

Fakultät Informatik

My very fancy thesis title.

Bachelorarbeit im Studiengang Informatik

vorgelegt von

Lukas Welsch

Matrikelnummer 3201095

Erstgutachter:	Prof. Dr. Korbinian Riedhammer
Zweitgutachter:	Prof. Dr. Jens Albrecht
Betreuer:	Dipl. Ing. Andreas Sachs
Unternehmen:	Sopra Financial Technology GmbH

© 2020

Dieses Werk einschließlich seiner Teile ist **urheberrechtlich geschützt**. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Autors unzulässig und strafbar. Das gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen sowie die Einspeicherung und Verarbeitung in elektronischen Systemen.

Prüfungsrechtliche Erklärung der/des Studierenden

Angaben des bzw. der Studierenden:

Name: Welsch

Vorname: Lukas


Matrikel-Nr.: 3201095

Fakultät: Informatik



Studiengang: Informatik



Semester: Wintersemester  2020/21

Titel der Abschlussarbeit:

Analyse und Visualisierung der Datenqualität innerhalb eines Data Warehouse

Ich versichere, dass ich die Arbeit selbständig verfasst, nicht anderweitig für Prüfungszwecke vorgelegt, alle benutzten Quellen und Hilfsmittel angegeben sowie wörtliche und sinngemäße Zitate als solche gekennzeichnet habe.

Muss noch ausgefüllt

Ort, Datum, Unterschrift Studierende/Studierender

Erklärung zur Veröffentlichung der vorstehend bezeichneten Abschlussarbeit

Die Entscheidung über die vollständige oder auszugsweise Veröffentlichung der Abschlussarbeit liegt grundsätzlich erst einmal allein in der Zuständigkeit der/des studentischen Verfasserin/Verfassers. Nach dem Urheberrechtsgesetz (UrhG) erwirbt die Verfasserin/der Verfasser einer Abschlussarbeit mit Anfertigung ihrer/seiner Arbeit das alleinige Urheberrecht und grundsätzlich auch die hieraus resultierenden Nutzungsrechte wie z.B. Erstveröffentlichung (§ 12 UrhG), Verbreitung (§ 17 UrhG), Vervielfältigung (§ 16 UrhG), Online-Nutzung usw., also alle Rechte, die die nicht-kommerzielle oder kommerzielle Verwertung betreffen.

Die Hochschule und deren Beschäftigte werden Abschlussarbeiten oder Teile davon nicht ohne Zustimmung der/des studentischen Verfasserin/Verfassers veröffentlichen, insbesondere nicht öffentlich zugänglich in die Bibliothek der Hochschule einstellen.

Hiermit ☒ genehmige ich, wenn und soweit keine entgegenstehenden Vereinbarungen mit Dritten getroffen worden sind,
☐ genehmige ich nicht,

dass die oben genannte Abschlussarbeit durch die Technische Hochschule Nürnberg Georg Simon Ohm, ggf. nach Ablauf einer mittels eines auf der Abschlussarbeit aufgebrachten Sperrvermerks kenntlich gemachten Sperrfrist

von Jahren (0 - 5 Jahren ab Datum der Abgabe der Arbeit),

der Öffentlichkeit zugänglich gemacht wird. Im Falle der Genehmigung erfolgt diese unwiderruflich; hierzu wird der Abschlussarbeit ein Exemplar im digitalisierten PDF-Format auf einem Datenträger beigelegt. Bestimmungen der jeweils geltenden Studien- und Prüfungsordnung über Art und Umfang der im Rahmen der Arbeit abzugebenden Exemplare und Materialien werden hierdurch nicht berührt.

Ort, Datum, Unterschrift Studierende/Studierender

Formular drucken

Datenschutz: Die Antragstellung ist regelmäßig mit der Speicherung und Verarbeitung der von Ihnen mitgeteilten Daten durch die Technische Hochschule Nürnberg Georg Simon Ohm verbunden. Weitere Informationen zum Umgang der Technischen Hochschule Nürnberg mit Ihren personenbezogenen Daten sind unter nachfolgendem Link abrufbar: <https://www.th-nuernberg.de/datenschutz/>

Kurzdarstellung

Kurze Zusammenfassung der Arbeit, höchstens halbe Seite. Deutsche Fassung auch nötig, wenn die Arbeit auf Englisch angefertigt wird.

