Projektbericht zum Modul Information Retrieval und Visualisierung Sommersemester 2022

Eine Analyse von Anzeigen auf einer deutschen Gebrauchtwagenplattform aus dem Jahr 2023

Johannes Göbel

17. Dezember 2023

1 Einleitung

Internetmarktplattformen gewannen zuerst im Jahr xxx mit Firmen wie e-bay, bla bla, asd, bedeutung. Heute, y Jahre später, interagieren Menschen so selbstverständlich mit solchen Plattformen so selbstverständlich, dass ... Gerade deshalb lohnt es sich einen Blick auf die Mechanismen von diesen Plattformen zu werfen. Beispielsweise könnte es ein besonderes Interesse zu sein nachzuvollziehen, wie Verkäufer ihre Angebote bewerben. Dies soll in dieser Arbeit am Beispiel von einer Gebrauchtwagenplattform geschehen. // Dabei sollen in dieser Arbeit aber auch Markteffekte wie Preisentwicklungen von PKWs mit unterschiedlichen Eigenschaften nicht außer Acht gelassen werden. Zu diesem Zwecke möchte ich die folgenden Forschungsfragen formulieren:

- Wie hängt der Angebotspreis mit den angegeben Merkmalen eines Gebrauchtwagens zusammen?
- Was sind die Werte für ein bestimmtes Land? Wo befindet es sich im Vergleich zu den anderen Ländern?
- Wie konstant sind die Werte von Autos? Verlieren Gebrauchtwagen konstant an Wert?

1.1 Anwendungshintergrund

Sie müssen genug Hintergrund bereitstellen, so dass die Lesenden sich ein Urteil bilden können, ob ihre Lösung funktioniert. Sie sollen die Lesenden jedoch nicht mit Anwendungsdetails so überschütten, dass der Fokus auf die Fragen zur Informationsvisualisierung untergehen.

Zu diesem Zweck soll eine - Problem: das sind immer Angebotspeise, wir kennen die Kaufpreise nicht

1.2 Zielgruppen

Beschreiben sie die Personengruppe oder Personengruppen, die das von ihnen benannte Anwendungsproblem lösen möchte. Auf welches Vorwissen können sie in dieser Gruppen von Anwenderinnen aufbauen? Welche Informationsbedürfnisse werden durch die Visualisierungen adressiert?

Die Zielgruppe dieser Arbeit sollen Gebrauchtwagenverkäufer sein. Diese Zielgruppe zeichnet sich durch eine hohe Fachkenntniss aus.
Im Markt profitert diese Gruppe von bestehenden Informationsasymetrien, die sich durch Marktintransparenzen bilden. Durch Internetmarktplattformen sind diese Marktintransparenzen seltener zu finden. Beispielsweise können Käufer Preise vergleichen und können damit besser abschätzen, wie hoch der reale Marktpreis ist.
Daher interessieren professionelle Marktteilnehmer, wie auch trotz dieser technologischen Entwicklungen, eine relevante Marge erzielen können. Dieser Informationsbedürfnisse möchte diese Arbeit mithilfe von Visualisierungen zu Preisentwicklungen und Bewerbungsstilen decken.

1.3 Überblick und Beiträge

In diesem Abschnitt geben sie einen kurzen Überblick über die Daten und verwendeten Visualisierungen. Dann benennen sie die Beiträge ihres Projekts. Diese Beiträge müssen sie in den hinteren Teilen des Berichts genauer ausführen und belegen.

