Richtlinie zum „Software-Sprint“

StoryWeb – Projekttitel

Schlussbericht

Zuwendungsempfänger:

Friedrich Lindenberg

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS22S42 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Kurze Darstellung der Aufgabenstellung und Motivation

Was war Deine Motivation? Welches Problem wolltest Du mit Deinem Projekt lösen? Wie war die geplante Vorgehensweise zur Problemlösung (auch Angabe der wichtigsten Meilensteine)?

Das Projekt StoryWeb sollte das Wissensmanagement im Journalismus verbessern. Denn neben interessierten Bürgern finden sich auch viele Gruppen, die Nachrichten aus beruflichem Interesse konsumieren: ForscherInnen aus Wirtschaft und Wissenschaft, aber vor allem auch andere JournalistInnen. Sie lesen die Texte vor allem, um bereits etablierte Zusammenhänge nachzuvollziehen.

Im Investigativjournalismus geht es dabei häufig um Akteursnetzwerke: welche Firmen, Verbände und Politikerinnen miteinander in Verbindung stehen, welche Mafiastrukturen über Ländergrenzen hinweg agieren, oder welche Seilschaften und Familienstrukturen die Vergabe von Posten beeinflussen.

Der Artikel als Darstellungsform ist dabei eine Hürde, weil er sich nicht für strukturierte Auswertung eignet. So lassen sich wichtige Fragen nicht beantworten: Welche Verbindungen gibt es zwischen den Akteuren in zwei Geschichten? Welche Mittelsmänner einer bestimmten Person des öffentlichen Interesses - etwa eines russischen Oligarchen - sind zuvor bereits bekannt geworden? Kommt einer Beteiligten eines alten Skandals in einem neuen Leak vor, das einem Rechercheteam nun vorliegt?

Journalistische Berichterstattung als Quelle für strukturierte Firmen- und Personeninformationen zu nutzen, ist keine neue Idee. Viele sogenannte Adverse-Media-Systeme konzentrieren sich jedoch darauf, die in einem Artikel genannten Entitäten mit bestimmten Ereignissen oder Fehlverhalten in Verbindung zu bringen. Ein Artikel über ein Bergbauunternehmen könnte beispielsweise Arbeitsrechtsverletzungen oder die Vertreibung indigener Völker erwähnen und dadurch zu einem negativen ESG-Risiko-Score für das erwähnte Unternehmen führen.

StoryWeb hingegen konzentriert sich auf eine andere Art der Analyse. Anstelle von bewertenden Einschätzungen (Sentiment Analysis) soll das Tool nach Konnektivität von Entitäten suchen. Entitäten im System – Personen, Unternehmen, Orte – sind im Sinne einer festen Taxonomie von

Verbindungstypen miteinander verknüpft: Sie können Familienmitglieder oder Geschäftspartner sein; eine Person könnte ein Unternehmen besitzen oder kontrollieren und so weiter.

Die Idee dahinter ist es auch, das Netzwerk sanktionierter und gesuchter Personen im Schwesterprojekt OpenSanctions auszuweiten: etwa um die zweitrangigen Akteure einzubeziehen, die sanktionierten Figuren dazu dienen, illegale Aktivitäten zu verbergen und zu verschleiern – Consiglieri, Offshore-Unternehmen, Bevollmächtigten - sogar alte Schulfreunde.

Beitrag des Projektes zu den Zielen der Förderinitative „Software-Sprint“

Welche Bezüge gibt es zu den Themenfeldern „Civic Tech“ und „Data Literacy“ des Software Sprints oder zu weiteren gesellschaftlich relevanten Zielen bzw. Lösungsansätzen?

Das Ziel von StoryWeb ist es, für JournalistInnen und andere AnalystInnen neue, digitale Herangehensweisen und Arbeitsformen zu erschließen. Das Projekt soll eine tiefgründigere und nachhaltigere Form der Berichterstattung und der Verwaltung journalistischen Wissens erforschen. Es soll Analysten auch anhalten, die eigenen Arbeitsergebnisse als Quelldaten für zukünftige Recherchen zu nutzen. Damit kann es einen Beitrag zur besseren Berichterstattung, insbesondere bei Cross-Border-Geschichten, leisten.

