Richtlinie zum „Software-Sprint“

OAS - OpenAudioSearch

Schlussbericht

Zuwendungsempfänger:

Moreira Veit Heinzmann Schumann GBR

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS21S16 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Kurze Darstellung der Aufgabenstellung und Motivation

Im Internet gibt es große und schnell wachsende Bestände an interessanten Radio-Features und Podcasts, die unter freien Lizenzen veröffenticht sind. Ein gutes Beispiel ist unser Partner in diesem Projekt, das community broadcasting archive (cba) oder auch freie-radios.net. Auf diesen Plattformen ist eine Vielzahl an relevanten Audio-Inhalten archiviert, aber die einzelnen Beiträge und Features sind oft nur schwer wiederaufzufinden. Insbesondere für ehrenamtlich getragene Projekte bleibt oft wenig Zeit, um Beschreibungstexte, Metadaten oder gar Transkripte einzupflegen, nachdem die eigentliche Produktion abgeschlossen ist – das gilt umso mehr für Inhalte, die vor zwanzig Jahren produziert wurden und nun im Rahmen von Digitalisierungsprojekten online gestellt werden.

Hier wollen wir mit Open Audio Search Abhilfe schaffen, einer intelligenten Suchmaschine für Audiodateien. Über automatische Spracherkennung wird gesprochene Sprache in Text umgewandelt, der dann in einer Volltextsuchmaschine indiziert wird. Die Engine kann RSS-Feeds abonnieren, um Inhalte von verschiedenen Quellen zu importieren. Über eine webbasierte Oberfläche können Nutzer*innen die generierten Transkripte von Radiosendungen und Podcasts durchsuchen und direkt von der Zeitmarke der Suchergebnisse abspielen. Open Audio Search wird als kollaboratives Open-Source-Projekt entwickelt und kann leicht auf einem beliebigen Server installiert werden.

Mit Open Audio Search wollen wir dazu beitragen, eine zukunftssichere und gut wartbare Open-Source-Plattform für *community media discovery* herzustellen. Wir hoffen, dass wir damit Community-Media-Produzent*innen mit mehr Reichweite und einem breiteren Publikum versorgen können. Denn wenn Inhalte nicht so schnell nach ihrer Produktion wieder verschwinden, sondern stattdessen Teil eines lebendigen und leicht durchforstbaren Archivs werden, dann gehen diese Inhalte auch nicht so schnell wieder verloren, wie es derzeit oft der Fall ist.

Die Meilensteine für das Projekt waren:

- Implementierung des Grundsystems: Python-Webserver mit Job-Queue für ASR-Extraction, Anbindung an Elastic Search, grundlegendes Datenmodell, REST-API
- Design und Implementierung des Such-Indexes: Schema, Queries, Facets, Aggregations
- Domänenspezfische Analyzer für Elastic Search
- Design und Implementierung eines User Interfaces für Volltext-Suche und Filter (Facets)
- Import-Pipeline für mindestens zwei Anbindungen

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

Es gibt bereits kommerzielle Lösungen zur Spracherkenung, aber diese sind für freie Radios und andere nichtkommerzielle Medienprojekte nicht gut verfügbar: Ein minutenbasiertes Preismodell für die Sprachtranskription ist schlicht nicht finanzierbar oder zumindest sehr unattraktiv. Zwar sind Open-Source-Tools sowohl zur Spracherkennung als auch zur Volltextsuche vorhanden, aber diese sind bislang schwer aufzusetzen und nicht integriert - was einen großen händischen Aufwand bedeutet, um zu nutzbaren Ergebnissen zu kommen.

Open Audio Search kann als Tool von Community-Media-Produzent*innen für die Recherche genutzt werden, oder um relevante Beiträge zu verknüpfen. Das Tool erleichtert das Auffinden dieser Inhalte für Interessierte und erhöht so die Reichweite zivilgesellschaftlich erstellter Inhalte. Außerdem werden durch das Bereitstellen einer einfachen Nutzer*innen-Oberfläche mit intelligenten Suchfunktionen KI-Technologien für diesen Sektor zugänglich gemacht. Damit wird ein Beitrag geleistet zur Diskussion um ethische Nutzung von KI und zivilgesellschaftlicher Teilhabe am politischen Diskurs. Das Verwenden von Open-Source-Infrastruktur macht unabhängiger von großen Plattformen und unterstützt damit auch Content- und Datensouveränität.

