environmental Health Sciences. Retrieved from: https://www.niehs.nih.gov/research/resources/software/biostatistics/pvca

5. OpenAI. ChatGPT. Version 3.5. Retrieved from https://chat.openai.com