**Projektbericht zum Modul**

**Information Retrieval und Visualisierung**

Sommersemester 2022

**Bearbeitetes Thema:**

**Netflix TV Shows and Movies**

Viet-Anh Do

Matrikelnummer: 220228228

GitHub Repository: <https://github.com/do46/IR_Netflix_TVShows_Movies_VAD>

Inhaltsverzeichnis

[1. Einleitung 3](#_Toc113186941)

[1.1. Anwendungshintergrund 3](#_Toc113186942)

[1.2. Zielgruppe 3](#_Toc113186943)

[1.3. Überblick und Beiträge 4](#_Toc113186944)

[2. Daten 5](#_Toc113186945)

[2.1. Technische Bereitstellung der Daten 6](#_Toc113186946)

[2.2. Datenaufbereitung 6](#_Toc113186947)

[3. Visualisierung 8](#_Toc113186948)

[3.1. Analyse der Anwendungsaufgaben 9](#_Toc113186949)

[3.2. Anforderungen an di Visualisierung 9](#_Toc113186950)

[3.3. Präsentation der Visualisierung 9](#_Toc113186951)

[3.3.1. Visualisierung Eins 9](#_Toc113186952)

[3.3.2. Visualisierung Zwei 9](#_Toc113186953)

[3.3.3. Visualisierung Drei 9](#_Toc113186954)

[3.4. Interaktion 10](#_Toc113186955)

[4. Implementierung 10](#_Toc113186956)

[5. Anwendungsfälle 10](#_Toc113186957)

[5.1. Anwendung Visualisierung Eins 10](#_Toc113186958)

[5.2. Anwendung Visualisierung Zwei 10](#_Toc113186959)

[5.3. Anwendung Visualisierung Drei 10](#_Toc113186960)

[6. Verwandte Arbeiten 10](#_Toc113186961)

[7. Zusammenfassung und Ausblick 10](#_Toc113186962)

[Literatur 10](#_Toc113186963)

# Einleitung

Streaming-Dienste wie Netflix, Amazon Prime oder Disney+ dominieren heutzutage den TV-Markt. Hinter der Erfolgsgeschichte dieser Video-on-Demand Plattformen stehen nicht nur ihre innovative Programme, sondern auch die Ausnutzung von Technologien, um ihre Kunden zu verstehen und ihnen maßgeschnittene TV Shows bzw. Filme zu produzieren bzw. vorzuschlagen. Mit weltweit über 200 Millionen Abonnements bis zum 2. Quartal 2022 ist Netflix neben Amazon Prime die Marktführer im Streaming-Dienst-Bereich. JustWatch, eine Firma mit Sitz in Berlin, verfügt inzwischen über die Datenbank mit über 3.600 Filme und 1.800 Serien [2]. In Rahmen dieser Arbeit wird dieser genannte Datensatz von JustWatch genauer betrachtet, analysiert und schließlich aufbereitet für die Visualsierung, wobei unterschiedliche Visualisierungstechniken angewendet werden.

Das Hauptziel dieser Arbeit ist es, die individuellen Fragen zum Thema Netflix Filme und Serien von der Zielgruppen dieser Arbeit mithilfe einem Visualisierungsprogramm zu beantworten. Um dieses Ziel zu erfüllen, werden unterschiedliche Eigenschaften des Datensatzes durch zwei- oder mehrdimensionale Darstellungen gegenübergestellt. Zuerst einmal werden lediglich zwei Eigenschaften der Filme und/oder Serien in einem einfachen Scatterplot visualisiert. Weiterhin sollte die Anzahl der Dimensionen nicht nur auf zwei beschränkt werden. Um ein tieferes Verständnis des Datensatzes zu erhalten, werden die Daten mittels parallelen Koordinaten dargestellt werden, wobei die Achsen ebenfalls beliebig modifizieren lassen. Schließlich um eine Überblick über die gesamte Datenstruktur des Datensatzes zu gewinnen, werden die Daten durch einen Iconplot präsentieren.

