CAIDAS Workshop

Datenanalyse für die Digital Humanities:
Projekte, Methoden, Einsichten

Zeit: 06.-08. Februar 2023 (Montag-Mittwoch mittag)

Ort: Hörsaal Z6, Zentrales Hörsaalgebäude,
Campus Hubland Süd, 97074 Würzburg

Organisation: Schweitzer, Frank, Prof. Dr. Dr. (ETH Zürich)
Scholtes, Ingo, Prof. Dr. (JMU Würzburg)

Synopsis

Der Workshop bringt 20 Wissenschaftlerinnen und Wissenschaftler aus den Sozial- und Geisteswissenschaften (Geschichte, Theologie, Literatur, Kultur) und den Computer- und Komplexitätswissenschaften (Datenanalyse, maschinelles Lernen, Netzwerkanalyse) zusammen, um die Herausforderungen für die "Digital Humanities" zu diskutieren. Drei unterschiedliche Perspektiven werden thematisiert:

"Projekte" geben einen Überblick über verfügbare Daten, ausgewählte Korpora, die teilweise digitalisiert, transkribiert und klassifiziert wurden (Briefkorrespondenzen, aber auch Sammlungen von Artefakten),

"Methoden" zeigen auf, wie solche Korpora mit Hilfe von Netzwerkwerkzeugen analysiert werden können,

"Einblicke" illustrieren anhand von Beispiele, wie Wissen durch die Anwendung digitaler Methoden generiert wurde, zeigen aber auch die Probleme und Gefahren bei der Interpretation.

Die Forschung der "Digital Humanities" ist zu oft in disziplinären Aktivitäten isoliert. Der Workshop soll einen neuen Diskurs ermöglichen, der über die Grenzen der wissenschaftlicher Disziplinen hinaus reicht.

CAIDAS Workshop
Datenanalyse für die Digital Humanities: Projekte, Methoden, Einsichten

Vorläufiges Programm

Montag, 06 Februar 2023

08:45	**Frank Schweitzer** (Chair of Systems Design, ETH Zürich) *Begrüssung und Eröffnung*
09:00	**Fotis Jannidis** (JMU Würzburg) *Modellierungsprobleme in den Computational Literary Studies* Anhand zweiter Beispiele, der Modellierung von ‚Plot' und der Modellierung von ‚Gewalt' werden typische Probleme bei der formalen Beschreibung von Phänomenen aus dem Bereich der Literatur diskutiert: die Abstraktheit und Kontextabhängigkeit vieler literaturwissenschaftlichen Begriffe macht deren Operationalisierung für die automatische Detektion schwierig. ‚Plot' ist ein solcher Begriff: er die reduziert die Handlung eines längeren Textes auf eine kurze Zusammenfassung, die die wesentlichen Ereignisse enthält. Die Aufgabe ist zur Zeit gleich in mehrfacher Hinsicht eine Herausforderung fürs NLP: Längere Texte wie etwa Romane können aufgrund der Architekturen nicht in einem Stück verarbeitet werden. Das Konzept des Ereignisses ist kaum operationalisierbar. Und ‚Relevanz' ist eine Kategorie, in die viel Welt-, Literatur- und Genrewissen eingeht. Der Vortrag diskutiert Lösungsmöglichkeiten.
09:45	**Malte Vogl** (MPI Wissenschaftsgeschichte, Berlin) *ModelSEN: Modellierung historischer Wissenssystem mit sozio-epistemischen Netzwerken* Das BMBF-geförderte Projekt ModelSEN befasst sich mit der Modellierung von und Methodenentwicklung für historische Wissenssysteme mittels sozio-epistemischer Netzwerke. Ein weiterer wichtiger Fokus ist der Ausbau von Kooperationen sowohl im technischen als auch inhaltlichen Bereich, z.B. über externe Fallstudien. In diesem Vortrag wird das Projekt mit seinen Grundlagen vorgestellt. Anhand von zwei Schwerpunkten werden aktuelle Forschungsergebnisse aus dem Bereich der Scientometrics und dem Agenten-basierten Modellieren dargestellt und Anknüpfungspunkte zu anderen Projekten aufgezeigt. Das Projektziel eines Kompendiums wird mittels zweier weiterer Beispiele verdeutlicht.
10:30	**Kaffeepause**
11:00	**Kaspar Gubler** (Uni Bern) *Repertorium Academicum (REPAC): Digitale Rekonstruktion akademischer Wissens- und Kommunikationsräume im vormodernen Europa* Das REPAC ist ein Forschungsprojekt der Digital History, das am Historischen Institut der Universität Bern betrieben wird. Die Datenbasis des Projekts bilden die Matrikellisten der europäischen Universitäten im Zeitraum von 1250 bis 1550. Die Matrikeln enthalten für gewöhnlich die Namen und Herkunftsorte der Studenten sowie das Datum der Immatrikulation. Diese Ausgangsdaten werden mit biographischen Daten zu studierten Fachrichtungen, beruflichen Tätigkeiten und schriftlichen Werken angereichert. Ziel ist es, eine wissensbasierte prosopographische Grundlage für die Forschung zur Wirkungsgeschichte der Gelehrten im europäischen Raum zu schaffen. Hierzu gehört die Rekonstruktion von Wissens- und Kommunikationsräumen, die mit Karten- und Netzwerkvisualisierungen sowie Zeitreihen auf verschiedenen Ebenen (Personen - Institutionen - Werke) analysiert werden.

