CIS9665 Applied Natural Language Processing

Spring 2025

Professor Vinayak Javaly

Team Wikiminds
Yuliia Pylpenko & Yashasvi Bhati

The Proposal : Wikiminds

Discovering Movies and Books Through Text

INTRODUCTION

We want to help users quickly find movies or books that match what they are looking for. Our system will compare the words in plot summaries or titles that users provide with our large database of information. By focusing on the text itself, we can suggest items that share similar themes or topics.

DATA SOURCES

- **Movies**
  tmdb_5000_credits.csv
  tmdb_5000_movies.csv
  Wikipedia Movie Plots (Kaggle)
  The Movies Dataset (Kaggle)
- **Books**
  Books Dataset (Kaggle)
  Books Details Dataset (Kaggle)
  Books Dataset (Kaggle)
  Goodreads Book Descriptions (Hugging Face)

PREPARING THE TEXT

Because the user will only enter a short summary or a single title, most of our data cleaning will take place on these large databases. We may remove punctuation, convert words to lowercase, and ignore very common words like “the” or “and.” The exact steps might change depending on which method we choose, but the idea is to highlight the most meaningful words. A big challenge will be making sure we **match** the user’s input to the right entries in our datasets so we can recommend similar movies or books.

TURNING SUMMARIES INTO NUMBERS

Once the text is ready, we need to represent each summary as a set of numbers (vectors) so we can measure how close two items are. Some methods we might try include:

- **TF-IDF**, which values important words more highly

- **Word2Vec or Doc2Vec (Vec2Doc)**, which looks at how words or entire documents relate to each other
- **Topic Modeling (for example, LDA)**, which groups text into themes

We will test a few methods to see what works best.

COMPARING AND RECOMMENDING

With numerical representations of each summary, we can figure out how similar any two items are. For large datasets, we might use Approximate Nearest Neighbor or Locality Sensitive Hashing (LSH) to speed up the search. When the user types a short description or a title, our system will compare it to the database and suggest the items that are the closest matches.

WHY THIS MATTERS?

By focusing on the text, users can discover movies and books they might not have heard of before, simply by describing what they want or typing a title they already know. This method does not rely on star ratings or user reviews. Instead, it highlights the words and themes in each summary, making it easier to find closely related content.

Project Progress Report

CURRENT ROLES

Most roles of the team stayed the same, with a high level of collaboration and adaptability. We all work on similar tasks and help each other as availability allows. Zoom links were set up by different members, as well as upkeep of notes, initiation of communication in different channels, etc.

MEETING NOTES

All meeting documentation has been recorded in the Meeting Notes section. Please refer to that for detailed updates, action items, and team discussions.

FINDINGS FROM THE EDA

We began our exploration by diving into 24 columns filled with useful data points like 'title', 'overview', 'tagline', 'genres', and 'keywords'. To stay aligned with our target audience—primarily English speakers—we filtered out non-English titles and content that may not have easily accessible dubbed versions.

We also removed unreleased, canceled, or rumored titles to ensure that users only receive recommendations they can actually watch. Nobody wants to get excited over a movie that doesn’t exist yet!

Some columns, like 'vote_average', had too many missing values to be useful, so we chose to drop them to keep our dataset clean and reliable.

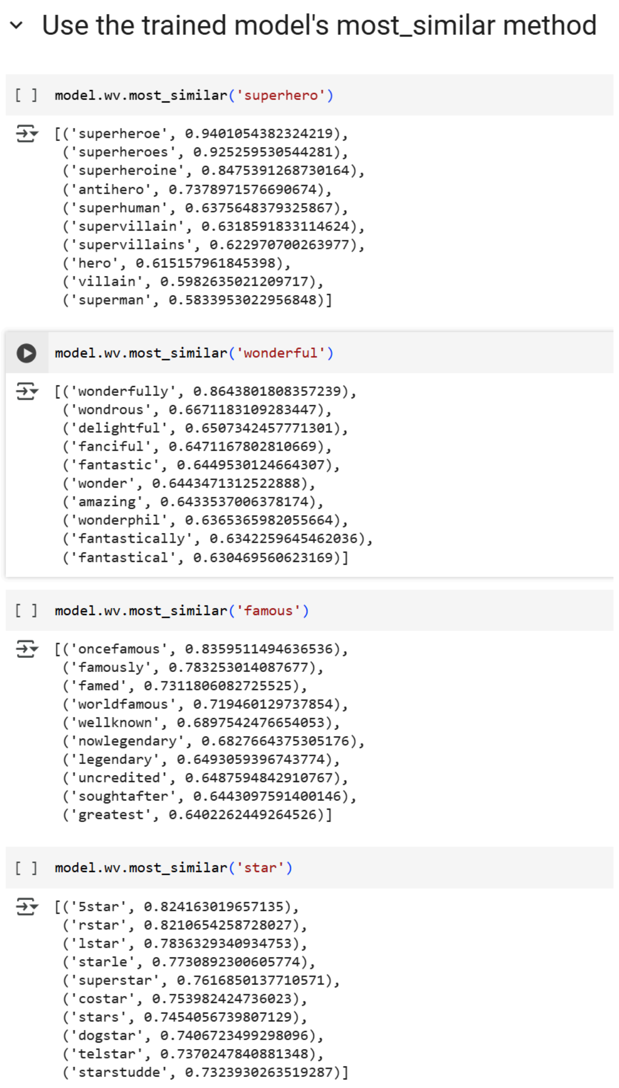
TASKS COMPLETED SO FAR

We've completed several key milestones:

- A full data overview and cleaning
- Initial experimentation with FastText, a model introduced by Facebook AI in 2016 as an upgrade to Word2Vec.

After trimming our dataset down from over 1.2 million entries to around 4,500 English-language US movies with complete data, we trained FastText and used the most_similar method along with cosine similarity to identify closely related movie titles.

For testing, we created a list of movie queries and printed the top 10 most similar titles for each, including their similarity scores.



Use the trained model's most_similar method

```
[ ] model.wv.most_similar('superhero')
```

```
[('superheroe', 0.9401054382324219),
 ('superheroes', 0.925259530544281),
 ('superheroine', 0.8475391268730164),
 ('antihero', 0.7378971576690674),
 ('superhuman', 0.6375648379325867),
 ('supervillain', 0.6318591833114624),
 ('supervillains', 0.622970700263977),
 ('hero', 0.615157961845398),
 ('villain', 0.5982635021209717),
 ('superman', 0.5833953022956848)]
```

```
model.wv.most_similar('wonderful')
```

