Richtlinie zum „Software-Sprint“

CO2-Datenbank

Schlussbericht

Zuwendungsempfänger:

Lars Wissler Elias Philipp GbR

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS19S26 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Kurze Darstellung der Aufgabenstellung und Motivation

Der Klimawandel ist die größte Gefahr für die Menschheit und die Erde – wir alle müssen unseren Teil dazu beitragen, die globale Erwärmung zu reduzieren. Viele bemühen sich bereits um ein nachhaltiges Leben, doch wer nicht gut informiert ist, kann keine wissensbasierten Entscheidungen treffen. Die EU hat dieses Problem bereits erkannt und unter anderem die Direktive 014/95/EU ins Leben gerufen, welche große Unternehmen dazu verpflichtet, regelmäßig über ihre Nachhaltigkeitsbestrebungen und die Aufschlüsselung ihrer CO2-Produktion zu informieren. Leider sind die generierten Informationen ohne zentrale Sammel- und Anlaufstelle nicht hilfreich. Erst mit einer solchen können die Daten aufbereitet sowie verglichen werden und das ökologische Verhalten großer Unternehmen transparent gemacht werden. Hierfür sollte eine Webplattform inklusive Web-Crawler und Parser entwickelt werden, die Unternehmensberichte nach Veröffentlichungen von CO2-Emissionsdaten durchsucht, strukturiert und frei zugänglich anbietet.

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

Kernzielgruppe sind Nutzer, die in ihrem professionellen Umfeld umfassende, gut belegte und vergleichbare Daten zu CO2-Emissionen von Unternehmen benötigen. Insbesondere zielt das Projekt auf Journalisten, denen die Themenfindung und Berichterstattung zum Einfluss großer Unternehmen auf das Klima vereinfacht wird. Aus diesem Grund bietet das Portal direkt Integration in die populärste Datenvisualisierungsumgebung der Branche, Datawrapper.de, an. Jeder Datenpunkt der ethidat-Datenbank ist mit der entsprechenden Quelle hinterlegt und kann auch als CSV-Datei abgefragt werden. So wird transparenter, welche Unternehmen wirklich klimaneutraler agieren als zuvor und die Berichterstattung kann von harten Fakten untermauert werden.

Ausführliche Darstellung der Ergebnisse

Um unser Ziel zu erreichen, haben wir einen Web-Crawler entwickelt, der ausgehend von einer Liste von 25.000 Unternehmenswebseiten das Internet nach Nachhaltigkeitsberichten und Jahresabschlüssen durchsucht. Als Grundlage für den Crawler haben wir Nutch, die Open Source Crawling Library von Apache verwendet, die weitreichende und skalierbare Funktionalitäten bereitstellt für Crawling bereitstellt. Ausgehend von einer Seed-Liste an URLs werden alle URLs, die auf einer Seite vorkommen extrahiert und aufgerufen. Um mit den gecrawlten Seiten später arbeiten zu können, müssen die Ergebnisse gespeichert werden. Das heißt, Inhalt, URL und Metadaten wie Zeitpunkt des Aufrufs müssen in einer Weise in eine Datenbank aufgenommen werden, sodass performant nach Schlüsselwörtern auf den Homepages gesucht werden kann, ohne zu viel Festplattenspeicher zu verwenden. Dieser Prozess nennt sich Indexierung. Apache hat für diesen Zweck die Bibliothek Solr entwickelt, die optimal mit Nutch zusammenarbeitet, die Indexierungs-Aufgaben übernimmt und per API aus verschiedenen Sprachen, in unserem Fall Java mit SolrJ, abgefragt werden kann. Der Aufbau der Crawler-Architektur war der erste Meilenstein des Projektes, der wiederholte Verbesserungen erforderte, um die Indexierung angemessen schnell und zielgerichtet durchführen zu können. Der Crawler lief seit Mai konstant und konnte in diesem Zeitraum mehr als 50 Millionen Webseiten in einer Solr-Datenbank von 3 TB indexieren.