11:45	**Stefan Aderhold** (Heidelberger Akademie der Wissenschaften) *„Ein Chamäleon“ oder der „Papst zu Tübingen“? Erkenntnisse und Grenzen der Netzwerkanalyse am Beispiel des Theologenbriefwechsels Jakob Andreaes* Die Forschungsstelle „Theologenbriefwechsel im Südwesten des Reichs in der frühen Neuzeit“ hat es sich zum Ziel gesetzt, das Desiderat von digitalisierten Briefquellen in der an die Reformationszeit anschließenden Konfessionalisierungsphase (etwa 1550-1620) aufzuarbeiten. Hierfür digitalisiert und verschlagwortet sie Briefe von und an 200 Theologen aus drei Territorien, die eine tragende Rolle für die Konfessionsbildung dieser Zeit spielten. Inzwischen verfügt die Forschungsstelle über eine ansehnliche Datenbasis, steht aber mit der Nutzung quantitativer Methoden noch ganz am Anfang. Der Vortrag wird die Möglichkeiten und Grenzen der Netzwerkanalyse des Theologenbriefwechsels anhand des schillernden Briefkorpus des lutherischen Theologen Jakob Andreae ausleuchten und diskutieren.
12:30	**Mittagspause**
14:00	**Holger Meyer** (Universität Rostock) *Graphbasierte Techniken zur Auswertung von Sagen und Legenden* Wer hat wann und wo Hexen- und Werwolfsagen erzählt in Mecklenburg? Gibt es vorherrschende Motive einzelner Erzähler? Was erzählen Frauen, was Männer? Diese und weitere Fragen lassen sich graphbasiert beantworten. Dazu stellen wir Techniken vor, die vorliegende Texte (in XML) auf Graphen (im Property Graph Modell) abbilden und Graph mining-Techniken zuführen. Am Beispiel der Sagensammlung des Wossidlo-Archives, die als digitale Forschungsammlung nutzbar ist, werden Ergebnisse anhand ausgewählter Mining-Techniken wie Summarizing und Clustering gezeigt.
14:45	**Frank Puppe** (JMU Würzburg) *Herleitung und Auswertung von Figurennetzwerken in Märchen mit Szeneneinteilung* Märchen haben einen klaren Aufbau und erzählen in kompakter Form eine in sich abgeschlossene Geschichte. Funktionale Ansätze wie das Schema von Vladimir Propp zu deren Charakterisierung werden allerdings ihrer Vielfältigkeit nicht gerecht. Wir stellen ein einfacheres Schema vor, bei dem Märchen durch Netzwerke der Figuren mit ihren Eigenschaften und Relationen sowie Aufteilung in Szenen entsprechend Verbesserungen bzw. Verschlechterungen des Zustandes des Protagonisten dargestellt werden. Deren semiautomatische Herleitung umfasst Aufgaben wie Named Entity Recognition, Co-Reference Resolution, Analyse von Dialogen, Erkennung von Eigenschaften wie Figurentyp, Altersklasse, Geschlecht, Sentiment, Herkunft, Magische Fähigkeit, usw. sowie Relationen wie Familienbeziehungen und positive und negative Beziehungen, Verwandlungen usw., die sich in verschiedenen Szenen auch ändern können. Wichtig ist auch die Visualisierung und die Erklärbarkeit durch explizite Textreferenzen für alle Charakterisierungen.
15:30	**Kaffeepause**
16:00	**Christian Zingg** (ETH Zürich) *Kommunikationsnetzwerke als Wissensgraphen: Die zeitabhängige Bedeutung von Personen* Wie breiten sich Ideen durch schriftliche Kommunikation in großen Gemeinschaften aus? Die Netzwerktheorie erlaubt uns, Personen und ihre Interaktionen als Netzwerk zu repräsentieren und wichtige Verbreitungsrouten von Ideen anhand von zeitabhängigen Zentralitätsmaßen zu identifizieren. Wir zeigen, wie solche Messwerte zusammen mit anderen personenbezogenen Informationen in einem Wissensgraphen aggregiert werden können, der dann effizient ausgewertet werden kann. Die Erstellung und Analyse eines Wissensgraphen wird am Beispiel des Briefaustausches der frühen Neuzeit illustriert.
16:45	**Tagesabschluß**

