Projektbericht zum Modul Information Retrieval und Visualisierung Sommersemester 2021

Titel des Dokuments

Marcus Gagelmann

28. August 2021

1 Einleitung

Afrika bildet mit einer Größe von mehr als 30 Millionen Quadratkilometern den zweitgrößten Kontinent der Erde. Insgesamt leben dort über 1,3 Milliarden Menschen, was etwa 17,2% der Weltbevölkerung ausmacht. Wo jedoch Land und Menschen aufeinander treffen, dort sind auch Konflikte nicht weit entfernt. Seid vielen Jahren fällt Afrika solchen sowohl politischen als auch kriegerischen Auseinandersetzungen zum Opfer. Ursachen hierfür könnten die verbreitete Armut, misswirtschaftende Regierungen sowie die wertvollen Ressourcen des Kontinents sein, um nur einige mögliche Gründe zu nennen. Seit 1997 werden diese Konflikte durch das ACLED-Projekt dokumentiert. ACLED steht hierbei für „Armed Conflict Location and Event Data“. Die gesammelten Daten werden von ACLED auf ihrer Website (`https://acleddata.com/#/dashboard`) frei zugänglich zur Verfügung gestellt.

Nach nun beinahe 25 Jahren an Konfliktaufzeichnungen sind mittlerweile sehr große Datenmengen entstanden, welche einem außenstehenden Tabellenbetrachter kaum einen Überblick über das gesamtgeschehen seit dem Aufzeichnungsbeginn geben können. Noch unwahrscheinlicher ist es, dass ein solcher Betrachter allein mithilfe der unaufbereiteten Daten Rückschlüsse auf mögliche Zusammenhänge zwischen den vielen Konflikten ziehen kann. Mit dieser Ausgangslage ist ein Verstehen der tatsächlichen Lage in Afrika anhand der Daten unmöglich. Hierdurch könnten mögliche Maßnahmen zur Verbesserung der Situation des Kontinents weniger effektiv ausfallen, als wenn das volle Potenzial der Daten ausgeschöpft werden würde.

Aus diesem Problem ergibt sich die Fragestellung, ob der gewählte Datensatz hilfreiche Aussagen zu Zusammenhängen ermöglicht, welche die Opfer der Konflikte der letzten 25 Jahre

betreffen. Das Ziel dieser Arbeit ist deshalb die Bereitstellung einer Anwendung, welche eine Analyse der Todesopferanzahlen der Afrikakonflikte ermöglicht.

1.1 Anwendungshintergrund

Mithilfe der im Rahmen dieser Arbeit entwickelten Anwendung werden einzelne Konflikte für einen Gesamtüberblick über eine oder mehrere Regionen Afrikas zusammengefasst, sowie nach einzelnen Jahren sortiert und hierbei auch mehrdimensional visualisiert. Insgesamt umfasst das Programm zwei verschiedene Ansichten, welche jeweil eine Menge von Konflikten auf unterschiedliche Art und Weise darstellen. In der Hauptansicht, dem Scatterplot, ist es dem Nutzer möglich einen Überblick über die Konfliktdichte und die Opferzahlen zu unterschiedlichen Zeitpunkten seit Aufzeichnungsbeginn zu erhalten. Die Zweitansicht, welche sich der Technik paralleler Koordinaten bedient, zeigt immer eine Teilmenge der Konflikte der Hauptansicht in jeweils drei Dimensionen. Hier kann der Nutzer einen genaueren Einblick erhalten, in welchem Monat eines Jahres welche Art von Konflikt wie viele Todesopfer forderte. Die Basismenge an Konflikten, welche für die Visualisierungen verwendet wird, ergibt sich aus der Menge aller Konflikte des Datensatzes, welche mithilfe eines Regionenfilters gekürzt wird. Dieser Filter wird vom Nutzer über eine dritte Ansicht in Form eines interaktiven Baumes gesteuert. Hier können sowohl die Konflikte größerer Regionen als auch die Konflikte einzelner Länder Afrikas in die visualisierte Gesamtmenge aufgenommen werden.