Ausführliche Darstellung der Ergebnisse

Open Audio Search steht beim ersten Preview-Release, ein größeres Beta-Release folgt bald. Mit heutigem Stand sind die folgenden Features implementiert und nutzbar:

- RSS-Feeds können von beliebigen Quellen importiert und in eine lokale Datenbank gespeichert werden
- Alle in den RSS-Feeds verlinkten Audiodateien können automatisch über unsere Speech-To-Text-Pipeline transkribiert werden
- Die Zuordnung von Metadaten-Feldern in den RSS-Feeds kann angepasst werden
- Umfassende Volltext-Suche in sowohl den Metadaten als auch den automatisch erstellten Transkripten
- Ein responsives, webbasiertes User Interface, über das alle importierten Medien abgespielt und alle indizierten Daten durchsucht und gefiltert werden können
- Wenn die Suche Ergebnisse aus einem Transkript findet, dann kann der zugehörige Audiotrack direkt von der Zeitmarke des Suchergebnisses im Transkript abgespielt werden
- Eine (bislang minimalistische) Admin-UI, um die eingebundenen RSS-Feeds zu verwalten
- Eine HTTP-API mit Authentifizierung, über die Anbindungen an Drittsysteme entwickelt werden können

Zielgruppe, Nutzen und mögliche Weiterentwicklungen

Da wir selbst einen Hintergrund in Community-Media-Netzwerken haben und unsere Motivation aus diesem Bereich stammt, besteht unsere Zielgruppe im ersten Schritt aus gemeinnützigen Anbieter*innen deutschsprachiger Audio-/audiovisueller, selbstgehosteter Inhalte. Z.B. Freie Radios im DACH-Raum, individuelle Podcaster*innen, Open GLAM Initiativen, Museen oder Bildungsinitiativen wie Wikimedia.

Wir haben schon einige konkrete Ideen, woran wir in der näheren Zukunft arbeiten wollen:

- Mit unseren Partner*innen von cba und freie-radios.net wollen wir in den nächsten Monaten eine Open Audio Search Produktivinstanz für deutschsprachige Community Media in Betrieb nehmen.
- In einem Schwesterprojekt, gefördert durch netidee, setzen wir unsere eigene Trainings- und Evaluationspipeline auf, um die Spracherkennung zu verbessern. Außerdem wollen wir durch Natural Language Processing und das Matchen mit einer Wissensdatenbank wie Wikidata bessere und bedeutsamere Suchergebnisse erzielen.
- Nächstes Jahr wollen wir ein Replikationssystem zwischen Open Audio Search Instanzen implementieren. Das ist der nächste Schritt in Richtung eines europäischen Inhalte-Netzwerks für Community Media. Mit Unterstützung der European Cultural Foundation und Partner*innen in Irland und Spanien wollen wir in absehbarer Zukunft eine erste Pilotversion dieses *european cultural backbone* starten.

Kurze Darstellung der Arbeiten, die zu keiner Lösung geführt haben

Wir haben zu Beginn unser Backend als Proof of Concept in Python entwickelt. Während der Förderzeit haben wir entschieden, das Backend stattdessen in Rust zu implementieren. Bis diese Entscheidung finalisiert war, sind ein paar Stunden in später dann obsolete Weiterentwicklung des Python-Proof-Of-Conecpts geflossen. Der Rust-Port hat sich in der Zwischenzeit als die richtige Entscheidung erwiesen.

Kurze Angabe von Präsentationsmöglichkeiten für mögliche Nutzer

- Wir haben ein funktionsfähiges Setup, das in unserer Demo ausprobiert werden kann. Diese Demo-Seite indiziert derzeit einigen aktuellen Inhalt von cba und freie-radios.net. Es handelt sich um ein Alpha-Release, deswegen ist mit Bugs und fehlenden Features zu rechnen. Die Demo wird in unregelmäßigen Abständen aktualisiert und/oder zurückgesetzt.
- Open Audio Search hat eine Website: openaudiosearch.org
- Der vollständige Quelltext zu Open Audio Search ist auf GitHub veröffentlicht.
- Die Anleitung zu Installation und Administration findet sich in der README.
- Eine Dokumentation der technischen Architektur ist hier in unserem Repository zu finden.
- Alle weiteren Ressourcen (z.B. Dokumentation der HTTP/REST-API sowie der internen Rust-API) sind in der README verlinkt.

Kurze Erläuterung zur Einhaltung der Arbeits- und Kostenplanung

Die Arbeits- und Kostenplanung war überwiegend realistisch und musste nicht angepasst werden.

Kurze Darstellung von etwaigen Ergebnissen bei anderen Stellen

Der regelmäßige Austausch mit unseren Projektpartner*innen vom Community Broadcasting Archive (cba) und die Ausrichtung auf eine baldige Production-Instanz war sehr fruchtbar für die gute Zielorientierung der Entwicklung. Durch eine Förderung der netidee, die vom cba akquiriert werden konnte, wird derzeit eine eigene Trainingspipeline zur Verbesserung der Spracherkennungs-Modelle entwickelt, sowie die Sprachanalyse (NLP) zur automatischen Verschlagwortung verbessert.