Dienstag, 07 Februar 2023

09:00	**Albin Zehe** (JMU Würzburg) *Analyse und Strukturierung von Plot-Elementen* Ein zentrales Ziel der computergestützten Literaturwissenschaften ist die Analyse des „Plots“ einer Geschichte. Da der Begriff Plot sehr abstrakt und schwer zu fassen ist, fokussieren sich aktuelle Techniken häufig auf die Analyse verschiedener Plotelemente. Dieser Vortrag beschäftigt sich mit zwei wesentlichen Aspekten: Der Extraktion und Analyse von Figurennetzwerken und der Strukturierung der Handlung in sinntragende Abschnitte, konkret Szenen. Im ersten Teil wird eine Analyse von Tolkiens Legendarium vorgestellt, wobei Figurennetzwerke automatisch extrahiert und anschließend mittels Graph Neural Networks analysiert werden. Da diese Analyse den Text bisher als statisch behandelt, also Veränderungen im Verlauf der Geschichte ignoriert, werden im zweiten Teil dann Szenen als eine Möglichkeit zur Strukturierung von literarischen Texten eingeführt: Diese ermöglichen eine Unterteilung der Handlung in sinntragende Abschnitte, welche eine gute Grundlage für dynamische Analysen auf den Texten liefern. Dabei werden sowohl die Schwierigkeit einer automatischen Szenenerkennung als auch der aktuelle State-of-the-Art beschrieben.
09:45	**Philiph Ströbel** (Universität Zürich) *Bullinger Digital - Texterkennung in einem reformatorischen Briefwechselkorpus* Automatisierte Handschriftenerkennung fokussierte sich in der Vergangenheit vorwiegend auf individualisierte Trainings- und Erkennungs-Methoden einzelner Handschriften. Typischerweise finden sich in den Geisteswissenschaften jedoch mehrere Hände umfassende Datensätze. Bei der Entwicklung von Erkennalgorithmen sollte man deshalb nicht nur die Fähigkeit nachweisen, einzelne Hände mit hoher Qualität erkennen zu können, sondern auch Wert auf den Umgang mit unterschiedlichen Handschriften legen. Ein frühneuzeitlicher Briefwechsel aus dem 16. Jahrhundert ist ein interessanter Fall, um traditionelle und neuere Ansätze bezüglich ihrer Flexibilität und Adaption zu vergleichen. Wir nutzen dafür Daten des Projekts Bullinger Digital, um Aussagen über Erkennqualität sowie Potenziale in der automatischen Erkennung zu machen.
10:30	**Kaffeepause**
11:00	**Audric Wannaz** (Universität Basel) *Familienbriefe des griechisch-römischen Ägyptens mithilfe von NLP untersuchen: Messungen von Netzwerken und Sprechakten.* Noch steckt die Anwendung von NLP in der Typologie griechischer Papyri in den Kinderschuhen. Dabei bietet dieser computergestützte Ansatz einzigartige quantitative (aber auch qualitative) Ergänzungen zum bisherigen Forschungsbild. Dieser Beitrag wird Zwischenergebnisse diesbezüglich am Beispiel eines spezifischen Korpus, Privatbriefe mit einem besonderen sozialen Fokus (eng. social letters), liefern. Messungen werden vorgestellt, die diese Briefe als Netzwerke oder als Sammlung von schriftlichen Sprechakten betrachten.
11:45	**Elena Suárez Cronauer** (Akademie der Wissenschaften Mainz) *Korrespondenzen der Frühromantik: Überlegungen zu Netzwerken und Methoden* Das DFG-geförderte Projekt „Korrespondenzen der Frühromantik. Edition – Annotation – Netzwerkforschung“ untersucht den Briefwechsel der um 1800 vor allem in Jena und Berlin aktiven Frühromantiker*innen und deren Wirkungskreis. Ein momentan im Aufbau befindlicher und über die Projektlaufzeit hinweg kontinuierlich weiterzuentwickelnder Knowledge Graph stellt die Datengrundlage des Projekts dar. Auf dieser Basis werden Netzwerkmodelle entwickelt und analysiert, um zu neuen Erkenntnissen über Kommunikation, Wissenstransfer und Wissensdiffusion innerhalb des Briefnetzwerks der Frühromantiker*innen zu gelangen sowie über bekannte Forschungsfragen neu nachzudenken. Der Beitrag gibt einen Einblick in die Werkstatt mit ersten Überlegungen zur Netzwerkmethodologie sowie Gedanken zu neuen Forschungsfragen für die Frühromantik durch historische Netzwerkanalyse.
12:30	**Mittagspause**

