

**FOM - Hochschule für Oekonomie & Management
Hamburg**

**Master-Studiengang Big Data & Business Analytics
2. Semester**

**Development of a solution for genetic analysis of
ALL genomes by implementing Latent Dirichlet allocation**

Betreuer:

Prof. Dr. Martin Münstermann „“

Autor:

Jacqueline Franßen

Matrikel-Nr: 496804

2. Fachsemester

Hamburg, den 25.07.2019

Contents

1	Abstract	1
2	Introduction	2
3	Related work	3
4	Latent Dirichlet Allocation (LDA)	4
4.1	General description	4
4.2	Examples and possible uses cases	4
4.3	Python package 'Gensim'	4
5	Acute Lymphoblastic Leukemia	5
5.1	Types of Leukemia and its causes	5
5.2	Examples for Genome Analysis: Next Generation Sequencing (NGS) . . .	5
5.3	Data sources: National Center for Biotechnology Information (NCBI) and Ensembl genome browser 96	5
6	Development of a solution for genetic analysis of Acute Lymphoblastic Leukemia (ALL) genomes by implementing LDA	6
6.1	Problems and challenges of genetic analysis	6
6.2	First steps: Draft of developed solution	6
6.3	Proposed solution	6
6.4	Results	6
7	Conclusion and Outlook	7
7.1	Lessons learned	7
7.2	Conclusion	7
7.3	Outlook	7

List of Figures

7.1	xxx	7
-----	---------------	---

List of Tables

1 Abstract

1

2 Introduction

3 Related work

analysis of genetic mutations which cause breastcancer [Jurca et al. 2016]

4 LDA

4.1 General description

LDA was developed by David Blei et. al in the year 2003 and is a clustering algorithm for text mining.

4.2 Examples and possible uses cases

images and schemes

4.3 Python package 'Gensim'

5 Acute Lymphoblastic Leukemia

5.1 Types of Leukemia and its causes

5.2 Examples for Genome Analysis: NGS

5.3 Data sources: NCBI and Ensembl genome browser 96

6 Development of a solution for genetic analysis of ALL genomes by implementing LDA

6.1 Problems and challenges of genetic analysis

6.2 First steps: Draft of developed solution

diagrams, schemes

6.3 Proposed solution

6.4 Results

7 Conclusion and Outlook

7.1 Lessons learned

7.2 Conclusion

7.3 Outlook

Figure 7.1: Bildunterschrift

Ein Zitat

□

Literaturverzeichnis

Jurca, Gabriela et al. (Dec. 2016). "Integrating text mining, data mining, and network analysis for identifying genetic breast cancer trends". en. In: *BMC Research Notes* 9.1. ISSN: 1756-0500. DOI: 10.1186/s13104-016-2023-5. URL: <http://bmcrsnotes.biomedcentral.com/articles/10.1186/s13104-016-2023-5> (visited on 04/06/2019).

Rechtsquellenverzeichnis

Bundesdatenschutzgesetz, BDSG, 1990, zuletzt geändert 2009

Gesetz gegen unlauteren Wettbewerb, UWG, 2004, zuletzt geändert 2013

Richtlinie 95/46/EG des Europäischen Parlaments und des Rates vom 24. Oktober 1995
NI. L 281/31 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener
Daten und zum freien Datenverkehr

Telemediengesetz, TDG, 2007, zuletzt geändert 2010

Ehrenwörtliche Erklärung

Hiermit versichere ich, dass die vorliegende Arbeit von mir selbstständig und ohne unerlaubte Hilfe angefertigt worden ist, insbesondere dass ich alle Stellen, die wörtlich oder annähernd wörtlich aus Veröffentlichungen entnommen sind, durch Zitate als solche gekennzeichnet habe. Ich versichere auch, dass die von mir eingereichte schriftliche Version mit der digitalen Version übereinstimmt. Weiterhin erkläre ich, dass die Arbeit in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde / Prüfungsstelle vorgelegen hat. Ich erkläre mich damit nicht einverstanden, dass die Arbeit der Öffentlichkeit zugänglich gemacht wird. Ich erkläre mich damit einverstanden, dass die Digitalversion dieser Arbeit zwecks Plagiatsprüfung auf die Server externer Anbieter hochgeladen werden darf. Die Plagiatsprüfung stellt keine Zurverfügungstellung für die Öffentlichkeit dar.

Ort, Datum (Vorname Nachname)