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

List of Figures

List of Tables

1 Abstract

1

2 Introduction

LDA provides many advantages, such as that it is suitable for large data and that it performs very well in extracting topics for Indonesian text documents [1].
This article is ordered as follows: section 1 3 will explore ... section 2 4 presents ... and section 3 5. Section 4 6 focusses on the developed implementation and classification algorithm.

[1] [Twinandilla et al. 2018]

3 Related work

In this section, some examples of using text mining techniques in biology or medicine are described. Zhao et al.[1] describe how topic modeling can be used to analyze NGS. Generally, by implementing topic modelling, text corpus are generated.
In the beginning of every genome analysis, there are several important questions to ask. Jurca et al.[2] recommend to ask the following questions: What are the top studied genes in breast cancer? What are the regulations and limitations of blood cancer research in every country? Which countries have studied the largest number of breast cancer? Which are the popular genes mentioned together by countries every year? Where do key genes lie in the soft clusters?
Jurca et al. describe a process to use large-scale text analysis of biomedical abstracts in order to generate new hypothesis about cancer biomarkers[3]. The target is to develop a data mining methodology to find out the genes associated with cancer. By analyzing disease-specific gene expression, experimental data is being checked. The key question is whether a gene has indeed been upregulated or downregulated with respect to a disease.
According to Xu et al.[4], micro Ribonucleic Acid (MIRNA)s build a class of 17-27 nucleotides single-stranded Ribonucleic Acid (RNA) molecules that regulate gen expression post-transcriptionally. In the described text-mining process, Xu et al. identified nine MIRNAs in bladder cancer and adopted protein-protein interaction sites between these miRNAs and target genes. The results of the analyzation process lead to two relationship types between bladder cancer and its MIRNA: casual and unspecified.
Topic modelling is not only used to analyze relationships between genomes but also to

[1] [Zhao et al. 2016]
[2] [Jurca et al. 2016]
[3] [Jurca et al. 2016]
[4] [Xu et al. 2013]

improve diagnoses for stroke disease. Djatna et al. [5] describe an 'Intuitionistic Fuzzy Based Decision Tree' to diagnose different types of stroke disease. To be precise, the different types of stroke diseases can be calculated by a Hamming distance. The term 'Fuzzy logic' means logic that underlies the reasoning of data using precise estimates. It is the fastest way to map input space into output space using a degree of membership.

Lloret et al.[6] built an automatic summarization algorithm for literature. It can includes three steps: First, topic identification, second topic interpretation and third summary generation. While describing the process of textual analyzation, Lloret et al. mention a specific term: term frequency inverse document frequency (TFIDF) which is important for topic modelling. In addition to topic-based approaches, there are graph-based approaches and discourse-based approaches. Graph-based approaches implicate nodes that represent text elements and the edges/links refer to synonymy[7]. Discourse-based approaches include Rhethorical Structure Theory (RST), Hidden-Markov-Models (HMM) or Bayesian models (BM).

Yang et al.[8] describe the process of 'constructing a database for relations between human copy number variant (CNV)s and human genetic disease via systematic text mining'. In general, CNV can cause disease, e.g. by manipulating gene dosage, disruption, fusion or other genetic position effects. To be more precise, there can CNV can lead to two types of autosomal variants: They can either cause deletion or amplification of the long or broken arm region of chromosomes 1-22 or can build multiples of chromosomes 1-21 (e.g. as in disease trisomy 21).