Ausführliche Darstellung der Ergebnisse

Welche konkreten Ergebnisse hast Du erzielt? Konnten alle Meilensteine erreicht werden? Welche zusätzlichen Erkenntnisse hast Du aus der Projektarbeit gewonnen, auch im Hinblick auf die Begleitung durch die Open Knowledge Foundation?

Noch vor der Projektphase hatte ich eine Zusammenarbeit einer Gruppe von JournalistInnen eingeleitet, die unter dem Namen “Bad Will Hunting”[1] ein sehr ähnliches Projekt durchführen wollten. Drei Medienhäuser hatten geplant, im Rahmen des Journalism AI-Fellowships der LSE gemeinsam eine Software zu entwickeln, die aus Nachrichtentexten Wissensgraphen (“Knowledge Graphs”) extrahieren sollte. Die Zusammenarbeit mit “Bad Will Hunting” stellte sich als hervorragendes Diskussionsforum für die konzeptionelle Verfeinerung der Story Web-Idee heraus. Leider wurde den beteiligten Entwicklern vom Guardian und dem niederländischen Recherche-Team Follow The Money nur sehr wenig Zeit für die Arbeit an dem Projekt eingeräumt, wodurch die Umsetzung von Bad Will Hunting eher im konzeptionellen Stadium verblieb.

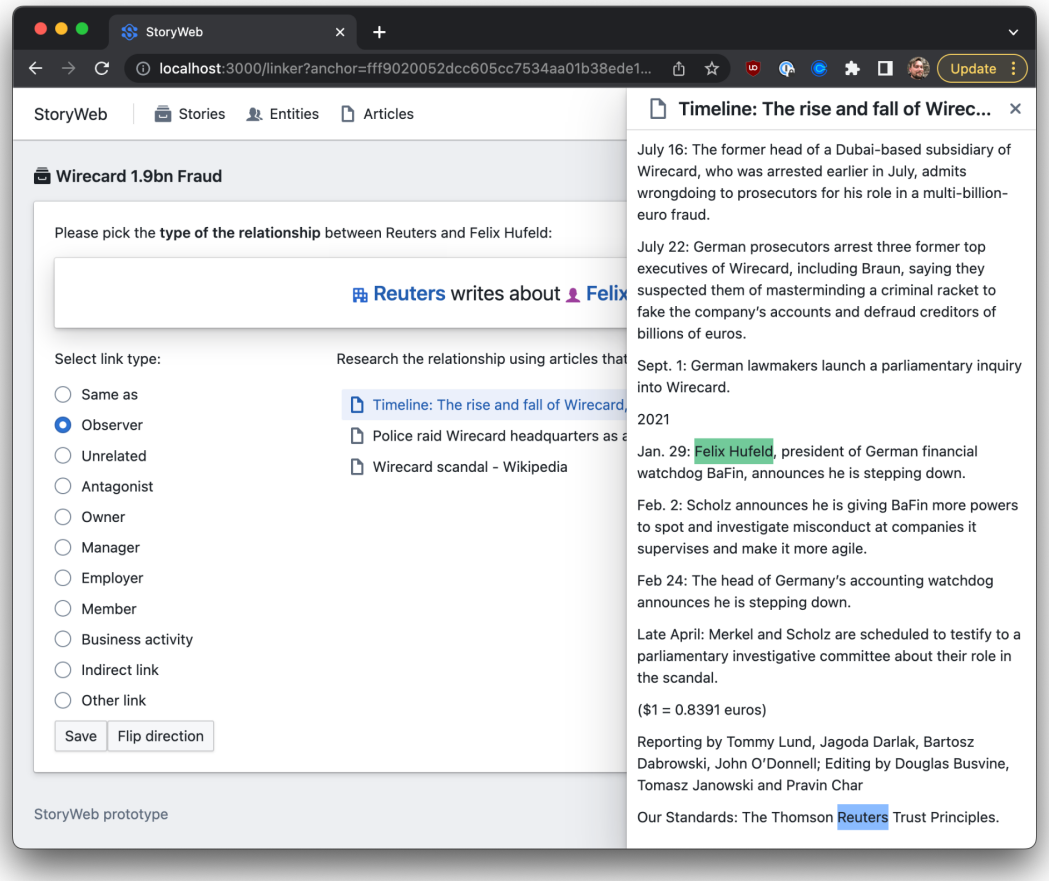
Als ersten Arbeitsschritt nach Projektstart habe ich mich um das Herunterladen und Bereinigen von Nachrichtenartikeln gekümmert (MS1). Dazu habe ich das Softwarepaket mediacrawl implementiert, welches Nachrichtenseiten durchforstet und alle verfügbaren Artikel in einer Datenbank abspeichert. mediacrawl wird durch eine YAML-basierte Konfigurationsdatei gesteuert, die für jede Nachrichtenseite Einstellungen vorgibt, was z.B. die Abrufgeschwindigkeit und die Auswahl von Artikeln, die gespeichert werden sollen, angeht. Mithilfe des “trafilatura”-Toolkits extrahiert mediacrawl dann aus den heruntergeladenen HTML-Seiten die verschiedenen Abschnitte des Artikels, etwa den Text, Titel, Veröffentlichungsdatum und AutorIn. Nach Ende des Crawl- und Extraktions-Vorgangs werden die extrahierten Artikel im articledata-Format gespeichert, einem