2 Daten

2.1 Datenverarbeitung

Der Datensatz auf der Webseite Kaggle unter dem Titel "Germany Used Cars Dataset 2023"(https://www.kag
used-cars-dataset-2023) bereitgestellt. Dieser enthält über 200 Tsd. Datenpunkte zu gescrapeten Daten von der Gebrauchtwagenplattform Autoscout24 aus dem Jahr 2023.
Dieser Datensatz wurde anschließend um einen Überblick zu bekommen in ein IPython Notebook geladen. Nach diesen ersten Überlegungen sollte der Datensatz bereinigt werden. Zum Beispiel Antriebsart electric, usw. Dazu wurden die Attribute bewertet, ob sie für eine spätere Visualisierung (1) elementar sind oder eher (2) sekundär. Darauf basierend wurde die Entscheidung getroffen, ob (1) fehlende Werte dieses Attributes zu einer Entfernung des Datums führen oder (2) die Werte aufgefüllt werden. Die Art und Weise einer Auffüllens wird anhand des Attributes entschieden. Das Ziel der Bereinigung ist es in jedem Feld einen verarbeitbaren Wert vorfinden zu können.
Zuerst wurde die unbenannte Indexspalte gelöscht. Ebenso alle Zeilen die keine Werte in den Spalten öffer$_d$escription öder"mileage$_i$n$_k$m"hatten.Anschlie„endwurdenallefehlendenWertederSpalte"col
WeilindiesemZusammenhangvieleunsinnigeDatenauftratenwurdenauchalleanderenSpaltenaufWerteg
ImAnschlussdaranwurdenalleDatengelöscht, wo"power$_p$söder"power$_k$w"nullist.

$Zuletztwurdeentschiedendie Spalte"fuel_consumption_gk"mßuentfernen, dadieseviele fehlendeWerteenthäl$
$Datenladenzukönnen.$
$Eineähnliche Umwandlungwurdeinder Spalte"registration_date"vollzogen, dieineinDatumsformatumgew$
$NachdiesenSchrittenwarenkeineleerenFeldermehrvorhanden.$
$DabeiwareinZieldieVergleichbarkeitvonFahrzeugenunterschiedlicherAntriebsartenzuGewährleisten, w$

Hier wurde der Datensatz auch um erste Merkmale wie offer_len, $diedieAnzahlderBuchstabendesAnzeigete$

Im Anschluss daran wurde der Datensatz im CSV-Format exportiert und in Elm geladen.
Dabei wurden drei Datensätze exportiert: (1) "data.csv": Der Standarddatensatz, der rein wie oben beschreiben erstellt wurde.
(2) ävg_star_data.csvünd(3)ßum_star_data.csv" : $DieDatensätzefürdenStarplot, welchedurcheinenzusätzlich$
$EineletzteBemerkungsollzueinemgrö„erenProblembeimLadenderDateninElmgewidmetsein.Leidertrate$
$NotebooksProblemebeiderKodierungderDatenauf, sodassunsichtbareCharactersinderCSVvorzufinden$

2.2 Eignung der Daten

In diesem Abschnitt soll diskutiert werden, inwieweit sich die nun vorliegenden Daten eignen um die vorher gestellten Forschungsfragen zu beantworten.
Grundsätzlich ist zu sagen, dass der hier verwendete Datensatz aus von einer Marktplatz-Webseite gescrapeten Daten besteht. Dies führt dazu, dass diese Daten ein realitätsnahes Bild dieses Marktplatzes geben können, wenn diese Daten Stichprobenartig erhoben wurden. Allerdings heißt das auch, dass die Übertragung der Erkenntnisse aus der Datenanalyse und -visualisierung eingeschränkt übertragbar auf andere Plattformen sind.
Des Weiteren ist zu bemerken, dass keine Kenntnis darüber besteht ob die hier vorliegenden Daten authentisch sind und tatsächlich und ohne eine Vorauswahl, die zu einer Verfälschung führen würde, gescrapet wurden. Andererseits bietet sich auch kein Grund an der Authentizität der Daten zu zweifeln.

3 Visualisierungen

3.1 Analyse der Anwendungsaufgaben

Analysieren sie die konkreten Anwendungsaufgaben, die die Lösung des Zielproblems durch die Anwender:innen bearbeitet werden müssen. Welche sinnvollen mentale Modelle helfen den Personen bei der Bearbeitung. Sind diese mentalen Modelle für sie notwendig, um die Aufgaben lösen zu können? Gehen sie bei ihrer Argumentation von den Anwendungsaufgaben aus und kommen sie dann zu den mentalen Modellen, deren Aufbau durch Visualisierungen unterstützt

wird.

