Big Data Analytics | Syllabus

Syllabus

Prof. Dr. Nicolas Meseth

18. Februar 2025

Termine

Sitzungen

Die verpflichtenden Sitzungen finden im **Raum OT1-2** statt. Die Adresse lautet: Oldenburger Landstraße, 49090 Osnabrück. Eine virtuelle Teilnahme via Zoom wird ermöglicht. Die Sitzungen werden aufgezeichnet und etwa eine Woche später bereitgestellt.

- Montag 07.04.2025 von 09:45 - 13 Uhr
- Montag 28.04.2025 von 09:45 - 13 Uhr
- Montag 05.05.2025 von 09:45 - 13 Uhr
- Montag 12.05.2025 von 09:45 - 13 Uhr
- Montag 19.05.2025 von 09:45 - 13 Uhr
- Montag 26.05.2025 von 09:45 - 13 Uhr

Online-Sprechstunde

Zusätzlich findet im Semester regelmäßig freitags von 09:30 - 11 Uhr eine Online-Sprechstunde ebenfalls via Zoom statt. Diese Sprechstunde dient der Klärung von Fragen zu den Seminarinhalten. Eine Anmeldung ist nicht erforderlich. Bitte beachten Sie, dass durch den aktivierten Wartebereich in Zoom mit Wartezeiten zu rechnen ist – Studierende werden in der Reihenfolge ihres Eintretens eingelassen.

Lernziele

Nach erfolgreicher Teilnahme am Seminar sind die Studierenden in der Lage, selbstständig Datenanalysen mit der Programmiersprache R auf beliebigen neuen Datensätzen durchzuführen. Sie können ihre Analyseergebnisse visuell darstellen und in einem Bericht auf den Punkt zusammenfassen:

- Die Studierenden können deskriptive und explorative Fragestellungen formulieren und mithilfe von R geeignete Analysen durchführen.
- Die Studierenden kennen passende Visualisierungsmethoden für verschiedene analytische Fragestellungen und setzen diese kompetent ein.
- Die Studierenden strukturieren ihre Analyseergebnisse, bereiten sie in einer pointierten Form auf und präsentieren sie zielgruppengerecht.
- Die Studierenden kennen verschiedene Arten von Daten und können diese anhand wichtiger Dimensionen unterscheiden. Dabei liegt der Fokus auf dem Umgang mit unterschiedlichen Datentypen und Skalen sowie der Unterscheidung zwischen strukturierten und unstrukturierten Daten.
- Die Studierenden kennen verschiedene Arten analytischer Fragestellungen und unterscheiden diese. Sie verstehen die Möglichkeiten und Grenzen dieser Fragestellungen sowie der zugehörigen Erhebungs- und Analyseverfahren und können sie anhand von Fallbeispielen bewerten.
- Die Studierenden verstehen probabilistische Modelle und können diese im Kontext von *Natural Language Processing* (NLP) zur Textanalyse *anwenden*.

Nicht Gegenstand des Moduls ist:

- Statistischen Methoden oder die statistische Absicherung von Analyseergebnissen (schließende Analysen / *inferential analysis*).
- Das *Training* von ML-Modellen.

Didaktisches Konzept

In diesem Seminar steht die **eigenständige Auseinandersetzung** der Studierenden mit den Methoden der explorativen Datenanalyse im Vordergrund. Die Studierenden lernen verschiedene Methoden kennen, die sie überwiegend durch Selbstlernmaterial anwenden und vertiefen. Das Lernmaterial umfasst Skripte mit Code-Beispielen, kurze Erklärvideos, schriftliche Anleitungen (Tutorials) und Übungsaufgaben – teilweise mit Musterlösungen.

Die Studierenden nutzen das bereitgestellte Material außerhalb der gemeinsamen hybriden Sitzungen zur Vor- und Nachbereitung. In den Sitzungen stellen die die Studierenden ihre

Ergebnisse zu den Übungsaufgaben vor und diskutieren offene Fragen. Zusätzlich hält der Dozent in ausgewählten Sitzungen kurze Impulsvorträge und demonstriert neue Methoden auf dem Bildschirm.

Werkzeuge

Im Rahmen des Seminars arbeiten wir intensiv mit R und dem RStudio. Für die Anwendungs von ML-Modellen verwenden wir Python. Für die Bereitstellung der Code-Beispiele setzen wir Git ein.

Prüfungsleistung

Prüfungsformat

Die Prüfung besteht aus der **Bearbeitung einer Fallstudie**. In Gruppen von 3-4 Personen wenden die Studierenden die gelernten Methoden an einem praktischen Fall innerhalb eines Zeitraums von 1-2 Wochen an.

Die Lösung wird als R-Code und zusammenfassender Bericht in Quarto eingereicht. Jede Gruppe präsentiert ihre Ergebnisse entweder in einem individuellen oder gemeinsamen Termin vor den Prüfenden. Nach der 15-minütigen Präsentation findet ein Fachgespräch mit den Gruppenmitgliedern statt.

Die Fallstudie besteht aus zwei Teilen:

1. Der erste Teil umfasst geschlossene Fragestellungen, die mithilfe geeigneter Analysen zu beantworten sind. Hier wird die Fähigkeit geprüft, die erlernten Methoden zur Beantwortung konkreter Fragen anzuwenden.
2. Der zweite Teil beinhaltet offene Fragestellungen. Die Gruppen entwickeln eigene Lösungsansätze und setzen diese mit den erlernten Methoden um. Dieser Teil prüft die Fähigkeit, komplexe Probleme zu strukturieren und eigenständige Lösungsstrategien zu entwickeln und umzusetzen.

Der Datensatz der Fallstudie wird erst bei Prüfungsbeginn bekannt gegeben.

Abgabe

Der Abgabemodus wird im Laufe des Semester bekanntgegeben.

Bewertung

Folgt bald.

Notwendige Vorkenntnisse

Es werden keine Vorkenntnisse für den Besuch dieses Seminars vorausgesetzt. Kenntnisse im Umgang mit R und dem Tidyverse sind hilfreich, ebenso wie Python-Kenntnisse. Grundsätzlich fällt das Seminar Studierenden mit Programmiererfahrung leichter.

Materialien

Die folgenden Materialien sind für dieses Seminar von Bedeutung. Während des Semesters können weitere Materialien hinzukommen.

Skript

Der Link zum Skript folgt bald.

Slides

- Exploratory Data Analysis with R and Python
- Searching and Transforming Text with R and stringr
- Rule-based Text Classification with R
- NLP - Text Representation

Weitere Slides folgen.

Codebeispiele

Alle Codebeispiele sind in diesem GitHub-Repository gespeichert:

`https://github.com/winf-hsos/data-analytics-code`

Literaturempfehlungen

- Wickham, Hadley, et al. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 2nd edition, O'Reilly Media, Inc, 2023. Online: https://r4ds.hadley.nz/
- Wilke, C. *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. First edition, O'Reilly Media, 2019. Online: https://clauswilke.com/dataviz/
- Huntington-Klein, Nick. *The Effect: An Introduction to Research Design and Causality*. CRC Press, Taylor & Francis Group, 2022. Online: https://theeffectbook.net/
- Ananthaswamy, Anil. *Why Machines Learn: The Elegant Math behind Modern AI*. Dutton, 2024.