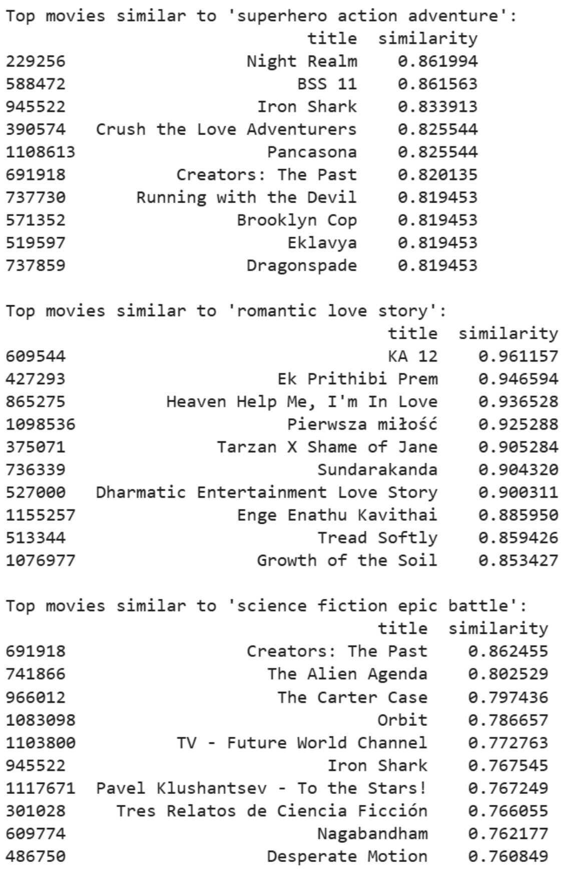
```
[('wonderfully', 0.8643801808357239),
 ('wondrous', 0.6671183109283447),
 ('delightful', 0.6507342457771301),
 ('fanciful', 0.6471167802810669),
 ('fantastic', 0.6449530124664307),
 ('wonder', 0.6443471312522888),
 ('amazing', 0.6433537006378174),
 ('wonderphil', 0.6365365982055664),
 ('fantastically', 0.6342259645462036),
 ('fantastical', 0.630469560623169)]
```

```
[ ] model.wv.most_similar('famous')
```

```
[('oncefamous', 0.8359511494636536),
 ('famously', 0.783253014087677),
 ('famed', 0.7311806082725525),
 ('worldfamous', 0.719460129737854),
 ('wellknown', 0.6897542476654053),
 ('nowlegendary', 0.6827664375305176),
 ('legendary', 0.6493059396743774),
 ('uncredited', 0.6487594842910767),
 ('soughtafter', 0.6443097591400146),
 ('greatest', 0.6402262449264526)]
```

```
[ ] model.wv.most_similar('star')
```

```
[('5star', 0.824163019657135),
 ('rstar', 0.8210654258728027),
 ('lstar', 0.7836329340934753),
 ('starle', 0.7730892300605774),
 ('superstar', 0.7616850137710571),
 ('costar', 0.753982424736023),
 ('stars', 0.7454056739807129),
 ('dogstar', 0.7406723499298096),
 ('telstar', 0.7370247840881348),
 ('starstudde', 0.7323930263519287)]
```



```
Top movies similar to 'superhero action adventure':
                              title  similarity
229256                  Night Realm    0.861994
588472                       BSS 11    0.861563
945522                   Iron Shark    0.833913
390574   Crush the Love Adventurers    0.825544
1108613                   Pancasona    0.825544
691918           Creators: The Past    0.820135
737730       Running with the Devil    0.819453
571352                 Brooklyn Cop    0.819453
519597                      Eklavya    0.819453
737859                  Dragonspade    0.819453

Top movies similar to 'romantic love story':
                                      title  similarity
609544                                KA 12    0.961157
427293                     Ek Prithibi Prem    0.946594
865275          Heaven Help Me, I'm In Love    0.936528
1098536                     Pierwsza miłość    0.925288
375071               Tarzan X Shame of Jane    0.905284
736339                         Sundarakanda    0.904320
527000   Dharmatic Entertainment Love Story    0.900311
1155257                Enge Enathu Kavithai    0.885950
513344                         Tread Softly    0.859426
1076977                  Growth of the Soil    0.853427

Top movies similar to 'science fiction epic battle':
                                     title  similarity
691918                  Creators: The Past    0.862455
741866                    The Alien Agenda    0.802529
966012                     The Carter Case    0.797436
1083098                              Orbit    0.786657
1103800          TV - Future World Channel    0.772763
945522                          Iron Shark    0.767545
1117671  Pavel Klushantsev - To the Stars!    0.767249
301028     Tres Relatos de Ciencia Ficción    0.766055
609774                         Nagabandham    0.762177
486750                    Desperate Motion    0.760849
```

PLAN FOR REMAINING TASKS

As a next step, we plan to evaluate how FastText stacks up against Word2Vec and potentially LDA (Latent Dirichlet Allocation). The goal is to determine which model delivers the most accurate and engaging recommendations for our end users.

Team Charter for Wikiminds

Purpose: To develop a recommendation model using Wikipedia and TMBD dataset that suggests movies based on user input which can be movie names/song lyrics/custom, etc	Goals: Build an effective algorithm that leverages text descriptions to generate accurate and meaningful recommendations.

Team Member Roles and Responsibilities

Name & email	Role	Task
Yashasvi	Organizer & Report Writer	Organizing meetings, taking notes, writing reports & managing drives and slides.
Yuliia	Meeting Coordinator & Writer	Taking meeting notes, writing reports, creating agendas, and managing G Drive.

Team Rules

Ground Rules

1. Meetings start on time; notify in advance if you can’t attend.
2. Agendas will be shared before each meeting.
3. Google Drive is our primary workspace.
4. Tasks will be assigned with clear deadlines.
5. Communicate early if you can’t meet a deadline.
6. Respect all contributions and collaborate professionally.
7. Decision will be made according to sociocratic regime (aka everyone can live with this decision)

Communication Protocol

Primary mode of communication:

WhatsApp

Frequency of communication:

Check messages in the group chat and respond within 1 day (~24 hours)

Response time:

Check messages in the group chat and respond within 1 day (~24 hours)

Who is the primary communication coordinator?

1. Yuliia; 2. Yash; (we will move this duty down the list if emergencies arise)

[e.g. shares all a reminder of approaching deadlines a day before; keeps team together on regular touch-bases]

Meeting Notes

MEETING 1: February 23, 2025 (Virtual on Zoom) | **ATTENDEES**: Yash, Hugo, Yuliia
Discussion Points:

- Reviewed the team charter.
- Discussed general expectations for group work and task distribution.
- Went over project requirements and proposal specifics.
- Action Items for Next Meeting (Tentatively: Thursday, 5:30 PM, Classroom).
    - Find datasets relevant to one of the following tasks:
        - Text Classification / Topic Modeling / Sentiment Analysis.
    - Clarify the final paper format.

MEETING 2: March 10, 2025 (Virtual on Zoom) **| ATTENDEES** : Yash, Hugo, Yuliia
Discussion Points:
Finalized the project topic:

- Content Recommendation System for books and movies.
- Based on Wikipedia descriptions of books/movies that a person liked.

Key Features Considered:

- Books must include the author.
- Movies must include the year of production.
- Tuning recommendations based on production year.
- Ensuring data consistency between books and movies.

Datasets considered:

- Movies
    - TMDB 5000 Credits
    - TMDB 5000 Movies
    - Wikipedia Movie Plots (Has plots)
    - The Movies Dataset (Has only cast & crew)
- Books
    - Books Dataset
    - Books Details Dataset (GoodReads)
    - Books Dataset (Amazon Books)
    - Goodreads Book Descriptions (GoodReads on Huggin Face)

Additional Discussions:

- Brainstormed ways to leverage Wikipedia for content recommendation.
- Discussed the approach and dataset finalization.
- Discussed ways to ensure data consistency for both books and movies.

MEETING 3: March 31, 2025 (Virtual on Zoom) | **ATTENDEES**: Yash, Hugo, Yuliia
Discussion Points:

- Finalized IMDb dataset for the project, containing over 1.2 million entries.

- Explored the option of using the cinemagoer Python package, but decided against it due to similar storage limitations.
- Chose to consult Professor Javaly regarding handling the large dataset.
- Agreed on next steps for data cleaning and basic EDA.
- Decided to maintain communication and share progress via the WhatsApp group.

MEETING 4: April 3, 2025 (Virtual on Zoom) | **ATTENDEES**: Yash, Hugo, Yuliia
Discussion Point:

- Met with Professor Javaly to discuss the challenges of working with large files.
    - Recommendation was to use sample data for initial tasks and testing on the full dataset later to avoid technicalities on Google Colab.
- The team briefly met after the class to discuss that FastText is an impressive model and would be a good fit for our project.

MEETING 5: March 8, 2025 (Virtual on Zoom) | **ATTENDEES**: Yash, Hugo, Yuliia
Discussion Point:

- Reviewed the project Progress utilizing the FastText model and first successful accomplishments.
- Discussed the possibility of using the Word2Vec model and comparing the results.
- How to evaluate the efficiency of the models is still in discussion.
- The deliverables of the Project Progress report were thoroughly discussed, and tasks were distributed within the team.
- Divided responsibilities for the Project Progress Report and clarified individual tasks.

MEETING 6: May 4, 2025 (Virtual on Zoom) | **ATTENDEES**: Yash, Yuliia
Discussion Point:

- Data cleaning
    - The team agreed on reviewing each column to decide whether to keep, remove, or merge.
    - It was recommended to create a data dictionary to document the cleaning process.
    - In addition to NULL checks, mode values will be analyzed to catch repetitive filler text.
    - Columns like 'taglines' and 'keywords' with NULLs will not be dropped but filled with empty strings to retain structure for model training.
- User input
    - Decided to allow multiple movie lookups using a looped input structure.
    - Function will be adjusted to search various fields (e.g., Plot, Overview, Story, Premise) to ensure coverage across movie types.
    - Input will be made case-insensitive for a smoother user experience.
    - If the Wikipedia page isn’t found, the system will suggest similar movies from the dataset and allow the user to choose from those options.
- Output ideas

    - Discussed adding optional user comments or generating a word cloud alongside results would be possible.
    - Brainstormed if sorting recommended movies by production date for better user context.

MEETING 7: May 10, 2025 (Virtual on Zoom) | **ATTENDEES**: Yash, Yuliia
Discussion Point:

- Discussed how FastText as a model can be hypertuned better to perform better and give better outputs.
- Went through the cleaning procedure of the dataset and the data dictionary.
- Discussed the potential of the model and some use cases and brainstormed on the concept of why we should just limit the user input to only movie names and why not let it be anything.

