

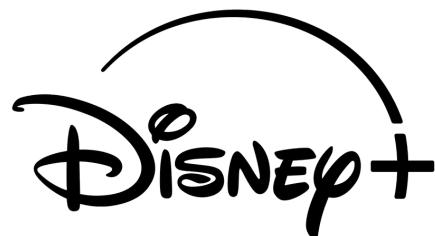
Data Exploration

DHBW Mannheim

Von Nico Dobberkau, Michael Greif und German Paul

Movie Recommender System durch
content-basiertes Filtering

Verwendete Datensätze:



Gliederung

1. Einleitung
2. Explanatory Data Analysis
3. Modelle
4. Recommender System
5. Wirtschaftlicher Nutzen
6. Kritische Würdigung und Fazit

Einleitung

Im Rahmen des Moduls "Data Exploration" an der Dualen Hochschule Baden-Württemberg (DHBW) Mannheim wurde uns die Aufgabe gestellt, uns mit einem selbst gewählten Thema mit Datenbezug auseinanderzusetzen. Unsere Wahl fiel dabei auf das Thema Recommender Systeme. Für die Versionskontrolle und Dokumentation haben wir GitHub verwendet: [Github Link](#)

Voraussetzungen für die erfolgreiche Umsetzung des Projekts

Für eine erfolgreiche Projektdurchführung sind folgende Voraussetzungen notwendig:

1. **Datensatz:** Ein umfangreicher Datensatz mit Filminformationen, der möglichst viele Details wie Regisseur, Titel, Genre und Beschreibung umfasst.
2. **Modell:** Ein Modell, das die Daten verarbeitet und auf Basis eines ausgewählten Films weitere ähnliche Filme empfiehlt.
3. **Business Use Case:** Ein validierter Anwendungsfall, der den gesellschaftlichen Nutzen unseres Projekts begründet.

Unser Ziel

Unser Ziel ist die Entwicklung eines content-basierten Recommender Systems, das basierend auf einem oder mehreren gesehenen Filmen Empfehlungen für weitere ähnliche Filme ausspricht. Zusätzlich soll das System in einem benutzerfreundlichen Interface präsentiert werden.

Explanatory Data Analysis

Die explorative Datenanalyse (EDA) stellt einen entscheidenden Schritt im Prozess der Datenwissenschaft dar, um Einblicke in die Struktur und Eigenschaften eines Datensatzes zu gewinnen. In unserem Projekt wurde eine umfassende EDA auf den Netflix, Disney+ und Prime Video-Datensatz angewendet, der eine breite Palette an Informationen über Filme und Serien auf der Streaming-Plattform enthält. Die Daten umfassen 19925 Einträge und 13 Spalten, die verschiedene Attribute wie Show-ID, Typ (Film oder TV-Show), Titel, Regie, Besetzung, Produktionsland, Hinzufügungsdatum auf der jeweiligen Plattform, Veröffentlichungsjahr, Alterseinstufung, Dauer, Genre, Plattform und eine kurze Beschreibung einschließen.

Diese umfassenden Datensammlungen stammen alle von Kaggle und ermöglichen ein umfangreiches Empfehlungssystem, das mehrere Streaming-Plattformen integriert.

Während des EDA-Prozesses wurden zuerst die grundlegenden Eigenschaften des Datensatzes untersucht, wie zum Beispiel die Anzahl der Zeilen und Spalten und eine Vorschau auf die ersten Einträge, um ein erstes Verständnis für die Daten zu entwickeln.

show_id	type	title	director	cast	country	date_added
0	s1	Movie	Dick Johnson Is Dead	Kirsten Johnson	NaN	United States September 25, 2021
1	s2	TV Show	Blood & Water	NaN Ama Qamata, Khosi Ngema, Gail Mabalane, Thaban...	South Africa	September 24, 2021
2	s3	TV Show	Ganglands	Julien Leclercq Sami Bouajila, Tracy Gotoas, Samuel Jouy, Nabi...	NaN	September 24, 2021
release_year	rating	duration	listed_in	description	platform	
2020	PG-13	90 min	Documentaries	As her father nears the end of his life, filmm...	Netflix	
2021	TV-MA	2 Seasons	International TV Shows, TV Dramas, TV Mysteries	After crossing paths at a party, a Cape Town t...	Netflix	
2021	TV-MA	1 Season	Crime TV Shows, International TV Shows, TV Act...	To protect his family from a powerful drug lor...	Netflix	