## Anwendungshintergrund

Die in dieser Projektarbeit verwendete Visualisierungstechniken sind die des Scatterplots, der Parallelen Koordinaten und der Iconplot. Damit ein besseres Verständlnis in diesen Techniken geschafft wird, werden sie im Folgenden kurz vorgestellt.

Bei einem Scatterplot geht es um eine Grafikdarstellung von beobachteten Wertepaaren zweier statistischer Merkmale mithilfe von Punkten auf Kartesischem Koordinatensystem. Aus diesen Punkten bilden sich eine Punktwolke, welche auch die Streuung der Daten auf dem

## Zielgruppe

Im Folgenden werden die Zielgruppen dieser Visualisierungsarbeit ausführlich beschrieben. Diese sind namentlich Verbrauchs- und Medienanalytiker und junge Konsumentinnen und Konsumenten.

Die Verbrauchs- und Medienanalytiker sind die Gruppe, die hauptsächlich mit Medienbewertung im Allgemeinen beschäftigen. Diese Visualisierungen können diese Zielgruppe eine Einblick in den Filmtrend über Jahre hinweg, in den Zusammenhänge zwischen ihren Eigenschaften sowie eine gesamte Überblick über allen Eigenschaften der gesamten Filmdatenbank von Netflix. Erkenntnisse aus dieser Visualisierung könnten sowohl für Filmproduzenten als auch für Streamingplattform vorteilhaft sein, denn sie können aus diesen Entdeckungen oder Trenderkennungen dem Zuschauern maßgeschneiderte Filme oder Serien produzieren bzw. in den Portfolio aufnehmen.

Die Gruppe der junge Konsumentinnen und Konsumenten können mithilfe der Visualisierung die Filmwelt selbst erkunden. Für viele junge Leute sind neben brandneue auch alte Filme und Serien von Interesse. Hierfür könnten die Bewertungen von beiden Datenbank eine gute Referenz für ihre Filmwahl.

## Überblick und Beiträge

Diese Arbeit ist ein Versuch, die gelernte Visualisierungstechniken auf ein echtes Anwendungsbeispiel anzuwenden. Der verwendete Datensatz befindet sich in der Webseite *kaggle.com*, eine zur Google gehörende Online-Community, die sich an Datenwissenschaftler richtet. Diese Datensatz liefert dabei Auskünfte über Netflix Filmen- und Serienportfolio, etwa Name, Produktionsjahr, verschiedene Bewertungsmetriken von zwei Filmdatenbanken, also IMDb und TMDb.

Zudem können die entsprechende Zielgruppen die Kategorien auswählen und dementsprechend die Filme auflisten für sie sich interessieren. Die Datensätze liefern neben der Kategorisierung der Filmen einen Überblick über die Bewertung der Filme. Insgesamt haben zwei Unternehmen die Filme bewertet. Mann kann auf dem ersten blick erkennen, dass die Bewertung der Filme unterschiedlich ausfallen. Die dargestellten visualisierungstechniken ermöglichen den Zielgruppen eine Möglichkeit die Produkte zu unterscheiden in filme oder Serien. Interessiert man nur für Serien kann man die Daten so filtern, dass man nur die Serien bekommt und je nach dem was man gerne anschauen möchte, bekommt man die entsprechenden dargestellten Daten. Zudem sind sowohl für die produzierende Filme als auch die Serien das Jahr des Veröffentlichung angegeben. Weiterhin kann man anhand der Daten den Besucheranzahl, die die Filme gestreamt haben, herauslesen.