14:00	**Patrick Andrist** (LMU München) *Prototyp einer Datenbank neuer Generation für das integrierte Studium von Handschriften* Mit dem wachsenden Interesse an der Handschriftenforschung und der rasch zunehmenden Zahl elektronischer Faksimiles steigt auch der Bedarf an einer neuen Generation von Datenbanken, die es ermöglichen, Handschriftendaten auf unterschiedliche und flexible Weise zu visualisieren und zu bearbeiten. Der in München entwickelte Prototyp einer solchen Datenbank ist ein erster Schritt in diese Richtung. In diesem Vortrag werde ich die Prinzipien des neuen Datamodells darlegen, sodann die Funktionsweise der im Kern des Systems stehenden "internal page ID" erklären und schließlich ein darauf gebautes Tool zur "massenhaften Beschreibung" des Inhalts vorstellen.
14:45	**Goran Glavaš** (JMU Würzburg) *Massively Multilingual Natural Language Processing* Multilingual language models (LMs) (e.g., multilingual BERT or XLM-R) have pushed the state of the art in multilingual NLP, yielding robust performance for various NLP tasks for languages with little or no task-specific training data. Multilingual LMs, however, suffer from a phenomenon known as curse of multilinguality: for a fixed model capacity, representations of individual languages deteriorate with inclusion of more languages into pretraining. The quality of text encodings thus varies drastically across languages, correlating highly with the size of the languages' pretraining corpora. I will present work on remedying the curse of multilinguality and improving NLP models for low-resource languages, including the approaches that leverage massively multilingual lexical resources (e.g., BabelNet, PanLex).
15:30	**Kaffeepause**
16:15	**CAIDAS Talk:** Frank Schweitzer (ETH Zürich) *Was können wir von Netzwerkanalysen erwarten?* Die "Digital Humanities" profitieren bereits von den Fortschritten der Informatik bei der Transkription handschriftlicher Dokumente (siehe READ-COOP) oder der Texterkennung von gedruckten Dokumenten. Damit wächst allerdings die Herausforderung, Methoden zur Gesamtanalyse dieser Kopora zu entwickeln. Der Netzwerkforschung basiert auf der Überzeugung, daß Wissen aus der Verknüpfung von Dokumenten anhand ihrer Metainformationen extrahiert werden kann, um das Studium von Einzeldokumenten zu ergänzen. In meinem Vortrag werde ich das Potenzial dieser Methodik anhand von Beispielen veranschaulichen. Wie können wir soziale Netzwerke extrahieren, die Bedeutung von Individuen messen und ihre freundschaftlichen oder gegnerischen Beziehungen erkennen? Wie können wir sinnvolle von zufälligen Interaktionen unterscheiden und Präferenzen für Interaktionen in Netzwerkmodellen berücksichtigen? Der Vortrag behandelt diese Fragen kursorisch, um die Relevanz der Netzwerkmethoden für künftige Forschungsprojekte aufzuzeigen.
17:30	**Tagesabschluß**