[1] https://blogs.lse.ac.uk/polis/2022/09/15/bad-will-hunting-the-story-so-far/

JSON-Datenformat das ich in Abstimmung mit dem “Bad Will Hunting”-Team zur Grundlage für StoryWeb-Importe gemacht hatte. Die mediacrawl-Software enthält zudem eine Beispielkonfiguration, mit der der Abruf von insgesamt acht Nachrichtenarchiven vorkonfiguriert ist (MS2 - von einer Auslieferung der mediacrawl-Ergebnisse, also der Artikelarchive - wurde aus urheberrechtlichen Gründen abgesehen).

Nachdem ich mit dem mediacrawl-Tool etwa 22,000 Medienberichte von spezialisierten Veröffentlichungen (Themen: Korruption, Geldwäsche, Betrug) gesammelt hatte, konnte ich die Entwicklung einer NLP-Verarbeitungskette in Angriff nehmen (MS3). Ich habe dazu ein einfaches SQL-Datenmodell entwickelt, in dem durch die NER-Software spaCy extrahierte Nennungen von Personen, Firmen und Orten in jedem Artikel gespeichert werden. Zunächst hatte ich gehofft, aus Urheberrechtsgründen nur Fragmente der Artikel (z.B. die Sätze, die mehr als eine Entität benennen) in der Datenbank zu speichern. Bei der Entwicklung der Benutzeroberfläche stellte sich dies jedoch als nicht praktikabel heraus: eine ordentliche Analyse der Verbindungsarten zwischen zwei Entitäten war nur mit dem gesamten Artikeltext möglich. Das führt dazu, dass die Veröffentlichung und Weitergabe einer StoryWeb-Datenbank im Gesamten nicht ohne weiteres möglich ist.

