Erkennung und Korrektur von Einlesefehlern in DNA-Sequenzen

Abschlussarbeit

zur Erlangung des akademischen Grades:

Bachelor of Science (B.Sc.)

an der

Hochschule für Technik und Wirtschaft (HTW) Berlin
Fachbereich 4: Informatik, Kommunikation und Wirtschaft
Studiengang *Angewandte Informatik*

1. Gutachter_in: Herr Prof. Dr. Christian Herta
2. Gutachter_in: Herr Dr. Christian Krumnow

Eingereicht von Fabian Vogt [Matrikelnr. s0570800]

10.12.2022

Danksagung

[Text der Danksagung]

Zusammenfassung

[Text der Zusammenfassung]

Abstract

[Summary of the thesis]

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Listings

.

1 Einleitung

Vorliegendes Template enthält exemplarisch (und damit unvollständig) Gliederungspunkte, Bestandteile und Hinweise für ein typisches Softwareentwicklungsprojekt, bei dem ein Prototyp erstellt wird. Es dient als Hilfestellung zu Ihrer weiteren Verwendung. Selbstverständlich müssen Sie selbst weitere Ergänzungen und Anpassungen vornehmen.

Viel Erfolg sowie gutes Gelingen bei Ihrer Abschlussarbeit!

Der Textteil beginnt hier und wird arabisch mit dieser Seite beginnend mit »1« arabisch nummeriert. Der Textteil gliedert sich in Kapitel und Unterkapitel. Soll jede Hierarchieebene benannt werden, dann ist folgende Terminologie üblich:

- 1. Hierarchieebene: Hauptkapitel
- 2. Hierarchieebene: Kapitel
- 3. Hierarchieebene: Unterkapitel
- 4. Hierarchieebene: Abschnitt

Der inhaltliche Aufbau einer Abschlussarbeit im Studiengang *Angewandte Informatik* hängt selbstverständlich vom Thema und vom Inhalt ab. Abweichungen von der diesem Template zu Grunde liegenden Gliederungsstruktur sind immer möglich, manchmal sogar zwingend notwendig. Stimmen Sie sich diesbezüglich immer mit Ihren Gutachter(inne)n ab.

Vergessen Sie niemals, all Ihre verwendeten Quellen anzugeben und korrekt zu zitieren[1]. Quellen können manuell referenziert und im Quellenverzeichnis eingetragen werden. Ergänzend bieten viele Textverarbeitungssssteme auch ausgelagerte Quellenverwaltungsdateien und - systeme an, über die mittels entsprechender Befehle im Textteil zitiert werden kann[2].

Visualisieren Sie im Textteil angemessen, z.B. mittels Abbildungen und Tabellen. Vorliegendes Template enthält beispielhaft eingebundene Abbildungen und eine Tabelle (vgl. f.), welche der Steinlausforschung[3] entnommen sind.

[1] Ergänzende Informationen können Sie auch in eine Fußnote auslagern. Hier wird die Fußnote dazu genutzt, um Ihnen bei Interesse am Thema Zitation vertiefende Quellen (z.B. [**balzert2011**] oder [**franck2013**]) anzubieten.

[2] Wie Sie hoffentlich feststellen werden, erfolgt die Literaturverwaltung in diesem Template mittels einer *.bib-Datei (diese enthält die verwendeten Quellen), welche die *.tex-Datei mittels Verwendung von biblatex und bibtex ergänzt.

[3] Analog zu Straube (In: [**pschy**]) handelt es sich bei der Steinlaus (*petrophaga lorioti*) um das »*kleinste einheimische Nagetier*«. Als stimmungsaufhellender Endoparasit erreicht es eine Größe von ca. 0,3 bis 3 mm und stammt aus der Familie der Lapivora. Die Steinlaus kommt ubiquitär vor und ist in der Regel apathogen.

