Datenbewusstsein

Wo, wie und wozu werden Daten gesammelt und verarbeitet? – Datenbewusstsein durch die Exploration von Empfehlungsdiensten

In diesem Dokument befinden sich ergänzende Informationen zu dem Unterrichtsmodul mit dem Titel „Wo, wie und wozu werden Daten gesammelt und verarbeitet? – Datenbewusstsein durch die Exploration von Empfehlungsdiensten im Kontext von Streamingdiensten“.

Steckbriefinformationen des Unterrichtsmoduls:

Titel:	Wo, wie und wozu werden Daten gesammelt und verarbeitet? – Datenbewusstsein durch die Exploration von Empfehlungsdiensten im Kontext von Streamingdiensten
Zielgruppe:	Informatik in Klasse 8 bis 10 (alle Schulformen)
Inhaltsfelder:	„Informatik, Mensch und Gesellschaft“, „Information und Daten“, „Informatiksysteme“ und „Künstliche Intelligenz und maschinelles Lernen“
Zeitlicher Umfang:	6-8 Unterrichtsstunden je 45 Minuten

Inhalt des Dokuments

1 Was meint ‚Datenbewusstsein‘?.....1

2 Beschreibungen ausgewählter Materialien und weiterführende Hintergrundinformationen:.....2

2.1 Empfehlungsdienste im Allgemeinen (in engl.: Recommender System).....2

2.2 Empfehlungsdienste bei Streamingdiensten.....2

2.3 Erhebung von Daten bei der Nutzung von Streamingdiensten.....4

2.4 Empfehlungsdienst in diesem Unterrichtsmodul.....5

2.5 Zusatzmaterialien und weitere Fakten.....8

3 Glossar relevanter Begriffe:.....8

3.1 Datenbegriff:.....8

3.2 Digitale Artefakte und datengetriebene digitale Artefakte:.....9

3.3 Architektur und Relevanz (Duale Natur digitaler Artefakte):.....9

3.4 Explizit und implizit erhobene Daten:.....9

3.5 Primäre und sekundäre Zwecke der Verarbeitung und Verwendung:.....9

3.6 Data Moves:.....10

1 Was meint ‚Datenbewusstsein‘?

Das Konzept Datenbewusstsein fasst für den Informatikunterricht das Ziel, ein Bewusstsein und Verständnis für die Erhebung, Verarbeitung und Verwendung persönlicher Daten während der Nutzung datengetriebener digitaler Artefakte[i] (s. Glossar in Abs. 3) bei Lernenden zu fördern. Die Erhebung persönlicher Daten während einer solchen Interaktion geschieht durch a) aktive Eingabe von Informationen

seitens des Nutzenden, b) durch Beobachtung und Tracking des Verhaltens sowie c) durch Verarbeitung bereits zuvor erhobener Daten. Dies kann unterteilt werden in die explizite Datenerhebung, also die mit der Handlung intendierte aktive und direkte Bereitstellung von Informationen durch den Nutzenden, und die implizite Datenerhebung, also durch nebenher zur eigentlichen Handlung ablaufende Prozesse, wie etwa Beobachtung, Tracking und Generierung durch Datenverarbeitung[ii] (detaillierte Beschreibung der Begrifflichkeiten in Abs. 3). Nutzende von datengetriebenen digitalen Artefakten sind sich oft der explizit erhobenen Daten bewusst, der implizit erhobenen Daten jedoch oft nicht bewusst. Die so erhobenen persönlichen Daten im Rahmen einer Interaktion mit einem datengetriebenen digitalen Artefakt können durch verschiedene Methoden verarbeitet werden, von einfachen Data Moves (s. Glossar in Abs. 3) bis zu Methoden des Maschinellen Lernens. Mit der Verarbeitung und Verwendung der Daten werden verschiedene Zwecke verfolgt. Dabei können erhobene Daten zum („technischen“) Betreiben von Funktionen des datengetriebenen digitalen Artefakts (primäre Zwecke) und/oder für darüberhinausgehende Zwecke oder etwa zur Untersuchung von Weiterentwicklungen des datengetriebenen digitalen Artefakts (sekundäre Zwecke) verarbeitet und verwendet werden (detaillierte Beschreibung der Begrifflichkeiten in Abs. 3). Primäre Zwecke sind dabei aus Sicht der Nutzenden zu verstehen und sekundäre Zwecke eher aus Sicht der Anbietenden (Was kann ein Anbieter mit den Daten sonst noch machen?). Im Sinne der verschiedenen Zwecke kann etwa ein digitaler Doppelgänger als modellhafte Repräsentation der Nutzenden konstruiert werden. Mit dem Konzept Datenbewusstsein sollte zum einen eine Aufmerksamkeit für datengetriebene digitale Artefakte sowie zum anderen ein Verständnis der Prozesse der expliziten und impliziten Erhebung und Generierung von Daten und der automatisierten Datenverarbeitungsprozesse zu primären oder sekundären Zwecken vermittelt werden. Die Schülerinnen und Schüler sollen also dazu befähigt werden, in einer Interaktion mit einem datengetriebenen digitalen Artefakt die Erhebung und Verarbeitung persönlicher Daten erkennen und analysieren zu können sowie daraufhin selbstbestimmte Entscheidungen für Interaktionen dieser Art treffen zu können.