Tabelle 1 Zeilenvorschau EDA

Die obige Abbildung zeigt die Ausgabe der ersten drei Reihen im Datensatz, welcher alle drei Plattformen umfasst.

Weiterführend wurden die Typen der Datenspalten inspiziert, um sicherzustellen, dass diese korrekt formatiert sind und um eventuelle Anpassungen vorzunehmen, die für die weiterführende Analyse benötigt werden. Die Untersuchung der deskriptiven Statistiken des Datensatzes lieferte zusätzliche Einblicke in die Verteilung quantitativer Merkmale, zum Beispiel das Veröffentlichungsjahr der Inhalte.

Ein wichtiger Schritt des EDA-Prozesses war der Umgang mit fehlenden Werten, da einige Spalten wie 'Director', 'Cast', 'Country', 'Rating' und 'Duration' unvollständige Daten aufwiesen.

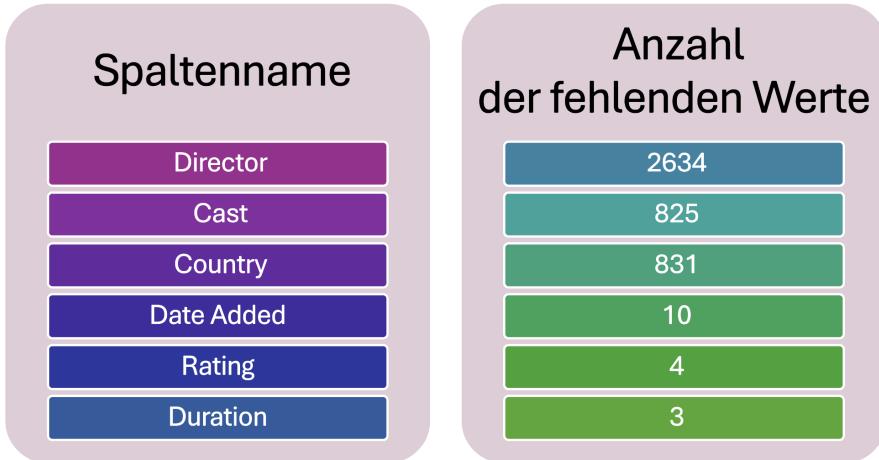


Abbildung 1 Fehlende Werte im Datensatz

Wie wir erkennen können, waren jedoch die fehlenden Werte unterschiedlich in Menge und Semantik. Es ist grundsätzlich wichtig, zu beachten, dass die Methode wie wir die Director Spalte behandeln, einen größeren Einfluss auf unser Model haben wir als die Duration Spalte. Da viele Einträge fehlen und die wichtigste Spalte für die Vorhersage der Titel und die Beschreibung ist und diese Werte in keiner Reihe fehlen, haben wir keine Reihe komplett entfernt (Ka et al. 2023).

