Richtlinie zum „Software-Sprint“

HubGrep – Meta-Suchmaschine für OpenSource Code Repositories

Schlussbericht

Zuwendungsempfänger:

Hartmann, Eliasson GbR

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01IS17S13 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Kurze Darstellung der Aufgabenstellung und Motivation

GitHub, betrieben von einer ClosedSource Software, in den Händen Microsofts, ist für viele die einzige Anlaufstelle für OpenSource Code Hosting. Teilweise wird GitHub (eine Hosting Platform) gar als Synonym zu git (einem Versionierungstool) benutzt. Dadurch sind viele Projekte, die nicht auf GitHub liegen, sondern beispielsweise auf selbst gehosteten GitLab oder Gitea Servern, praktisch unauffindbar.

Mit HubGrep erstellen wir einen Suchindex über viele Hosting Services, und bieten über eine Suchmaschine allen Repositories die Chance geben, gefunden zu werden, und die

Ergebnisse gleichwertig darstellen - egal von welcher Platform sie kommen.

Die Vorgehensweise war dabei wie folgt:

1. Auf Grundlage eines Kommandozeilen Prototyps wurde eine Meta-Suchmaschine als Webservice gebaut. Interessierte können ihre eigenen Hosting Instanzen hinzufügen (per Web Interface oder Pull Request), sofern der genutzte Service unterstützt wird.

2. Es wurde Dokumentiert, wie Interessierte eine eigene Instanz aufsetzen können, deployment Tools geschrieben.

3. Indexer und Crawler: Um die Limitierungen der Meta-Suche zu umgehen, haben wir Crawler geschrieben, die die Eckdaten der Repositories an einen Indexer schicken, um einen lokalen Index für schnellere Suche anzulegen und eventulles Rate Limiting der Dienste zu umgehen.

5. Die Meta-Suche (Meilenstein 1) wurde durch eine Suche auf Basis des Suchindexes ersetzt.

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

Durch eine Liste freier Hosting Services, die User selbst erweitern können, und dem entsprechenden Index aller Inhalte wurden Datensätze erzeugt, die es allen ermöglichen, vorher unzugängliche Projekte zu finden, oder die eigenen Projekte auffindbarer zu machen. Ausserdem können die veröffentlichten Rohdaten Grundlage für weitere Projekte sein.

Ausführliche Darstellung der Ergebnisse

Alle Meilensteine wurden erreicht.

Als Ergebnis des ersten Meilensteins entstand eine Metasuchmaschine (https://github.com/HubGrep/hubgrep_search/tree/legacy/metasearch), die benutzbar ist, und vorerst unter https://meta.hubgrep.io/ weiter betrieben wird.

Angefangen mit dem zweiten Meilenstein wurde das System dokumentiert. Die entstandene Dokumentation wird weiter gepflegt und ist unter https://docs.hubgrep.io/ erreichbar. Hier wird die Suchmaschine selbst, als auch das Ökosystem darum (Crawler und Indexer, Bedienung durch den Endnutzer) beschrieben.

Zum dritten Meilenstein wurden Crawler für drei Hosting Services geschrieben: Github, Gitea und Gitlab.

Ausserdem wurde ein "Indexer" als Backend geschrieben, in dem alle Daten gesammelt werden.

Als Ergebnis können hier die letzen Snapshots der Daten der Hoster abgerufen werden: https://indexer.hubgrep.io/

Der vierte Meilenstein befasste sich mit der Integration der gesammelten Daten in die Suchmaschine in Meilenstein 1. Der Plan war hier, die Suchergebnisse aus der Metasuche mit den Suchergebnissen aus dem eigenen Suchindex zu kombinieren. Da wir aber keine Möglichkeit haben, einen einheitlichen "Score" für externe Suchergebnisse und Suchergebnisse aus unserem eigenen Index zu bilden, wurde dies zugunsten der zuverlässigeren, schnelleren Suche im eigenen Index verworfen.