1.2 Zielgruppen

Als Zielgruppe für die Visualisierungsanwendung kommen Personen in Frage, welche einen Überblick über die Konflikt- und Todesopferdichte zu verschiedenen Zeitpunkten und in verschiedenen Regionen des afrikanischen Kontinents erlangen wollen, sowie einen genaueren Einblick in die Art der Konflikte benötigen. Das Programm ist für einen schnellen Einstieg und einen groben Überblick in der Thematik der Afrikakonflikte konzipiert. Aus diesem Grund benötigen Zielgruppen in der Regel kaum Vorwissen um die Anwendung sinnvoll zu nutzen. Politisches und geografisches Vorwissen kann in manchen Anwendungsgebieten für das weitere Verständnis sinnvoll sein, ist aber keines Falls Voraussetzung um Erkenntnisse aus den Daten mithilfe der Anwendung zu gewinnen.

Die Verwendung einer solchen Anwendung ist bei Hilfsorganisationen zur Planung zukünftiger Einsätze denkbar. Hiermit könnten zukünftig diejenigen Regionen Afrikas ermittelt werden, welche langfristig die größten Opfer zu beklagen haben und aufbauend auf diesen Fakt weitere Hilfsleistungen am dringensten benötigen. Weiterhin können Hilfskräfte leicht einen Eindruck von der Art der bisherigen Konflikte einer bestimmten Region bekommen und damit zukünftige Einsätze besser auf diese Art von Konflikten vorbereiten.

Ein weiteres Anwendungsgebiet könnte die Risikoeinschätzung für Reisen durch das Auswärtige Amt (Link) darstellen. Hier bietet die Anwendung die Möglichkeit, die Gefahr für deutsche Reisende in Konfliktbelasteten Regionen Afrikas besser beurteilen zu können. Gäbe es in den letzten Jahren eine hohe Dichte an politischen Unruhen oder Todesopfer durch Kampfhandlungen, so können hierdurch Urlaubs- und Durchreisende rechtzeitig vor Gefahren gewarnt werden.

1.3 Überblick und Beiträge

Dieses Projekt setzt sich zusammen aus dem Datensatz des ACLED-Projekts von 1997-2021 und drei unterschiedlichen Visualisierungstechniken, welche interaktiv miteinander verbunden wurden. Vom verwendete Datensatz werden insgesamt sechs Felder jedes Datenobjektes für die Umsetzung der Anwendung genutzt. Hierzu gehören das Jahr sowie das Datum des Konflikts, die Art des Konflikts, die Region sowie das Land in dem ein Konflikt vorgefallen ist und die Anzahl an Todesopfern, welche ein Konflikt forderte. Die in dem Programm verwendeten Visualisierungstechniken beschränken sich auf eine Hauptansicht mit einem Scatterplot und einer umschaltbaren Zweitansicht, welche parallele Koordinaten zur Umsetzung verwendet. Zusätzlich gibt es noch einen Regionsfilter, welcher mithilfe expliziter Bäume anschaulich umgesetzt wurde. Im Scatterplot findet weiterhin die sogenannte X-Ray-Technik Anwendung, welche dem Nutzer Überlagerungen von Punkten aufzeigen kann.

Die Beiträge dieses Projekts belaufen sich auf die Auswahl, die Verarbeitung sowie die Visualisierung von Daten der Konflikte in Afrika in den Jahren 1997 bis 2021. Ganz konkret wird hierbei der Prototyp einer möglichen Anwendung diskutiert, welcher dem Nutzer die gewählten Daten und deren Zusammenhänge als interaktiv verbundenes System näher bringen soll. Hierbei werden die einzelnen Teile der Anwendung kritisch betrachtet, gelungene Ansätze hervorgehoben und Vorschläge zur weiteren Verbesserung geäußert.

2 Daten

Der Datensatz des ACLED-Projekt für die Afrikakonflikte von 1997 bis 2021 beinhaltet insgesamt 65443 einzelne Datenobjekte bzw. Datenzeilen, welche jeweils einen gesonderten Konflikt repräsentieren.