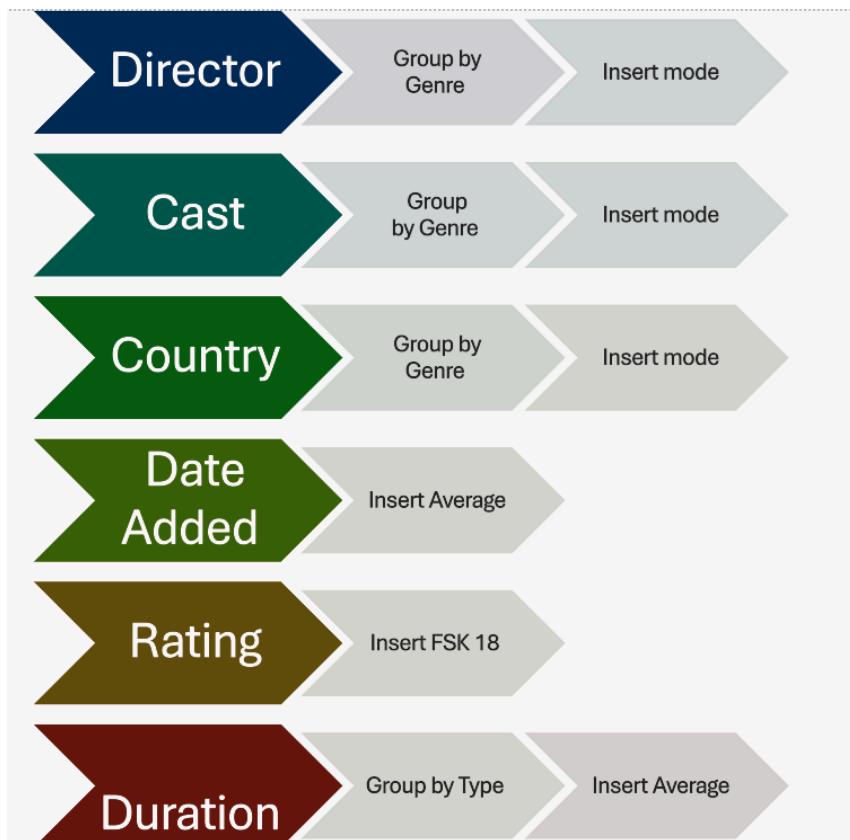


Abbildung 2 Umgang mit fehlenden Werten

Stattdessen haben wir, wie im obigen Schaubild zu sehen ist, beim Director, bei der Cast und dem Land nach Genres sortiert und davon den Mode genommen. Von den vielen möglichen Optionen sind wir davon ausgegangen, dass diese Vorgehensweise die besten Ergebnisse bringt, wenn es um den Ausgleich von Komplexität und Effektivität geht.

Ebenfalls haben wir uns dieses Vorgehen erlaubt, da in unserem Modell der Titel und die Beschreibung einen viel größeren Einfluss haben und somit die Vorhersage nur sehr minimal verzerrt wird. Beim Rating haben wir uns dafür entschlossen, es bei Fehlen als FSK 18 einzustufen, da es besser ist, dass Minderjährigen ein valider Film nicht empfohlen wird, als andersrum.

Im nächsten Schritt haben wir noch sichergestellt, dass im Datensatz keine Duplikate vorliegen.

Die visuelle Data Exploration war ebenfalls ein zentraler Bestandteil des EDA-Prozesses. Mittels verschiedener Grafiken und Tabellen konnten wir wichtige Muster und Erkenntnisse aus dem Datensatz in einfacher Form aufzeigen und somit auch für Laien, den Datensatz verständlicher zu machen. Unsere Visualisierungen umfassten zum Beispiel die Verteilung der Inhalte nach Typ (Film oder TV-Show), die Anzahl der Veröffentlichungen nach Jahren, die Top-Produktionsländer und die durchschnittliche Dauer der Inhalte.

Die gesamte Explanatory Data Analysis wurde in Python durchgeführt, da wir mit dieser Programmiersprache viel Erfahrung in vorangegangenen Semestern sammeln konnten und als Version-Control-Tool haben wir Git verwendet, um die Zusammenarbeit im Team zu erleichtern.

Modelle

Allgemein

Im Rahmen des Projekts haben wir uns zwei Modelle zur Generierung von Filmempfehlungen verwendet. Im Fokus standen dabei BERT (Bidirectional Encoder Representations from Transformers) und TF-IDF (Term Frequency-Inverse Document Frequency), die jeweils auf unterschiedliche Weise die Herausforderungen der Textanalyse angehen. Diese Methoden wurden ausgewählt, da beide Ihre Vorteile bei der Textanalyse mit sich bringen und durch die verschiedenen Vorgehensweisen der Modelle möglicherweise zu anderen Ergebnissen führen. Somit kann der User selbst entscheiden, welche Resultate ihm besser gefallen (Sahu et al. 2022).