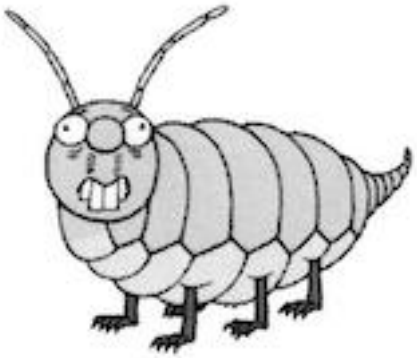

Abbildung 1.1: *Beispielgrafik: Steinlaus; Bildquelle [***loriot***]*

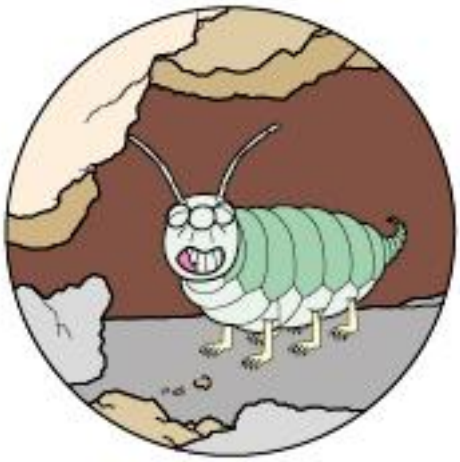

Abbildung 1.2: *Beispielgrafik: Fressende Steinlaus; Bildquelle [***loriot2***]*

Tabelle 1.1: *Übersicht: Untersuchte Steinläuse*

Untersuchte Objekte mit Lokation des Habitats		
ID (nickname)	Ort	Größe/Länge (in mm)
1 (Rosalinde)	Berlin, Mauerpark	1.4
2 (Devil in disguise)	Brandenburg, BER-Airport	2.8
3 (Hannes)	Berlin, Olympia-Stadion	2.1
4 (Her Majesty)	Berlin, Humboldt-Forum	2.0

1.1 Motivation

Bei der DNA-Sequenzierung wird eine Folge von Basen aus einem vorliegenden DNA-Sample extrahiert. Die vier Basen lauten: Adenin, Guanin, Thymin und Cytosin. Sie werden üblicherweise durch ihre Anfangsbuchstaben A, T, G, C abgekürzt. Dabei entstehen Fehler mit einer bestimmten Wahrscheinlichkeit, die von der Art des verwendeten Sequenzierungsverfahren abhängt. Wenn beim Einlesen einer einzelnen Base ein Fehler passiert, wird diese mit einer anderen Base verwechselt. Zum Beispiel wird ein A plötzlich zu einem G.

Beim gängigsten Sequenzierungsverfahren, dem Next-Generation-Sequencing *NGS*, treten Fehler mit einer Wahrscheinlichkeit von 0.1% auf (source). Meistens haben solche Fehler nur wenig Einfluss auf eine Diagnose, die aus der fehlerhaften Sequenzierung entstanden ist (source). Wenn aus einem fehlerhaften DNA-Read jedoch eine Fehldiagnose entsteht, können die Konsequenzen verherrend sein (source). Vor allem, wenn in Folge auf die Diagnose die falschen Medikamente verabreicht werden.

Daher ist es wichtig, dass fehlerhafte DNA-Reads korrigiert werden. Entweder muss der Fehler direkt korrigierbar sein, oder er muss wenigstens erkannt werden, sodass der Teil der Sequenz noch einmal eingelesen werden kann. [Beschreibung des groben Kontextes der Arbeit; im Detail sollten Sie dies im Grundlagenteil darstellen]

1.2 Problem- und Zielstellung (Scope)

Es wird untersucht, ob fehlerhafte DNA-Reads mit Hilfe eines Reinforcement-Learning Algorithmus korrigiert werden können. Anstatt die DNA-Sequenz direkt als Input zu nutzen, wird das ProcBert Transformer Model genutzt: Die fehlerhafte Sequenz wird in das Bert-Model gefüttert und anschlieÃend werden die hidden states des BERT-Models extrahiert. Diese hidden states enthalten nicht nur Informationen über die aktuelle Base, sondern auch über Gesamtsequenz. Es ist die codierte Repräsentation der Base im Kontext der Gesamtsequenz. In der Theorie könnte ein einziger hidden state bereits genug Informationen enthalten, um entscheiden zu können ob dahinter eine fehlerhafte Base steckt.