2 Beschreibungen ausgewählter Materialien und weiterführende Hintergrundinformationen:

2.1 Empfehlungsdienste im Allgemeinen (in engl.: Recommender System)

Ein Empfehlungsdienst verfolgt das Ziel die Menge aller vorhandenen Items (z.B. Filme, Musiktitel, Shopping-Produkte, …) auf eine Vorauswahl (Empfehlungen) einzuschränken, um den Nutzer:innen bei der Entscheidungsfindung zu unterstützen. Dem Nutzer/der Nutzerin sollten also nicht alle Items angezeigt werden, sondern nur eine Auswahl an Items, für die sich der Nutzer potenziell interessieren könnte, um eine Informationsüberflutung zu umgehen. Die Anbietenden des Dienstes zielt damit auf eine Gewinnmaximierung ab, indem der Nutzer/die Nutzerin „neue und interessante“ Items „entdeckt“. Dadurch werden die Nutzer:innen zu längeren und häufigeren Zugriffen (Steigerung der Nutzungszeit) angeregt, wodurch sie mehr Daten hinterlassen und womöglich der Umsatz durch Käufe oder Werbungen gesteigert werden kann.

Im Wesentlichen gibt es inhaltsbasierte (content-based), kollaborative (collaborative) und hybride Methoden zum Filtern der Items. Beim kollaborativen Filtern werden ähnliche Nutzer:innen identifiziert, um dann Empfehlungen basierend auf deren Daten (bspw. Filmbewertungen) zu ermitteln (hier etwa: Mittelwerte der Bewertungen der ähnlichen Nutzer:innen). Beim inhaltsbasierten Filtern werden Daten herangezogen, welche inhaltliche Informationen über die Produkte enthalten bzw. zumindest operationalisieren (z.B. Tags, Genres, Wortvorkommen in Textbeschreibungen). Das hybride Filtern verbindet verschiedene Methoden des kollaborativen und inhaltsbasierten Filterns – i.d.R. nacheinander.

2.2 Empfehlungsdienste bei Streamingdiensten

Zwei Zitate zu Empfehlungsdiensten bei Netflix und Spotify:

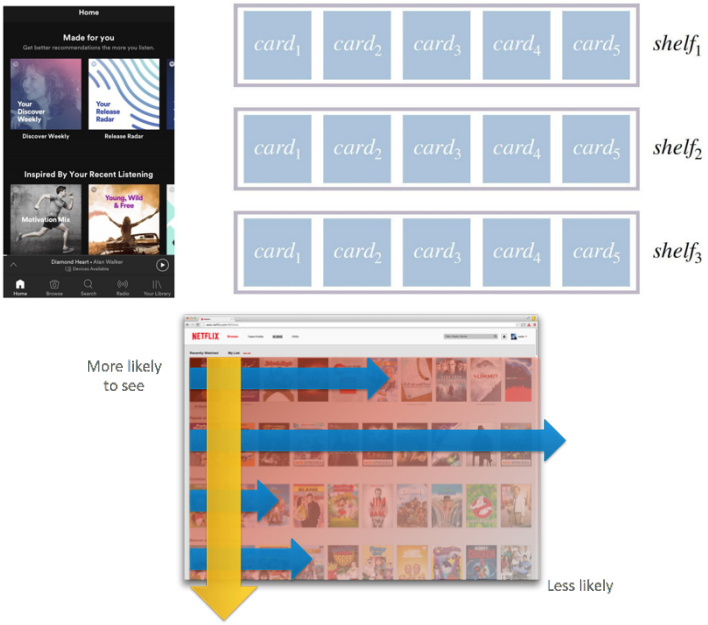
“*A problem we face is that our catalog contains many more videos than can be displayed on a single page and each member comes with their own unique set of interests. Thus, a general algorithmic challenge*

becomes how to best tailor each member's homepage to make it relevant, cover their interests and intents, and still allow for exploration of our catalog." - Nextflix TechBlog (https://netflixtechblog.com/learning-a-personalized-homepage-aa8ec670359a)

"*Spotify has created engines to control and manage everything from your personal best home screen to carefully chosen and organized playlists like Discover Weekly, and continues to explore new ways to understand music, and why people listen to one song or genre over another. All this is achieved with a combination of different recommender systems.*" (https://www.linkedin.com/pulse/how-spotify-recommender-system-works-daniel-roy-cfa/)

In diesem Unterrichtsmodul liegt ein Fokus auf Streamingdienste, wodurch vor allem Plattformen wie Netflix und Spotify im Mittelpunkt stehen. Trotzdem ist dieser Markt stetig am Wachsen: Amazon Prime Video, Disney+, Apple Music oder Amazon Music sind nur einige der Wettbewerber. Schon seit dem Aufkommen von modernen Streamingdiensten arbeiten diese an folgender Frage: Wie kann einem Kunden/einer Kundin möglichst maßgeschneiderte (personalisierte) Produktempfehlungen angeboten werden und wie können somit möglichst viele Kunden zum Bezahlen im jeweiligen Preismodell motiviert werden? Für die Berechnung solcher Empfehlungen spielen verschiedene Faktoren eine entscheidende Rolle: Der grundsätzliche Aufbau (Layout) der Plattform, die Möglichkeiten zur Interaktion für den Kunden/die Kundin und beispielsweise auch das soziale Umfeld, welches sich über verschiedene Benutzer:innen hinweg bildet.