MEETING 9: May 11, 2025 (Virtual on Zoom) | **ATTENDEES**: Yash, Yuliia
Discussion Point:

- After working with data, a new discovery was made that the title data has movies that contain just one symbol and do not refer to actual movies.
- Additional discoveries that were discussed, such as non-Latin characters in the title and in the movie description, were able to removed those.

MEETING 10: May 11, 2025 (Virtual on Zoom) | **ATTENDEES**: Yash, Yuliia
Discussion Point:

- BERT embedding takes ~4 minutes per thousand movies, which would not be sustainable for 750K of what our dataset looks like at the moment. It would take 50 hours to train that. Therefore, some other options were explored.
- MiniLM helped bring down the time to 2 minutes per thousand movies. Researching some other options, it seems that batching embeddings can speed up the process. After trying 32, 64, 128 batches, the performance was not linear. Smaller batches seem to overwhelm the memory, while larger batches are hard to process

MEETING 11: May 12, 2025 (Virtual on Zoom) | **ATTENDEES**: Yash, Yuliia
Discussion Point:

- The expected time to train 750K dataset has to be ~13 hours, however, it did not train in 16 hours, so we made a strategic decision to reduce the size of the model.
- We kept all the cleaning practices and limited the dataset to 15% top movies with the most reviews posted. The new dataset is ~34K movies takes ~40 minutes to train BURT and about 20 minutes to train FastText using the ‘en_core_web_lg’ spaCy model. This is a large model with higher accuracy and full word vectors.

MEETING 12: May 14, 2025 (Virtual on Zoom) | **ATTENDEES**: Yash, Yuliia
Discussion Point:

- Discussed ideas for the final presentation ideas, on how to present our model in the best way possible.

WikiMinds - The Final Report

Introduction

This project focuses on developing a movie recommendation system using Natural Language Processing (NLP) techniques. The goal was to enable users to receive personalized movie suggestions by either entering a Wikipedia movie title or providing their own plot summary. Throughout the project, we experimented with two different models—FastText and BERT. While FastText offered a solid baseline, the BERT-based model (MiniLM v64) delivered more accurate and contextually relevant recommendations. After careful comparison, we selected BERT MiniLM v64 as our final model. This report outlines the overall methodology, data processing steps, model selection process, and key insights gained during development.

Our Motivation

The motivation behind this research question stems from a desire to move beyond traditional recommendation systems, which typically rely on structured metadata like genres, taglines, or user ratings. While these methods are effective to a degree, they often miss the deeper narrative or thematic similarities between titles. By leveraging NLP and models like BERT, we aim to understand the actual content and context of a story through its description.

Our approach allows for a more nuanced recommendation engine—one that captures the essence of a plot rather than just its labels. We also wanted to explore the potential of creating a "Venn space" where anything with a narrative—whether it's a movie, show, book, or even a podcast, could be compared and recommended across formats. While we are currently focusing on movies, the framework we’ve built has the flexibility to expand into other forms of media, opening up exciting possibilities for cross-content recommendations.

Our Why

This is an important question because traditional recommendation systems often miss deeper story elements by relying on genres or ratings. By focusing on plot-level understanding, we can offer more meaningful, personalized, and cross-media recommendations that reflect what users actually enjoy.

Challenges Faced During Development

- Inconsistent or missing plot data, especially for lesser-known titles.
- Difficulty in capturing subtle differences between similarly worded plot descriptions.
- Handling flexible user inputs, such as varying title formats or missing sections like "Plot" vs. "Premise."
- Building logic to handle incomplete or ambiguous Wikipedia pages.
- Comparing and selecting between FastText and BERT, balancing accuracy with performance before finalizing on MiniLM v64.
- Dealing with unseen noise in the dataset such as unlisted plots, single-character movie titles, special characters.

Key Learnings

Throughout this project, we gained a deeper understanding of how language models process text. We learned the importance of thorough data cleaning, especially when preparing text for NLP tasks. Model tuning played a key role in improving recommendation quality, and we saw firsthand how different models behave—eventually selecting BERT MiniLM for its contextual strength.

We also learned to handle errors gracefully, especially when dealing with inconsistent or incomplete user inputs. Most importantly, we learned to think from a machine's perspective, structuring language in a way the model can interpret and developed patience while waiting for models to train and improve.

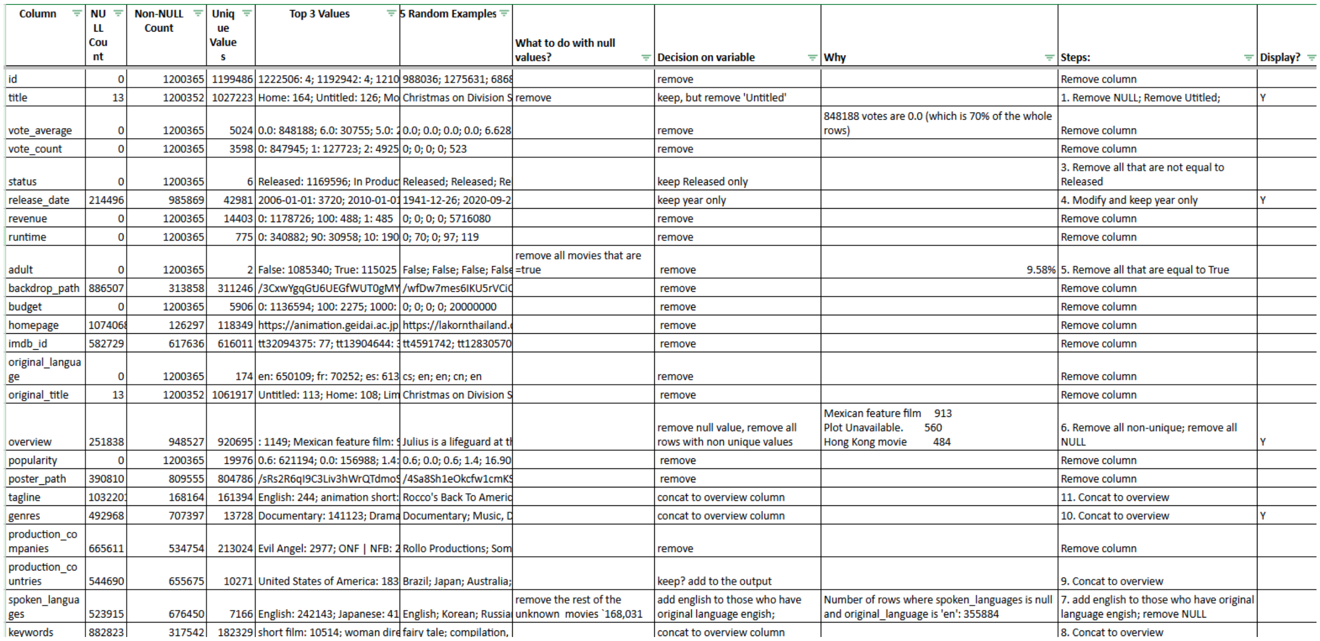
Helpful Resources

- Feedbacks from Prof. Vinayak Javaly
- Javaly, V. (2024). Code Notebooks adapted from CIS9665: Applied Nlp. Baruch College, CUNY, Spring 2025 semester.
- OpenAI. (2025). ChatGPT-generated code snippets and troubleshooting support via GPT-4. Accessed periodically throughout the project.
- Wikipedia Package – for extracting and understanding Wikipedia content structure.
- Hugging Face – for implementing and experimenting with BERT MiniLM models.
- Streamlit - for brainstorming and testing to launch an interactive website
- FastText – for understanding and applying word embedding techniques.

Description of Data Preparation

Working with such a big dataset has been a great learning experience. Due to the fact that there is no data dictionary, we had to develop one ourselves, however, we had to do much more additional research to get a good handle on understanding the dataset. During this process, we learned that there is much more that we have to be aware of, beyond checking for NULL values. WE found a large number of duplicates that would not be an empty cell, but a filler, like "Untitled" for the titles, and not a helpful description of the movie plot like "Short film" (and many similar ones). Developing the table below helped us to flag certain things and collaboratively make decisions on which columns to keep, remove, or merge for the NLP model training and use to display for the user's convenience.

We took an approach of preserving as many movie records as we could. Something that was interesting to find out is that even when you control for all of that work, after further inspection we learned that there are some non-Latin symbols and descriptions in different languages, as well as titles that consist of one letter and do not seem to be real.