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

Abstract

Only if thesis is written in English.

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

Inhaltsverzeichnis

1. Einleitung	1
1.1. This is an Important Section	1
1.1.1. And an even more important subsection	1
2. Grundlagen	3
2.1. Definition von Datenqualität	3
2.1.1. Datenqualitätsdimensionen	3
2.1.2. Definition / Metriken der Dimensionen	4
2.1.3. Visualisierungen / Charts zur Darstellung der Metriken	4
3. Daten	5
4. Methoden	7
4.1. Metriken	7
4.2. Bestehende Verfahren	7
4.3. Stakeholderanalyse	7
5. Experimente	9
6. Ausblick	11
7. Fazit	13
A. Supplemental Information	15
Abbildungsverzeichnis	17
Tabellenverzeichnis	19
Auflistung	21
Literaturverzeichnis	23
Glossar	25

Kapitel 1.

Einleitung

You can also write footnotes.¹

Wie kann die Daten Qualität aufbereitet und visualisiert werden, sodass Entscheider schnell und gezielt Maßnahmen einleiten können? -> Wie visualisiert man am besten? -> Wie kann man die DQ messen? (Am besten ohne großen Aufwand durch die Experten, also möglichst viel automatisch machen) ->

versch. Arten visualisieren und dann selbst entscheiden bzw Entscheidungen von Experten mit einfließen lassen Anpassung der Adressen aus den letzten Jahren (Durchschnitt wie oft wird Kundenstammdaten aktualisiert)

1.1. This is an Important Section

It is possible to reference glossary entries as library as an example.

1.1.1. And an even more important subsection

¹Footnotes will be positioned automatically.

Kapitel 2.

Grundlagen

2.1. Definition von Datenqualität

Datenqualität wird in der Literatur auf verschiedene Arten definiert. Viele Datenqualitätsstudien verwenden Korrektheit als einziges (oder als Haupt-) Datenqualitätsmerkmal. [Wang 96] Datenqualität umfasst jedoch nicht nur die Korrektheit von Daten, sondern auch anderen Dimensionen. Ein Paper von 2011 zeigt den Einfluss von anderen Dimensionen, wie zb. Vollständigkeit auf die Erkennungsrate von Klassifikatoren [Espí 11] Die Autoren schlagen vor den Aufbereitungsprozess und vor allem Datenqualität nicht nur auf die Korrektheit von Daten zu beziehen sondern auch auf die anderen Datenqualitätsdimensionen zu achten.

2.1.1. Datenqualitätsdimensionen

Die Dimensionen Richtigkeit, Vollständigkeit, Konsistenz und Aktualität werden in den meisten Veröffentlichungen genannt, allerdings gibt es keinen Standard, weder im Bezug auf die verwendeten Dimensionen, noch die Definition der Dimensionen. [Scan 02] **Begründung finden! In dieser Arbeit werden diese Definition von Datenqualität verwendet, da es sich am Besten anbietet?** Jede dieser Datenqualitätsdimensionen umfasst eine Facette der Datenqualität.

Richtigkeit Richtigkeit wird als Gleichheit zwischen zwei Werten definiert, sodass die Daten die Wirklichkeit korrekt repräsentieren. Beispiel: Eine Person mit dem Namen Max Mustermann ist als Mx Mustermann abgespeichert. Die gespeicherten Daten repräsentieren nicht die Wirklichkeit, sie sind somit nicht richtig.

Aktualität / temporäre Richtigkeit Als Spezialisierung der Richtigkeit kann die temporäre Richtigkeit gesehen werden, die richtige Repräsentation zu einem bestimmten Zeitpunkt oder Zeitspanne beschreibt. Die temporäre Richtigkeit zum aktuellen Zeitpunkt wird auch als Aktualität bezeichnet.