According to their article, Yang et al.[9] used a CNV database which linked the CNV information to the NCBI Gene and Ontology database. Yang et al. mention three steps in the text mining process. First, during the pre-processing step, unstructured fields are split into separated sentences by using Natural Language Toolkit (NLTK), a python package[10]. After that, in the named entity recognition (NER) step, all disease mentions within the DNorm system, such as MeSH IDs are recognized. In the third

[5] [Djatna, Hardhienata, and Masruriyah 2018]
[6] [Lloret and Palomar 2012]
[7] [Lloret and Palomar 2012]
[8] [Yang et al. 2018]
[9] [Yang et al. 2018]
[10] [*Natural Language Toolkit — NLTK 3.4.1 documentation* 2019]

step, Relation extraction (RE), the positions in sentences and entities are compared to generate instances that constist of two candidate entities within one single sentence. Yang et al. mention two more processing methods: Parallel Processing and Post Processing which includes data cleaning and statistics. The term 'data cleaning' is explained as 'de-duplicating data after each step of the process to reduce repetitive operations and prevent statistical errors'. This is a very useful step in biomedicine since biomedical databases may contain errors. For that reason, users can give feedback through a feedback mechanism to improve the quality of the databases.
Lu et al.[11] used multi-channel LDA to model healthcare data. In fact, by creating a learned latent variable model, the likelihood of a set of diagnosis, medications, contextual information in a patient's record can be evaluated. This can help to identify outliers and improve medical data quality. Furthermore, disease groups can be identified, or missing medication or diagnosis can be predicted. Lu et al. used Association Rule Mining (ARM) and supervised learning to predict missing medications. The topic model defines how words in a document are generated through the control of latent topics.

[11] [Lu, Wei, and Hsiao 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. It counts to the most popular topic modelling algorithms[1]. According to Zhao et al., topic modelling requires of a number documents which represent each of them a mixture of latent topics. Moreover, each topic is expressed by a distribution of words. During LDA, two relationships are analyzed: First, the relationship between documents and words, also called 'per-document topic distributions'. Second, the relationship between words and topics ('per-topic word distributions'). To measure the relationships exactly and to make inference about topics and documents for text mining, probability matrices are calculated.

Park et al.[2] define topic models as follows: Documents are no longer a collection of words, but a collection of topics. Furthermore, LDA is a generative topic model which uses a dirichlet parameter (also called dirichlet prior) to model documents. By changing the dirichlet prior, the number of topics that the model assigns to each word and document can be controlled. To be more precisely, a small dirichlet prior means a small number of topics assigned to each word. By increasing the dirichlet prior, the distribution of topics to each word rises. Moreover, the dirichlet parameter is obtained for each document and can be fitted using a maximum or estimated likelihood. If the dirichlet parameter does not fit, the gain in computational efficiency is obtained. Otherwise, there is no advantage (when dirichlet parameter fits well).

According to Jurca et al.[3] the text mining process can be divided into four steps: First, the information has to be retrieved by user queries (Information Retrieval (IR)). Sec-

[1] [Zhao et al. 2016]
[2] [Park and Ramamohanarao 2009]
[3] [Jurca et al. 2016]

ond, different vocabularies and ontologies have to be integrated (NER). Third, during Information extraction (IE), relationships between biological entities in the texts are extracted by either using co-occurence processing or Natural Language Processing (NLP). Last, there has to be gained biologically meaningful knowledge about how biological entities are related by implementing Knowledge Discovery (KD) methods. Moreover, there can be distinguished between three types of clustering: hard clustering, hierarchical clustering and soft clusternig. Hard clustering describes the process of separating items into distinct groups where each item is exactly in one cluster. Hierarchical clustering implicates creating single-link clusters (how similar the items are to one another) and complete-link (how dissimilar the items are). Soft clustering means that items cannot be distinctly separated into clusters and partly are member of two or more clusters at a time.

Besides, Djatna et al.[4] mention data mining techniques, such as Classification and Regression Tree (CART), Iterative Dichotomized 3 (ID3), Decision Tree (DT), Principal Component Analysis (PCA) and LDA.

Lu et al.[5] define topic models as a text mining approach that assumes observed word co-occurences which are governed by latent variables. LDA includes the identification of latent topics from a set of documents, analyzing long-term topic trends and modelling words and references in documents.

According to Hoffman et al., LDA is a probabilistic (Bayesian) model of text documents[6]. The idea of LDA is to define a document as a collection of k topics. Each topic defines a multinominal distribution over a vocabulary which is drawn from a dirichlet.