Jedes Datenobjekt des Datensatzes liegt in insgesamt 29 Dimensionen vor, welche im einzelnen in Abbildung 1 eingesehen werden können. Verwendet werden von diesen Dimensionen in der Anwendung EVENT_DATE, YEAR, EVENT_TYPE, REGION, COUNTRY und FATALITIES. Diese gewählten Dimensionen beinhalten alle für das Programm benötigten Daten und bieten den Vorteil, unkodiert und fast ausschließlich in direkt verwendbarer Form vorzuliegen. Direkt verwendbar bedeutet hierbei, dass die Dimensionen keine weitere Verarbeitung benötigen und direkt in einfach Elm-Datenstrukturen (Int und String) umgewandelt werden können. Lediglich die Dimension EVENT_DATE bedarf einer zusätzlichen Übersetzerfunktion, welche

1. ISO
2. EVENT_ID_CNTY
3. EVENT_ID_NO_CNTY
4. EVENT_DATE
5. YEAR
6. TIME_PRECISION
7. EVENT_TYPE
8. SUB_EVENT_TYPE
9. ACTOR1
10. ASSOC_ACTOR_1
11. INTER1
12. ACTOR2
13. ASSOC_ACTOR_2
14. INTER2
15. INTERACTION
16. REGION
17. COUNTRY
18. ADMIN1
19. ADMIN2
20. ADMIN3
21. LOCATION
22. LATITUDE
23. LONGITUDE
24. GEO_PRECISION
25. SOURCE
26. SOURCE_SCALE
27. NOTES
28. FATALITIES
29. TIMESTAMP

Abbildung 1: Dimensionen des ACLED-Datensatzes
Quelle: Eigene Darstellung

genutzt werden kann, um den erhaltenen Event-Date-String auf den zugehörigen Monat zu mappen.

Es bleibt die Frage, in wie weit sich der Datensatz des ACLED-Projekts für die Zielgruppen dieser Arbeit eignet. Die eindeutigen Daten zu Zeitpunkt, Name des Landes und Todesopfern eignen sich gut für Hilfsorganisationen, um zu ermitteln, welche Länder aktuell am meisten unter den Konflikten leiden. Hierdurch kann eine Vorauswahl getroffen werden, welche Länder momentan Hilfsleistungen benötigen könnten. Mithilfe der Art des Events (EVENT_TYPE) kann es diesen Organisationen weiterhin gut möglich gemacht werden, die Problemlage im Land besser zu verstehen und einen potenziellen Hilfseinsatz präziser auf eine bestimmte Art von Konflikten vorzubereiten. Bei der Aufgabe einer Gefahreneinschätzung für Reisende durch das Auswärtige Amt kann mithilfe der gewählten Dimensionen des Datensatzes durchaus auch geholfen werden. Informationen wie die Konfliktdichte, die Todesopferdichte und auch die Konfliktarten in einer bestimmten Region lassen eine erste Einschätzungen zu der Situation eines Landes und zur Gefahr für Reisende zu.
Eine Verwendung der vorhandenen Geodaten zur Erweiterung der Anwendung durch eine interaktive Karte wäre durchaus denkbar und würde den Zielgruppen beim Verstehen von Zusammenhängen zwischen den Geodaten und den Konfliktdaten zusätzlich helfen. Eine solche Anwendung der Geodaten würde jedoch aufgrund von Problemen bei der Datenaufbereitung (siehe **??**) sowie der nötigen Ergänzung des Projektes durch Polygondaten für Länderumrisse

den Rahmen dieser Arbeit sprengen.
Es ist an dieser Stelle wichtig anzumerken, dass die Daten, welche im Programm Anwendung finden, nicht allein ausreichen um die genannten Aufgaben der Zielgruppen zu erfüllen. Da sowohl die Planung von Hilfseinsätzen, als auch die Einschätzung von Gefahrenländern Auswirkungen auf die Sicherheit von Menschenleben haben kann, ist eine Ergänzung des Datensatzes durch zusätzliche detaillierte Informationen zur Politik, Mentalität, Religion und anderen landes- und regionsspezifischen Informationen abseits der aufgezeichneten Konflikte unabdingbar. Beispielsweise kann offen gezeigte Homosexualität momentan in der Republik Kongo zu willkürlichen Verhaftungen wegen angeblich sittenwidrigen Verhaltens führen (`https://www.auswaertiges-amt.de/de/aussenpolitik/laender/kongorepublik-node/kongorepubliksicherheit/208542#content_1`), was unbedingt bei einer Gefahreneinschätzung des Landes für westliche Reisende berücksichtigt werden sollte. Diese Information könnte jedoch nicht aus dem gewählten Datensatz ermittelt werden, da sich dieser lediglich auf größere Konflikte und nicht auf einzelne Verhaftungen innerhalb der Republik Kongo stützt.