Mittwoch, 08 Februar 2023

09:00	**Maximilian Schich** (Tallinn University) *Kunst- und Kulturverstehen: Von Netzwerken zu Bedeutungsräumen und zurück* Ausgehend von Wissensgraphen der 1990er, beschäftigt sich die Kunst- und Kulturwissenschaft seit 20 Jahren mit komplexen Netzwerken. Ältere Ansätze werden dabei systematisch erweitert zur Analyse multidimensionaler Netzwerke von Netzwerken. Doch das Paradigma hat Grenzen. Will man zum Beispiel Bilder vollständig verstehen, so kommen mehr kontinuierliche Bedeutungsräume ins Spiel, wie sie im Deep Learning seit 2013 große Erfolge feiern. Aber, wie diskrete Netzwerke wollen auch diese impliziten Bedeutungsräume selbst untersucht und verstanden werden. Dieser Vortrag stellt eine Methode vor, die es erlaubt verstehbare Bedeutungsräume ästhetischer Artefakte zu erzeugen. Dabei schließt sich der Kreis zu den Netzwerken im Problem der visuellen Projektion.
09:45	**Christof Weiß** (JMU Würzburg) *Computergestützte Musikanalyse: Szenarien, Methoden und Evaluationsprobleme* Die Analyse großer Musikkorpora bietet großes Potential für die Musikwissenschaft. Um hierfür auch Audiodaten erschließbar zu machen, müssen Signalverarbeitungs- und Machine-Learning-Techniken unter Berücksichtigung musiktheoretischer Kenntnisse eingesetzt werden. Deep-Learning-Verfahren stoßen in diesem Bereich schnell an Grenzen: Oft sind Datensätze sind nicht im nötigen Umfang verfügbar, Aufgabenstellungen nur schwammig definiert und Annotationen hochsubjektiv. Dieser Vortrag stellt typische Szenarien und Datensätze vor, gibt einen Einblick in technische Methoden und zeigt Fallstricke bei deren Evaluierung und Bewertung auf. Um diesen Herausforderungen zu begegnen, werden Evaluationsstrategien vorgestellt, die die Verfügbarkeit verschiedener Einspielungen klassischer Werke gezielt ausnutzen (Cross-Version Experimente).
10:30	**Kaffeepause**
11:00	**Ramona Roller** (ETH Zürich) *Netzwerkanalyse zur Komposition und Evolution innerprotestantischer Gruppen im 16. Jahrhundert* Die europäische Reformation war geprägt von innerprotestantischen Konflikten zwischen verschiedenen Ideologien wie Lutheranern, Zwinglianisch-Reformierten und Baptisten. Um diese Konflikte besser zu verstehen, untersuchen wir die Zusammensetzung und Entwicklung ideologischer Gruppen mithilfe von community detection in einem Briefkorrespondenznetzwerk. Wir setzen die Gruppenzugehörigkeit der Reformatoren im Netzwerk in Zusammenhang mit ihrer Biografie und historischen Ereignissen in Europa des 16. Jahrhunderts. Desweiteren erklären wir die Bildung von Gemeinschaften mit soziologischen Gruppenbildungsprozessen. Wir liefern neue Erklärungen für die Entwicklung ideologischer Gruppen, z. B., inwiefern gemäßigte Reformatoren zum Zusammenschluss von ideologisch heterogenen Gruppen beitragen.
11:45	**Ingo Scholtes** (CAIDAS, JMU Würzburg) *Abschluß und Verabschiedung*