Allgemeines Vorgehen bei den Modellen



Abbildung 3 Vorgehen bei den Modellen

Datenverarbeitung

Zunächst haben wir die Daten in das richtige Format gebracht. Dabei wurden drei Schritte durchgeführt:

1. Konvertierung aller Wörter in Kleinschreibung
2. Entfernung von Stoppwörtern
3. Anwendung von Stammwortbildung (Stemming)

Da "Apfel" und "apfel" dasselbe Wort darstellen, haben wir zunächst alle Wörter in ein einheitliches Format gebracht, indem wir sie in Kleinschreibung konvertiert haben.

Anschließend haben wir Stoppwörter entfernt. Stoppwörter sind häufig vorkommende Wörter wie "und", die keine signifikante Bedeutung tragen und die Analyse verfälschen könnten. Das Entfernen dieser Wörter verhindert, dass Filme aufgrund der Häufigkeit solcher irrelevanten Wörter fälschlicherweise als ähnlich eingestuft werden.

Zuletzt haben wir die Wörter auf ihre Wortstämme zurückgeführt, ein Prozess, der als Stemming bekannt ist. Dabei wird beispielsweise "driving" zu "drive" reduziert. Dies hilft, verschiedene Formen eines Wortes zu vereinheitlichen und die Analyse zu präzisieren.

Embeddings und Kosinusähnlichkeit

Beim Embedding werden Wörter in einem gegebenen Text zu Vektoren konvertiert.

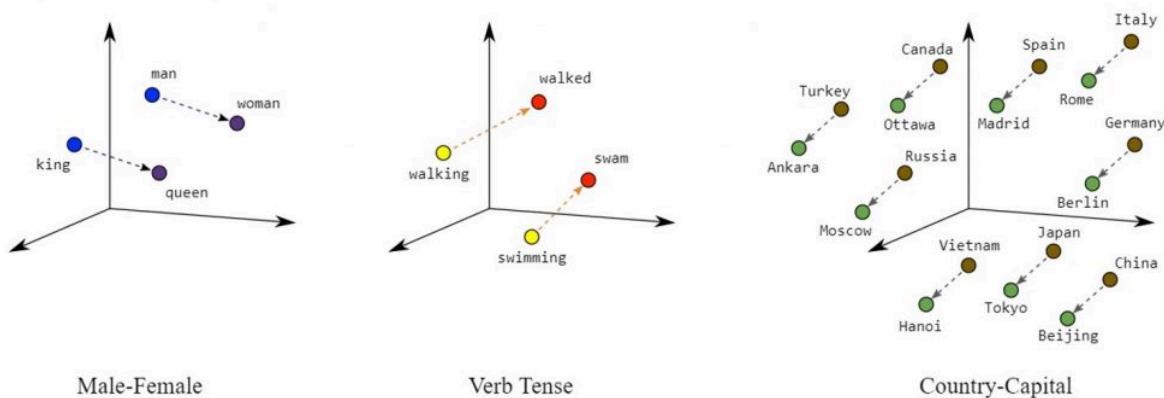


Abbildung 4 Vektor Embeddings (Google Cloud 2024)

Diese vektorielle Darstellung ermöglicht subsequente Berechnungen von Ähnlichkeitsmaßen zwischen den einzelnen Vektoren, wie in der obigen Abbildung illustriert. Die von uns implementierten Modelle nutzen dieses Prinzip, um die semantische Ähnlichkeit zwischen den Filmbeschreibungen quantitativ zu erfassen und zu vergleichen.

