Übung: Strukturierte und unstrukturierte Daten erkunden

Am Beispiel der Tagesschau-Daten

Prof. Dr. Nicolas Meseth

Nachrichtendatensätze enthalten eine interessante Mischung: Neben klar strukturierten Variablen wie Datum, Ressort oder Wortanzahl gibt es vollständig unstrukturierte Variablen wie Artikeltexte und Überschriften. Diese Übung führt euch durch eine systematische Erkundung dieses Datensatzes von der ersten Orientierung über die Analyse einzelner Variablen bis hin zu einfachen Methoden, wie man auf unstrukturierten Text eine Struktur projiziert.

Schritt 1: Projekt aufsetzen und Daten einlesen

1. Erstellt ein neues Projekt für diese Übung und öffnet es in unserer Entwicklungsumgebung Positron. Stellt sicher, dass ihr eine *virtuelle R-Umgebung* für das Projekt erstellt und aktiviert habt.

2. Installiert die folgenden Pakete in eurer virtuellen Umgebung: `tidyverse`, `janitor` und `skimr`. Verwendet das Metapaket `pacman`, damit ihr die Pakete mit einem einzigen Befehl installieren und laden könnt.

3. Ladet den Datensatz `tagesschau.zip` auf euren Computer herunter und speichert ihn in einem Unterordner `data/` in eurem Projektverzeichnis.

4. Erzeugt einen neuen Ordner `scripts/` in eurem Projektverzeichnis und erstellt eine neue R-Skriptdatei `explore_news.R` in diesem Ordner.

5. Der Datensatz liegt als komprimierte ZIP-Datei vor. Schaut euch die Datei genauer an und prüft, wie das Format aussieht (z. B. Trennzeichen, Zeichenkodierung). Lest die Daten dann mit einer geeigneten Funktion aus `readr` ein und speichert den Code in eurer

Skriptdatei.

Schritt 2: Erster Blick auf den Datensatz

6. Recherchiert den Hintergrund der Daten: Woher stammen sie, wie wurden sie erhoben, und welchen Zeitraum decken sie ab? Was bedeuten die einzelnen Variablen? Haltet eure Erkenntnisse stichpunktartig fest.

7. Verschafft euch mit R einen ersten Überblick. Nutzt dafür mindestens zwei verschiedene Funktionen (z. B. `glimpse()`, `skim()`, `head()`). Beantwortet dabei: Wie viele Zeilen und Spalten hat der Datensatz? Welche Datentypen wurden beim Einlesen automatisch zugewiesen?

8. Jede Beobachtung in einem Datensatz sollte eindeutig identifizierbar sein. Welche Variable könnte hier als eindeutiger Identifikator dienen? Überprüft eure Vermutung mit R, z. B. mithilfe von `n_distinct()` oder `get_dupes()` aus dem `janitor`-Paket. Was macht ihr, wenn keine bestehende Variable geeignet ist?

Schritt 3: Strukturierte Variablen - Skalenniveaus und Datentypen

In diesem Abschnitt konzentriert ihr euch auf die Variablen, die klare Messeigenschaften besitzen.

9. Betrachtet die folgenden Variablen: `ressort`, `tag`, `language`, `word_count`, `date_time`. Ordnet jeder Variable ein Skalenniveau zu (nominal, ordinal, intervall oder verhältnisskaliert). Begründet eure Entscheidung in einem kurzen Satz pro Variable.

10. Prüft, ob die von R automatisch zugewiesenen Datentypen zu den Skalenniveaus passen. Bei welchen Variablen gibt es eine Abweichung? Welche Variablen sollten als Faktor (`factor`) gespeichert werden, und warum?

11. Wandelt `ressort`, `tag` und `language` in Faktoren um. Nutzt dafür `mutate()` in Kombination mit `as_factor()` oder `factor()`. Schaut euch anschließend mit `levels()` an, welche Ausprägungen jeweils vorhanden sind. Wie viele Ressorts gibt es insgesamt?

12. Die Variable `tag` enthält viele fehlende Werte (`NA`). Ermittelt den Anteil der fehlenden Werte in Prozent. Was könnte ein fehlender `tag`-Wert inhaltlich bedeuten? Ist er wirklich "unbekannt", oder könnte er auch "kein Tag vergeben" bedeuten? Wie würdet ihr damit umgehen?