Netflix steht mit seinen über 15000 Filmen und Serien[1] und über 200 Millionen zahlenden Abonnenten weltweit[2] vor der Herausforderung, dass es seinen Nutzer:innen nur eine begrenzte Anzahl an Film- und Serientiteln vorschlagen und auf der Startseite anzeigen kann. Ein vergleichbares Bild ergibt sich bei Spotify: mit 70 Millionen Songs[3] und 365 Millionen monatlich aktive Nutzer:innen[4] findet sich dort zwar ein anderes Produkt wieder, der Aufbau Plattform-Seite und die Datenbasis sind jedoch ähnlich. Die folgenden Illustrationen verdeutlichen die Ähnlichkeit der beiden Plattformen. Auf der linken Seite befindet sich das Layout von Spotify's Oberfläche, das rechte Bild beschreibt selbigen Aufbau bei Netflix:



[1] https://www.comparitech.com/blog/vpn-privacy/netflix-statistics-facts-figures/
[2] https://de.statista.com/statistik/daten/studie/196642/umfrage/abonnenten-von-netflix-quartalszahlen/
[3] https://de.statista.com/statistik/daten/studie/378806/umfrage/anzahl-der-verfuegbaren-songs-auf-spotify/
[4] https://de.statista.com/infografik/13769/monatlich-aktive-nutzer-und-zahlende-abonnenten-von-spotify-weltweit/

Bei diesen Startseiten werden sowohl Reihen als auch Spalten entsprechend an den Nutzer/die Nutzerin angepasst sortiert. Beide Übersichten verdeutlichen, wie wertvoll der Platz auf der Startseite ist und wie relevant damit der Einsatz eines Empfehlungsdienstes wird. Das Geschäftsmodell von Netflix als auch von Spotify ist abonnementbasiert. Dabei gibt es jeweils verschiedene Abo-Pläne, welche im Einzelnen jedoch nicht das Angebot erweitern, sondern z.B. die Anzahl der Nutzenden pro Account verändern. Das Ziel ist es also, die Anzahl der zahlenden Kunden zu maximieren und Abonnements immer wieder zu verlängern. Dies wird dadurch erreicht, möglichst passende Vorschläge für Filme/Musik zu realisieren. Neben Streamingdiensten mit vollem Zugriff auf das komplette Angebot nach Zahlung einer monatlichen Rate gibt es auch andere Geschäftsmodelle, wie etwa bei Amazon Prime Video. Diese Plattform stellt nach einem Abonnement ebenfalls einen Teil seines Film- und Serienangebots zur Verfügung (ca. 13000 Titel). Darüber hinaus gibt es jedoch Medien, welche nach wie vor durch die Zahlung eines einmaligen Betrages freigeschalten werden müssen (ca. 25000 Titel). Dabei kann es sein, dass die ersten Staffeln einer Serie im Abonnementpreis enthalten sind, ab einer bestimmten Staffel jedoch eine zusätzliche Zahlung fällig wird.

2.3 Erhebung von Daten bei der Nutzung von Streamingdiensten

Das Nutzungsverhalten der Nutzer:innen von Streamdiensten ist essenziell, um das Geschäftsmodell und damit den Umsatz stetig zu optimieren. Dabei spielen grundsätzlich jegliche Arten von Interaktionen der Nutzer:innen eine Rolle. Dies fängt bei einfachen Feedbackmechanismen an, wie etwa der “Gefällt mir”-Button, und geht weiter zum Nutzungsverhalten über angeschaute Filme und angehörte Musik. Dies ist nur der Anfang von dem, wie Streamingdienste Daten erheben und verarbeiten. Ein Ziel ist es, einen digitalen Doppelgänger eines jeden Nutzenden zu konstruieren. Auf Basis dessen können dann Empfehlungen und weitere Verarbeitungen und Verwendungen dessen erfolgen.

Ein detailliertes Beispiel der Datenverwendung ist die sogenannte “Completion Rate” (deutsch: Abschlussrate)[5] einer Serie. Dabei kann folgende Frage beantwortet werden: “Wie viele Nutzer:innen, die Serie XY angefangen haben zu schauen, schauten diese bis zu Staffel 3?”. Fällt die Antwort auf diese Frage z.B. auf 30% aus, öffnet sich ein weiterer Pool an Fragen: “An welchem Punkt haben die meisten Nutzer:innen die Serie abgebrochen?” und “Wie lange und mit wie großen Pausen zwischen den Folgen haben die 30% die Serie bis Staffel 3 angesehen?”. Wenn nun ein gewisser Anteil eine Serie zu Ende geschaut hat, liegt es nahe, dass der Serienproduzent (wie inzwischen etwa Netflix) eine weitere Staffel produzieren wird – und im anderen Fall eventuell eine neue Serie produziert, die das Feedback der Nutzer:innen aufgreift, die die Serie nicht zu Ende gesehen haben.

Generell gilt: Bei allen Streamingdiensten werden zunächst alle verfügbaren (persönliche) Daten explizit und implizit erhoben und zu einem gegebenen Zeitpunkt verarbeitet, um das Geschäftsmodell weiter zu optimieren. Darüber hinaus können potenziell auch einzelne Daten an Dritte weitergegeben werden (z.B. Facebook oder andere Werbepartner).[6] Es ist jedoch wichtig anzumerken, dass gerade bei Netflix die monetäre Einnahmequelle auf den Abonnements und nicht auf der Weitergabe von Daten basiert. Diese Entscheidung wurde getroffen, um einen möglich Abgang von Nutzer:innen zu verhindern.[7]

Empfehlungsdienste finden sich heute in fast allen bekannten Online-Plattformen wieder. So nutzt Google solche Systeme beispielsweise bei der Google Suche zum Anzeigen der Suchergebnisse oder auf der Plattform YouTube zum Erzeugen einer personalisierten Startseite bzw. geben personalisierter Videoempfehlungen. Um jeweils zum Nutzer/zur Nutzerin passende Produkte vorzuschlagen, nutzt auch Amazon solchen Modellen. Auch die Bibliothek des KIT in Karlsruhe arbeitet mit ähnlichen Verfahren auf Basis dessen Literaturempfehlungen ausgegeben werden können.