Column	NULL Count	Non-NULL Count	Unique Values	Top 3 Values	5 Random Examples	What to do with null values?	Decision on variable	Why	Steps:	Display?
id	0	1200365	1199486	1222506: 4; 1192942: 4; 1210	988036; 1275631; 6868		remove		Remove column	
title	13	1200352	1027223	Home: 164; Untitled: 126; Mo	Christmas on Division S	remove	keep, but remove 'Untitled'		1. Remove NULL; Remove Utitled;	Y
vote_average	0	1200365	5024	0.0: 848188; 6.0: 30755; 5.0: 2	0.0; 0.0; 0.0; 0.0; 6.628		remove	848188 votes are 0.0 (which is 70% of the whole rows)	Remove column	
vote_count	0	1200365	3598	0: 847945; 1: 127723; 2: 4925	0; 0; 0; 0; 523		remove		Remove column	
status	0	1200365	6	Released: 1169596; In Produc	Released; Released; Re		keep Released only		3. Remove all that are not equal to Released	
release_date	214496	985869	42981	2006-01-01: 3720; 2010-01-01	1941-12-26; 2020-09-2		keep year only		4. Modify and keep year only	Y
revenue	0	1200365	14403	0: 1178726; 100: 488; 1: 485	0; 0; 0; 0; 5716080		remove		Remove column	
runtime	0	1200365	775	0: 340882; 90: 30958; 10: 190	0; 70; 0; 97; 119		remove		Remove column	
adult	0	1200365	2	False: 1085340; True: 115025	False; False; False; False	remove all movies that are =true	remove	9.58%	5. Remove all that are equal to True	
backdrop_path	886507	313858	311246	/3CxwYgqGtJ6UEGfWUT0gMY	/wfDw7mes6IKU5rVCiC		remove		Remove column	
budget	0	1200365	5906	0: 1136594; 100: 2275; 1000:	0; 0; 0; 0; 20000000		remove		Remove column	
homepage	1074068	126297	118349	https://animation.geidai.ac.jp	https://lakornthailand.c		remove		Remove column	
imdb_id	582729	617636	616011	tt32094375: 77; tt13904644: 3	tt4591742; tt12830570		remove		Remove column	
original_language	0	1200365	174	en: 650109; fr: 70252; es: 613	cs; en; en; cn; en		remove		Remove column	
original_title	13	1200352	1061917	Untitled: 113; Home: 108; Lim	Christmas on Division S		remove		Remove column	
overview	251838	948527	920695	: 1149; Mexican feature film: 9	Julius is a lifeguard at th		remove null value, remove all rows with non unique values	Mexican feature film 913 Plot Unavailable. 560 Hong Kong movie 484	6. Remove all non-unique; remove all NULL	Y
popularity	0	1200365	19976	0.6: 621194; 0.0: 156988; 1.4:	0.6; 0.0; 0.6; 1.4; 16.90		remove		Remove column	
poster_path	390810	809555	804786	/sRs2R6qI9C3Liv3hWrQTdmoS	/4Sa8Sh1eOkcfw1cmKS		remove		Remove column	
tagline	1032201	168164	161394	English: 244; animation short:	Rocco's Back To Americ		concat to overview column		11. Concat to overview	
genres	492968	707397	13728	Documentary: 141123; Drama	Documentary; Music, D		concat to overview column		10. Concat to overview	Y
production_companies	665611	534754	213024	Evil Angel: 2977; ONF | NFB: 2	Rollo Productions; Som		remove		Remove column	
production_countries	544690	655675	10271	United States of America: 183	Brazil; Japan; Australia;		keep? add to the output		9. Concat to overview	
spoken_languages	523915	676450	7166	English: 242143; Japanese: 41	English; Korean; Russia	remove the rest of the unknown movies `168,031	add english to those who have original language engish;	Number of rows where spoken_languages is null and original_language is 'en': 355884	7. add english to those who have original language engish; remove NULL	
keywords	882823	317542	182329	short film: 10514; woman dire	fairy tale; compilation,		concat to overview column		8. Concat to overview	

NLP Algorithms Used & Results

BERT, FastText, MiniLM

Why did you choose this NLP algorithm(s)?

BERT seemed to be an interesting option to create a thoughtful project with good recommendations. However, in practice, it seemed too computationally expensive and was 4 times slower than MiniLM. On the other hand, it seems that MiniLM has less parameters, increasing the speed significantly, while providing decent search.

How did you choose this NLP algorithm(s)?

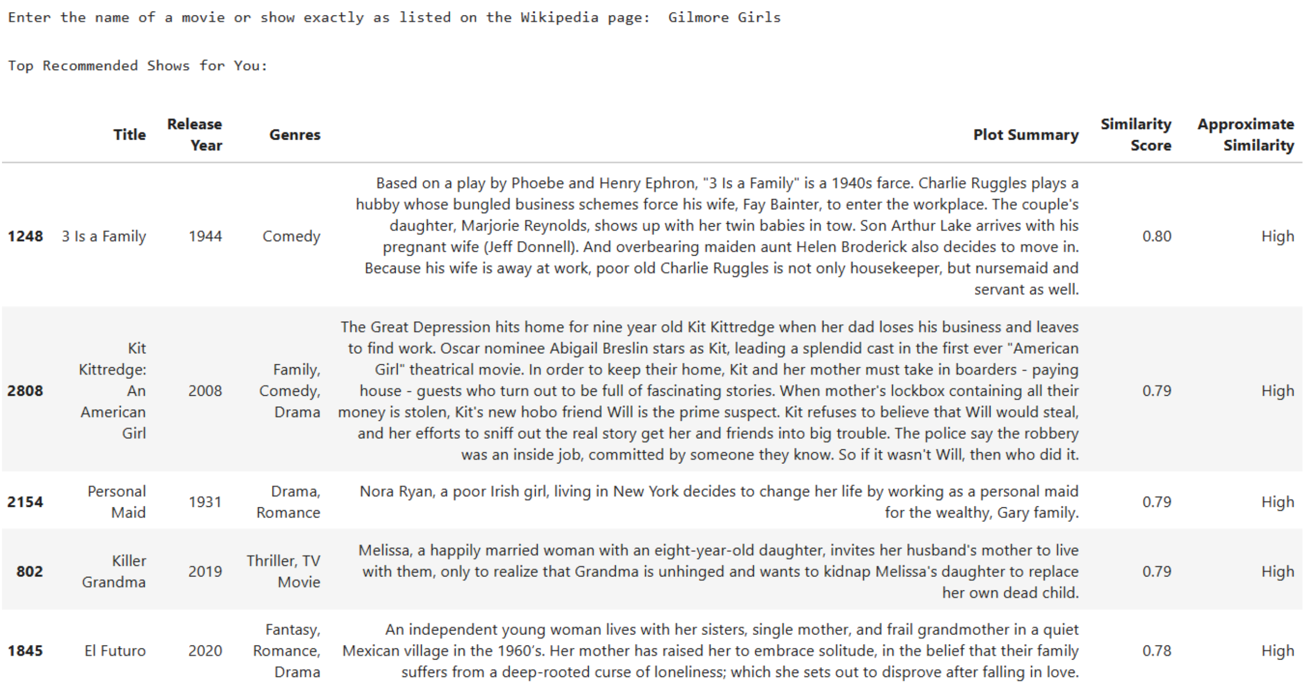
While we initially planned to use BERT, we pivoted to MiniLM due to performance considerations, as both models share a similar architecture. We performed text cleaning separately and used MiniLM to embed the movie database. For the input, whether from Wikipedia or custom user text, we applied the same embedding function. Finally, we computed similarities between the input and the database embeddings to generate the top 5 most relevant movie recommendations.

What are the results?

BERT:

BERT results on a subset of 3000 random movies, after cleaning procedures, we had ~ 2000 movies.

