Cross Domain Search Recommendation Engine - CS 410 Final Project

Aarush Shah
aarushs2@illinois.edu
University of Illinois Urbana Champaign

Jay Sunil Goenka
jgoenka2@illinois.edu
University of Illinois Urbana Champaign

Yash Jain
yash7@illinois.edu
University of Illinois Urbana Champaign

Yash Mandavia
yashm4@illinois.edu
University of Illinois Urbana Champaign

ABSTRACT

To encapsulate our project in a nutshell, this project introduces a Cross-Domain Recommendation Engine for Personalized Storytelling, which connects users' favorite movies with tailored anime suggestions. By leveraging natural language processing (NLP) techniques, such as KeyBERT, Sentence-BERT, and LDA our model captures thematic, emotional, and narrative details to provide meaningful recommendations. Challenges included dataset refinement, integrating sentiment analysis with embedding models, and achieving cross-domain relevance. The project provided insights into blending computational methods with artistic narratives, enriching our understanding of content personalization and advanced text processing.

ACM Reference Format:
Aarush Shah, Jay Sunil Goenka, Yash Jain, and Yash Mandavia. 2018. Cross Domain Search Recommendation Engine - CS 410 Final Project. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

In recent years, **anime content** has witnessed an extraordinary surge in global popularity, captivating audiences through its diverse storytelling, rich visual styles, and complex emotional arcs. To meet the growing interest in more personalized **content recommendations**, we decided to work on a **cross-domain recommendation engine** designed to bridge the gap between mainstream movies and anime.

By accepting movies that users love as input, this system conducts an in-depth analysis of **plot structures**, thematic elements, and emotional trajectories, ultimately providing users with a carefully curated and ranked list of anime that align with their narrative preferences. This platform offers new users a unique way to discover anime based on movies they already enjoy, and avid anime fans a fresh way to explore titles that resonate with their specific tastes.

2 PROJECT DESCRIPTION

The proposed recommendation system leverages personalized input from users, combined with advanced natural language processing techniques, to understand not just the surface-level attributes but the core emotional and thematic nuances of each story. Below is an overview of the system's components:

- Users enter their favorite movies. By analyzing these selections in depth, the system identifies the distinctive aspects of each story that contribute to the appeal for the user.
- Moving beyond traditional genre-based recommendation engines, this system employs sophisticated sentiment and narrative analysis to decode the emotional journey within each story. Through these insights, it discerns whether the narrative concludes with hope, bittersweet reflection, or perhaps a sense of introspection, allowing for more meaningful matches based on emotional resonance and narrative depth.
- For instance, users who appreciate narratives centered on redemption, resilience, or personal growth will receive suggestions that prioritize similar narrative arcs and thematic undertones, ensuring the recommendations are genuinely reflective of their tastes.
- The system compiles a ranked list of anime recommendations most closely aligned with the user's preferences. Each recommendation is accompanied by a brief contextual summary, providing users with insight into the content.

3 METHODOLOGY

In our project, we employ **KeyBERT** and **Sentence-BERT** to create a personalized anime recommendation system based on the movies a user enjoys. With **KeyBERT**, we extract key phrases from each movie summary to capture essential elements like **themes**, **genres**, tone, and mood, as well as character dynamics, setting, and target audience. For instance, a sci-fi movie summary might yield phrases such as "space mission," "alien life," and "future technology," while also highlighting character relationships ("father-son bond") or setting details ("dystopian future"). **KeyBERT** will work through all the suggested movies, compiling these important relations. These key phrases will then form a profile of the user's preferences, helping us understand not only the broad genre of each movie but also subtler aspects like tone and narrative elements.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

To find anime with similar content, we use **Sentence-BERT** to generate semantic embeddings for each movie and anime summary, allowing us to measure overall content similarity. By calculating **cosine similarity** between these embeddings, we rank anime based on how closely they match the user's input movies in terms of themes, character dynamics, mood, and intended audience. Combining these models, we can bridge the unique content of movies and anime in a way that captures the key details in a way that has not been done before.