---

[5] https://neilpatel.com/blog/how-netflix-uses-analytics/
[6] z.B. https://www.spotify.com/de/legal/privacy-policy/
[7] https://geschaeftsmodell-workshop.de/geschaeftsmodell/beispiele/netflix-business-model#Das_Business_Model_Canvas_von_Netflix_als_Powerpoint_und_PDF

Dabei spielen im Grunde immer ähnliche Daten eine Rolle. Auf der einen Seite die verschiedenen Produkte, welche auf der Plattform angeboten werden (Filme, Videos, Bücher, ...). Auf der anderen Seite stehen die Nutzer:innen und ermöglichen die Erhebung und Generierung wertvoller Daten durch die Interaktion mit den jeweiligen digitalen Artefakten (Schreiben von Rezensionen, Ansicht von Produkten, Verbindungen zu anderen Nutzer:innen, ...).

2.4 Empfehlungsdienst in diesem Unterrichtsmodul

2.4.1 Explizite und implizite Bewertungen

Bewertungen für Produkte, wie bspw. Filme, können explizit oder implizit vorliegen. Explizit sind Bewertungen dann, wenn der Nutzer/die Nutzerin das Produkt direkt beurteilt, bspw. über ein Gefällt-mir-Button oder eine Sternebewertung. Dadurch gibt der Nutzer/die Nutzerin i.d.R. seine Meinung von dem Produkt bzw. sein Interesse an dem Produkt aktiv zum Ausdruck. Implizite Bewertungen werden nicht von dem Nutzer/der Nutzerin direkt angegeben. Das bedeutet, dass bestimmte Daten erhoben, generiert und verarbeitet werden, welche bspw. als Operationalisierung für das Interesse an dem Produkt dienen können. Beispiele für implizite Bewertungen sind: Hat der Nutzer/die Nutzerin das Produkt gekauft? Hat er oder sie den Film vollständig geschaut oder früher beendet? Wurde der Film mehrmals geschaut? Wurde sich das Produkt gemerkt (Merklisten)?

In dem Datensatz, welcher im Unterricht eingesetzt wird (Beschreiben siehe Abschnitt 2.4.2), wurden neben den expliziten Bewertungen zusätzlich implizite Beurteilungen generiert, um die beiden Konzepte zu veranschaulichen. Dabei sind implizite Filmbewertungen im Filmdatensatz die binäre Antwort auf die Frage, ob Nutzer:innen einen Film zu Ende geschaut hat oder nicht. Dieses neue Attribut wurde künstlich, jedoch auf Basis der vorhandenen Bewertungen angelegt. Mit einer Wahrscheinlichkeit von 85% wurden Bewertungen mit mehr als vier Sternen auf den Status “1” (Film zu Ende angeschaut) gesetzt. Liegt die Bewertung unter vier Sterne fand dies nur in 40% der Fälle statt. Alle übrigen Bewertungen erhielten den Status “0” (Film nicht zu Ende angeschaut). Im Anschluss wurden 25% der expliziten Bewertungen entfernt, um die Relevanz der impliziten Bewertungen darzustellen.

2.4.2 Genutzte Ratingdaten und Aufbereitung dieser für das Unterrichtsmodul

In diesem Unterrichtsmodul ziehen wir reale Nutzungsdaten von Nutzer:innen der Plattform MovieLens (movielens.org) heran. Auf der Plattform angemeldete Nutzer:innen können dort u.a. Filme bewerten und Filmempfehlungen bekommen. Es ist also ein Empfehlungsdienst eingebettet. Die Betreiber haben Bewertungsdaten öffentlich zugänglich gemacht[8]. Für das Unterrichtsmodul haben wir diese Daten aus Performancegründen verkleinert, sodass wir lediglich ca. 50000 Bewertungen von ca. 5000 Usern zu insgesamt ca. 600 Filmen nutzen.

Die Filme, die bewertet werden können, wurden manuell nach einer subjektiven Einschätzung des Bekanntheitsgrades sowie unter Einbezug von IMDB-Hitlisten ausgewählt. Filme, welche unter den möglichen Empfehlungen erscheinen, haben eines Mindestanzahl an Bewertungen erhalten. Die Nutzer:innen in dem Datensatz wurden so ausgewählt, dass sie alle mindestens einen der Filme, welche über das Empfehlungsmodul bewertet werden können, selbst bewertet haben. Grundsätzlich lag das Hauptaugenmerk bei der Datengenerierung auf der Balance zwischen der Performance des Modells und den für die Berechnungen verfügbaren Hardware-Ressourcen.

