Richtlinie zum „Software-Sprint“

UwaziML – Machine Learning als Rettungsring in der Dokumentenflut

---

Schlussbericht

Zuwendungsempfänger:

Natalie Widmann

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS17S34 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)?

Gemeinnützige Organisationen haben oft nur begrenzte Mittel um aus der Flut an Dokumenten, relevante Informationen zu filtern, Erkenntnisse aus verschiedenen Quellen zu verbinden und Muster von Diskriminierung, Korruption oder Ausbeutung zu entdecken. Dies ist jedoch notwendig um Regierungen und Menschenrechtsverletzer zur Rechenschaft zu ziehen.

Große Dokumentensammlungen zu strukturieren und im Detail zu analysieren ist mit einem großem Zeit- und Kostenaufwand verbunden, da die Dokumente einzeln und per Hand bearbeitet werden müssen.

Machine Learning und *Natural Language Processing* eröffnen neue Möglichkeiten um die Analyse von Textdokumenten so effizient und intuitiv wie möglich zu gestalten. Zum Beispiel können die Namen von Personen, Organisationen oder Ländern, automatisch extrahiert und so mit Informationen aus anderen Dokumenten kombiniert werden.

Ziel des Projektes ist es, diese Methoden in Uwazi, einem open-source Projekt zur kollaborativen Bearbeitung und Veröffentlichung von Dokumentensammlungen,

zu integrieren. Dadurch können Nutzer, auch ohne Programmierkenntnisse, maschinelle Lernalgorithmen erstellen, sodass die für sie relevante Information in noch unbearbeiteten Dokumenten automatisch identifiziert wird.
Eine einfache Korrektur der Ergebnisse ermöglicht es die Algorithmen immer weiter anzupassen und gibt dem Nutzer die Kontrolle über die verschiedenen Modelle.

Eingebunden in Uwazi sind diese Methoden für zivilgesellschaftliche Organisationen zugänglich und erlauben eine intuitive und effiziente Nutzung.

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

Wer ist die Zielgruppe für Deine Lösung? Wie profitiert sie von den Ergebnissen Deines Projekts? Welche Bezüge gibt es zu den Themenfeldern und Zielen des Software Sprints?

Das Projekt richtet sich in erster Linie an Menschenrechtsorganisationen, deren Arbeit auf der Zugänglichkeit und dem Verständnis von Dokumentensammlungen basiert. Durch die automatische Kategorisierung von Dokumenten, wird den Nutzern mühsame Arbeit abgenommen und sie können sich auf ihre Kerntätigkeiten fokussieren.

Aber auch andere Gruppen, wie beispielsweise Investigativjournalisten oder Wissenschaftler, profitieren von UwaziML, da sie Dokumente für ihre Recherchen schneller strukturieren und analysieren können.

Damit unterstütz UwaziML zivilgesellschaftliche Projekte die zur Stärkung der Transparenz öffentlicher Einrichtungen beitragen.
Als Open Source Projekt ist die Nutzung der Software jedem frei zugänglich.

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?

Durch die Förderinitiative „Software-Sprint“ konnte ein maschineller Lernalgorithmus in Uwazi integriert werden. Genauer gesagt, ein Convolutional Neural Network (CNN), zu deutsch faltendes neurales Netzwerk, wird für die Kategorisierung von Sätzen verwendet. Dies ermöglicht relevante Information aus Dokumenten zu extrahieren und Sätze als Evidenz dafür zu benutzen.
Aufgrund dieser Trainingdaten extrahiert der Algorithmus ähnliche Sätze aus noch unbearbeiteten Dokumenten und gibt dem Nutzer einen Einblick in deren Inhalt.
Durch die Integration eines Feedbackloops, der Möglichkeit Vorschläge

anzunehmen oder abzulehnen, kann das Machine Learning Modell jederzeit korrigiert und verbessert werden.

Ein weiteres konkretes Ergebnis der Arbeit während der Projektinitiative „Software-Sprint“ beinhalten die Generalisierung der genutzten Algorithmen, sodass sie für unterschiedlichste Dokumentensammlungen und Projekte verwendbar sind. Die Architektur des neuronalen Netzes ermöglicht es Sätze, egal ob kurz oder lang, zu kategorisieren. In Verbindung mit einem trainierten Word Embedding Model, einer mathematischen Methode die Einblick in die semantische Ähnlichkeit zweier Wörter und deren Kontext gibt, ist es möglich Synonyme von relevanten Schlüsselwörtern zu identifizieren und dadurch die machinell vorgeschlagenen Ergebnisse zu verbessern.

Aber auch die Nutzeroberfläche ist ein wichtiger Bestandteil des Projekts, da sie die Schnittstelle zwischen maschinellem Lernen und den Nutzern wie zum Beispiel Menschenrechtsorganisationen oder Investigativjournalisten, bildet. Sie ermöglicht nicht nur eine intuitive Benutzung und ein effizientes Training der Algorithmen, sondern ist auch besonders wichtig um Vertrauen aufzubauen und Informationen über die Zuverlässigkeit der Ergebnisse zu erhalten.

Des Weiteren ist für dieses Projekt Transparenz und Kontrolle ein relevanter Aspekt. Die Nutzer sollen nicht mit automatischen Entscheidungen der Algorithmen konfrontiert werden, sondern Vorschläge erhalten um effizienter an für sie interessantes Wissen zu kommen.
Um jederzeit die Kontrolle über die im Algorithmus verwendeten Daten zu haben, wurde in Uwazi ein neues Interface erstellt. Es zeigt alle verwendeten Sätzen, deren zugeordnete Kategorie und gibt die Möglichkeit diese zu ändern.