Testing

Resultat für den Film "Jaws":

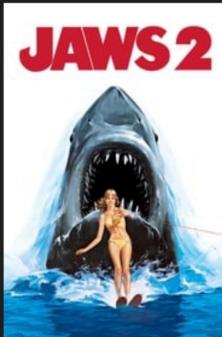
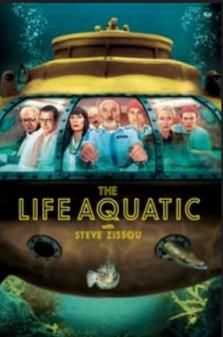
Jaws 2	The Life Aquatic With Steve Zissou	Baby Shark & More Kids Songs (little Treehouse)	Rise Of The Great White Shark	World's Most Dangerous Shark?
				
Director: Jeannot Szwarc	Director: Wes Anderson	Director: Usp Studios	Director: Andy Casagrande	Director: Kevin Bachar
Country: United States	Country: -	Country: -	Country: -	Country: -
Genre: Dramas, Horror Movies, Thrillers	Genre: Action, Adventure, Comedy	Genre: Animation, Kids	Genre: Documentary, Special Interest	Genre: Animals & Nature, Documentary
Platform: Netflix	Platform: Amazon_prime	Platform: Amazon_prime	Platform: Amazon_prime	Platform: Disney_plus

Abbildung 5 Recommendations für "Jaws"

Resultat für den Film "Frozen":

Frozen Fever	Olaf's Frozen Adventure	Lego Disney Frozen	Frozen 2	Arendelle Castle Yule Log
				
Director: Chris Buck, Jennifer Lee	Director: Stevie Wermers-skelton, Kevin Deters	Director: -	Director: Chris Buck, Jennifer Lee	Director: -
Country: United States	Country: United States	Country: United States	Country: United States	Country: United States
Genre: Animation, Family, Fantasy	Genre: Animation, Comedy, Family	Genre: Action-adventure, Animation, Comedy	Genre: Action-adventure, Animation, Family	Genre: Animation, Family
Platform: Disney_plus	Platform: Disney_plus	Platform: Disney_plus	Platform: Disney_plus	Platform: Disney_plus

Abbildung 6 Recommendations für "Frozen"

Die Analyse des Modells zeigt eine deutliche Fokussierung auf die Hauptfiguren in Filmen. Bei der Betrachtung des Films "Frozen" fällt eine Diskrepanz auf: Während ein thematisch passender Film empfohlen wird, ist die zugehörige visuelle Darstellung des Filmcovers inkorrekt. Da diese Information über eine externe API bezogen wird, sind solche Ungenauigkeiten hinzunehmen.

Die Evaluierung des Modells offenbart sowohl Stärken als auch Schwächen. Zu den Stärken zählt die präzise Identifikation von Filmen mit ähnlichen Hauptfiguren oder verwandten Handlungssträngen. Exemplarisch sei hier die Empfehlung von fünf haibasierten Filmen für

"Jaws" und fünf Filmen mit der Figur Elsa für "Frozen" genannt. Auch Fortsetzungen wie "Jaws 2" werden akkurat erkannt.