Die Anwendungshintergrund der ausgewählten Filme und Serien wird in diesem Kapitel angerissen und kurz erklärt und in Kapitel 3 wird den Lesern ein genauer Überblick gegeben. Der Auswahl des Scatterplots für diesen Datensätzen kann so argumentiert werden dass anhand dessen viele Informationen über diese Sreamingsdienst gewonnen können. Dabei ist zu beachten dass je nach dem welche Zielgruppe man gehört, kann man dementsprechend die Achsen auswählen um besondere Erkenntnisse hinsichtlich der Streaming-Angebote aus einem Blick zu erkennen. Der Parallelen Koordinaten ermöglichst die Zielgruppen mehr als zwei Eigenschaften miteinander zu vergleichen um ein besseres Verständnis hinsichtlich der Datensätze zu bekommen. Diesbezüglich können auch Erkenntnisse bezüglich Trends und Besonderheiten sowohl der Filme als auch den Serien den Interessierten ermöglichen. Mithilfe der dritten Visualisierungstechnik können die Zielgruppen Informationen bei den verschiedenen Merkmalen der Filmen hinsichtlich der zwei angegebenen Bewertungsunternehmen in x und x Achse abgebildet werden.

# Daten

Wie bereits erwähnt stammt der verwendete Datensatz einem Nutzer von der Plattform *kaggle.com*, wobei dessen Ursprung nach seiner Angabe aus der Film- und Seriendatenbank JustWatch zurückzuführen ist. Dieser Datensatz enthält 15 Spalten mit 5.978 Film- und Serieneinträge.

Der Datensatz fängt an mit der Spalte „id“, wobei sich die IDs von Filmen und Serien durch die Präfix „tm“ für Filmen und „ts“ für Serien unterscheiden. Neben „title“ für Titel, „type“ für Typen und „release\_year“ für Erscheinungsjahr beinhaltet der Datensatz noch die folgenden Spalten: „description“, „age\_certification“, „runtime“, „genres“, „production\_countries“, „seasons. Eine kurze Beschreibung erhält man durch die Spalte „description“, während die Spalte „age\_certification“ die Auskunft über Altersbeschränkung liefert. „runtime“ gibt die Dauer des einzelnen Film oder die Standarddauer einer Serienfolge an und „genres“ umfasst alle „tags“ von dem Film und Serien, welche sich selbstverständlich auf deren Genres bezieht. „production\_countries“ gibt die Information über Produktionsländer sowohl der Filmen als auch der Serien an, wohingegen in der Spalte „seasons“ nur die Angaben über die Anzahl an Staffeln der Serien eingetragen werden.

Anschließend befinden sich die Informationen bezüglich der zwei Datenbanken, IMDb und TMDb. „imdb\_id“ gibt an, mit welcher ID der Film oder die Serie in der IMDb bezeichnet wird. Die IMDb-Bewertung befindet sich in der Spalte „imdb\_score“. Registrierte IMDb-Nutzer können für jeden veröffentlichten Titel in dieser Datenbank eine Stimme (von 1 bis 10) abgeben. Einzelne Stimmen werden dann aggregiert und zu einer einzigen IMDb-Bewertung zusammengefasst. Nutzer können ihre Stimmen unbegrenzt aktualisieren. Die neue Bewertung wird die vorherige aber dann überschreibt, sodass eine gültige Stimme lediglich pro Titel und pro Nutzer ist [2]. Die dazugehörige Anzahl der abgegebenen Stimme findet man in der Spalte „imdb\_votes“. Von seitens TMDb enthält der Datensatz zwei Spalten, namentlich „tmdb\_popularity“ und „tmdb\_score“. Diese beiden Metriken werden anhands zahlreichen Daten kontinuerlich fortgeschrieben und täglich um 7 Uhr UTC aktualisiert. Zu diesen Daten gehören beispielweise die Anzahl der Stimme pro Tag, Anzahl der Abrufe pro Tag, Bewertung des vorherigen Tags, Anzahl aller abgegebenen Stimmen, usw. [3]

Die Datenqualität dieser Daten eignet sich für beide Zielgruppen der Arbeit gut. Dieser Datensatz verfügt sich über eine klare Struktur, die einheitliche Formatierung des Zellenwertes innerhalb einer Spalte und die Vollständigkeit der Einträge, was die Qualität der Daten auch nach der Datenaufbereitung garantiert. Weiterhin mit über 5.000 Einträge bietet der Datensatz eine angemessene Tiefe sowohl für die Forschungs- als auch für die Selbsterkundungszwecke geeignet. Somit hat der Datensatz die Anforderungen der beiden Zielgruppen erfüllt.

