

Master-Thesis

Identifikation und Vergleich von Autorenangaben zu Software
zwischen verschiedenen Datenquellen

Eingereicht am: 16. September 2024

von: Kevin Jahrens
geboren am 05.08.1999
in Bad Oldesloe

Matrikelnummer: 480592

Betreuer: Prof. Dr. -Ing. Frank Krüger
Zweitbetreuer: M.A. Stephan Druskat

Master-Thesis

für: Herr **Kevin Jahrens**

Identifikation und Vergleich von Autorenangaben zu Software zwischen verschiedenen Datenquellen

Identifikation and comparison of authors of software across different data sources

Disposition

Software spielt eine zentrale Rolle in der Wissenschaft und sollte daher in wissenschaftlichen Arbeiten zitiert werden. Insbesondere für Autoren wissenschaftlicher Software ist die Zitation wesentlicher Bestandteil der wissenschaftlichen Anerkennung, sodass diese auch zunehmend in wissenschaftlichen Lebensläufen genannt werden und Beachtung finden. Anders als bei wissenschaftlichen Publikationen ist bei wissenschaftlicher Software aktuell noch unklar, welcher Anteil an der Entwicklung zu einer Nennung als Autor führt. Darüber hinaus existieren in verschiedenen Datenquellen widersprüchliche Angaben für Zitationsvorschläge bzgl. der Autoren einer Software.

Ziel dieser Masterarbeit ist es zu untersuchen inwieweit sich die Angaben von Autoren für Open Source Software unterscheiden. Dazu sollen öffentlich verfügbare Repositorien mit R und Python Paketen – als Stellvertreter für wissenschaftliche Software – hinsichtlich ihrer Autorenangaben untersucht werden. Insbesondere sollen die angegebenen Metadaten in den Repositorien (z.B. citation.cff) mit den Metadaten in Paketdatenbanken (<https://pypi.org/> und <https://cran.r-project.org/>) und den Entwicklungsanteilen automatisch verglichen werden.

1. Literaturrecherche Autorenrolle in Open Source Software und zur Disambiguierung von Autorennamen
2. Datensammlung: Identifikation und Download verfügbarer Metadaten zu „wichtigen“ Softwarepaketen
3. Automatische Auflösung und Abgleich der Autorennennungen aller Datenquellen
4. Analyse von Unterschieden in der Nennung von Autoren
5. Dokumentation der Ergebnisse in einer schriftlichen Master-Thesis

Startdatum: 16.09.2024
Abgabedatum: 17.03.2024



Prof. Dr. rer. nat. Litschke
Chairman of the Examination Committee



Prof. Dr.-Ing. Krüger
Supervisor

Abstract

Maximal eine halbe Seite.

Inhaltsverzeichnis

1	Einleitung	5
1.1	Motivation	5
1.2	Vorgehen	5
1.3	Gliederung	5
2	Grundlagen	6
2.1	Versionsverwaltung	6
2.2	Paketverwaltung	6
2.3	Zitationsformate	6
2.3.1	Citation File Format	6
2.3.2	BibTeX	6
2.4	Named Entity Recognition	6
2.5	Entity Resolution/ Author name disambiguation	6
2.6	Fuzzy suche	6
3	Methodik	7
3.1	Datenbeschaffung	7
3.2	Limitierungen	7
3.3	Abgleich	7
4	Ergebnisse	8
5	Diskussion	9
6	Fazit und Ausblick	10
6.1	Fazit	10
6.2	Ausblick	10
	Anhang A Beispielanlage	11
	Abbildungsverzeichnis	12
	Tabellenverzeichnis	13
	Algorithmenverzeichnis	14
	Quellcodeverzeichnis	15
	Abkürzungsverzeichnis	16
	Selbstständigkeitserklärung	18

1 Einleitung

1.1 Motivation

1.2 Vorgehen

1.3 Gliederung

2 Grundlagen

2.1 Versionsverwaltung

2.2 Paketverwaltung

2.3 Zitationsformate

2.3.1 Citation File Format

2.3.2 BibTeX

2.4 Named Entity Recognition

2.5 Entity Resolution/ Author name disambiguation

2.6 Fuzzy suche

3 Methodik

3.1 Datenbeschaffung

3.2 Limitierungen

3.3 Abgleich

4 Ergebnisse

5 Diskussion

6 Fazit und Ausblick

6.1 Fazit

6.2 Ausblick

A Beispielanlage

Beispieltext.

Abbildungsverzeichnis

Tabellenverzeichnis

Algorithmenverzeichnis

Quellcodeverzeichnis

Abkürzungsverzeichnis

DoS Denial of Service. *Glossar*: Denial of Service

Datenträger

```
/.....Wurzelverzeichnis
├── OrdnerA..... Ein Ordner auf dem Datenträger
│   ├── OrdnerB..... Ein Unterordner auf dem Datenträger
│   │   └── datei.xyz..... Eine Datei
│   └── thesis.pdf..... PDF-Datei dieser Bachelor-Thesis
```

Im Unterverzeichnis `tools` des Projekts findet sich das Perl-Skript `dirtree.pl`, mit welchem Inhalte für das `dirtree`-Environment (siehe oberhalb) semiautomatisch erstellt werden können.

Die Nutzung aus der Kommandozeile ist wie folgt:

```
perl dirtree.pl /path/to/top/of/dirtree
```

Quelle des Skripts:

<https://texblog.org/2012/08/07/semi-automatic-directory-tree-in-latex/>

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die anderen Quellen im Wortlaut oder dem Sinn nach entnommen wurden, sind durch Angaben der Herkunft kenntlich gemacht. Dies gilt auch für Zeichnungen, Skizzen, bildliche Darstellungen sowie für Quellen aus dem Internet.

Ich erkläre ferner, dass ich die vorliegende Arbeit in keinem anderen Prüfungsverfahren als Prüfungsarbeit eingereicht habe oder einreichen werde.

Die eingereichte schriftliche Arbeit entspricht der elektronischen Fassung. Ich stimme zu, dass eine elektronische Kopie gefertigt und gespeichert werden darf, um eine Überprüfung mittels Anti-Plagiatssoftware zu ermöglichen.

A handwritten signature in black ink, reading "Kevin Zehner". The signature is written in a cursive style with a large initial 'K' and a long, sweeping underline.

Wismar, den 16. September 2024

Ort, Datum

Unterschrift