App Development

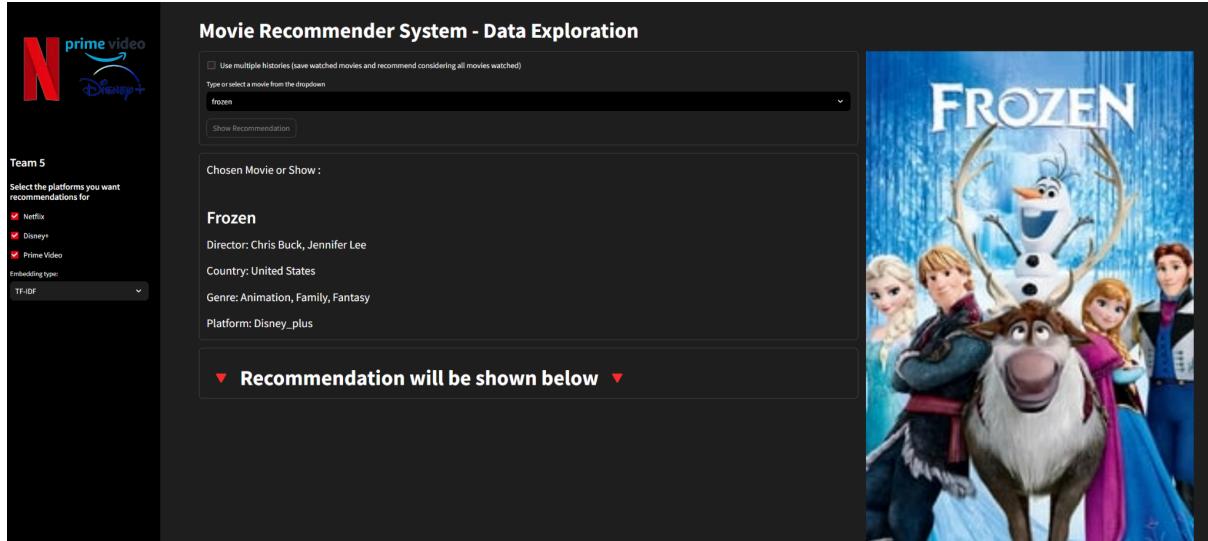


Abbildung 7 Gesamte Web-Applikation

Die entwickelte Web-Applikation bietet diverse Funktionalitäten zur Personalisierung der Filmempfehlungen. In der Seitenleiste können Nutzer die zu berücksichtigenden Abonnements selektieren, wobei standardmäßig alle drei verfügbaren Optionen aktiviert sind. Zusätzlich besteht die Möglichkeit, zwischen zwei Modelltypen zu wählen: BERT oder TF-IDF. Diese Auswahl beeinflusst den zugrundeliegenden Algorithmus der Empfehlungsgenerierung.

Im Hauptbereich der Applikation, unterhalb der Überschrift, findet sich eine Option zur Aktivierung der Historien-Funktion. Bei Aktivierung werden alle Filme, für die eine Empfehlung generiert wird, als "gesehen" markiert. Dies hat zwei wesentliche Auswirkungen auf die Funktionsweise des Systems:

1. Exklusion aus zukünftigen Empfehlungen: Bereits gesehene Filme werden in nachfolgenden Empfehlungen nicht mehr berücksichtigt, um Redundanzen zu vermeiden.
2. Inklusion in Ähnlichkeitsberechnungen: Die als gesehen markierten Filme werden in die Berechnung von Ähnlichkeiten für zukünftige Empfehlungen einbezogen, was potenziell die Präzision der Empfehlungen erhöhen kann.

Unterhalb dieser Funktion befindet sich ein Suchfeld, das es den Nutzern ermöglicht, spezifische Filme zu suchen. Nach der Auswahl eines Films werden relevante Informationen sowie eine visuelle Darstellung des Filmcovers präsentiert. Diese Funktion dient der detaillierten Informationsbereitstellung und unterstützt die Nutzer bei der Entscheidungsfindung.

Die Integration dieser verschiedenen Funktionalitäten zielt darauf ab, ein personalisiertes und interaktives Nutzererlebnis zu schaffen, das sowohl die Präferenzen des Nutzers als auch sein bisheriges Sehverhalten berücksichtigt.

BERT

BERT (Bidirectional Encoder Representations from Transformers) ist ein fortschrittliches, auf tiefem Lernen basierendes Modell, das kontextuelle Nuancen im Text erfasst und eine hervorragende Leistung bei einer Vielzahl von NLP-Aufgaben zeigt. Es ist besonders geeignet für Aufgaben, die ein Verständnis des Kontexts und feiner Sprachnuancen erfordern, aufgrund seines Ansatzes, der kontextbezogene Einbettungen verwendet. Diese Technologie erwies sich als besonders nützlich bei der Analyse von Filmbeschreibungen und Serieninhalten, da sie in der Lage ist, komplexe Zusammenhänge und Charakterinteraktionen zu verstehen. Der Vorteil von BERT ist ganz klar, dass das Modell auch den Kontext einer Filmbeschreibung erkennen kann und somit die einzelnen Wörter auch verschieden gewichtet, je nach Wichtigkeit im Satz. Im Beispielsatz: "She plays the violin beautifully", würde BERT die Violine höher gewichtet, da es einen wesentlichen Einfluss auf den Inhalt des Satzes hat (Pal, Leuski und Traum 2023).

TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) hingegen ist eine traditionellere statistische Methode, die die Bedeutung von Wörtern in Dokumenten bewertet, ohne Kontext oder Wortreihenfolge zu berücksichtigen. Obwohl es mehr Speicherplatz und Rechenzeit erfordert, was es weniger geeignet für Online-Anwendungen oder umfangreiche Verarbeitung macht, bleibt es effektiv für Dokumentenabruf und einige Textklassifizierungsaufgaben, insbesondere bei längeren Texten. Ein weiterer Vorteil von TF-IDF ist, dass es kein Vortraining benötigt, im Gegensatz zu BERT. Dies macht es zu einer schnelleren und einfacheren Methode zur initialen Textanalyse und -verarbeitung.

Folgende Formel haben wir zur Empfehlungsgenerierung verwendet:

```
self.combined_scores = 50 * similarity_scores + 1 * overlap_director + 2 * overlap_cast + 0.5 * overlap_country + 2 * overlap_genre
```

Je höher der combined_score desto höher ist die Ähnlichkeit zum vorliegenden Film. Es ist wichtig, zu erwähnen, dass die Gewichte anpassbar sind und sich nur bei uns bewiesen haben. In einem anderen Projekt mit anderen Daten könnten andere Gewichte zu besseren Vorhersagen führen (Stober 2005).

Abschließend zu den Modellen

Für die Umsetzung dieser Modelle nutzten wir die Python-Bibliotheken Scikit-learn, Torch und Sentence Transformers. Scikit-learn wurde hauptsächlich für die Implementierung von TF-IDF verwendet, während Torch und Sentence Transformers für die Nutzung und Feinabstimmung von BERT zuständig waren. Diese Bibliotheken ermöglichen es uns, sowohl traditionelle als auch moderne Methoden der Textverarbeitung effizient zu implementieren und zu vergleichen.

In unserer App haben Benutzer die Möglichkeit, zwischen den beiden Modellen zu wechseln, um deren Leistungsfähigkeit auf den gleichen Daten zu evaluieren. Dies erlaubt eine Demonstration und Bewertung der Ergebnisse beider Ansätze in Echtzeit.

Wirtschaftlicher Nutzen

Man könnte sich fragen, warum unser System genutzt werden sollte, obwohl bereits etablierte Dienste wie Netflix, Disney+ und Prime Video existieren. Der entscheidende Unterschied liegt in der plattformübergreifenden Recommendation. Die genannten Streaming-Dienste haben nur begrenzte Angebote und können daher nur Filme und Serien aus ihrem eigenen Portfolio vorschlagen. Dadurch könnten relevante Filmempfehlungen verloren gehen. Beispielsweise könnte ein perfekt passender Film auf Netflix nicht angezeigt werden, nachdem man etwas auf Disney+ geschaut hat.

Hier kommt unser Recommender-System ins Spiel, das diese plattformübergreifenden Empfehlungen ermöglicht. Unser System kann beispielsweise eine Netflix-Serie empfehlen, die hervorragend zu einem Film passt, den der Nutzer auf Disney+ gesehen hat. Dies erhöht den Mehrwert für den Nutzer erheblich und bietet ein umfassenderes und unvoreingenommenes Empfehlungserlebnis.

Darüber hinaus neigen Streaming-Anbieter dazu, ihre eigenen Produktionen zu bevorzugen, um deren Zuschauerzahlen zu steigern. Unser System hingegen vergleicht die Inhalte vollständig unvoreingenommen. Dies bedeutet, dass die Empfehlungen einzig und allein auf der Qualität und Relevanz der Inhalte basieren, völlig unabhängig von deren Herkunft.

