Projektpraktikum Wintersemester 2024/25

Web Science

Prof. Dr. Matthias Thimm, Isabelle Kuhlmann
Artificial Intelligence Group, FernUniversität in Hagen

1 Überblick

Web Science ist eine interdisziplinäre Forschungsrichtung zur Analyse von sozio-technischen Systemen, insbesondere dem World Wide Web. Forschungsfragen in Web Science beziehen sich oft auf den Zusammenhang von mikroskopischen Aktionen (wie Freundschaftsanfragen in sozialen Netzwerken) und makroskopischen Beobachtungen (wie Netzwerkstrukturen von sozialen Netzwerken). Die wissenschaftliche Methodologie benutzt dabei Methoden der sozialen Netzwerkanalyse, der statistische Analyse, sowie der Data Science. In diesem Projektpraktikum werden insbesondere praktische Erfahrungen mit maschinellen Lernmethoden erprobt, die zuvor im Kurs „Einführung in Maschinelles Lernen“ gelehrt wurden.

2 Themen

Das konkrete Ziel dieses Projektpraktikums ist die praktische Anwendung von Methoden des maschinellen Lernens auf einen konkret vorgegebenen Datensatz aus dem Kontext der Web Science. Dies beinhaltet insbesondere die folgenden Teilaufgaben:

1. Analyse des vorgegeben Datensatzes und Identifizierung einer geeigneten Problemstellung, die mit Methoden des maschinellen Lernens angegangen werden kann.
2. Aufbereitung und Vorverarbeitung des Datensatzes.
3. Anwendung von *drei* klassischen Methoden des maschinellen Lernens (aus Kapitel 2 und 3 des Kurses „Einführung in Maschinelles Lernen“) zur Beantwortung der Problemstellung.
4. Anwendung von wenigstens *einem Deep-Learning*-Ansatz auf die Problemstellung. Dies beinhaltet eine entsprechende Literaturrecherche und Diskussion im Abschlussbericht.
5. Interpretation und Diskussion der bisher erzielten Ergebnisse. Aufbauend darauf soll *eine* neue Idee für einen passenden Ansatz (beispielsweise eine neue Architektur für ein neuronales Netzwerk) konzipiert, implementiert und angewendet werden.

Benutzen Sie bei der Bearbeitung den aus der Veranstaltung „Einführung in Data Science“ bekannten *Data Science Life Cycle*.

Die in diesem Projektpraktikum zur Verfügung stehenden Datensätze (=Themen) sind wie folgt:

1. Stimmungsanalyse mit Twitter
   `https://www.kaggle.com/datasets/jp797498e/twitter-entity-sentiment-analysis`
2. Erkennung von *Hate Speech* mit Twitter
   `https://www.kaggle.com/datasets/arkhoshghalb/twitter-sentiment-analysis-hatred-speech`
3. Qualität von Wikipedia-Artikeln
   `https://www.kaggle.com/datasets/urbanbricks/wikipedia-promotional-articles`
4. COVID-19-Risikoerkennung
   `https://www.kaggle.com/datasets/meirnizri/covid19-dataset`

Eine Hinzunahme weiterer für die gewählte Problemstellung passender Datensätze ist ausdrücklich erwünscht. Bitte beachten Sie dabei evtl. vorhandene Nutzungsrechte.

3 Weitere Informationen

Es folgen weitere Informationen zum Inhalt und Ablauf des Projektpraktikums, die genauen Daten finden Sie weiter unten:

- Eine Einführung in die Thematik des Projektpraktikums und seines Ablaufs wird in einem Online-Meeting gegeben.
- Machen Sie sich frühzeitig mit den zur Verfügung stehenden Datensätzen/Themen (siehe oben) vertraut. Falls Sie Präferenzen zu Themen haben, so teilen Sie uns diese vor der Einführungsveranstaltung mit.
- Die Studierenden werden nach der Einführungsveranstaltung in Gruppen von jeweils 5 Personen eingeteilt, wobei zuvor abgegebene Präferenzen zu den Themen bestmöglich beachtet werden. Jede Gruppe bearbeitet eines der oben genannten Themen unabhängig von den anderen Gruppen.
- Die Teilnehmenden organisieren die interne Gruppenarbeit selbst, hier noch einige Hinweise dazu:
  - Jede Gruppe benennt eine(n) Teilnehmende(n) als Moderator(in)/Leiter(in) der Gruppe. Diese Personen kümmern sich um die interne Organisation und sind der/die erste Ansprechpartner(in). Es wird empfohlen, für die interne Organisation auch eine entsprechende Kommunikationsplattform (z. B. Slack) zu nutzen.
  - Es wird keine spezifische Methodologie oder Programmiersprache festgelegt; entwickelte Softwareartefakte sollten aber nach entsprechenden Richtlinien der gewählten Methodologie/Programmiersprache konzipiert, implementiert und dokumentiert werden.
  - Nutzen Sie bitte Git (beispielsweise über GitHub oder eine GitLab-Installation) zur Versionierung von allen innerhalb der Gruppe erstellten Artefakte, insbesondere Softwareartefakte, Präsentationen und Berichte. Geben Sie uns bitte direkt bei der Erstellung einen Lesezugriff auf Ihr Repository und stellen Sie sicher, dass das Repository wenigstens 3 Monate nach Praktikumsende noch lesbar ist.
- In einer Zwischenpräsentation stellen alle Gruppen in einer jeweils 10-minütigen Präsentation ihre bisherige Vorgehensweise und Ergebnisse den anderen Gruppen und der Praktikumsleitung vor. Die Präsentation einer Gruppe sollte von nur einer Person gehalten werden.
- Das Praktikum schließt mit einer finalen Präsentation der Praktikumsteilnehmenden ab. Hier berichten alle Gruppen in einem Vortrag von jeweils 20 Minuten über ihre Ergebnisse. Dieser Vortrag sollte von maximal zwei Vertretern/-innen der Gruppe gehalten werden.
- Zusätzlich reicht jede Gruppe einen Abschlussbericht ein, der die Arbeit der Gruppe dokumentiert und 12 Seiten (ohne Referenzen) nicht überschreiten darf (dies ist eine harte Grenze). Neben der schriftlichen Ausarbeitung sind auch die entwickelten Softwarekomponenten Teil der finalen Abgabe. Benutzen Sie für den Abschlussbericht die in Abschnitt 5 vorgegebene Gliederung.
- Wir empfehlen und ermutigen Sie, uns bei jeglichen Fragen jederzeit via E-Mail anzuschreiben. Bei Bedarf können wir dann auch gerne einen persönlichen (Online-)Termin vereinbaren.

Insbesondere empfehlen wir auch, uns in regelmäßigen Abständen über den Verlauf der Arbeit zu informieren (dies liegt in der Verantwortung des/der gewählten Moderators/-in der jeweiligen Gruppe).

- Für die notwendigen Experimente im Praktikum können Sie Rechnerressourcen vom Lehrgebiet bekommen. Bitte schreiben Sie dazu eine E-Mail an Alexander Zock[1].

Für die Erstellung von Präsentationen und Berichten benutzen Sie bitte die LaTeX-Vorlagen des Lehrgebiets (`https://github.com/aig-hagen/aig-templates`) und beachten die allgemeinen Richtlinien, die unter `http://mthimm.de/teaching/general/guidelines.pdf` verfügbar sind.

4 Wichtige Daten

Alle Veranstaltungen finden online via Zoom[2] statt.

- Abgabe Präferenzen zu Themen (Top-3) via E-Mail[3]: 01.10.2024
- Einführungsveranstaltung: 02.10.2024, 15:00-16:00 Uhr
- Zwischenpräsentation: 17.12.2024, 16:00-17:30 Uhr
- Abschlusspräsentation: 18.03.2025, 14:00-17:00 Uhr
- Abgabe Praktikumsbericht: 31.03.2025

Beachten Sie bitte, dass die Teilnahme sowohl bei der Zwischen- als auch bei der Abschlusspräsentation verpflichtend ist.

5 Gliederung des Abschlussberichts

1. Einleitung
2. Aufgabenverteilung
   - Ein Abschnitt pro Teammitglied
   - Kurze Übersicht, was die Person innerhalb des Praktikums beigetragen hat, insbesondere entwickelter Code, Beitrag zum Abschlussbericht, organisatorischer Beitrag, Beitrag zur Abschlusspräsentation, etc.
3. Teaminterne Organisation
   - Wie wurde innerhalb des Teams kommuniziert?
   - Welche Programmiersprache? Warum?
   - Welche Tools/Techniken wurden verwendet?
   - etc.
4. Datensätze und Problemstellung
   - Kurze Beschreibung der Datensätze
   - Welches Problem soll mithilfe der Datensätze gelöst werden?
5. Ansätze
   - Welche klassischen, *Deep Learning*-, und eigenen Ansätze wurde verwendet?

[1] `alexander.zock@fernuni-hagen.de`
[2] `https://e.feu.de/thimm-zoom`
[3] `matthias.thimm@fernuni-hagen.de`, `isabelle.kuhlmann@fernuni-hagen.de`

- Kurze Beschreibung neuer Techniken und Ideen

6. Experimente
   - Wie ist der Experimentaufbau, welche Evaluationsmetriken betrachten Sie?
   - Beschreibung und Interpretation der Ergebnisse
7. Ausblick
   - Was wurde nicht geschafft bzw. hat nicht funktioniert? Warum?
   - Was kann noch verbessert werden? Wie?
8. Zusammenfassung und Fazit