Schritt 4: Wertebereich und deskriptive Statistiken

13. Welchen Zeitraum deckt der Datensatz ab? Ermittelt das früheste und das späteste Datum in `date_time`. Extrahiert außerdem das Erscheinungsjahr mit `lubridate::year()` und erstellt eine neue Variable `year`. Wie hat sich die Anzahl der Artikel pro Jahr entwickelt?

14. Berechnet für `word_count` passende deskriptive Kenngrößen. Welche Kenngrößen sind für eine verhältnisskalierte Variable sinnvoll? Berechnet Minimum, Maximum, Median, Mittelwert und Standardabweichung. Vergleicht Median und Mittelwert - was sagt euch der Unterschied beider Werte über die Verteilung?

15. Erstellt für `word_count` ein einfaches Histogramm mit `ggplot2`. Wählt eine sinnvolle Anzahl an Balken (`bins`). Beschreibt die Verteilung: Ist sie symmetrisch, linksschief oder rechtsschief? Gibt es Ausreißer?

16. Für nominalskalierte Variablen ist der Modus die geeignete Maßzahl. Ermittelt die häufigste Ausprägung (*Modus*) für `ressort` und für `tag`. Stellt die Häufigkeitsverteilung der Ressorts anschließend als Balkendiagramm dar. Sortiert die Balken nach Häufigkeit.

17. Berechnet die mittlere Wortanzahl getrennt nach `ressort`. Welches Ressort produziert im Schnitt die längsten Artikel? Stellt das Ergebnis als horizontales Balkendiagramm dar.

Schritt 5: Unstrukturierte Variablen - Texte erkunden

18. Betrachtet die Variablen `title`, `shorttext` und `text`. Was unterscheidet sie von den bisher analysierten Variablen? Warum lassen sich auf diese Variablen keine klassischen deskriptiven Statistiken wie Mittelwert oder Median anwenden?

19. Eine einfache Form von Struktur ist die Länge eines Textes. Berechnet mit `str_length()` die Zeichenlänge der Variablen `title` und speichert sie als neue Variable `title_length`. Berechnet dann Minimum, Median und Maximum. Was ist der kürzeste und was der längste Titel im Datensatz?

20. Nutzt `str_count()` mit einem Leerzeichen als Muster (`" "`), um die Anzahl der Wörter in `title` zu schätzen. Speichert das Ergebnis als `title_word_count`. Wie viele Wörter hat ein typischer Tagesschau-Titel im Median?

21. Texte lassen sich auch durch das Suchen nach bestimmten Mustern (Schlüsselwörter) strukturieren. Erstellt mit `str_detect()` eine logische Variable `mentions_ukraine`, die angibt, ob der Titel das Wort "Ukraine" enthält. Wie viele Artikel erwähnen die Ukraine im Titel? In welchem Jahr gab es die meisten solchen Artikel? Wie ist die zeitliche Entwicklung?

22. Untersucht auf dieselbe Weise ein zweites Thema eurer Wahl (z. B. "Klimawandel", "Corona", "USA", "Bundestagswahl", "Iran"). Zählt die Treffer pro Jahr und stellt die zeitliche Entwicklung als Liniendiagramm dar. Wie könnt ihr beide Themen in einem Diagramm vergleichen? Welche Erkenntnisse könnt ihr daraus ziehen?

Schritt 6: Reflexion

23. Fasst den Unterschied zwischen strukturierten und unstrukturierten Variablen in eigenen Worten zusammen. Welche Analyseschritte waren für beide Variablentypen möglich, und welche nur für strukturierte?

24. Welche Einschränkungen hat die einfache Textanalyse mit `str_detect()` im Vergleich zu einer vollwertigen inhaltsanalytischen Auswertung? Nennt mindestens zwei Probleme, die bei der Schlüsselwortsuche auftreten können.

25. Betrachtet die Variable `ressort`. Sie wurde beim Einlesen als `character` gespeichert, aber ihr habt sie in einen Faktor umgewandelt. Welchen praktischen Unterschied macht das bei der Analyse und Visualisierung in R?

26. Der Datensatz enthält Artikel aus fast zwei Jahrzehnten. Welche Probleme könnten bei einem Vergleich von Artikeln aus dem Jahr 2006 mit solchen aus dem Jahr 2024 entstehen? Denkt dabei auch an mögliche Veränderungen im Datensatz oder der Datenquelle selbst.