Steps:

- **Data Preprocessing:** The dataset was cleaned and structured in a consistent format by removing rows with missing or incomplete data and normalizing column values. Text preprocessing techniques, including tokenization, removal of stop words, special characters, and lowercasing, were applied to optimize the textual data for subsequent analysis.
- **Exploratory Data Analysis (EDA):** Detailed exploration of the dataset was conducted to uncover key insights. Visualization tools, such as WordCloud, were used to analyze text frequency distributions and identify recurring patterns. These analyses provided a foundational understanding of the dataset, informing feature engineering and model configuration.
- **Model Implementation:** Advanced pre-trained models from the Hugging Face ecosystem, including KeyBERT and Sentence-BERT, were used for keyword extraction and semantic analysis. Text data was passed to the models in manageable batches to maintain state consistency and optimize memory usage. A predefined lexicon of genre keywords was curated to enhance the accuracy of genre classification.
- **Semantic Similarity Search with Word2Vec:** To complement the recommendations, Word2Vec embeddings from the Google News corpus (via Gensim) were utilized for semantic similarity computation. Vector representations of the text summaries were generated, and cosine similarity scores were calculated to rank anime based on their semantic closeness to the input movie summaries. The anime were ranked in descending order of relevance to create a personalized recommendation list.

4 CONCEPTS INVOLVED

The project involves several advanced concepts in **Natural Language Processing (NLP)**, as well as text mining, machine learning, and information retrieval. The key concepts include:

- **Word Embeddings and Word2Vec:** - **Word2Vec** is a neural network-based technique for learning distributed representations of words, where words with similar meanings have close vector representations in a continuous vector space. In our project, we use **Word2Vec** to compute semantic similarities between movie and anime summaries, aiding in ranking recommendations.
- **Cosine Similarity:** - A metric used to measure the semantic similarity between vector representations of text. By calculating the cosine of the angle between two embedding vectors, we rank anime based on how closely their content matches the user's favorite movies.
- **Key Phrase Extraction with KeyBERT:** - **KeyBERT** is a keyword extraction tool that uses transformer-based embeddings to identify the most relevant phrases in a document. We use **KeyBERT** to extract key phrases from movie summaries, capturing themes, genres, tone, and narrative elements to build user preference profiles.
- **Semantic Embeddings with Sentence-BERT:** - **Sentence-BERT** is a fine-tuned BERT model for generating high-quality sentence embeddings. By embedding both movie and anime summaries into a shared semantic space, we measure cross-domain content similarity to provide personalized recommendations.
- **Latent Dirichlet Allocation (LDA):** - **LDA** is a probabilistic model for uncovering latent topics in a collection of documents. In our project, LDA aids in thematic analysis, helping to classify anime and movies into relevant categories based on their content.
- **Sentiment Analysis:** - Sentiment analysis evaluates the emotional tone of text to identify underlying sentiments such as positivity, negativity, or neutrality. We integrate sentiment analysis with semantic embeddings to ensure that recommendations align with the user's preferred emotional tone and mood.
- **Information Retrieval and Ranking:** - By combining embeddings, similarity measures, and thematic analysis, our system performs information retrieval to identify and rank anime recommendations. This integration ensures cross-domain relevance, bridging the gap between movie narratives and anime storytelling.

5 EVALUATION

This section outlines the evaluation process of our project, focusing on technical challenges, lessons learned, results, and potential improvements.

5.1 Technical Challenges

- **Dataset Preparation:** Preparing the dataset required balancing diversity and computational efficiency. This involved reducing the dataset size while preserving key elements essential for meaningful recommendations.
- **Model Selection:** Choosing models that could effectively capture semantic and thematic similarities across movies and anime was challenging. Rigorous testing and fine-tuning were necessary to ensure both accuracy and scalability.
- **Integration of Techniques:** Combining **KeyBERT**'s keyword extraction capabilities with **Sentence-BERT**'s semantic embeddings presented difficulties in maintaining coherence and relevance. Aligning these techniques required extensive adjustments to achieve consistent outputs.