2.1 Technische Bereitstellung der Daten

Bereitgestellt werden die verwendeten Daten von Kaggle (urlhttps://www.kaggle.com/), einer Plattform auf der unter anderem Datensätze für Data-Science-Projekte geteilt werden können. Eine andere Möglichkeit für den Download der Daten bietet die Seite des ACLED-Projekts (urlhttps://acleddata.com/data-export-tool/). In dieser Arbeit wird der Datensatz von Kaggle verwendet, da die dort bereitgestellten Daten, anders als auf der ACLED-Seite, bereits nach Konflikten im Raum Afrika gefiltert sind. Läd man den Datensatz direkt von Kaggle herunter, so erhält man diesen im CSV-Dateiformat. Als Seperator werden hierbei Semikolons verwendet. Die größe des Datensatzes beläuft sich auf etwa 30,5 MB (30.565.061 Bytes) im Standard CSV-Format, welche sich jedoch auf 62,4 MB (65.467.818 Bytes) mehr als verdoppelt nach einer Umwandlung der Daten in ein JSON-Dateiformat. Die Notwendigkeit für eine solche Umwandlung des Formats wird im Kapitel 2.2 genauer besprochen.
Jede Dimension des Datensatzes ist grundsätzlich entweder als Integer-Wert (Ziffern ohne Komma oder Anführungszeichen) oder als String-Wert (Zeichenfolgen in Anführungszeichen) gegeben. Hierzu gehören auch die Werte LATITUDE und LONGITUDE, welche eigentlich Float-Werte darstellen sollten. Dies stellt ein Problem, welches vor der eigentlichen Verwendung der Daten beim Einlesen durch eine Aufbereitung behoben werden müsste. Die Angaben in LATITUDE und LONGITUDE liegen als String vor, welcher unterschiedlich lange Ziffernfolgen enthalten kann. Hierbei ist keine Kommastelle angegeben, wodurch die Koordinaten des Konflikts nicht eindeutig wiedergegeben werden. Beispielsweise existiert ein Konflikt mit dem Breitengrad 36672 und dem Längengrad 2789. Mit diesen Daten könnten die Koordinaten (3.6672, 27.89) gemeint sein, welche eine Position im Wald in der Demokratischen Republik Kongo bestimmen. Aufgrund der Größe des Kontinents könnten hier jedoch auch die Koordinaten (36.672, 2.789) in Frage kommen, welche eine Position in der Stadt Koléa in Algerien beschreiben. In diesem Fall

ist die zweite Möglichkeit korrekt. Es wird jedoch ersichtlich, dass die Position des Kommas in den Breiten- und Längengraden nicht trivial bestimmt werden kann. Einige weitere Dimensionen des Datensatzes sind ebenfalls nicht sehr leicht zugänglich. Die Dimension INTERACTION liegt beispielsweise, neben einigen anderen, nur als numerischer Code vor. Um die tatsächliche Art der Interaktion der betroffenen Parteien ermitteln zu können, müsste hierfür vorher ein Übersetzer mithilfe des Codebooks des ACLED-Projekts (`https://reliefweb.int/sites/reliefweb.int/files/resources/ACLED_Codebook_2017FINAL%20%281%29.pdf`) gebaut werden.

-einige daten nur als numerischer code angegeben -> bei nutzung übersetzung notwendig (einlesen oder als funktion) -geodaten als int, nicht als float -> erstmal uneindeutig -metadaten wie timestamp, geoprecision, timeprecision, etc ignoriert

Wie sind die Daten zugänglich? Welche Formate werden genutzt. Gibt es Besonderheiten beim Lesen der Formate?

2.2 Datenvorverarbeitung

-einmalig umgewandelt in json datei -> lesbarkeit während der entwicklung, vorsortierte strukturen welche von elm beim einlesen direkt genutzt werden können -vorfilterung nach regionen, initial nach ghana -> filterConflicts erklären -datum auf monat reduziert -> ausreichend für visualisierung der parallelen Koordinaten -> jahr als intervall genauer aber schwer zu realisieren und übersteigt den nutzen -Regionsnamen komplett und ländernamen bei zu großer länge gekürzt um in baumknoten zu passen -> bei regionen kein afrika, bei ländern auf anfangsbuchstaben reduziert (hover anzeige für übersicht vorhanden) -Parallele k.: Konflikttyp auch mit subtyp bezeichenbar -> weggelassen für übersichtlichkeit da obertyp immernoch notwendig für gesamtverständnis -> sonst zu viel auf einmal angezeigt für schnellen überblick -> potenzial für höheren detailgrad

Welche Datenvorverarbeitungsschritte sind notwendig? Beschreiben Sie die einzelnen Schritte und begründen sie sie, z.B. warum werden manche Daten weggelassen, über welche Mengen werden Durchschnitte berechnet, warum sind die so berechneten Werte aussagekräftiger als andere Werte.