3.2 Anforderungen an die Visualisierungen

Bei diesem Visualisierungsprojekt sind klare Anforderungen an die Darstellung von Daten von entscheidender Bedeutung. Um potenziellen Käufern eine hilfreiche Entscheidungsgrundlage zu bieten, sollten die Visualisierungen Aspekte wie Fahrzeugpreise, Kilometerstand, Baujahr, etc. einschließen. Eine intuitive Benutzeroberfläche ist unabdingbar, damit die Interessenten mühelos durch die Vielzahl von Angeboten navigieren können. Die Visualisierungen sollten daher die Möglichkeit bieten, Fahrzeuge anhand verschiedener Parameter zu filtern und zu vergleichen. Ein interaktiver Ansatz mit Filteroptionen und Hover-Funktionen könnte dabei helfen, detaillierte Informationen zu einzelnen Fahrzeugen abzurufen. Die visuelle Darstellung sollte sowohl informativ als auch ansprechend sein, um die Benutzererfahrung zu optimieren und eine effektive Entscheidungsfindung zu ermöglichen.
Basierend auf diesen Metaanforderungen werden konkrete User Stories formuliert, die im Rahmen des Projektes addressiert werden sollen.
Konkrete Visualisierungsanforderungen:

1. Als Nutzer möchte ich einen interaktiven Scatterplot haben, der es mir ermöglicht, Fahrzeuge anhand verschiedener Parameter wie Preis, Kilometerstand und Baujahr zu vergleichen, um schnell einen Überblick über die verfügbaren Optionen zu erhalten.

2. Als Nutzer wünsche ich mir einen benutzerfreundlichen Parallelplot, der alle relevanten Informationen zu einzelnen Fahrzeugen auf einen Blick darstellt. Damit kann ich die Zusammenhänge zwischen den verschiedenen Parametern besser verstehen und gezielt nach meinen Präferenzen filtern.

3. Als Nutzer interessiere ich mich für einen Star Plot, der mir eine visuelle Zusammenfassung der wichtigsten Merkmale verschiedener Fahrzeugmarken bietet. Dadurch kann ich schnell erkennen, wie sich die Marken in Bezug auf verschiedene Parameter unterscheiden.

Allgemeine Anfoderungen an die Oberfläche:

1. Als Anwender möchte ich eine ansprechende und intuitive Oberfläche erleben, die meine Navigation durch die Gebrauchtwagenanzeigen erleichtert. Die Visualisierungen sollten informativ und leicht verständlich sein, um mir eine effiziente Entscheidungsfindung zu ermöglichen.

2. Als Nutzer erwarte ich, dass die Visualisierungen eine hohe Benutzerfreundlichkeit bieten, einschließlich interaktiver Elemente wie Filteroptionen und Hover-Funktionen. Dadurch

kann ich gezielt nach meinen Anforderungen suchen und zusätzliche Details zu einzelnen Fahrzeugen abrufen.

3. Als Anwender möchte ich die Möglichkeit haben, die Visualisierungen individuell anzupassen, indem ich verschiedene Parameter auswähle oder abwähle. Dadurch kann ich meine Suche und Analyse personalisieren und mich auf die für mich relevanten Informationen konzentrieren.