Für das Unterrichtsmodul nutzen wir die Daten in Form von Datentabellen (DataFrames). Diese sind in den nachfolgenden Bildern dargestellt. Sie umfassen in der ersten Datentabelle Informationen über die Filme

[8] Referenz zum Projekt: https://dl.acm.org/doi/10.1145/2827872; Daten: https://grouplens.org/datasets/movielens/

(Titel, Genre, Erscheinungsjahr) und in der zweiten Datentabelle gerade die explizit und implizit erhobenen *Bewertungen* der Nutzer:innen (Ids, Datum, Uhrzeit, Fertig_Angeschaut, Rating). Sowohl die Nutzer:innen als auch die Filme bekommen eine ID zugewiesen, mit der sie eindeutig identifiziert werden.

	movieId	title	genres	Erscheinungsjahr
0	121364	Memories of My Melancholy Whores (2011)	Comedy|Drama|Romance	2011
1	177261	Jackals (2017)	Horror|Thriller	2017
2	188769	Polish Legends: Twardowsky (2015)	Sci-Fi	2015
3	203831	Anabolic Life (2017)	Thriller	2017
4	190203	Kaleidoscope (2017)	Thriller	2017
...	...	...	...	...
5666	170993	Mini's First Time (2006)	Comedy|Crime|Drama	2006
5667	191137	La Désintégration (2012)	Drama	2012
5668	46850	Wordplay (2006)	Documentary	2006
5669	115680	Time Lapse (2014)	Crime|Drama|Sci-Fi|Thriller	2014
5670	198503	The Isle (2018)	Horror	2018

5671 rows × 4 columns

	userId	movieId	Datum	Uhrzeit	Fertig_Angeschaut	rating
0	1	29	18-01-2017	16:56:07	1	4.5
2	1	77561	13-08-2015	15:40:40	1	4.0
3	1	72998	13-08-2015	16:03:43	1	4.0
4	1	70293	13-08-2015	15:58:02	1	4.0
5	1	68952	18-08-2019	03:18:04	1	2.5
...	...	...	...	...	...	...
1460217	9080	3979	27-05-2018	15:55:41	1	1.5
1460218	9080	3949	19-05-2018	09:02:01	1	4.5
1460219	9080	3418	28-05-2018	15:09:28	1	4.0
1460220	9080	3354	19-05-2018	09:38:53	0	2.0
1460222	9080	195163	21-01-2019	16:04:36	0	2.0

1210706 rows × 6 columns

2.4.3 Jupyter Notebooks im Allgemeinen

Jupyter Notebooks ermöglichen das Ausführen von Pythoncode in Echtzeit mit Hilfe sogenannter Codezellen. Die Ergebnisse werden jeweils unter der aktuellen Zelle angezeigt. Erläuterungen zu Aufgaben zwischen den einzelnen Codezellen können auf Basis der Auszeichnungssprache Markdown realisiert werden. Der gesamte Code kann dabei jederzeit manipuliert werden, was das spielerische Herantasten an Programmierung ermöglicht. Zu beachten ist, dass die Ausführung im Falle des Empfehlungsdienstes in diesem Modul nicht auf der lokalen Maschine, sondern auf dem zentralen Server der Universität Paderborn stattfindet.

2.4.4 Vorbereitetes Jupyter Notebook

Für dieses Unterrichtsmodul haben wir Bibliotheken entwickelt und ein Jupyter Notebook für den Unterricht vorbereitet. In diesem Jupyter Notebook (*Empfehlungsdienst für Filme.ipynb*) werden zunächst die Daten automatisiert eingelesen und ein Empfehlungsdienst am Beispiel von Netflix beschrieben. Anschließend ermitteln die Lernenden nach Eingabe von eigenen Bewertungen eigene Filmempfehlungen über einen bereits implementierten Empfehlungsdienst. Dieser basiert auf dem k-Nearest-Neighbor Algorithmus (Erklärung siehe unten) und nutzt als Basis seiner Vorschläge die vorgefilterten Bewertungsdaten. In der Standardeinstellung arbeitet der Dienst ausschließlich mit expliziten Bewertungen. Über einen Schalter im Code kann dieser jedoch die Empfehlungen auch basierend auf impliziten Bewertungen berechnen. Im nächsten Schritt wird die Frage behandelt, welche Daten erhoben wurden. Dafür können die Lernenden eine User-Movie-Tabelle (Erklärung siehe unten) aufrufen. Danach beschäftigt sich das Notebook mit der Frage, wie personalisierte Empfehlungen automatisiert berechnet werden können. Zur Visualisierung wird ein 2-dimensionales Koordinatensystem herangezogen. Darin

können zwei Filme ausgewählt werden. Zu diesen Filmen werden dann alle vergebenen Bewertungen visualisiert. Somit sind einfache Analysen in Relation zur eigenen Bewertung möglich.

Alle nötigen Befehle werden in gelben Hinweisboxen erklärt. In blauen Boxen werden die Aufgaben detailliert formuliert und es werden grüne Einführungs- bzw. Erklärboxen eingeschoben.