3 Visualisierungen

3.1 Analyse der Anwendungsaufgaben

Analysieren sie die konkreten Anwendungsaufgaben. Welche Visualisierungen helfen den Personen, die die Software verwenden, sinnvolle mentale Modelle aufzubauen. Sind diese mentalen Modelle für sie notwendig, um die Aufgaben lösen zu können?

3.2 Anforderungen an die Visualisierungen

Leiten sie Anforderungen an das Design der Visualisierungen ab, die sich durch ihre Analyse des Zielproblems ergeben.

3.3 Präsentation der Visualisierungen

Präsentieren sie die visuelle Abbildungen und Kodierungen der Daten und Interaktionsmöglichkeiten. Sie müssen begründen, warum und wiegut ihre Designentscheidungen die erstellten Anforderungen erfüllen. Weiterhin müssen sie begründen, warum die gewählte visuelle Kodierung der Daten für das zulösenden Problem passend ist. Typische Argumente würden hier auf Wahrnehmungsprinzipien und Theorie über Informationsvisualisierung verweisen. Die besten Begründungen diskutieren explizit die konkrete Auswahl der Visualisierungen im Kontext von mehreren verschiedenen Alternativen. Diskutieren sie die Expressivität und die Effektivität der einzelnen Visualisierungen.

Die eben beschriebenen Präsentationen und Begründungen sollen für jede der drei folgenden Visualisierungen durchgeführt werden.

3.3.1 Visualisierung Eins

3.3.2 Visualisierung Zwei

3.3.3 Visualisierung Drei

3.4 Interaktion

-checkboxes als filter nicht umgesetzt -> bessere übersicht über geografische verhältnisse

Erklären sie die möglichen Interaktionen mit den einzelnen Visualisierungen und die möglichen Verknüpfungen zwischen ihnen. Begründen Sie warum die konkreten Interaktionen umgesetzt wurden und welche Zwecke für die Anwenderinnen mit ihnen unterstützt werden. Begründen sie ebenfalls warum sie andere Interaktionsmöglichkeiten nicht umgesetzt haben.

4 Implementierung

Beschreiben Sie die Implementierung ihrer Visualisierungsanwendung in Elm. Stellen die Gliederung ihres Quellcodes vor. Haben Sie verschiedene Elm-Module erstellt. Was war aufwändig umzusetzen, was ließ sich mit dem vorhanden Code aus den Übungen relativ einfach umsetzen?

Wie sieht die Elm-Datenstruktur für das Model aus, in dem die verschiedenen Zustände der Interaktion gespeichert werden können.

5 Anwendungsfälle

Präsentieren sie für jede der drei Visualisierungen einen sinnvollen Anwendungsfall in dem ein bestimmter Fakt, ein Muster oder die Abwesenheit eines Musters visuell festgestellt wird. Begründen sie warum dieser Anwendungsfall wichtig für die Zielgruppe der Anwenderinnen ist. Diskutieren sie weiterhin, ob die oben beschriebene Information auch mit anderen Visualisierungstechniken hätte gefunden werden können. Falls dies möglich wäre, vergleichen sie die den Aufwand und die Schwierigkeiten ihres Ansatzes und der Alternativen.

5.1 Anwendung Visualisierung Eins

5.2 Anwendung Visualisierung Zwei

5.3 Anwendung Visualisierung Drei

6 Verwandte Arbeiten

Führen sie eine kurze Literatursuche in der wissenschaftlichen Literatur zu Informationsvisualisierung und Visual Analytics nach ähnlichen Anwendungen durch. Diskutieren sie mindestens zwei Artikel. Stellen sie Gemeinsamkeiten und Unterschiede dar.

7 Zusammenfassung und Ausblick

Fassen sie die Beiträge ihre Visualisierungsanwendung zusammen. Wo bietet sie für die Personen der Zielgruppe einen echten Mehrwert.

Was wären mögliche sinnvolle Erweiterungen, entweder auf der Ebene der Visualisierungen und/oder auf der Datenebene?

Anhang: Git-Historie