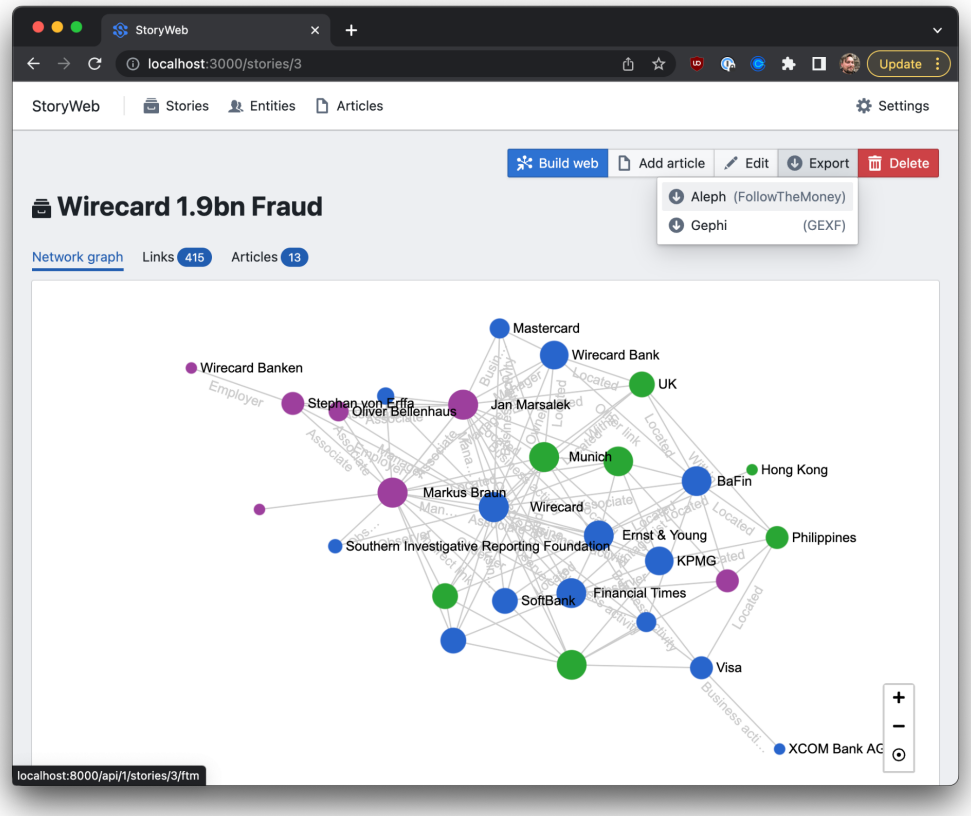
Die so entstandene SQL-Datenbank enthält insgesamt etwa eine halbe Million Nennungen von Entitäten. Durch die Entwicklung von verschiedenen SQL-Anfrage-Strategien (siehe “QUERIES.md”) konnte ich damit eine erste Weboberfläche entwickeln, die gemeinsame Nennungen von Namen darstellt. Im nächsten Schritt wurde das Modell um eine Cluster-Funktionalität erweitert, um verschiedene Nennungen der gleichen Entität in unterschiedlichen Artikeln zu aggregieren.



Zu jedem Cluster können dann ebenfalls in der Weboberfläche vom Nutzer die Beziehungstypen zu anderen Clustern abgefragt werden, die häufig gemeinsam in Artikeln vorkommen (MS4). Dazu habe ich eine Ontologie von Entitäts- und Verbindungstypen entwickelt, die sich gegenseitig bedingen: eine Person kann mit einer anderen Person verwandt sein, aber nicht mit einer Firma oder einem Ort, etc. Neben inhaltlich relevanten Beziehungstypen (Familie, Beschäftigung, Eigentum, Zusammenarbeit oder Konflikt zwischen Personen und Firmen) verfügt die Ontologie auch über Einträge, die sich eher auf die journalistische Struktur der Inhalte beziehen: der Satz “Reuters konnte die Firma X nicht für eine Stellungnahme erreichen”, etwa, stellt eine textliche Meta-Beziehung dar, die im Datenmodell als “observer” dokumentiert wird: Reuters schreibt Texte über X. Ebenso verfügt das Modell über die speziellen Beziehungstypen “same as” (President Trump = Donald Trump) und “unrelated” (Donald Trump und Nelson Mandela).

Die so entstandene Software enthielt ein relativ vollständiges Datenmodell, in dem Entitäten (bzw. Cluster) und deren Verbindungstypen modelliert werden konnten - es fehlte jedoch eine sinnvolle Benutzerführung mit einem Anfang und einem Ergebnis. Aus diesem Grund habe ich schließlich Ende November eine neue Abstraktion eingeführt: eine “Story”. Stories stellen einen thematischen Komplex dar: etwa einen Skandal, oder eine größere Geschichte mit Nachrichtenwert. Jeder Story können dann verschiedene Artikel zugeordnet werden, die diese im Teil oder als Ganzes beschreiben. Dadurch werden die in StoryWeb enthaltenen Artikel in handhabbare Einheiten aufgespalten. Innerhalb dieser Einheiten können dann gezielter die Verbindungstypen der erwähnten Akteure abgefragt werden. Eine “Story” zum Wirecard-Skandal, etwa, die insgesamt 17 Artikel umfasst, lässt dann schließen: 16 der 17 Artikel erwähnen die Wirecard-Firma zusammen mit einem ihrer ehemaligen Top-Manager - und dies sind die ersten Verbindungen, die abgefragt werden.