5.2 Results

The evaluation of the system highlighted both strengths and limitations:

- The models occasionally produced similar outputs for multiple rows due to high textual semantic similarity in the input data. This limitation was particularly evident in datasets

where summaries shared overlapping themes or narrative elements. This was attributed as a limitation of the models we used as the state of the model would hinder or impact the outputs.
- Incorporating the **Word2Vec** model for semantic search based on emotions and sentiments improved the recommendations. This added a new dimension to the results, enabling deeper alignment with user preferences by capturing emotional and contextual nuances.
- The use of pre-trained models like **Sentence-BERT** effectively bridged the semantic gap between movies and anime. However, its performance was sometimes impacted by the variability in summary quality, emphasizing the importance of refined preprocessing.
- Overall, the system achieved meaningful cross-domain recommendations, successfully connecting user movie preferences with tailored anime suggestions by leveraging advanced NLP techniques. We were able to effectively rank the anime recommendations based on cosine similarity performing the essnetial task of any recommendation engine.

5.3 Lessons Learned

- Incorporating emotional and sentiment analysis improves the quality of recommendations by aligning them more closely with user preferences.
- Semantic embeddings provide a robust foundation for cross-domain content matching, effectively identifying thematic and narrative connections.
- Dataset refinement and model tuning are critical for ensuring high performance while maintaining computational efficiency.

6 DISCUSSION

The evaluation process revealed valuable insights into the system's capabilities, strengths, and areas that warrant further exploration. While the project successfully captures thematic and narrative similarities between movies and anime, several enhancements could significantly improve its performance and broaden its utility:

- **Expanding the Dataset:** Incorporating additional media types, such as books, TV series, or video games, can diversify recommendations and enable the system to cater to a broader audience with varied interests. A more diverse dataset would also provide richer narrative structures and thematic variations, enhancing the model's ability to capture unique cross-domain similarities. Furthermore, integrating multilingual datasets could allow the system to recommend content across different languages and cultural contexts, further expanding its applicability.
- **Enhancing Sentiment Analysis:** The current approach utilizes pre-trained sentiment analysis models, which, while effective, may not fully capture the nuanced emotional arcs and complex themes present in some narratives. Due to resource and time constraints, we could not train a custom sentiment analysis model. However, training a model specifically on a dataset containing detailed emotional and thematic annotations for movies and anime could significantly enhance the system's understanding of user preferences. Such a model could better identify subtle emotional shifts and underlying tones, improving the personalization and relevance of recommendations.
- **Incorporating Feedback Loops:** Introducing a mechanism for collecting and incorporating user feedback could enable continuous learning and dynamic adaptation of the recommendation system. Feedback loops would allow the system to refine its understanding of user preferences over time by learning from explicit inputs, such as user ratings or rankings, and implicit signals, such as viewing patterns or interactions with recommendations. This iterative improvement process would lead to more accurate and personalized suggestions, making the system more user-centric and adaptable to changing tastes.
- **Improving Computational Efficiency:** Scaling the system to handle larger datasets and a higher volume of users may require optimizations in computational efficiency. Techniques such as dimensionality reduction, distributed processing, and caching frequently used embeddings could reduce latency and improve response times. These improvements would make the system more robust and capable of handling real-world deployment scenarios.
- **Enhancing Interpretability and Transparency:** Providing users with explanations for recommendations can build trust and engagement. For example, showing users which themes, genres, or emotional elements from their favorite movies influenced a particular anime recommendation can help them better understand and appreciate the system's suggestions. Incorporating visualizations or intuitive interfaces could further enhance user experience.
- **Addressing Bias in Recommendations:** As with any AI-driven system, ensuring fairness and avoiding unintended biases in recommendations is crucial. Expanding the diversity of the dataset, fine-tuning models to account for minority genres or themes, and periodically auditing recommendations can help mitigate potential biases and ensure inclusivity.

These enhancements, taken together, offered us an idea on how to build upon this project. Moreover gave us ideas on how to provide more accurate, diverse, and personalized recommendations. By combining improvements in data diversity, model sophistication, user engagement, and computational efficiency, the system can evolve into a highly versatile and impactful cross-domain recommendation engine.