Vollständigkeit Die Daten sind umfangreich genug für die jeweilige Aufgabe. [Wang 96] Da ein Data Warehouse aus relationalen Datenbanken besteht, wird die Vollständigkeit im

Folgenden auf relationale Datenbanken bezogen. Die Vollständigkeit wird von [Pipi 02] in drei Klassen kategorisiert.

- Schema: Grad der Daten, die nicht im Schema fehlen
- Spalte: Anzahl der fehlenden Werte innerhalb einer Spalte
- Population: Wenn eine Spalte alle Institute (36,70..) beinhalten sollte, aber welche fehlen, dann herrscht Populationsunvollständigkeit

Konsistenz

2.1.2. Definition / Metriken der Dimensionen

Um diese eher abstrakten Definitionen messbar zu machen haben sich einige Verfahren etabliert. Diese lassen sich grob in die Kategorien subjektiv und objektiv aufteilen. [Pipi 02] Bei den subjektiven Verfahren wird die vorherrschende Datenqualität durch Experten geschätzt. Dieses Verfahren ist allerdings nicht nur zeitaufwendig für die Experten, sondern ist auch fehleranfällig. **HIERZITATANGEBEN** []

Eine weitere Möglichkeit besteht darin objektive Verfahren zu verwenden, bei denen mit Hilfe von mathematischen Funktionen die Datenqualität berechnet oder geschätzt wird.

Des weiteren gibt es ein kombiniertes Verfahren, bei denen Grenzwerte von Experten geschätzt und anschließend durch mathematische Funktionen abgeglichen werden.

2.1.3. Visualisierungen / Charts zur Darstellung der Metriken

Sollte zum Beispiel sehr gut interpretierbar sein. Genaue Vorgaben zu den Visualisierungen müssen allerdings von den Stakeholdern erfragt werden

Kapitel 3.

Daten

Die in der Arbeit verwendeten Daten stammen aus dem Data Warehouse einer deutschen Bank. Aufgrund des Datenschutzes wurden personenbezogene Daten pseudonymisiert und anonymisiert.

Ein- und Ausschlusskriterien für die Daten:

Kontaktdaten StammdatenKunden

Kreditkartenkonten Sparkonten -> wie viele Prozent würden den Vorgaben entsprechen? -> welche Dimensionen eignen sich für verschiedene Datensätze? -> Experte muss sagen, wie schnell veraltet irgendwas für die Aktualität

Visualisierung auf zwei Arten: Major critical (Sonar) 1. Visualisierung der Metriken (Ergebnisse) 2. Visualisuerng der Daten selbst, zb Nul Values gelb markiert

Kapitel 4.

Methoden

Um die nachfolgenden Methoden zur Analyse und Visualisierung von Datenqualität, zu vergleichen werden sowohl qualitative als auch quantitative Untersuchungen durchgeführt.

Nach [Pipi 02] besteht die Schwierigkeit nicht darin die Metriken zu formulieren, sondern die Datenqualitätsdimension zu definieren, die auf den spezifischen Anwendungsbereich des Unternehmens passt.

Metriken:

4.1. Metriken

Simple Ratio. Desired outcomes to total outcomes -> geeignet für Richtigkeit, Vollständigkeit, Konsistenz

4.2. Bestehende Verfahren

Richtigkeit Um die Richtigkeit zu berechnen gibt es zwei verschiedene Möglichkeiten.

4.3. Stakeholderanalyse

Nach [Pipi 02] gibt es drei Stakeholder: the collector, custodians and consumers of data products

Es besteht die Möglichkeit durch einen Fragebogen die vorherrschende Datenqualität mit Hilfe der Stakeholder zu berechnen. [Pipi 02]

In this chapter, we're actually using some code!

```
1 x = 1
2 if x == 1:
3     # indented four spaces
4     print("x is 1.")
```

Listing 4.1: This is an example of inline listing

You can also include listings from a file directly:

```
1 x = 1
2 if x == 1:
3     # indented four spaces
4     print("x is 1.")
```

Listing 4.2: This is an example of included listing

Kapitel 5.