Schließlich stellen solche Stories dann auch einen Kontext dar, innerhalb dessen eine Visualisierung der Daten Sinn ergibt. Mithilfe des Visualisierungs-Tools Sigma/Graphology habe ich deshalb eine Netzwerkdarstellung der Akteure innerhalb einer Story umgesetzt (MS6).



Neben der Netzwerkdarstellung besteht auch die Option, die Daten in zwei unterschiedliche Formate zu exportieren: GEXF (Graph Exchange Format) wird von verschiedenen Netzwerkanalyse-Softwares wie Gephi und NetworkX unterstützt, während das FollowTheMoney-Format die Weiternutzung in der Recherche-Software Aleph und den Abgleich mit OpenSanctions-Daten ermöglicht.

Auf der Basis dieser Webanwendung bestand nun die Möglichkeit, das System mit konkreten Geschichten zu erproben. Dafür habe ich mehrere Artikel-Sammlungen zu bestimmten Themenkomplexen im StoryWeb-Repository bereitgestellt (siehe “STORIES.md”) und insgesamt ca. 1100 Entitätenlinks manuell bestimmt. Daraus ergibt sich auf der Basis des Artikelarchivs ein Trainingskorpus von ca. 5100 Sätzen, in denen zwei Entitäten erwähnt werden, deren Beziehungsart durch manuelle Klassifizierung bekannt ist. Auf Basis dieser Trainingsdaten habe ich einige Konzepte zur automatischen Klassifikation von Beziehungen erprobt - etwa ein Bayes-Classifier und eine einfache token-basierte Regelengine - jedoch ohne einen davon letztendlich in der StoryWeb-Software vollständig zu implementieren.

Eine weitere Unterstützung bei der Entwicklung einer automatischen Klassifizierung sind zudem einfach Heuristiken, z.B.: Wenn ein Entity, etwa “Reuters”, im Großteil aller Fälle nur einen Verbindungstypen eingeht (“observer”, so.), dann steigt die Wahrscheinlichkeit, dass eine neue Beziehung dieses Entities den gleichen Typ besitzt.

Zielgruppe, Nutzen und mögliche Weiterentwicklungen

Welcher Nutzen ergibt sich für die Zielgruppe aus den Ergebnissen Deines Projekts? Welche weitergehenden Effekte ergeben sich aus der Open-Source-Stellung der Ergebnisse? Gibt es Ideen für die Weiterentwicklung Deiner Lösung und Pläne zu deren Umsetzung?

Mit StoryWeb können JournalistInnen und andere Recherchierende in einem grafischen Prozess eine Reihe von Artikeln auswerten, ein Netzwerk generieren und dieses dann in geeigneten Formaten in anderen Tools weiter Analysieren (z.B. Aleph, Gephi). Daneben sind durch das Projekt zwei weitere Open Source-Programme entstanden (mediacrawl, articledata), die zur Weiternutzung geeignet sind.

Die StoryWeb-Software steht nach dem Projektende der Öffentlichkeit auf GitHub zur Verfügung - es besteht die Möglichkeit einer einfachen Installation via Docker-Image.

StoryWeb hat leider erst gegen Ende der Projektlaufzeit eine ausreichende Bedienreife erreicht, um von nicht-technischen Nutzern produktiv eingesetzt werden zu können. Die Erprobung mit unerfahrenen Usern wird deshalb wohl erst nach dem Projektzeitraum richtig beginnen - und damit auch das Sammeln von Ideen für die nächste Version.

Klar ist schon jetzt: die weitere Automatisierung des Verbindungserkennung wird eine große Rolle spielen (evtl. via ChatGPT?), und auch die Reihenfolge, in der dem Nutzer mögliche Verbindungen vorgeschlagen werden, kann noch erheblich optimiert werden.