Die resultierenden Dokumente, meist PDFs, wurden durch eigens entwickelte Parser geschickt und die CO2-Daten nach Scopes und Unternehmen strukturiert ausgelesen und abgespeichert. Zu diesem Zweck haben wir eine Java-Applikation entwickelt. Mir der Bibliothek SolrJ wurde Java mit der Solr-Datenbank verbunden und nach Schlüsselwörtern wie „sustainability report“ oder „climate change report“ mögliche Seiten mit Emissionsdaten abgefragt. Diese Seiten wurden mittels der URL nun neu aufgerufen, denn der Solr-Index speichert nur den rohen Text ohne Metainformationen wie HTML oder PDF, die für den Parser nötig sind. War das Ziel der URL ein PDF, wurde der Text zeilenweise mit PDFBox, der Apache Library zum Lesen von PDF Dokumenten , ausgelesen. Für HTML-Dokumente wurde JSoup verwendet. Die gefundenen Dokumente waren „Echttext“-Dokumente, sodass zum Glück keine Texterkennung mittels OCR notwendig war. Die meisten relevanten Daten waren in Tabellen zu finden, sodass ein Großteil der Arbeit an dem Parser in die Entwicklung eines robusten Tabllenerkennungsfunktion geflossen ist, was insbesondere bei PDF Dokumenten eine anspruchsvolle Aufgabe ist. PDF Tabllen werden gezeichnet und sind nicht als Tabellen markiert. Das heißt eine Tabelle besteht letztlich nur aus Wörtern mit vielen Leerzeichen zwischen ihnen und einigen horizontalen und vertikalen Linien um die Wörter herum. Dabei sieht natürlich die Tabelle jedes Unternehmens verschieden aus, was die Erkennung sehr schwierig macht. Glücklicherweise haben viele der Tabellen eine ähnliche Struktur, insofern sie eine Entwicklung über mehrere Jahre in der Horizontalen und Dimensionen der Daten in der Vertikalen darstellen. Ohne diese grundlegende Struktur wäre ein robuster Parser kaum realisierbar gewesen. Ein zweiter wichtiger Punkt für den Parser war die Standardisierung des Emssions-Reportings in die Klassen direkte, indirekte und Dritt-Emissionen (Scope 1, Scope 2, Scope 3).

Dank dieser beiden Punkte war es möglich, Tabellen auch in PDF-Nachhaltigkeitsberichten mit einer Trefferquote von über 70% auszulesen und zu strukturieren. Ungefähr 20 % der Dokumente konnten von vornherein nicht abgefragt werden, da die Homepage des Unternehmens die automatische Anfrage blockierte. In etwa 50 % aller Nachhaltigkeitsberichte deutscher Unternehmen beinhalteten trotz der eingangs genannten EU-Richtlinien keine verwertbaren Informationen zu ihren Treibhaus-Emissionen. In diesen Fällen konnte auch ein Mensch die entsprechenden Daten nicht auslesen.

Im Laufe der Projektlaufzeit konnten wir die Daten von 300 deutschen und internationalen Unternehmen auslesen, die wir auf www.ethidat.de zur Ansicht und zu Download bereitstellen. Im Durchschnitt konnten wir pro Unternehmen eine Historie von 3 Jahren Emissionen auslesen, genug um erste Entwicklungen darzustellen. Subjektiv können wir feststellen, dass die Emissionen über diesen Zeitraum tendenziell konstant geblieben sind, was bei einem anhaltenden Wachstum letztlich auf eine leichte Verringerung der Emissionen pro Produktionseinheit schließen lässt. Auf Ethidat können diese Ergebnisse selbst eingesehen und entsprechende Schlüsse gezogen werden. Für die Seite haben wir den Mean-Stack (MongoDB, Express, Angular, NodeJS) verwendet. Ein Backend stellt API-Funktionalität für die Kommunikation mit der MongoDB bereit. Über diese API speicherte der Java Parser die in JSON strukturierten Ergebnisse ab und das Frontend greift so diese Daten ab. Wir haben eine Änderungsfunktion in die Webseite integriert, um schnell und unkompliziert in einem Community-Effort zu ermöglichen, fehlerhafte Datenpunkte zu korrigieren oder weitere hinzuzufügen und so die Möglichkeit geschaffen, die automatische Datenextraktion bei Bedarf manuell zu korrigieren. Für jeden Datenpunkt kann mit zwei Klicks eine Änderung angeregt werden mit dem korrigierten Wert und unter Angabe einer Quelle. Über einen solchen Antrag werden wir automatisch per Email benachrichtigt, können die Änderung prüfen und freigeben oder ablehnen.