Zielgruppe, Nutzen und mögliche Weiterentwicklungen

Unsere Zielgruppe hat nun Zugriff auf eine Suchmaschine, die über viele Hoster hinweg OpenSource Repositories finden kann. Nutzer können ihre eigenen, oder ihnen bekannte Hosting Services hinzufügen, um ihre eigenen Projekte auffindbar zu machen.

Außerdem gibt es die Möglichkeit, auf Basis des gesammelten Datensatzes andere Projekte aufzubauen.

Durch die Open-Source-Stellung der Software können Nutzer selbst Crawler für neue Hosting Service Typen schreiben, Bugs fixen, neue Funktionen hinzufügen, oder den kompletten Stack selbst betreiben, um beispielsweise möglicher Zensur vorzubeugen.

Zur Weiterentwicklung gibt es eine Reihe von Möglichkeiten:

Da wir gerade erst angefangen haben, unser eigenes Such-Backend zu nutzen, liegt unsere Priorität auf dem Verbessern der Gewichtung und dem Sortieren der Ergebnisse. Eine andere naheliegende Erweiterung wäre, mehr Hosting-Services zu unterstützen und zu durchsuchen - beispielsweise Bitbucket.

Einige APIs stellen mehr Informationen bereit als andere, deshalb kennen wir beispielsweise die Lizenz oder Programmiersprache nicht von allen Projekten. Alle Projekte zu klonen, zu analysieren, und um die fehlenden Metadaten zu erweitern wäre eine ressourcen- und zeitintensive Aufgabe, aber wir haben eine Datenbasis, auf der wir aufbauen könnten. Mit den geklonten Projekten hätten wir dann, über die API Metadaten hinaus, auch alle Commit IDs und könnten über alle Hoster hinweg Forks und Mirrors von Projekten erkennen, und in den Ergebnissen gruppieren.

Kurze Darstellung der Arbeiten, die zu keiner Lösung geführt haben

Registrierungspflicht

Die erste veröffentlichte Version von HubGrep hatte eine Registrierungspflicht: um Hoster hinzuzufügen und später zu bearbeiten, musste ein Benutzerkonto erstellt werden. Der Grund dafür war, dass wir geplant haben, dass die Administratoren ihre Hoster selbst hinzufügen sollten.

In der Praxis war eine der ersten Beschwerden, dass dieses Vorgehen "nicht besonders Datensparsam" sei. Dazu kommt, dass für die Crawler, ausgenommen die Github Crawler, keine API Keys oder ähnliches benötigt werden.

Diese Funktion wurde also wieder entfernt, und jeder kann beliebige Services hinzufügen, solange sie von HubGrep unterstützt werden.

Suchproxy

Die erste Version von HubGrep war ein Proxyservice zu den Suchfunktionen der einzelnen Hoster. Da wir zwei Meilensteine später einen eigenen Suchindex aufgebaut hatten, und die Integration der Proxysuche mit der Suche im eigenen Index komplexer war als gedacht, und der eigene Index, selbst mit dem großen Github Datensatz, gut funktioniert, haben wir beschlossen ausschließlich den eigenen Suchindex zu nutzen.

Github Crawler

Aufgrund des Aufbaus der Github APIs war es nötig, verschiedene Varianten des Crawlers zu bauen.

Kurze Angabe von Präsentationsmöglichkeiten für mögliche Nutzer

Eine Instanz der Suchmaschine kann unter https://hubgrep.io aufgerufen und genutzt werden. Die Dokumentation zur Nutzung und zum Betrieb einer Instanz sind unter https://docs.hubgrep.io zu finden.

Der development-Blog befindet sich unter https://blog.hubgrep.io.

Der Code für die einzelnen Projekte und Dokumentation ist in der Github Organisation unter https://github.com/hubgrep zu finden.

Kurze Erläuterung zur Einhaltung der Arbeits- und Kostenplanung

Jan konnte nicht so viele Stunden wie geplant für das Projekt leisten. Stattdessen hat Henrik mehr Stunden geleistet als ursprünglich geplant, um dies zu kompensieren. Dadurch konnte der Zeitplan eingehalten werden.

Kurze Darstellung von etwaigen Ergebnissen bei anderen Stellen

Es gab keine Entwicklungen Dritter die Einfluss auf das Ergebnis hatten.