Das Problem wird modeliert, indem (unverfälschte) menschliche DNA-Reads künstlich verfälscht werden. Daraus entsteht ein Trainingsdatensatz aus fehlerhaften DNA-Sequenzen und ihren jeweils unverfälschten Originalen.

Dieser Trainingsdatensatz wird genutzt, um Reinforcement-Learning Agenten zu trainieren.

Als Ergebnis soll eine Beurteilung entstehen, ob und wie weit sich ein Reinforcement-Learning-Ansatz zur Lösung des Problems eignet.

[Beschreibung der Problemstellung sowie der sich daraus ergebenden Teilprobleme,-ziele und Forschungsfrage(n), welche Sie mit Ihrer Arbeit addressieren]

1.3 Aufbau der Arbeit

Die Arbeit beginnt mit der Erarbeitung der theoretischen Konzepte. Es werden grundlegende Fragen beantwortet: Was ist DNA-Sequnzierung? Wie können dabei Fehler

entstehen? Wie funktioniert das Transformer Model? [Beschreibung des Aufbaus der Arbeit]

2 Theoretische Grundlagen

[Beschreibung des Kontextes der Arbeit mit allen durch die Problemstellung tangierten Bereichen, Methoden, Theorien, Erkenntnissen, Technologien, ...]

2.1 Kontext

2.1.1 Domain

2.1.2 Technologien

2.1.3 Methoden und Konzepte

2.2 DNA-Sequenzierung

2.2.1 NGS

2.2.2 Entstehung von Fehlern

2.3 Machine Learning für DNA

2.3.1 NLP

Das Natural Language Processing ist ein Teilgebiet des Maschinellen Lernens, dass sich mit sequenziellen Daten im Bereich Sprachen beschäftigt. Dazu werden meist Text- und Audiodaten verwendet. Es spielt jedoch keine Rolle, in welchem Format die Daten eingespeiÃt werden, solange sie bestimmte Sprachmuster aufweisen.

In jeden Fall handelt es sich beim Input für NLP um sequentielle Daten. Einzelteile der Sequenz (Wörter) stehen in Beziehung zu anderen Wörtern. Sie sind mit einer gewissen Syntax angeordnet und ergeben zusammen einen Satz.

2.3.2 Transformer

Der Transformer ist eine

2.3.3 ProcBert

2.4 Reinforcement Learning

2.4.1 Grundlagen

2.4.2 Bekannteste Anwendungsbeispiele

2.4.3 Anwendung zur DNA-Fehlererkennung

2.4.4 DNA-Korrektur Environment

2.4.5 Unterschiedliche Ansätze bei der Korrektur

2.5 ...

2.5.1 ...

2.5.2 ...

3 Methodologie

[Beschreibung des geplanten Vorgehens(-modells) zur Lösung der Problemstellung; umfasst u.a.:

- Anforderungserhebung und -analyse
- Konzeption, Entwurf
- Umsetzung (Implementierung)]

3.1 Ergebnisartefakte

[Beschreibung der Ergebnisse / Ergebnistypen, welche Sie im Rahmen der Problemlösung generieren / erzielen wollen, z.B. Algorithmus, Prototyp einer Software(komponente), ...]

3.2 Datenschutzaspekte

[Beschreibung von Aspekten des Datenschutzes im Zusammenhang mit Ihrer Abschlussarbeit]

3.3 Ethische Aspekte

[Beschreibung von Aspekten der Ethik[1] im Zusammenhang mit Ihrer Abschlussarbeit]

[1] vgl. hierzu ergänzend allgemeine Codizes (z.B. [**acm**], [**ieee**] oder [**gi**]) sowie auch domain-spezifische Normen und Verfahrensweisen im Rahmen einer kritischen Reflektion.

4 Anforderungserhebung und -analyse

[Beschreibung der Erhebung, Granularisierung und Priorisierung der zu Grunde liegenden Anforderungen]

4.1 Nutzer- und Systemanforderungen

4.1.1 Funktionale Anforderungen

Obligatorisch (MUSS)

Fakultativ (Kann)

4.1.2 Nicht-funktionale Anforderungen

Obligatorisch (MUSS)

Fakultativ (Kann)

4.2 ...