Besonders spannend finde ich die Idee, StoryWeb als Werkzeug auch auf Datensätze anzuwenden, die nicht aus medialer Berichterstattung stammen. Zum Beispiel könnte die Software von JournalistInnen bei der Arbeit mit einer gerichtlichen Ermittlungsakte genutzt werden, um die Beteiligten des Falls und deren Beziehungen herauszuarbeiten.

Hat die Arbeit in dem Projekt Dich in Deiner persönlichen, fachlichen Weiterentwicklung unterstützt?

Besonders spannend fand ich die Auseinandersetzung mit der Disambiguierung und Identifizierung von nur namentlich identifizierten Entitäten: Ist Wladimir Putin im Artikel A die gleiche Person, um die es auch im Artikel B geht? Ist dies wiederum die gleiche Person, die vom Wikidata Item Q7747 beschrieben wird?

Das Problem von “Entity Linking” gilt in der Literatur selbst dann als schwer, wenn eine komplette Liste aller gesuchten Entitäten vorliegt - bei StoryWeb ist es aber offensichtlich das Ziel, neue, bisher unbekannte Entitäten durch die Textanalyse zu beschreiben. Anders als beim Projekt OpenSanctions muss der Prozess dabei lediglich mit Namen und Kontextinformationen stattfinden. Die Literatur- und Code-Recherche zu dem Thema hat mir einen neuen Respekt für das Problem beschert: die neuesten Ansätze aus der Forschung zu Graph Neural Networks scheinen immer noch funktional begrenzt, erfordern jedoch erhebliche Rechenressourcen.

Kurze Darstellung der Arbeiten, die zu keiner Lösung geführt haben

Gab es Arbeiten bzw. Lösungsansätze, die nicht weiter verfolgt wurden? Was waren die Hintergründe, und wie bist Du alternativ vorgegangen?

Ich habe viel Zeit damit zugebracht, Ansätze zur Disambiguierung von Firmen- und Personennamen in Texten zu recherchieren. Dabei geht es etwa darum, eine Person eindeutig dadurch zu identifizieren, mit welchen anderen Personen oder Firmen sie regelmäßig gemeinsam erwähnt wird (“Der Markus Braun, der immer wieder im Zusammenhang mit der Wirecard-Bank und Jan Marsalek erwähnt wird - und nicht der gleichnamige Automechaniker aus Kassel”). Freundlicherweise konnte ich dabei auch auf den Rat anderer Prototypefund-Teilnehmer Tobias Sterbak und Sergei Levich zurückgreifen.

Letztlich wird das Problem in StoryWeb allerdings weniger durch eine algorithmische Herangehensweise gelöst, denn durch die Einführung der “Story” als Gruppierung von Artikeln wird es einfacher, anzunehmen, dass alle Erwähnungen eines gleichen Namens in den Artikeln innerhalb einer Story auch auf die gleiche Person oder Firma verweisen.

Das “Story”-Konzept führte letztlich auch dazu, dass ich mich entschlossen habe, den direkten Import von einzelnen Artikeln via Benutzeroberfläche direkt in eine Story zu implementieren. Das führte letztlich dazu, dass die Relevanz des Artikel-Archivs auf der Basis des mediacrawl-Tools wesentlich weniger relevant in der Nutzung des Tools ist, als zunächst gedacht.

Kurze Angabe von Präsentationsmöglichkeiten für mögliche Nutzer

Wo können sich Interessenten detailliert über Deine Projektergebnisse informieren (z.B. Webseite, GitHub, Veröffentlichungen)?

Interessenten können das Projekt auf GitHub einsehen:

- https://github.com/opensanctions/storyweb - Die primäre StoryWeb-Anwendung, Dokumentation, Installationsanweisungen und Beispieldaten.
- https://github.com/opensanctions/mediacrawl - Das Teilprojekt zum Herunterladen und Bloßstellen von Nachrichtenartikeln

- https://github.com/pudo/articledata - Eine Minimalspezifikation für Artikel-Daten, die von storyweb und mediacrawl verwendet wird, um Text vom einen ins andere System zu transportieren.

Eine einfache Projektwebseite ist unter https://storyweb.opensanctions.org/ abrufbar.