Here are the results. The results for “**Gilmore Girls**

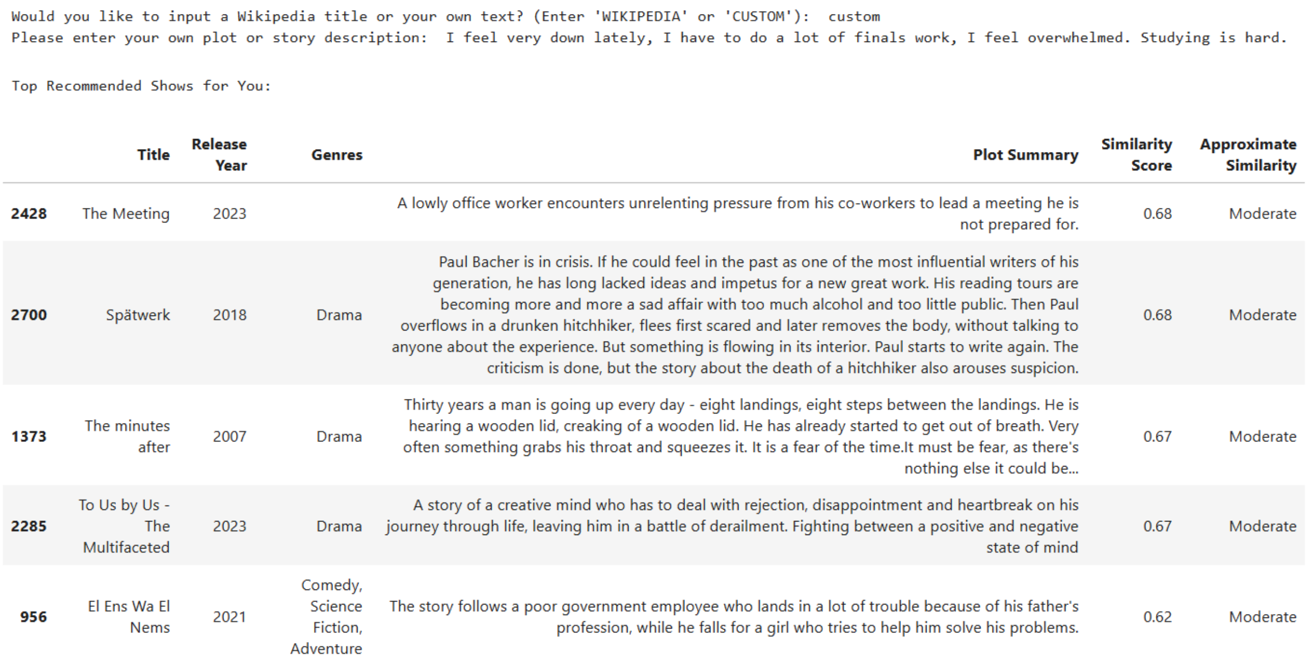


```
Enter the name of a movie or show exactly as listed on the Wikipedia page:  Gilmore Girls

Top Recommended Shows for You:
```

	Title	Release Year	Genres	Plot Summary	Similarity Score	Approximate Similarity
1248	3 Is a Family	1944	Comedy	Based on a play by Phoebe and Henry Ephron, "3 Is a Family" is a 1940s farce. Charlie Ruggles plays a hubby whose bungled business schemes force his wife, Fay Bainter, to enter the workplace. The couple's daughter, Marjorie Reynolds, shows up with her twin babies in tow. Son Arthur Lake arrives with his pregnant wife (Jeff Donnell). And overbearing maiden aunt Helen Broderick also decides to move in. Because his wife is away at work, poor old Charlie Ruggles is not only housekeeper, but nursemaid and servant as well.	0.80	High
2808	Kit Kittredge: An American Girl	2008	Family, Comedy, Drama	The Great Depression hits home for nine year old Kit Kittredge when her dad loses his business and leaves to find work. Oscar nominee Abigail Breslin stars as Kit, leading a splendid cast in the first ever "American Girl" theatrical movie. In order to keep their home, Kit and her mother must take in boarders - paying house - guests who turn out to be full of fascinating stories. When mother's lockbox containing all their money is stolen, Kit's new hobo friend Will is the prime suspect. Kit refuses to believe that Will would steal, and her efforts to sniff out the real story get her and friends into big trouble. The police say the robbery was an inside job, committed by someone they know. So if it wasn't Will, then who did it.	0.79	High
2154	Personal Maid	1931	Drama, Romance	Nora Ryan, a poor Irish girl, living in New York decides to change her life by working as a personal maid for the wealthy, Gary family.	0.79	High
802	Killer Grandma	2019	Thriller, TV Movie	Melissa, a happily married woman with an eight-year-old daughter, invites her husband's mother to live with them, only to realize that Grandma is unhinged and wants to kidnap Melissa's daughter to replace her own dead child.	0.79	High
1845	El Futuro	2020	Fantasy, Romance, Drama	An independent young woman lives with her sisters, single mother, and frail grandmother in a quiet Mexican village in the 1960's. Her mother has raised her to embrace solitude, in the belief that their family suffers from a deep-rooted curse of loneliness; which she sets out to disprove after falling in love.	0.78	High

Results for the custom entry: “**I feel very down lately, I have to do a lot of final work, I feel overwhelmed. Studying is hard.**”

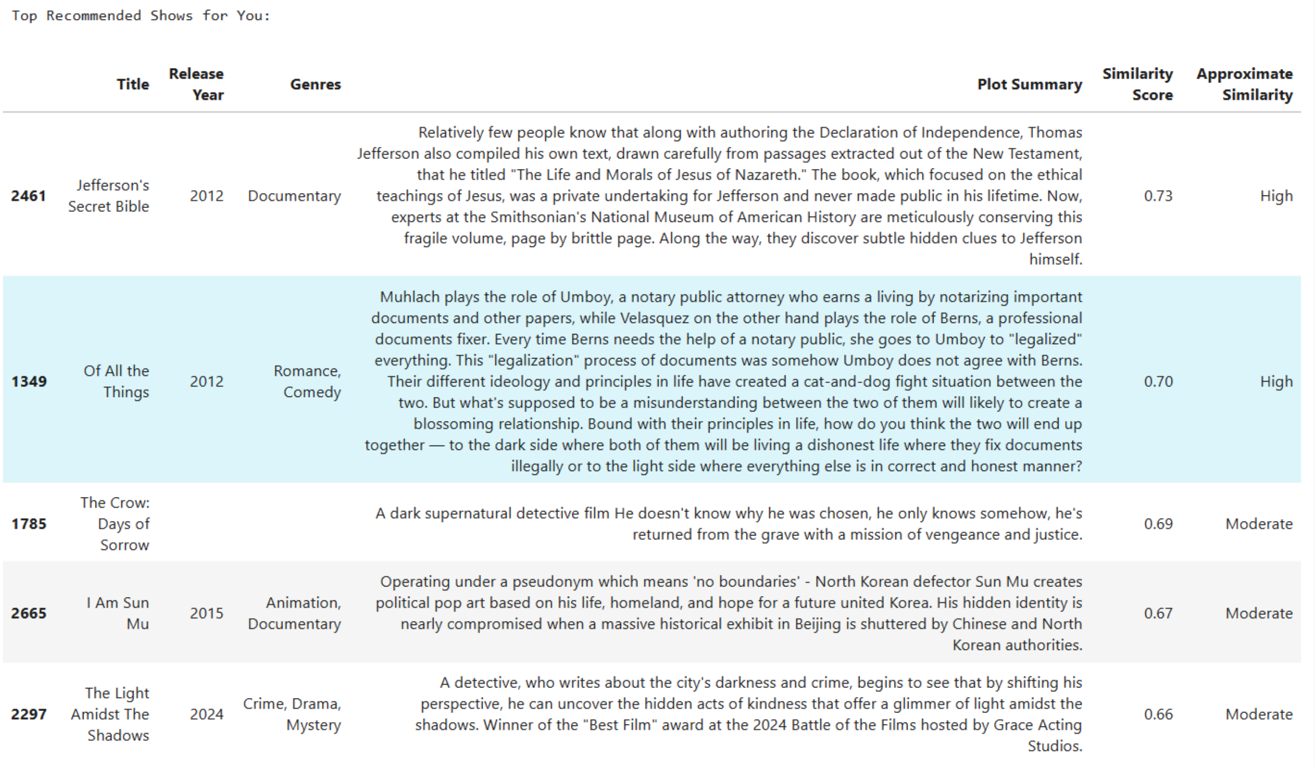


```
Would you like to input a Wikipedia title or your own text? (Enter 'WIKIPEDIA' or 'CUSTOM'):  custom
Please enter your own plot or story description:  I feel very down lately, I have to do a lot of finals work, I feel overwhelmed. Studying is hard.

Top Recommended Shows for You:
```