## Technische Bereitstellung der Daten

Die technische Bereitstellung der Daten erfolgt mithilfe des GitHubs Repository, in dem das Visualisierungsprojekt umgesetzt worden ist. Alle Daten, sowohl im Original („titles.csv“) als auch die Aufbereiteten, befinden sich im Ordner „Data“, wobei die letzteren im Unterordner „AufbereiteteDaten“ gespeichert wird. Der Prozess der Datenaufbereitung wird im anschließenden Unterkapitel 2.2 detailiert erklärt.

Das Dateiformat ist „comma separated values“ mit der Dateinamenserweiterung *.csv*, wobei die Werte der nebeneinanderliegenden Spalten mit einem Kommazeichen separiert werden. Da der Delimiter in diesem Fall „Komma“ ist, soll das Dezimaltrennzeichen ein Punkt sein, um Mehrdeutigkeit zu vermeiden. Leerfelder sind allerdings nicht zu vermeiden und Einträge mit Leerfelder werden auch bei der Datenaufbereitung entfernt.

## Datenaufbereitung

Die originalen Daten werden im Python mit dem Package „pandas“ aufbereitet. Visual Studio Code und Jupyter Notebook finden auch Einsatz, um das Code besser zu visualisieren. Die Nutzung von Excel wird hier vermieden, da diese möglicherweise die Datenstruktur ändern könnte. Hinzukommt, dass Excel eigentlich für die Interaktion zwischen Maschine und Menschen gedacht wurde. In diesem Fall fungiert das Aufbereitungstool allerdings um eine Schnittstelle zwischen .csv Datei und Elm-Programmierung, darum ist die Nutzung von Excel nicht geeignet.

Aus dem großen Datensatz, die nach der Verarbeitung „titleslesslessdf.csv“ umbenannt wird, entstehen dazu noch zwei kleinere Datensätze, nämlich „moviedf.csv“ und „showdf.csv“. Der erste enthält Daten zu den Filmen, während in dem zweiten die Daten zu den Serien zu finden sind. Somit gibt es insgesamt drei Datensätze, ein gesamter Datensatz, der eine für Filmen und der andere für Serien. Die Unterteilung der Daten wird durch die Spalte „type“ ermöglicht, da diese genau angibt, ob es sich bei dem Eintrag um einen Film („MOVIE“) oder um eine Serie („SHOW“) handelt.

Wie bereits erwähnt sollen alle Einträge mit Leerfelder entfernt werden, damit alle Einträge gleichbehandelt werden können. Nach der Bereinigung reduziert sich die Anzahl der Einträge bei dem großen Datensatz von 5806 auf 5041 Einträge. Auch zu beachten ist die Daten bezüglich der Anzahl des Staffels, welche beim Filmen immer mit Leerfelder versehen und im Python mit „na“ (not assigned) gekennzeichnet werden. Die Filmeneinträge sollen daher nicht aufgrunddessen gelöscht werden. Die Funktion „fillna()“ sorgt dafür, dass die Spalte „seasons“ von Filmen durch einen bestimmten Wert ausgefüllt werden (hier den Wert „0“).

Weiterhin wird eine weitere Spalte namens „numberTags“ generiert und am Ende der Tabelle zugefügt. Warum wird diese Spalte generiert, wird ausführlich im Kapitel … erklärt. Diese Spalte ist der Spalte „genres“ zurückzuführen, wobei der Wert der generierten Spalte entspricht die Listenlänge der originalen Spalte. In Betracht kommen nach der Datenaufbereitung nur noch die folgenden Spalten: „id“, „title“, „type“, „release\_year“, „runtime“, „genres“, „seasons“, „imdb\_score“, “imdb\_votes“, „tmdb\_popularity“, „tmdb\_score“ und „numberTags“.