Kurze Erläuterung zur Einhaltung der Arbeits- und Kostenplanung

Gab es im Projektverlauf Ereignisse, die eine Anpassung der Planung erforderlich machten – z.B. Mehr- oder Minderaufwand bei der Bearbeitung von Teilaufgaben?

Die Arbeits- und Kostenplanung für das Projekt war ausreichend zur Umsetzung der Software - entstanden ist jedoch ein Prototyp, der eine ausführliche Erprobung und Anpassung noch vor sich hat.

Ich habe etwa die ersten sechs bis acht Wochen des Förderzeitraums damit verbracht, das mediacrawl-Werkzeug zum Sammeln von Nachrichtentexten zu entwickeln, und das damit Artikel-Archiv durch "named entity extraction" experimentell zu analysieren. Danach habe ich in schneller Folge zwischen der Entwicklung des Web-Frontends, des Datenmodells und des Software-Backends gewechselt, um verschiedene denkbare Datenmodelle für die Anwendung direkt im User Interface zu erproben.

Ich hatte gehofft, früher im Projektverlauf mit JournalistInnen zusammenzuarbeiten. Nachdem die zu Anfang geplante Kooperation mit drei Medien ("Bad Will Hunting") nicht zu konkreten Ergebnissen führte, habe ich mich von den technischen Herausforderungen des Projekts ablenken lassen. Ich hoffe, nach dem Projektende noch einige der Kontakte zu Journalisten aufzugreifen, um das Tool weiter zu erproben und Feedback für eine zweite Iteration zu sammeln.

Kurze Darstellung von etwaigen Ergebnissen bei anderen Stellen

Gab es Entwicklungen anderer Personen oder Institutionen, die Einfluss auf Deine Arbeiten und die Zielsetzung hatten? Wenn ja, worin bestand dieser und wie bist Du damit umgegangen?

Das Projekt ist durch eine Vielzahl von bestehenden Projekten und Gedanken inspiriert. Dazu gehören unter anderem verschiedene kommerzielle Produkte aus dem "adverse media"- und ESG-Bereich, die aus Nachrichtentexten strukturierte Daten über Firmen und Personen extrahieren wollen.

Eine weitere Inspiration für StoryWeb war das DIGAS-System, das von Spiegel und Axel Springer gemeinsam betrieben wird, und Nachrichtentexte der letzten Jahrzehnte verschlagwortet und mit Akteursinformationen verknüpft.

Weniger produktiv als gedacht verlief leider die Zusammenarbeit mit "Bad Will Hunting", einem Konsortium aus Guardian, Follow The Money und Daily Maverick, die parallel zu StoryWeb an ähnlichen Techniken arbeiten wollten. Ich hatte gehofft, dass StoryWeb als Benutzeroberfläche für die Machine Learning-basierte Herangehensweise von BWH zum Einsatz kommen könnte.

Besonders relevant ist auch das "trafilatura"-Projekt von Adrien Barbaresi, das die besonders komplexe Aufgabe löst, Artikeltext aus dem HTML-Gewust moderner Nachrichtenseiten heraus frei zu stellen. Erst dadurch werden die Texte für die Extraktion von Personen- und Firmennamen (in

unserem Fall durch die Software “spaCy”) zugänglich. Im Projektzeitraum veröffentlichte auch Microsoft einen experimentellen Entity Linker[2], der auf Basis von spaCy funktioniert. Auch die Firma hinter spaCy, Explosion, scheint an einem neuen Entity-Linking-System zu arbeiten, zum Teil inspiriert durch die Arbeit der o.g. “Bad Will Hunting”-Gruppe[3].

Schließlich fiel der Projektzeitraum mit der Veröffentlichung des von OpenAI entwickelten ChatGPT-Systems zusammen. ChatGPT erscheint als Machine Learning-System besonders für die durch StoryWeb umgesetzte Beziehungsanalyse geeignet, was auch der im Feburar veröffentlichte Prototyp GraphGPT[4] eindrucksvoll demonstriert.

[2] https://github.com/microsoft/spacy-ann-linker
[3] https://explosion.ai/blog/guardian
[4] https://github.com/varunshenoy/GraphGPT