Experimente

Aufbau Experimente: Ziele* Aufbau* Ergebnisse* Interpretation* Threats*to*Validity (Seite 93 <https://userpages.uni-koblenz.de/laemmel/esecourse/slides/perf.pdf>)

Ideen: - Komplexe Funktionen mit Stakeholdern basteln, zb wenn verheiratet dann Alter > 18 - Daten für zb Aktualität müssen definiert werden, ob sie beispielsweise überhaupt verfallen können. Zb Geburtsdatum ändert sich nie; Alter schon - Ist es möglich solche Regeln mit Hilfe von ML abzuleiten oder funktioniert das gar nicht? - Daten vor einem Monat berechnen, wie viele sich ändern müssten (aufgrund von zb Timeliness, correctness) und dann nachschauen wie viele sich tatsächlich geändert haben - Big Data Quality A Quality Dimension evaluation hat zwei konkrete Experimente, dort kann man sich gute Ideen holen. Es wird auch ein Experte zu Rate gezogen, der beispielsweise angibt, welche Daten gelöscht werden können (zb wenn 80% der Attribute fehlen). Es hat auch einige Visualisierungen - Mit SQL: <https://dataform.co/blog/advanced-data-quality-testing>

Fragen: - Woher richtige Daten bekommen?

Auf die verschiedenen Ebenen Aktualität, Richtigkeit, Vollständigkeit und Konsistenz eingehen! Oft ist es besser die Daten nachzufordern, anhand eines möglichen Fehlers kann nicht der Originalzustand wiederhergestellt werden

- Welche Visualisierungen bieten sich an? - Gibt es evtl Visualisierungen, die DQ-Probleme aufzeigen?

Kapitel 6.

Ausblick

Kapitel 7.

Fazit

Anhang A.

Supplemental Information

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

Das hier ist der zweite Absatz. Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

Und nun folgt – ob man es glaubt oder nicht – der dritte Absatz. Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

Nach diesem vierten Absatz beginnen wir eine neue Zählung. Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

Dies hier ist ein Blindtext zum Testen von Textausgaben. Wer diesen Text liest, ist selbst schuld. Der Text gibt lediglich den Grauwert der Schrift an. Ist das wirklich so? Ist es gleichgültig, ob ich schreibe: „Dies ist ein Blindtext“ oder „Huardest gefburn“? Kjift – mitnichten! Ein Blindtext bietet mir wichtige Informationen. An ihm messe ich die Lesbarkeit einer Schrift, ihre Anmutung, wie harmonisch die Figuren zueinander stehen und prüfe, wie breit oder schmal sie läuft. Ein Blindtext sollte möglichst viele verschiedene Buchstaben enthalten und in der Originalsprache gesetzt sein. Er muss keinen Sinn ergeben, sollte aber lesbar sein. Fremdsprachige Texte wie „Lorem ipsum“ dienen nicht dem eigentlichen Zweck, da sie eine falsche Anmutung vermitteln.

Abbildungsverzeichnis

Tabellenverzeichnis

Auflistung

4.1. This is an example of inline listing	8
4.2. This is an example of included listing	8

Literaturverzeichnis

- [Espi 11] R. Espinosa Oliva, J. Zubcoff, and J.-N. Mazón. *A Set of Experiments to Consider Data Quality Criteria in Classification Techniques for Data Mining*. June 2011.
- [Pipi 02] L. L. Pipino, Y. W. Lee, and R. Y. Wang. “Data Quality Assessment”. *Communications of the ACM*, Vol. 45, No. 4, pp. 211–218, Apr. 2002.
- [Scan 02] M. Scannapieco and T. Catarci. “Data Quality under a Computer Science Perspective”. *Journal of The ACM - JACM*, Vol. 2, Jan. 2002.
- [Wang 96] R. Y. Wang and D. M. Strong. “Beyond Accuracy: What Data Quality Means to Data Consumers”. *Journal of Management Information Systems*, Vol. 12, No. 4, pp. 5–33, March 1996.

Glossar

library A suite of reusable code inside of a programming language for software development. i, 1

shell Terminal of a Linux/Unix system for entering commands. i