Ein weiterer Vorteil unserer App ist die einfache Handhabung, bestimmte Filme aus den Empfehlungen auszuschließen. Bei den üblichen Anbietern werden alle geschauten Filme verwendet, um zukünftige Empfehlungen zu bestimmen. Bei uns jedoch haben Nutzer die Kontrolle und können spezifische Filme ausschließen, damit diese nicht in die Empfehlungen einfließen.

Kritische Würdigung und Fazit

Im Verlauf unseres Projekts haben wir einige wichtige Erkenntnisse gewonnen. Erstens berücksichtigt unser Recommender-System keine aktuellen Neuerscheinungen, was die Relevanz der Empfehlungen beeinträchtigen kann. Der neuste Film im Datensatz ist vom 9. September 2021.

Zweitens wird kein Vergleich zwischen Nutzern durchgeführt, sodass Filme, die innerhalb einer Nutzergruppe beliebt sind, möglicherweise nicht vorgeschlagen werden, da unser System keinen kollaborativen Filter nutzt.

Eine weitere signifikante Limitation des Modells besteht darin, dass Handlungselemente, die nicht explizit in der Beschreibung erwähnt werden, unberücksichtigt bleiben. Dies kann potenziell zu suboptimalen Empfehlungen führen. Ein möglicher Lösungsansatz für dieses Problem könnte in der Implementierung von Large Language Models (LLMs) liegen. Diese könnten beispielsweise zur Analyse von Filmuntertiteln eingesetzt werden, um detailliertere

und umfassendere Filmbeschreibungen zu generieren. Eine solche Erweiterung hätte das Potenzial, die Qualität der Empfehlungen signifikant zu verbessern.

Trotz dieser Herausforderungen bietet unser System signifikante Vorteile. Die plattformübergreifende Integration von Daten aus Netflix, Disney+ und Prime Video ermöglicht umfassendere und unvoreingenommene Empfehlungen. Darüber hinaus können Nutzer bestimmte Filme aus ihren Empfehlungen ausschließen, was die Relevanz der Vorschläge erhöht. Insgesamt stellt unser Projekt einen bedeutenden Schritt in Richtung besserer Film- und Serienempfehlungen dar. Die gewonnenen Erkenntnisse und identifizierten Verbesserungspotenziale bieten eine solide Grundlage für die Weiterentwicklung unseres Systems.

Literaturverzeichnis

Pal, D., Leuski, A. und Traum, D. R. (2023): "Comparing Statistical Models for Retrieval based Question-answering Dialogue: BERT vs Relevance Models." The International FLAIRS Conference Proceedings

Ka, N. G., Durga, K. T. V., Hrishita, N., Ramsankar, R. und Panda, M. (2023): "A Cross-Platform Movie Filtering and Recommendation System Using Big Data Analytics." Procedia Computer Science, vol. 00.

Google Cloud (2024): "Meet AI's Multitool: Vector Embeddings." [online] Verfügbar unter: <https://cloud.google.com/blog/topics/developers-practitioners/meet-ais-multitool-vector-embeddings?hl=en> [abgerufen am 11.07.2024]

Stober, S. (2005): "Kontextbasierte Web-Navigationsunterstützung mit Markov-Modellen." Diplomarbeit, Otto-von-Guericke-Universität Magdeburg.

Sahu, S., Kumar, R., Pathan, M. S., Shafi, J., Kumar, Y. und Ijaz, M. F. (2022): "Movie Popularity and Target Audience Prediction Using the Content-Based Recommender System." IEEE Access, 10, 42030-42046

Tabellenverzeichnis

Tabelle 1 Zeilenvorschau EDA

Abbildungsverzeichnis

Abbildung 1 Fehlende Werte im Datensatz

Abbildung 2 Umgang mit fehlenden Werten

Abbildung 3 Vorgehen bei den Modellen

Abbildung 4 Vektor Embeddings (Google Cloud 2024)

Abbildung 5 Recommendations für "Jaws"

Abbildung 6 Recommendations für "Frozen"

Abbildung 7 Gesamte Web-Applikation