Ein Meilenstein der während der Projektlaufzeit nicht erreicht wurde, ist das Testen der Anwendung mit Uwazi-Nutzern. Für eine Erklärung siehe Punkt Einhaltung der Arbeits- und Kostenplanung.

Die Zusammenarbeit mit der Open Knowledge Foundation war sehr wichtig, da durch regelmäßige Treffen nicht nur ein Austausch über die Projekte, sondern vor allem auch über Arbeitsweisen, Methoden oder neue Anwendungen zustande kam. Vor allem das Feedback in den Coaching Sessions hat UzwaiML gezeigt wie wichtig es ist den Nutzer in den Mittelpunkt der Arbeit zu stellen.

Zielgruppe, Nutzen und mögliche Weiterentwicklungen

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

Die Zielgruppe hat mithilfe der Uwazi Software die Möglichkeit ohne Fachwissen oder einen Experten im Bereich Machine Learning, diese Technologie als Unterstützung der manuellen Kategorisierung von Dokumenten zu nutzen.

Durch die Veröffentlichung des Codes ist es möglich auch ohne Uwazi auf die Funktionen zuzugreifen. Eine API ermöglicht techversierten Nutzern die Modelle auf eigenen Daten zu trainieren und die Ergebnisse Projekt spezifisch aufzubereiten und darzustellen.

Ideen für die Weiterentwicklung beinhalten: die Nutzung von FastText als Word Embeddings, Anpassung der Modelle an unterschiedliche Metadatenformate (Datum, Länder, Organisationen, Dokumententypen,...), Umstellung von Binary auf Multi-Class-Classification.
Die Umsetzung erfolg in Zusammenarbeit mit dem Huridocs Team.

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

Vor allem die enge Zusammenarbeit mit den Huridocs Designern, Entwicklern und Projektmanagern hat mich persönlich aber auch fachlich in vieler Hinsicht weitergebracht. Ich hab ein besseres Verständnis für die Architektur eines Open-Source Software Projektes und habe die unterschiedlichen Arbeitsprozesse kennengelernt. Es ist wichtig im Team über Ideen zu diskutieren um unterschiedliche Meinung einzuholen und verschiedene Perspektiven zu reflektieren.
Auch im Bereich maschinelles Lernen habe ich Techniken verwendet über die ich vorher nur konzeptionelles Wissen hatte. Während der Projektlaufzeit habe ich mithilfe von Tensorflow diese Algorithmen implementiert und mit ihnen experimentiert um bessere Ergebnisse zu erzielen.

Kurze Darstellung der Arbeiten, die zu keiner Lösung geführt haben

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

In einer Dokumentensammlung kann eine Information mehrere unterschiedliche Werte annehmen, beispielsweise kann ein Gesetzentwurf nicht nur angenommen oder abgelehnt werden, sondern auch zur weiteren Bearbeitung an unterschiedliche Behörden geleitet werden. Dies erfordert eine flexible Anpassung der möglichen Kategorien, was bisher aufgrund der technischen Umsetzung an der Schnittstelle zwischen Uwazi und der maschinellen Lernalgorithmen nicht weiter verfolgt wurde.
Alternativ wird pro Kategorie ein neuer Algorithmus trainiert. Diese Lösung ist nicht ganz so effizient wie eine Multi-Class Kategorisierung, aber in dem derzeit geplanten Umfang der Dokumente und zu extrahierenden Information praktikabel und beugt anderen Schwierigkeiten, wie einer ungleichmäßigen Verteilung der Trainingsdaten für die unterschiedlichen Kategorien vor.

Kurze Angabe von Präsentationsmöglichkeiten für mögliche Nutzer

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

Das, mit Hilfe des Prototype Funds entstandene, Video unter https://prototypefund.de/project/uwaziml/ gibt einen guten Einblick in die Projektidee und das erzielte Ergebnis.
Des Weiteren gibt es unter uwazi.io generelle Informationen über Uwazi und mit den folgenden Links ist der Source Code für Uwazi und den projektspezifische Machine Learning Teil einsehbar:
https://github.com/huridocs/uwazi
https://github.com/Tilana/Classification

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 Integration der maschinen-gestützten Algorithmen in Uwazi und das Überprüfen der korrekten Funktionsweise hat mehr Zeit als Vorhergesehen in Anspruch genommen.
Ergebnisse der maschinellen Lernalgorithmen sind für Nutzer und Programmierer manchmal schwer nachvollziehbar und da die Nutzung einer Software auf Vertrauen basiert haben wir uns entschieden während der Projektzeit keine Nutzertest mit Menschenrechtsorganisationen durchzuführen.
Dafür haben wir die internen Tests zur Überprüfung der Funktionsweise mit unterschiedlichen Dokumentensammlungen und Kategorien ausgeweitet.

Generell wurde, bis auf zwei Stunden extra, die geplante Anzahl an Arbeitsstunden eingehalten.

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?

UwaziML ist ein Projekt an der Schnittstelle zwischen den Bereichen Maschinellem Lernen, Softwareentwicklung, aber auch Menschenrechten. Den größten Einfluss auf meine Arbeit hatten wissenschafltiche Publikationen, deren Aubereitung in Blogposts, sowie Open Source Code. Maschinelles Lernen ist ein breit gefächertes Feld, in dem es wichtig ist von der Forschung anderer zu lernen um effiziente und akkurate Ergebnisse zu erzielen. Dadurch haben sich in UwaziML vor allem die Methoden für die Kategorisierung von Dokumenten geändert, nicht jedoch die Zielsetzung.