	Title	Release Year	Genres	Plot Summary	Similarity Score	Approximate Similarity
2428	The Meeting	2023		A lowly office worker encounters unrelenting pressure from his co-workers to lead a meeting he is not prepared for.	0.68	Moderate
2700	Spätwerk	2018	Drama	Paul Bacher is in crisis. If he could feel in the past as one of the most influential writers of his generation, he has long lacked ideas and impetus for a new great work. His reading tours are becoming more and more a sad affair with too much alcohol and too little public. Then Paul overflows in a drunken hitchhiker, flees first scared and later removes the body, without talking to anyone about the experience. But something is flowing in its interior. Paul starts to write again. The criticism is done, but the story about the death of a hitchhiker also arouses suspicion.	0.68	Moderate
1373	The minutes after	2007	Drama	Thirty years a man is going up every day - eight landings, eight steps between the landings. He is hearing a wooden lid, creaking of a wooden lid. He has already started to get out of breath. Very often something grabs his throat and squeezes it. It is a fear of the time.It must be fear, as there's nothing else it could be...	0.67	Moderate
2285	To Us by Us - The Multifaceted	2023	Drama	A story of a creative mind who has to deal with rejection, disappointment and heartbreak on his journey through life, leaving him in a battle of derailment. Fighting between a positive and negative state of mind	0.67	Moderate
956	El Ens Wa El Nems	2021	Comedy, Science Fiction, Adventure	The story follows a poor government employee who lands in a lot of trouble because of his father's profession, while he falls for a girl who tries to help him solve his problems.	0.62	Moderate

Results for “**L(Death Note)**”



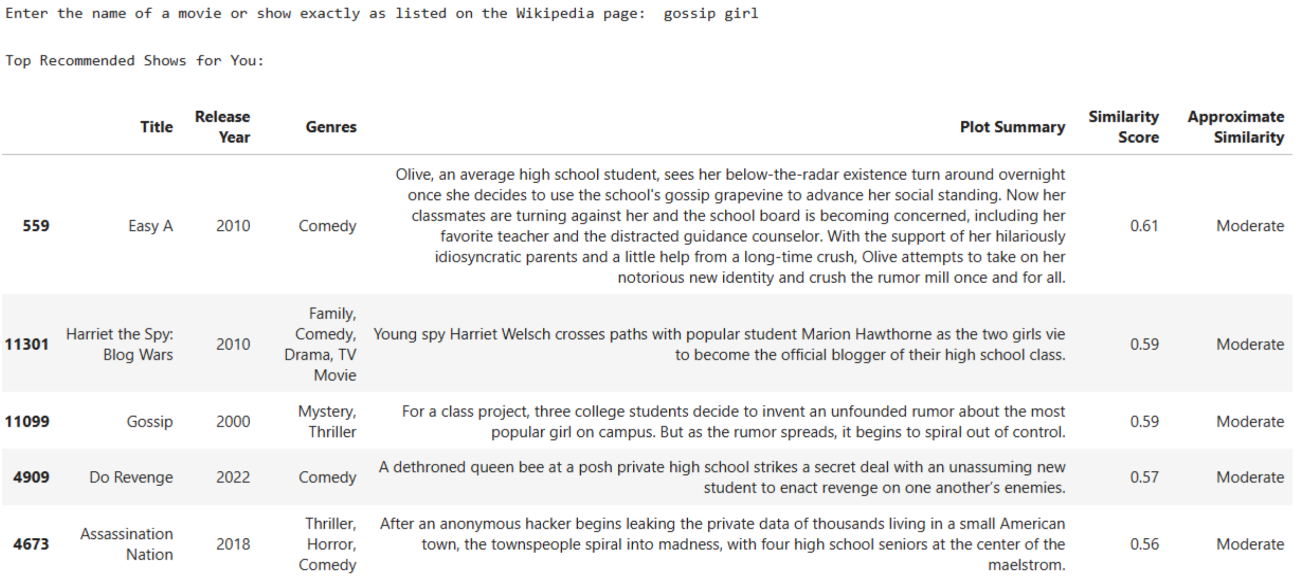
```
Top Recommended Shows for You:
```

	Title	Release Year	Genres	Plot Summary	Similarity Score	Approximate Similarity
2461	Jefferson's Secret Bible	2012	Documentary	Relatively few people know that along with authoring the Declaration of Independence, Thomas Jefferson also compiled his own text, drawn carefully from passages extracted out of the New Testament, that he titled "The Life and Morals of Jesus of Nazareth." The book, which focused on the ethical teachings of Jesus, was a private undertaking for Jefferson and never made public in his lifetime. Now, experts at the Smithsonian's National Museum of American History are meticulously conserving this fragile volume, page by brittle page. Along the way, they discover subtle hidden clues to Jefferson himself.	0.73	High
1349	Of All the Things	2012	Romance, Comedy	Muhlach plays the role of Umboy, a notary public attorney who earns a living by notarizing important documents and other papers, while Velasquez on the other hand plays the role of Berns, a professional documents fixer. Every time Berns needs the help of a notary public, she goes to Umboy to "legalized" everything. This "legalization" process of documents was somehow Umboy does not agree with Berns. Their different ideology and principles in life have created a cat-and-dog fight situation between the two. But what's supposed to be a misunderstanding between the two of them will likely to create a blossoming relationship. Bound with their principles in life, how do you think the two will end up together — to the dark side where both of them will be living a dishonest life where they fix documents illegally or to the light side where everything else is in correct and honest manner?	0.70	High
1785	The Crow: Days of Sorrow			A dark supernatural detective film He doesn't know why he was chosen, he only knows somehow, he's returned from the grave with a mission of vengeance and justice.	0.69	Moderate
2665	I Am Sun Mu	2015	Animation, Documentary	Operating under a pseudonym which means 'no boundaries' - North Korean defector Sun Mu creates political pop art based on his life, homeland, and hope for a future united Korea. His hidden identity is nearly compromised when a massive historical exhibit in Beijing is shuttered by Chinese and North Korean authorities.	0.67	Moderate
2297	The Light Amidst The Shadows	2024	Crime, Drama, Mystery	A detective, who writes about the city's darkness and crime, begins to see that by shifting his perspective, he can uncover the hidden acts of kindness that offer a glimmer of light amidst the shadows. Winner of the "Best Film" award at the 2024 Battle of the Films hosted by Grace Acting Studios.	0.66	Moderate

BERT Conclusion: We found these descriptions amazing, very specific, especially for a very limited random sample of 2000 rows.

MiniLM

While scores for MiniLM were significantly lower than for BERT, we found the model performing well. Please see the results with our comments, The results for “**Gossip Girl**”



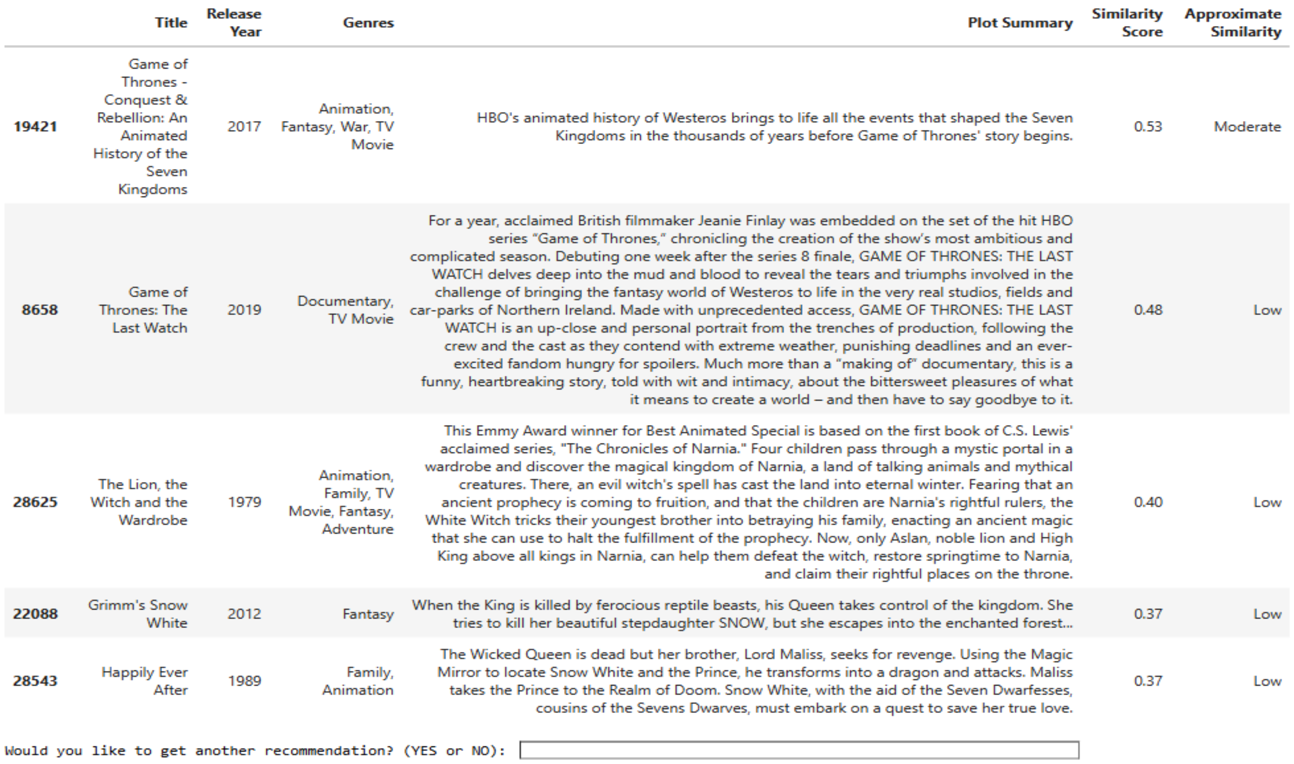
```
Enter the name of a movie or show exactly as listed on the Wikipedia page:  gossip girl

Top Recommended Shows for You:
```