Mit diesen Ergebnissen wurden alle geplanten Ziele erreicht. Im Rahmen der Projektarbeit konnten wir zusätzlich das Nutzerprofil deutlich schärfen bzw. verschieben. Im Laufe der Zeit und im Rahmen des projektinternen Austauschs und den Coachings zeichnete sich deutlich eine Kernzielgruppe von professionellen Nutzern wie Journalisten, Politikern oder NGOs, die beruflich Daten benötigen, anstatt, wie initial angedacht, den „normalen“ Bürger direkt ansprechen und informieren zu wollen. Deshalb wurde Kontakt mit Freien Journalisten u.A. aus dem Projektumfeld aufgenommen und deren Bedürfnisse direkt in das Portal eingearbeitet. Eine wichtige Erkenntnis war, dass eine fremde graphische Aufarbeitung für Journalisten nur begrenzt hilfreich ist, da jeder seine eigenen gewohnten Tools und persönliches Layout nutzt. Eine weit verbreitete Variante der graphischen Präsentation in der Branche ist www.datawrapper.de, bei der mittels API und CSV-Import Graphiken und Charts erstellt werden können und die unter anderem von der New York Times, dem Spiegel und der Zeit verwendet werden. Dementsprechend haben wir eine Anbindung dorthin bereitgestellt und uns nicht wie zuvor geplant auf die eigene exportierbare Aufbereitung konzentriert.

Zielgruppe, Nutzen und mögliche Weiterentwicklungen

Daten zur Entwicklung von Emissionsdaten von Unternehmen im Rahmen von Scope 1, 2 & 3 Emissionen zu bekommen wird deutlich vereinfacht. Anstatt die Daten manuell auf den Homepages oder in den Berichten der Unternehmen suchen zu müssen, gibt es nun eine zentrale Anlaufstelle, auf der die Daten graphisch aufbereitet gesammelt präsentiert werden und sie direkt als CSV oder in Datawrapper exportiert werden können. Die Ergebnisse und Methoden können von der Community weiterentwickelt werden. Insbesondere wurde das Portal so konzipiert, dass Nutzer neue Daten hinzufügen oder korrigieren können, belegt durch Quellen und nach einem Review. So kann sich im Wikipedia-Stil durch Crowdsourcing eine umfassende Datenbasis geschaffen werden. Perspektivisch soll die Datenbank nicht nur CO2-Daten beinhalten, sondern auch Daten zu Abfall und Recycling, Wasserverbrauch, Frauenquote oder Zertifizierungen und so zu einer allgemeinen Ethik-Datenbank anwachsen. Persönlich konnten wir unsere Programmierfähigkeiten erweitern. Highlights dabei waren die Extraktion von Tabellen aus PDF-Dateien und die Umsetzung einer durchklickbaren Tour auf der Homepage.

Kurze Darstellung der Arbeiten, die zu keiner Lösung geführt haben

Ursprünglich sollte deutlich breiter nach Unternehmensberichten gesucht werden. Nachdem die Größe des Internets klar wurde, mussten die Ausgangspunkte auf bekannte Unternehmenshomepages beschränkt werden, damit Größe und Dauer des Indexierungsprozesses in einem akzeptablen Rahmen blieb.

Kurze Angabe von Präsentationsmöglichkeiten für mögliche Nutzer

Auf www.ethidat.de oder unter www.demoweek.prototypefund.de finden sich alle Ergebnisse und Dokumente.

Kurze Erläuterung zur Einhaltung der Arbeits- und Kostenplanung

Die initial gemeinsam mit Prototype Fund erstellte „Road-Map“ der Software konnte in allen Meilensteinen erfüllt werden. Daher konnte der geplante Arbeitsaufwand und die geplanten Kosten eingehalten werden.

Kurze Darstellung von etwaigen Ergebnissen bei anderen Stellen

Uns sind bisher keine Entwicklungen in der Art von ethidat / Co2-Datenbank von anderen Personen bekannt. Auch von institutionelle Entwicklungen wurden wir wir nicht beeinflusst. Falls es die hier entwickelte Software institutionell gibt, ist diese bisher nicht einer breiten Öffentlichkeit zugänglich gemacht worden.