Technologisches Gewerbemuseum
Höhere Technische Lehranstalt für Informationstechnologie
Schwerpunkt Systemtechnik

Diplomarbeit

TextminR - Webscraping & API-Bereitstellung

Viele Texte einfach zugänglich gemacht

API-Bereitstellung
Yusuf Akalin 5BHIT

Textextrahierung aus Dateien
Benjamin Kissinger 5BHIT

Webscraping von Textdaten
Alexander Hauser 5BHIT

Betreuer: Prof. Dominik Dolezal
Ausgeführt im Schuljahr 2023/24

Abgabevermerk:
19.09.2023 Übernommen von:

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Ort, Datum	Yusuf Akalin
Ort, Datum	Benjamin Kissinger
Ort, Datum	Alexander Hauser

Kurzfassung

Es kann nicht daran gedacht werden, das Ziel zu erreichen, „auf die Sterne zu zielen." Im übertragenen wie im wörtlichen Sinne ist es eine Aufgabe, die Generationen zu beschäftigen. Und egal, wie viel Fortschritt man macht, es gibt immer den Nervenkitzel, nur anzufangen.

Es kann nicht daran gedacht werden, das Ziel zu erreichen, „auf die Sterne zu zielen." Sowohl im übertragenen Sinne als auch buchstäblich ist es eine Aufgabe, die Generationen zu beschäftigen. Und egal, wie viel Fortschritt man macht, es gibt immer den Nervenkitzel, nur anzufangen.

Für diejenigen, die die Erde aus dem Weltraum gesehen haben, und für die Hunderte und vielleicht Tausende mehr, die dies tun, verändert die Erfahrung ganz sicher Ihre Perspektive. Die Dinge, die wir in unserer Welt teilen, sind viel wertvoller als die, die uns trennen.

Abstract

There can be no thought of finishing for „aiming for the stars.“ Both figuratively and literally, it is a task to occupy the generations. And no matter how much progress one makes, there is always the thrill of just beginning.

There can be no thought of finishing for „aiming for the stars.“ Both figuratively and literally, it is a task to occupy the generations. And no matter how much progress one makes, there is always the thrill of just beginning.

For those who have seen the Earth from space, and for the hundreds and perhaps thousands more who will, the experience most certainly changes your perspective. The things that we share in our world are far more valuable than those which divide us.

Inhaltsverzeichnis

Kapitel 1

Vorwort

Die Diplomarbeit ist kein Aufsatz! Auch wenn sie interessant gestaltet werden sollte, ist sie unpersönlich und im passiv zu schreiben. Besonders sind die Quellenangaben, welche entsprechend gewählt und referenziert werden müssen. Innerhalb dieser Vorlage existieren 2 Dateien, die zu genau diesem Zweck erstellt wurden. Die Datei `bibliography.bib` beinhaltet alle Quellenangaben und verwendete Literatur, `glossaries.tex` alle Definitionen von Begriffen und Akronymen, welche in der Arbeit selbst nicht genauer erklärt werden.

1.1 Quellen

Das richtige zitieren spielt innerhalb der wissenschaftlichen Arbeit eine wichtige Rolle. Die Vorlage nutzt zur Verwaltung von Literatur ein Programm mit dem Namen `biblatex`. Mit diesem werden alle Einträge, welche sich in der Datei `bibliography.bib` befinden verarbeitet und können in der Arbeit selbst über das Kommando `\cite{key}` referenziert werden.

Als kleines Beispiel findet sich hier nun ein Zitat über Schall, aus dem ersten Phsyik Lehrbuch der Autoren, Schweitzer, Svoboda und Trieb.

> „Mechanische Longitudinalwellen werden als Schall bezeichnet. In einem Frequenzbereich von 16 Hz bis 20 kHz sind sie für das menschliche Ohr wahrnehmbar. Liegen die Frequenzen unter diesem Bereich, so bezeichnet man diese Wellen als Infraschall, darüber als Ultraschall." [2, S. 145]

Eine Referenz wie diese ist möglich, wenn der entsprechende Eintrag in der dafür vorgesehenen Datei vorhanden ist. In diesem Fall sieht die Definition der Quelle wie folgt aus:

```
@book{ physik1,
    title = {Physik 1},
    author = {Christian Schweitzer, Peter Svoboda, Lutz Trieb},
    year = {2011},
    subtitle = {Mechanik, Thermodynamik, Optik},
    edition = {7. Auflage},
    publisher = {Veritas},
    pages = {140, 145-150},
    pagetotal = {296}
}
```

Auflistung 1.1: Eintrag einer Buchquelle in BibLatex

Bei einem direkten Zitat empfiehlt es sich auch die Seitenzahl anzugeben. Dies kann über die Option des Kommandos `\code[S. Zahl]{key}` bewerkstelligt werden.