	Title	Release Year	Genres	Plot Summary	Similarity Score	Approximate Similarity
559	Easy A	2010	Comedy	Olive, an average high school student, sees her below-the-radar existence turn around overnight once she decides to use the school's gossip grapevine to advance her social standing. Now her classmates are turning against her and the school board is becoming concerned, including her favorite teacher and the distracted guidance counselor. With the support of her hilariously idiosyncratic parents and a little help from a long-time crush, Olive attempts to take on her notorious new identity and crush the rumor mill once and for all.	0.61	Moderate
11301	Harriet the Spy: Blog Wars	2010	Family, Comedy, Drama, TV Movie	Young spy Harriet Welsch crosses paths with popular student Marion Hawthorne as the two girls vie to become the official blogger of their high school class.	0.59	Moderate
11099	Gossip	2000	Mystery, Thriller	For a class project, three college students decide to invent an unfounded rumor about the most popular girl on campus. But as the rumor spreads, it begins to spiral out of control.	0.59	Moderate
4909	Do Revenge	2022	Comedy	A dethroned queen bee at a posh private high school strikes a secret deal with an unassuming new student to enact revenge on one another's enemies.	0.57	Moderate
4673	Assassination Nation	2018	Thriller, Horror, Comedy	After an anonymous hacker begins leaking the private data of thousands living in a small American town, the townspeople spiral into madness, with four high school seniors at the center of the maelstrom.	0.56	Moderate

Excellent results overall, while the match score is moderate, we would consider them well matched.

The results for “**Game of Thrones**”

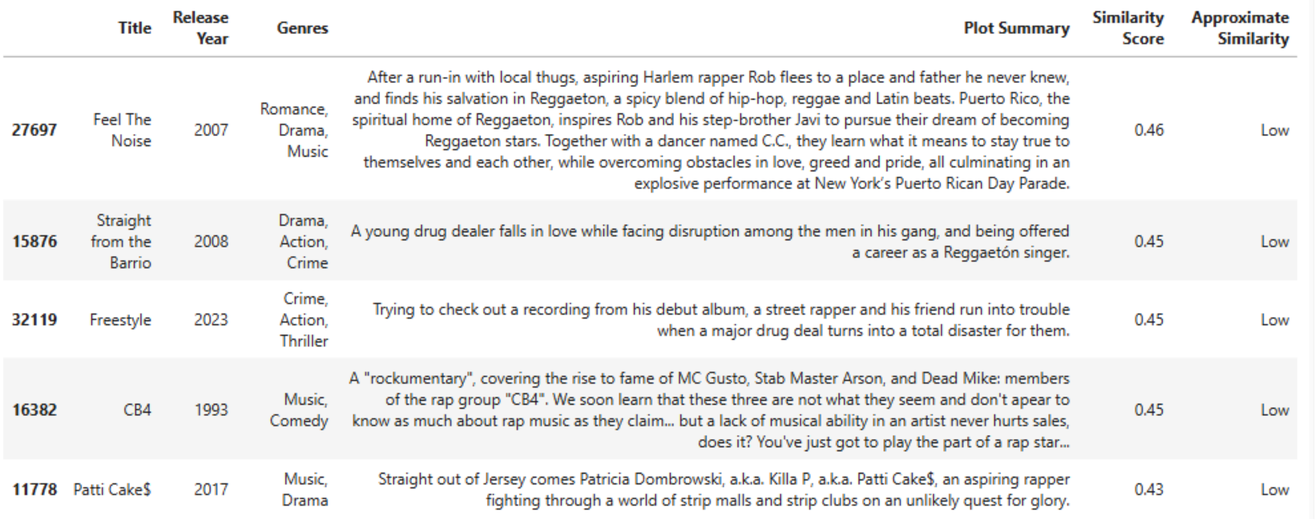


	Title	Release Year	Genres	Plot Summary	Similarity Score	Approximate Similarity
19421	Game of Thrones - Conquest & Rebellion: An Animated History of the Seven Kingdoms	2017	Animation, Fantasy, War, TV Movie	HBO's animated history of Westeros brings to life all the events that shaped the Seven Kingdoms in the thousands of years before Game of Thrones' story begins.	0.53	Moderate
8658	Game of Thrones: The Last Watch	2019	Documentary, TV Movie	For a year, acclaimed British filmmaker Jeanie Finlay was embedded on the set of the hit HBO series “Game of Thrones,” chronicling the creation of the show’s most ambitious and complicated season. Debuting one week after the series 8 finale, GAME OF THRONES: THE LAST WATCH delves deep into the mud and blood to reveal the tears and triumphs involved in the challenge of bringing the fantasy world of Westeros to life in the very real studios, fields and car-parks of Northern Ireland. Made with unprecedented access, GAME OF THRONES: THE LAST WATCH is an up-close and personal portrait from the trenches of production, following the crew and the cast as they contend with extreme weather, punishing deadlines and an ever-excited fandom hungry for spoilers. Much more than a “making of” documentary, this is a funny, heartbreaking story, told with wit and intimacy, about the bittersweet pleasures of what it means to create a world – and then have to say goodbye to it.	0.48	Low
28625	The Lion, the Witch and the Wardrobe	1979	Animation, Family, TV Movie, Fantasy, Adventure	This Emmy Award winner for Best Animated Special is based on the first book of C.S. Lewis' acclaimed series, "The Chronicles of Narnia." Four children pass through a mystic portal in a wardrobe and discover the magical kingdom of Narnia, a land of talking animals and mythical creatures. There, an evil witch's spell has cast the land into eternal winter. Fearing that an ancient prophecy is coming to fruition, and that the children are Narnia's rightful rulers, the White Witch tricks their youngest brother into betraying his family, enacting an ancient magic that she can use to halt the fulfillment of the prophecy. Now, only Aslan, noble lion and High King above all kings in Narnia, can help them defeat the witch, restore springtime to Narnia, and claim their rightful places on the throne.	0.40	Low
22088	Grimm's Snow White	2012	Fantasy	When the King is killed by ferocious reptile beasts, his Queen takes control of the kingdom. She tries to kill her beautiful stepdaughter SNOW, but she escapes into the enchanted forest...	0.37	Low
28543	Happily Ever After	1989	Family, Animation	The Wicked Queen is dead but her brother, Lord Maliss, seeks for revenge. Using the Magic Mirror to locate Snow White and the Prince, he transforms into a dragon and attacks. Maliss takes the Prince to the Realm of Doom. Snow White, with the aid of the Seven Dwarfesses, cousins of the Sevens Dwarves, must embark on a quest to save her true love.	0.37	Low

```
Would you like to get another recommendation? (YES or NO):
```

While the theme seems to be on point, the overall mood did not seem to be matched. We would not consider Happily Ever After a good match for someone who liked Games of Thrones, but since the theme is maintained, we would consider that a good enough result. Interestingly, the scores for identical shows like Game of Thrones: The Last Watch are low.