5 Konzeption & Entwurf

[Beschreibung des Entwurfs auf Basis der Methodologie / der geplanten Vorgehensweise zur Problemlösung im Kontext der Anforderungen (i.A. der Art der Arbeit)]

5.1 Prozess

5.2 Systemarchitektur

5.3 Softwarearchitektur

5.4 Schnittstellen

5.5 Datenmanagement

5.6 ...

6 Implementierung

[Beschreibung der Implementierung[1]auf Basis des Entwurfs und der Methodologie / der geplanten Vorgehensweise zur Problemlösung im Kontext der Anforderungen. Hier ist Raum für Listings, wie z.B. das nun Folgende:

```
1 object HelloWorld {
2   def main(args: Array[String]): Unit = {
3     println("Hello, world!")
4   }
5 }
```

Listing 6.1: *Ein Beispiel: Hello World (Scala)*

Umfangreicher Quell-Code sollte in den Anhang ausgelagert werden.]

[1]Beachten Sie bei der Implementierung und deren Dokumentation bitte Clean Code Empfehlungen (vgl. hierzu z.B. [**martin2008**]).

7 Test

[Beschreibung, wie Sie auf Basis des geplanten Testvorgehens was mit welchen Kriterien und Technologien getestet haben]

8 Darstellung und Bewertung der Ergebnisse

[Beschreibung der Ergebnisse aus allen voran gegangenen Kapiteln sowie der zuvor generierten Ergebnisartefakte mit Bewertung, wie diese einzuordnen sind]

9 Zusammenfassung

[Aggregierte retrograde Kurzbeschreibung der Arbeit]

9.1 Schlussfolgerungen

[Beschreibung der insgesamt zu konstatierenden Schlussfolgerungen im Zusammenhang mit der Arbeit]

9.2 Limitationen

[Beschreibung der Ergebnisse einer kritischen Reflektion und Begründung dessen, was die Arbeit nicht zu leisten vermag]

9.3 Ausblick

[Beschreibung und Begründung potenzieller zukünftiger Folgeaktivitäten im Zusammenhang mit Ihrer Arbeit (z.B. weitere Anforderungen, Theoriebildung, ...]

10 Abkürzungsverzeichnis

11 Glossar

A Appendix

A.1 Quell-Code

A.2 Tipps zum Schreiben Ihrer Abschlussarbeit

- Achten Sie auf eine neutrale, fachliche Sprache. Keine „Ich“-Form.
- Zitieren Sie zitierfähige und -würdige Quellen (z.B. wissenschaftliche Artikel und Fachbücher; nach Möglichkeit keine Blogs und keinesfalls Wikipedia[1]).
- Zitieren Sie korrekt und homogen.
- Verwenden Sie keine Fußnoten für die Literaturangaben.
- Recherchieren Sie ausführlich den Stand der Wissenschaft und Technik.
- Achten Sie auf die Qualität der Ausarbeitung (z.B. auf Rechtschreibung).
- Informieren Sie sich ggf. vorab darüber, wie man wissenschaftlich arbeitet bzw. schreibt:
  - Mittels Fachliteratur[2], oder
  - Beim Lernzentrum[3].
- Nutzen Sie LaTeX[4].

[1] Wikipedia selbst empfiehlt, von der Zitation von Wikipedia-Inhalten im akademischen Umfeld Abstand zu nehmen [**wikipedia2019**].

[2] Z.B. [**balzert2011**], [**franck2013**]

[3] Weitere Informationen zum Schreibcoaching finden sich hier: `https://www.htw-berlin.de/studium/lernzentrum/studierende/schreibcoaching/`; letzter Zugriff: 13 VI 19.

[4] Kein Support bei Installation, Nutzung und Anpassung allfälliger LaTeX-Templates!

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt durch meine Unterschrift, dass ich die vorstehende Arbeit selbstständig und ohne fremde Hilfe angefertigt und alle Stellen, die ich wörtlich oder annähernd wörtlich aus Veröffentlichungen entnommen habe, als solche kenntlich gemacht habe, mich auch keiner anderen als der angegebenen Literatur oder sonstiger Hilfsmittel bedient habe. Die Arbeit hat in dieser oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen.

_____

Datum, Ort, Unterschrift