Nach der Verwendung einer Quelle, wird diese auch im Literaturverzeichnis gelistet, welche sich am Ende des Dokuments befindet.

Kapitel 2

Danksagung

Kapitel 3

Einleitung

Zu Beginn wird die Ausgangslage beschrieben, wobei interessant ist woher das Projekt kommt und welche Ansätze an dessen Konzept beteiligt waren. Hier werden auch Ziele gesetzt und Probleme bestimmt, welche in der Arbeit selbst eine große Rolle spielen.

Kapitel 4

Studie

Um eine effiziente API, die den NutzerInnen Texte in sekundenschnelle liefern kann und verschiedene wichtige Funktionalitäten wie das Einfügen von eigenen Texten oder auch das Suchen nach bestimmten Attributen der gespeicherten Daten umzusetzen, ist die Verwendung bestehender Technologien sinnvoll und verringert den Programmieraufwand.

Das Projekt besteht aus drei Teilbereichen, die API, die Textextrahierung und das Webscraping. Bei jeder dieser gibt es mehrere bestehende Frameworks, die für das Realisieren dieser Teilbereiche genutzt werden können.

4.1 API-Frameworks

Die Umsetzung der API kann in unterschiedlichsten Arten und Weisen erfolgen. Wichtig ist hierbei die Auswahl der richtigen Programmiersprache und des passenden Frameworks. Im folgenden werden zwei Ansätze analysiert. Eines davon ist die Umsetzung mittels Java und dem Spring-Framework. Die andere Option ist die Umsetzung mittels Python und dem FastAPI-Framework.

Warum Java bzw. Python?
Dafür gibt es mehrere Gründe. Mike Koder erwähnt in seiner Masterarbeit: "[...] the ability to automatically infer OpenAPI documentation were detected to reduce manual repetitive work."[1]. Dies zeigt, dass die Möglichkeit, eine automatische OpenAPI (früher bekannt als Swagger) Dokumentation zu generieren, den Arbeitsaufwand reduziert. Zusätzlich wird in der erwähnten Masterarbeit bestätigt, dass FastAPI und Spring unter den Frameworks mit den meisten Features sind und dass die Programmiersprache keinen großen Einfluss auf die Produktivität hat, sondern dass das gewählte Framework wichtiger ist. Für dieses Projekt kommen daher nur die beiden genannten Frameworks, Spring und FastAPI, infrage, da nur Erfahrung in Java und Python besteht.

4.1.1 Java & Spring

4.1.2 Python & FastAPI

4.1.3 Datenbanl

https://ieeexplore.ieee.org/abstract/document/7433710

4.2 Textextrahierung

4.3 Webscraping

Das Entwicklung eines Web

4.4 Fazit

Das Fazit soll zeigen, wie das Projekt durchgeführt werden kann.

Kapitel 5

Konzept

Nach der Definition der Problemstellungen und Ziel soll recherchiert werden, wie diese erreicht, beziehungsweise gelöst werden können. Diese Studie beschäftigt sich mit möglichen Lösungen und Technologien und analysiert deren Eigenschaften um konkrete Vor- und Nachteile zu finden. Beendet wird dieser Abschnitt mit einem Fazit.

Nachdem die Studie abgeschlossen und der Weg bestimmt ist soll nun ein Konzept oder eher noch ein Ablauf zur Lösung beschrieben werden. Hier finden sich Diagramme, Skizzen, Drehbücher, Mockups, ..., welche als Basis für die eigentliche Entwicklung verwendet werden.

Kapitel 6

Implementierung

Hier wird die Umsetzung des Projekts beschrieben und auf Details zu den einzelnen Technologien eingegangen. Im Optimalfall werden die Lösungen und Wege zu den zuvor definierten Problemen und Zielen geschildert. Eine bestehende Dokumentation, welche während der Arbeit erstellt wurde kann hier von großem Vorteil sein!

Kapitel 7

Retrospektive

Kurz vor dem Ende wird der Verlauf des Projekts analysiert und geprüft, ob die Ziele erreicht und die Probleme gelöst wurden. Es wird auch auf Schwierigkeiten eingegangen, welche erst während der Arbeit zum Vorschein kamen und es können Verbesserungsvorschläge und Erkenntnisse vorgetragen werden. Außerdem kann auch auf den weiteren Verlauf in der Zukunft eingegangen werden.

Kapitel 8

Conclusio

Hier findet eine letzte Zusammenfassung der Arbeit statt.

Literaturverzeichnis

[1] Mike Koder. „Increasing Full Stack Development Productivity via Technology Selection“. In: (2021).

[2] Christian Schweitzer, Peter Svoboda und Lutz Trieb. *Physik 1. Mechanik, Thermodynamik, Optik*. 7. Auflage. Veritas, 2011, S. 140, 145–150. 296 S.

Auflistungsverzeichnis