Recommendations for the lyrics of the song ***Empire State of Mind Alicia Keys and Jay Z.***



	Title	Release Year	Genres	Plot Summary	Similarity Score	Approximate Similarity
27697	Feel The Noise	2007	Romance, Drama, Music	After a run-in with local thugs, aspiring Harlem rapper Rob flees to a place and father he never knew, and finds his salvation in Reggaeton, a spicy blend of hip-hop, reggae and Latin beats. Puerto Rico, the spiritual home of Reggaeton, inspires Rob and his step-brother Javi to pursue their dream of becoming Reggaeton stars. Together with a dancer named C.C., they learn what it means to stay true to themselves and each other, while overcoming obstacles in love, greed and pride, all culminating in an explosive performance at New York’s Puerto Rican Day Parade.	0.46	Low
15876	Straight from the Barrio	2008	Drama, Action, Crime	A young drug dealer falls in love while facing disruption among the men in his gang, and being offered a career as a Reggaetón singer.	0.45	Low
32119	Freestyle	2023	Crime, Action, Thriller	Trying to check out a recording from his debut album, a street rapper and his friend run into trouble when a major drug deal turns into a total disaster for them.	0.45	Low
16382	CB4	1993	Music, Comedy	A "rockumentary", covering the rise to fame of MC Gusto, Stab Master Arson, and Dead Mike: members of the rap group "CB4". We soon learn that these three are not what they seem and don't apear to know as much about rap music as they claim... but a lack of musical ability in an artist never hurts sales, does it? You've just got to play the part of a rap star...	0.45	Low
11778	Patti Cake$	2017	Music, Drama	Straight out of Jersey comes Patricia Dombrowski, a.k.a. Killa P, a.k.a. Patti Cake$, an aspiring rapper fighting through a world of strip malls and strip clubs on an unlikely quest for glory.	0.43	Low

We found it interesting to see matches like “Harlem”, “rap” etc. Excellent match!

On the other hand, we noticed that MiniLM definitely has lower score matches, which is a sacrifice we had to take to ensure that the model is less computationally expensive. Working with a 1/15th part of the full dataset, BERT had an incredible performance, while we did not notice much of an improvement by providing MiniLM with a much bigger dataset. Scores are on average half of what BERT provided.

FastText

While the FastText model, trained on approximately 33,000 rows, produced high similarity scores, the results were not always logically aligned with the input context. For example, when the input was the Wikipedia page for the popular TV show ***Friends***, three out of the five recommended titles were documentaries, an unexpected and somewhat irrelevant outcome. Similarly, for ***Martha Stewart***, most of the recommendations leaned toward comedy, whereas we would have expected themes related to crime or business.

In the case of custom plot inputs, the model performed slightly better. It was able to capture the emotional tone of a ***Diary Entry***, interpreting it as somber or emotionally heavy, and returned reasonably appropriate movie suggestions. For the ***Jay-Z song lyrics***, the model successfully picked up on keywords like “Staten Island,” “hustler,” and “New York City,” which were present in some of the recommendations. However, several other suggestions, despite having high similarity scores, lacked logical relevance.

Given these limitations, we found that the BERT MiniLM model produced far more contextually appropriate and meaningful results. As a result, we chose to finalize our system using the MiniLM model.

Model Evaluation

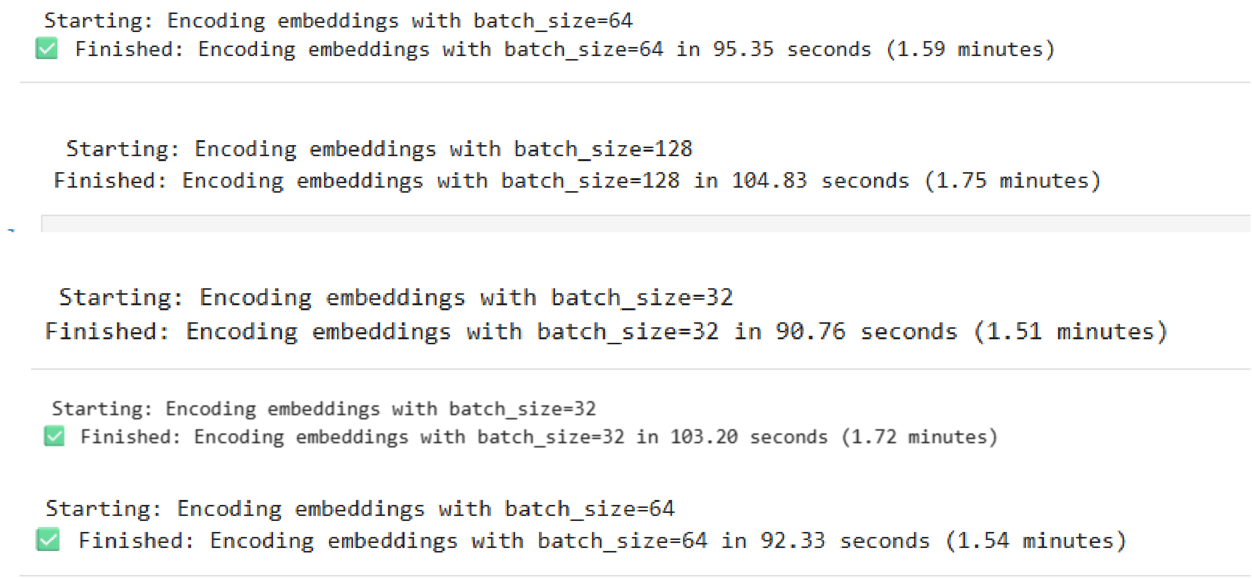
If unsupervised learning is used, how did you evaluate your model?

We checked for the shows that we know well and assessed the results based on keywords, and plots provided. We assessed both custom entries and entries that came from Wikipedia on our models.

Show model tuning process (e.g. hyperparameter tuning)

BERT

Researching alternative options revealed that batching embeddings can significantly impact processing speed; however, the performance gains were not linear. Testing batch sizes of 32, 64, and 128 showed that smaller batches like 32 tended to overwhelm the memory, while larger batches such as 128 became too demanding to process efficiently. The optimal performance was observed at a batch size of 64, beyond which the speed decreased rather than improved. Since this was tested on the small dataset it seemed to be the safest option to go with 64 per batch.



```
Starting: Encoding embeddings with batch_size=64
✅ Finished: Encoding embeddings with batch_size=64 in 95.35 seconds (1.59 minutes)

 Starting: Encoding embeddings with batch_size=128
Finished: Encoding embeddings with batch_size=128 in 104.83 seconds (1.75 minutes)

 Starting: Encoding embeddings with batch_size=32
Finished: Encoding embeddings with batch_size=32 in 90.76 seconds (1.51 minutes)

Starting: Encoding embeddings with batch_size=32
✅ Finished: Encoding embeddings with batch_size=32 in 103.20 seconds (1.72 minutes)

Starting: Encoding embeddings with batch_size=64
✅ Finished: Encoding embeddings with batch_size=64 in 92.33 seconds (1.54 minutes)
```

FastText

For FastText, following hyperparameters to optimize performance for semantic similarity across movie plots:

- **`vector_size=100`**: Balanced detail and efficiency
- **`window=5`**: Moderate context range
- **`min_count=1`**: Included all words, even rare ones
- **`sg=1`**: Used skip-gram for better semantic learning
- **`epochs=10`**: Allowed multiple passes for learning
- **`workers=4`**: Enabled parallel processing

This setup gave reasonable results, but BERT MiniLM offered better context and accuracy, making it our final choice.

Conclusions from findings

FastText performed well for basic similarity matching and was faster to train, but it struggled with deeper contextual understanding of plots. It’s suitable for lightweight applications with limited resources.

BERT MiniLM outperformed FastText by capturing nuanced meanings and relationships within the text. Its contextual awareness led to more accurate and relevant recommendations.

Overall, while FastText is efficient, BERT MiniLM is better suited for tasks involving detailed language understanding, making it the stronger choice for our recommendation system.

Practical implications

How do our results apply to the real world?

In the age of AI, personalized content has become the norm—your music playlist on Spotify, your Instagram feed, even your shopping suggestions are all tailored specifically to you. Yet, when it comes to streaming platforms like Netflix or Hulu, recommendations often remain static, trend-based, and impersonal. That’s no longer good enough. Entertainment should be as uniquely tailored as everything else in your digital life. Our model changes that—it allows users to input anything from a favorite video game plot or song lyric to a diary entry, and receive truly personalized movie and show recommendations. This approach empowers individuals, enhances user engagement, and opens new possibilities for content discovery beyond algorithms based solely on past clicks or broad trends.

How can a company, society or customers benefit from your results?

Companies can use this approach to increase user retention through better recommendations. Society benefits from greater content accessibility and diversity, especially by surfacing lesser-known works. Customers enjoy more relevant, narrative-driven suggestions that align closely with their interests.

What would you do differently the next time you are assigned a similar project?

Next time, we would spend more time up front really getting to know the dataset. With large text data, it's important to account for things like duplicates, overly short entries, special characters, and random symbols. We learned the value of continuously building a data dictionary to document these patterns and guide preprocessing decisions.

We’d also stay more flexible with our modeling strategy—being ready to switch models or optimize based on performance and scalability. Finally, we’d test more frequently and at smaller stages, especially for edge cases like how Wikipedia pages are structured or how special characters are handled. Continuous testing helped us uncover issues early and adapt our approach accordingly.

– end of this report –