Beim ersten Zugriff auf das Verzeichnis mit den Jupyter Notebooks muss man einen Login erstellen, mit dem zu einem späteren Zeitpunkt wieder an die letzte Bearbeitung angeschlossen werden kann. Andernfalls würden die Bearbeitungen nach schließen des Jupyter Notebooks gelöscht werden. Das Verzeichnis ist unter folgendem Link zu erreichen: https://ddi-jupyter.cs.upb.de/empfehlungsdienst/

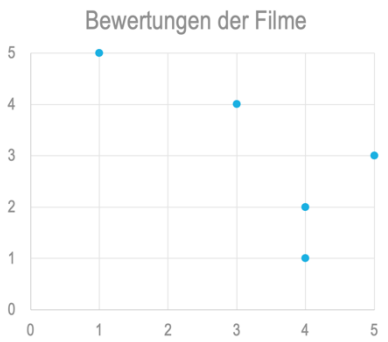
2.4.5 k-Nearest-Neighbor Algorithmus zur Erstellung eines Modells:

Der k-nearest-neighbor Algorithmus wird in dem vorbereitetem Jupyter Notebook mithilfe der Bibliothek sklearn zur Erstellung eines Modells verwendet. Dieses Modell kann anschließend zum Entscheiden von ähnlichen Nutzer:innen (eigentlich: nächsten Nachbarn) auf Basis von Daten aus z.B. einem Streamingdienst angewendet werden. Die konkrete Funktionsweise des Algorithmus wird im Unterricht nicht im Detail vermittelt, es soll lediglich die Idee der Vorgehensweise verstanden werden. An der Stelle der Modellerzeugung wird bewusst eine Black-Box gesetzt, um die im Rahmen dieser Unterrichtsreihe gesetzten Lernziele zu erreichen und keine Überforderung zu erzeugen.

Beispiel:

In der nebenstehenden Tabelle ist ein Minimalbeispiel gegeben. Es gibt Bewertungsdaten von fünf Nutzer:innen zu zwei Filmen. Anhand dieses Beispiels kann bereits das Suche nach den k nächsten Nachbarn erklärt werden. Gesucht sind zum Beispiel zwei Nutzer:innen, die ähnlich zum markierten User 5 sind. Das sind dann etwa die User 1 und 4, da diese die kleinste Abweichung in ihren Bewertungen der beiden Filme zu User 5 haben. Konkret heißt das, dass die Abstände zwischen der Tabellenzeile von User 5 und denen von User 1 und 4 am kleinsten sind, die Differenz also möglichst klein ist. (Randnotiz: Mathematisch nutzen wir in unserer Umsetzung die euklidische Metrik für die Bestimmung von Abständen.) Dies kann auch in dem nebenstehenden Koordinatensystem visualisiert werden. Die Bewertungen zu Film A entsprechenden den Werten auf der x-Achse und zu Film B denen auf der y-Achse. So stellt jeder Punkt im Koordinatensystem einen User da, der beide Filme bewertet hat. Mit dieser Vorgehensweise können zu einem gewählten User die k ähnlichsten Nutzer:innen einfach identifiziert werden. So können im Koordinatensystem etwa beliebig viele Nutzer:innen hinzugefügt werden. Um nun anhand dieser ähnlicher Nutzer:innen zu User 5 für einen dritten Film C herausfinden, ob dieser empfohlen werden sollte, wird eine Prediction ermittelt. Die Prediction wird etwa durch den Mittelwert der Bewertungen des Films C der ähnlichen Nutzer:innen ermittelt. In dem Beispiel der nebenstehenden Tabelle ist dies dann 4,5 (Mittelwert von 4 und 5). Das heißt, wenn User 5 den Film C schauen und bewerten würde, würde er wahrscheinlich eine Bewertung von 4,5 abgeben. Dem User 5 sollte der Film C also durchaus empfohlen werden. Dieses (hier stark reduzierte) Verfahren ist auf eine große Anzahl von Nutzer:innen und Filmen übertragbar.

	Film A	Film B
User 1	5	3
User 2	3	4
User 3	1	5
User 4	4	2
User 5	4	1



	Film A	Film B	Film C
User 1	5	3	4
User 2	3	4	
User 3	1	5	4
User 4	4	2	5
User 5	4	1	

2.4.6 User-Movie-Tabelle als hilfreiche Tabelle zur Ermittlung von Empfehlungen

Die user-movie-Tabelle ist eine Datentabelle, die in diesem Kontext Filmbewertungen (Zellen) von Nutzer:innen (Zeilenweise userIds) zu den jeweiligen Filmen (Spaltenweise Filmtitel) aufführt. Diese Tabelle ist für den Empfehlungsdienst recht zentral, anhand dieser wird beispielsweise das vom k-Nearest-Neighbor Algorithmus ermittelte Modell mit einer aus der Tabelle erstellten sparse-Matrix berechnet. Ähnliche User werden also mithilfe der Abstände zwischen den jeweiligen Zeilen in dieser Tabelle bestimmt.

2.5 Zusatzmaterialien und weitere Fakten

Bild einer Startseite von Netflix:

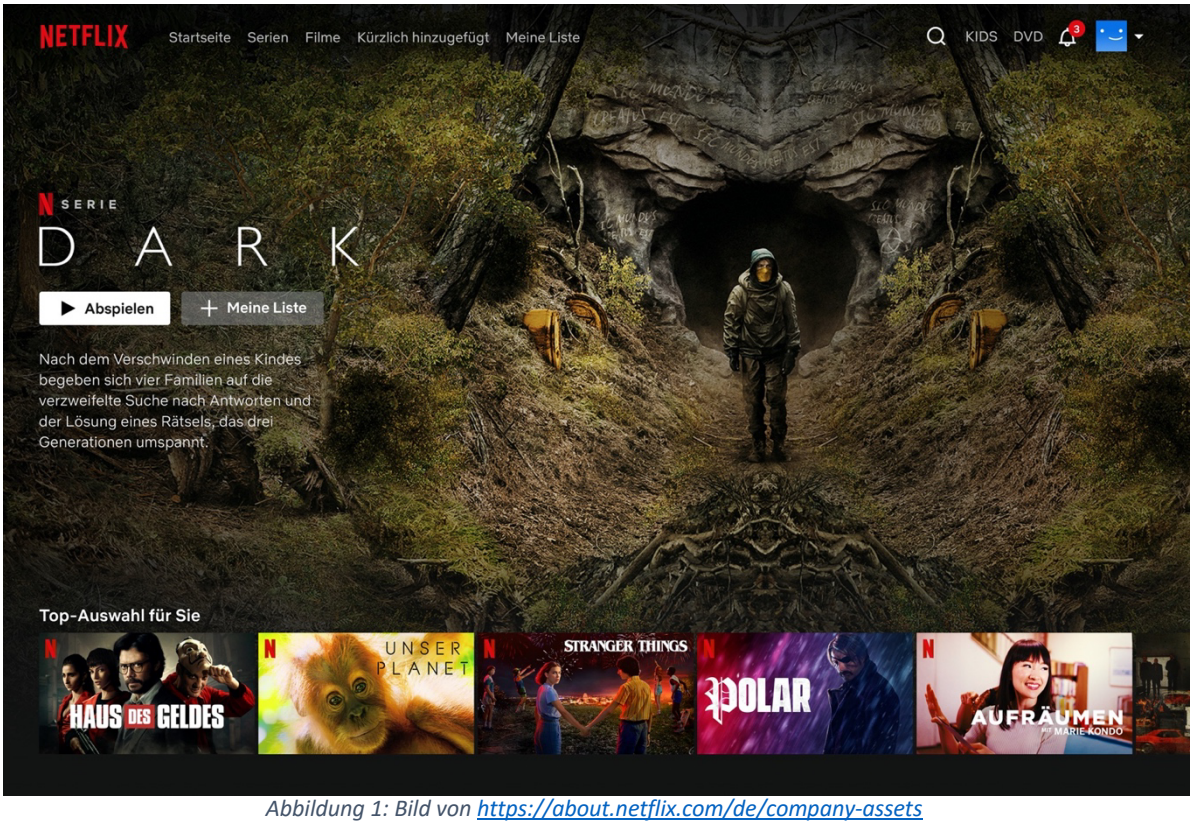


Abbildung 1: Bild von https://about.netflix.com/de/company-assets

Die Abbildung der Startseite wird als Motivation und Einstieg für die personalisierte Startseite verwendet. Weitere Informationen hierzu finden sich im Verlaufsplan.

3 Glossar relevanter Begriffe:

3.1 Datenbegriff:

Wie im Kernlehrplan NRW für Informatik in Klasse 5 und 6 dokumentiert (s. Inhaltsfeld „Information und Daten“) kann zwischen Information und Daten unterschieden werden: Daten sind (digital) repräsentierte Informationen und können etwa in Informatiksystemen gespeichert und verarbeitet werden. Hinsichtlich des Konzepts Datenbewusstsein ist besonders hervorzuheben, dass Kontexte, Phänomene oder etwa auch Interessen, Emotionen oder Handlungen einer Person anhand ausgewählter Merkmale modelliert werden. Gerade die persönlichen Daten, welche für Datenbewusstsein eine große Rolle spielen, entstammen einem Kontext, wo die jeweilige Person involviert ist oder war. Bei der Thematisierung von Daten sollte der Modellierungsaspekt nicht vernachlässigt werden, da die Kontexte, Phänomene oder Personen nicht

vollständig repräsentiert werden. Im Kontext des Datenbewusstseins bedeutet das gerade, dass die erhobenen, persönlichen Daten kein Abbild einer Person darstellen, sondern lediglich modellhaft anhand ausgewählter Merkmale repräsentiert. Dadurch kann auch eine verzerrte Repräsentation einer Person entstehen. Außerdem sollte beachtet werden, dass gewisse Informationen, wie etwa Emotionen oder Interesse, im Sinne der Merkmale für die erhobenen Daten operationalisiert werden, was oft nicht trivial ist (Was heißt es, wenn ein datengetriebenes digitales Artefakt das „Interesse" der Nutzenden speichert? – Was ist das? Woran wird das fest gemacht?)

3.2 Digitale Artefakte und datengetriebene digitale Artefakte:

Im Konzept Datenbewusstsein wurde der Begriff der *datengetriebenen digitalen Artefakte* (ddA) eingeführt. Dieser beschreibt eine spezielle Art von digitalen Artefakten. *Digitale Artefakte* sind ein Sammelbegriff für digitale Werkzeuge, Computersysteme aller Art, ihre Bestandteile, ihre Verbindung untereinander. Sie umfassen Sowohl Hardware, Software, Daten und Objekte sowie Algorithmen und Datenstrukturen. *Datengetriebene digitale Artefakte* sind dann digitale Artefakte, die sich selbst oder ihre Rückmeldung in der Interaktion mit diesem durch die Verarbeitung erhobener Daten verändert. Diese nutzen dann oft zum Beispiel auch Methoden des Maschinellen Lernens.

3.3 Architektur und Relevanz (Duale Natur digitaler Artefakte):

Die duale Natur digitaler Artefakte oder auch Dualität beschreibt eine analytische Trennung von Aspekten eines digitalen Artefakts (s.o.). Ein digitales Artefakt kann dieser Auffassung nach aus der Perspektive auf die Architektur und auf die Relevanz beschrieben werden[iii]. *Architektur* meint alle technologischen Strukturen und Mechanismen. *Relevanz* hingegen meint Intentionen, Funktionen, Meinungen, Interpretationen und der Kontext eines digitalen Artefakts.