3.3 Präsentation der Visualisierungen

Präsentieren sie die visuelle Abbildungen und Kodierungen der Daten und Interaktionsmöglichkeiten. Sie müssen begründen, warum und wie gut ihre Designentscheidungen die erstellten Anforderungen erfüllen. Weiterhin müssen sie begründen, warum die gewählte visuelle Kodierung der Daten für das zulösenden Problem passend ist. Typische Argumente würden hier auf Wahrnehmungsprinzipien und Theorie über Informationsvisualisierung verweisen. Die besten Begründungen diskutieren explizit die konkrete Auswahl der Visualisierungen im Kontext von mehreren verschiedenen Alternativen. Machen sie hier nicht den Fehler, einfach nur Visualisierung aus den vorgegebenen Bereichen zu diskutieren, weil das in der Regel nicht sinnvoll ist. Wenn sie sich für einen Scatterplot entschieden haben, ist ein Zeitreihendiagramm in der Regel keine Alternative. Diskutieren sie also nicht einfach Zeitreihendiagramme, weil sie in den Anforderungenen an das Projekt neben Scatterplots stehen, sondern suchen sie nach echten alternativen Visualisierungen, die zum Aufbau eines vergleichbaren mentalen Modells führen. Diskutieren sie die Expressivität und die Effektivität der einzelnen Visualisierungen.

Die eben beschriebenen Präsentationen und Begründungen sollen für jede der drei folgenden Visualisierungen durchgeführt werden.

3.3.1 Visualisierung Eins

3.3.2 Visualisierung Zwei

3.3.3 Visualisierung Drei

3.4 Interaktion

Die präsentierten Visualisierungstechniken müssen interaktiv zu einer Anwendung verknüpft werden. Die Interaktion mit einer Visualisierung soll in den anderen Visualisierungen zu einer Änderung führen. Erklären sie die möglichen Interaktionen mit den einzelnen Visualisierungen und die möglichen Verknüpfungen zwischen ihnen. Begründen Sie warum die konkreten Interaktionen umgesetzt wurden und welche Zwecke für die Anwenderinnen mit ihnen unterstützt werden. Begründen sie ebenfalls warum sie andere Interaktionsmöglichkeiten nicht umgesetzt haben. Wenn sie keine der geforderten Interaktionen umsetzen, erhalten Sie im gesamten Projekt deutlichen Punktabzug.

4 Implementierung

Beschreiben Sie die Implementierung ihrer Visualisierungsanwendung in Elm. Stellen die Gliederung ihres Quellcodes vor. Haben Sie verschiedene Elm-Module erstellt. Was war aufwändig umzusetzen, was ließ sich mit dem vorhanden Code aus den Übungen relativ einfach umsetzen?

Wie sieht die Elm-Datenstruktur für das Model aus, in dem die verschiedenen Zustände der Interaktion gespeichert werden können.

5 Anwendungsfälle

Präsentieren sie für jede der drei Visualisierungen einen sinnvollen Anwendungsfall in dem ein bestimmter Fakt, ein Muster oder die Abwesenheit eines Musters visuell festgestellt wird. Begründen sie warum dieser Anwendungsfall wichtig für die Zielgruppe der Anwenderinnen ist. Diskutieren sie weiterhin, ob die oben beschriebene Information auch mit anderen Visualisierungstechniken hätte gefunden werden können. Falls dies möglich wäre, vergleichen sie die den Aufwand und die Schwierigkeiten ihres Ansatzes und der Alternativen.

5.1 Anwendung Visualisierung Eins

5.2 Anwendung Visualisierung Zwei

5.3 Anwendung Visualisierung Drei

6 Verwandte Arbeiten

Führen sie eine kurze Literatursuche in der wissenschaftlichen Literatur zu Informationsvisualisierung und Visual Analytics nach ähnlichen Anwendungen durch. Diskutieren sie mindestens zwei Artikel. Stellen sie Gemeinsamkeiten und Unterschiede dar.

7 Zusammenfassung und Ausblick

Fassen sie die Beiträge ihre Visualisierungsanwendung zusammen. Wo bietet sie für die Personen der Zielgruppe einen echten Mehrwert.

Was wären mögliche sinnvolle Erweiterungen, entweder auf der Ebene der Visualisierungen und/oder auf der Datenebene?

Anhang: Git-Historie