# Visualisierung

In diesem Kapitel wird den Lesern einen Überblick verschaffen wie die Visualisierungsthe-chniken bei den oben genannten Zielgruppen Anwendung finden und ob die ausgewählten Visualisierungen die Zielgruppen aussagekräftige Informationen liefern können.

## Analyse der Anwendungsaufgaben

Für die Visualisierungstechniken wurden verschiedene Variable aus den Datensätze in Elm gefiltert um aussagekräftige Antworten auf die Fragen der unterschiedlichen Zielgruppen beantworten zu können. Somit wurde entschieden Variable wie Bewertungsdienste, Anzahl der Zuschauer, Anzahl der Filme und Serien und Produktionsjahr zu berücksichtigen. Da die erhobene Daten sehr umfangreich sind können auch spezifische fragen der Zielgruppen be-antworten können. Zudem muss man beachten dass die Visualisierungtechniken sind so dar-gestellt dass die Zielgruppen Vorkenntnisse in Streamingsdienste haben. Man ist davon aus-gegangen dass das Vorwissen der Zielgruppen dabei helfen kann aussagekräftige Informati-onen aus dem Daten zu ziehen.

Zudem kann man wie auch bereits in dem Kapitel 1.3 erwähnt wurde, Merkmale verglichen werden, um spezielle Fragen zu beantworten. Diese Fragen können mithilfe der Parallele Koordinaten beantworten werden. Zum bespiel kann man vergleichen welche filme aus wel-chen Jahr mehr Zuschauer erhalten haben oder wie die Filme oder die Serien bewertet wur-den. Diesbezüglich wäre interessant zu wissen wie die Bewertung der Serien ausfallen. D.h. kann man Rückschlüsse ziehen welche Serien hinsichtlich der Anzahl der folgen bewertet wurden oder wie der Anzahl der Besucher ausfällt. Das wäre zum Beispiel ein interessante Erkenntnis on die Kunden kürzere Serien oder lange Serien mit umfangreiche folgen gern anschauen. Weiterhin ist es möglich, mithilfe des Stickfigure Plots herauszulesen wie der Auswahl der fünf ausgewählten Merkmalen in Bezug der Bewertungsdiensten ausfällt.

## Anforderungen an die Visualisierung

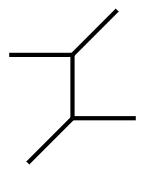
## Präsentation der Visualisierung

### Visualisierung Eins

### Visualisierung Zwei

### Visualisierung Drei

Für die Dritte Visualisierung werden die sogenannte Stickfigure-Technik angewendet. Im Unterschied zu der ersten Visualsierung werden Datenpunkte mithilfe eines Stickfigure in einer mehrdimensionalen Darstellung visualisiert. Bei dieser Technik werden zwei Attribute verwendt, um das Stickfigure auf das Koordinatensystem platziert. Das Stickfigure selbst wird durch die restlichen Attribute gezeugt, wobei einzelne Attribute durch einen Winkel des Stickfigure abgebildet wird. In der aktuellen Darstellung werden



## Interaktion

# Implementierung

# Anwendungsfälle

## Anwendung Visualisierung Eins

## Anwendung Visualisierung Zwei

## Anwendung Visualisierung Drei

# Verwandte Arbeiten

# Zusammenfassung und Ausblick

Literatur

1. IMDb.com Help Center. Ratings FAQ. https://help.imdb.com/article/imdb/track-movies-tv/ratings-faq/G67Y87TFYYP6TWAV#

2. Rabe L (2022) Anzahl der zahlenden Streaming-Abonnenten von Netflix weltweit vom 3. Quartal 2011 bis zum 2. Quartal 2022. https://de.statista.com/statistik/daten/studie/196642/umfrage/abonnenten-von-netflix-quartalszahlen/. Zugegriffen: 28. Juli 2022

3. themoviedb.org Getting Started. https://developers.themoviedb.org/3/getting-started/popularity