3.4 Explizit und implizit erhobene Daten:

Im Konzept Datenbewusstsein wurden die Begrifflichkeiten der *explizit* und *implizit erhobenen Daten* eingeführt. Diese stehen in der Regel in der Verbindung zum Nutzenden - oft stellen sie personenbezogene Daten dar. Die explizit erhobenen Daten sind jene, die der Nutzende mit seiner Handlung intendiert zu erstellen, also direkt und aktiv eingegeben bzw. erzeugt hat. Darüber sind sich Nutzende in der Regel bewusst. Dies sind zum Beispiel bei Social Media Plattformen gepostete Texte und Bilder, bei einer Suchmaschine etwa der Suchbegriff oder beim Telefonieren über das Mobilfunknetz die Telefonnummer desjenigen, den man anrufen möchte. Im Gegensatz dazu, werden implizit erhobene Daten indirekt durch Beobachtung (Tracking) oder Verarbeitung bereits erhobener Daten nebenher zur eigentlichen Handlung des Nutzenden erhoben und generiert. Dieser Datenerhebung sind sich Nutzende oft nicht bewusst. Im Beispiel der Social Media Plattform sind dies etwa Likes und Klicks, bei der Suchmaschine etwa Klicks auf Suchergebnisse oder beim Telefonieren über das Mobilfunknetz etwa Standortdaten der verbundenen Basisstationen.

3.5 Primäre und sekundäre Zwecke der Verarbeitung und Verwendung:

Im Konzept Datenbewusstsein wurden die Begrifflichkeiten der *primären* und *sekundären Zwecke* der Verarbeitung und Verwendung erhobener Daten eingeführt. Diese beziehen sich auf die Verarbeitung und Verwendung von Daten über einen Nutzenden, die bei der Nutzung von datengetriebenen digitalen Artefakten erhoben werden. *Primäre* und *sekundäre Zwecke* beziehen sich auf die Intention, mit der diese zuvor erhobenen Daten verarbeitet und verwendet werden. Primäre Zwecke umfasst, dass die erhobenen Daten dazu verarbeitet und verwendet werden, um das datengetriebenen digitalen Artefakten mit den Features anbieten zu können. Diese beziehen sich auf einer Nutzerperspektive auf die Verarbeitung und Verwendung: Die Daten werden verarbeitet und verwendet, um Nutzenden Features anbieten zu können. Im Beispiel der Suchmaschine ist dies etwa das Anzeigen von Suchergebnissen. Auch inbegriffen wäre, wenn die Suchergebnisse personalisiert geordnet werden. Im Sinne des Features für den Nutzenden würde dies bedeuten, dass der Nutzende gerade die Ergebnisse angezeigt bekommt, die für ihn idealerweise relevant

sind. Sekundäre Zwecke bedeutet, dass die Daten verarbeitet und verwendet werden, um andere/weitere Zwecke zu verfolgen – z.B. weitere wirtschaftliche oder wissenschaftliche Zwecke. Diese „Zweitverwertung" der Daten bezieht sich auf einer Anbieterperspektive auf die Verarbeitung und Verwendung der erhobenen Daten: Wozu kann ein Anbieter eines datengetriebenen digitalen Artefakts die erhobenen Daten noch nutzen? Im Kontext von Streamingdiensten (z.B. Spotify) könnte dies etwa umfassen, dass Nutzungsdaten (z.B. gehörte Musik) zur Analyse der Emotionen der Nutzenden verwendet werden.

3.6 Data Moves:

Mit Data Moves werden Datenoperationen beschrieben. Diese umfassen etwa folgende[iv]:

- *Filtern:* Bilden einer Teilmenge der Daten
- *Gruppieren:* Daten in Teilgruppen unterteilen
- *Zusammenfassen:* Aggregieren von Daten
- *Berechnen:* Neue Attribute ausgehend von existierenden Daten erstellen (z.B. Ausgehend von zwei Spalten eine dritte Spalte erzeugen)
- *Merging/Joining:* Datensätze zusammenführen
- *Reorganisieren:* zum Beispiel ändern der Darstellung der Daten

---

[i] Höper, L. & Schulte, C., (2021). Datenbewusstsein: Aufmerksamkeit für die eigenen Daten. In: Humbert, L. (Hrsg.), *INFOS 2021 – 19. GI-Fachtagung Informatik und Schule*. Gesellschaft für Informatik, Bonn. (S. 73-82). DOI: 10.18420/infos2021_f235

[ii] Höper, L. & Schulte, C., (2021). Datenbewusstsein im Kontext digitaler Kompetenzen für einen selbstbestimmten Umgang mit datengetriebenen digitalen Artefakten. In: Gesellschaft für Informatik (Hrsg.), *INFORMATIK 2021*. Gesellschaft für Informatik, Bonn. (S. 1623-1632). DOI: 10.18420/informatik2021-136

[iii] Schulte, C., & Budde, L. (2018). A Framework for Computing Education: Hybrid Interaction System: The need for a bigger picture in computing education. In *Proceedings of the 18th Koli Calling International Conference on Computing Education Research* (S. 1-10).

[iv] Erickson, T., Finzer, B., Reichsman, F., & Wilkerson, M. (2018). Data Moves: one key to data science at school level. In *Proceedings of the International Conference on Teaching Statistics (ICOTS-10)* (Vol. 6).