What is more, every term has a probabilistic relationship to every document. The topic model probabilities are stored as term relationships in thesaurus. The term frequencies are stored in the document index.

Twinandilla et al.[7] mention three variables to be defined before the LDA process: α (the diversity of sentence distribution), β (the diversity of topic distribution) and γ (the similarity between sentences and titles).

[4] [Djatna, Hardhienata, and Masruriyah 2018]
[5] [Lu, Wei, and Hsiao 2016]
[6] [Hoffman, Bach, and Blei 2010]
[7] [Twinandilla et al. 2018]

In their article 'Multi-document summarization using k-means and LDA-significant sentences, Twinandilla et al. [8] describe the research process by implementing the following six steps.

1.Step: Preprocessing

First, all words and sentences need to be simplified by using a bag of words as well a bag of sentences. In detail, this step includes case folding (putting all words into lower case), tokenization (cutting a document into an array of words or sentences and eleminating punctuation), stopword removal (deleting words that appear often without a particular meaning) and stemming (changing words in a document that appear often wothout a particular meaning).

2.Step: Calculate the number of clusters

Second, the number of clusters needs to be calculated by using k-means clustering.

3.Step: LDA

In this step, Twinandilla et al. distinguish between generative and inference LDA. Generative LDA forms a document from a collection of words whereas inference LDA only retrieves information from documents.

4.Step: Sentence LDA

Fourth, during sentence LDA, documents are represented as topic representation. Each topic is a sentence distribution that represents a sentence. This sentence has significant weight on a multi-document summarization.

5.Step: Summary formation

In this step, each document is sorted by a decreasing value of the final sentence weight. After that, the p percent of sentences with the highest value has to be chosen from each document. The p value is subsequently called 'summarization level'.

6.Step: Arrange selected sentences in a sequence

The last step includes the arrangement of all selected sentences in a sequence. This means putting them into a useful order to summarize all documents.

[8] [Twinandilla et al. 2018]

As reported by Blei et al.[9], LDA is a generative probabilistic model for collections of discrete data such as text corpora. In addition to that, LDA is represented as three level hierarchical Bayesian model in which each item of a collection is a finite mixture over an underlying set of topics. What is more, each topic is modeled as an infinite mixture over an underlying set of probabilities. Blei et al. define topic probabilities as an explicit representation of a document.

TFIDF TFIDF is a scheme through which a basic vocabulary of words or terms is chosen. For each document in the corpus a count (which represents the number of occurences for each word) is formed[10]. There are three terms to be distinguished: First, a word is a basic unit of descrete data, defined to be an item from the vocabulary indexed by 1...V. Second, a document is a sequence of N words. Third, a corpus is a collection of M documents. After a suitable normalization process, the term frequency count is compared to an inverse document frequency count. This leads to the total number of occurences of a word in the entire corpus. The result is a term-by-document matrix X whose columns contain TFIDF values for each document in the corpus.

LDA process The process of LDA can be briefly described with the following step[11]: First, there are M documents in the corpus. Second, each document j has N_j words. In the next step, the observed value w_{ji} describes the appearance of a word i in a document j. In addition to that, all words will be clustered into K topics which are defined as object classes. Finally, each topic k is modelled as a multinomial distribution over the codebook.

As reported by Wang et al., LDA is a language model which clusters co-occuring words into topics. Moreover, documents are described as 'bag of words'. Wang et al. describe a special form of LDA: Spatial LDA. It encodes spatial structure among visual words, assuming the partition of words into documents is known a priori.

[9] [Blei, Ng, and Jordan 2003]
[10] [Blei, Ng, and Jordan 2003]
[11] [Wang and Grimson 2008]

LDA extensions Extensions of LDA are author-topic model, dynamic-topic model and correlated topic model. Wang et al. refer to the dynamic-topic model while describing how visual words are clustered into topics which correspond to object classes.

4.2 Examples and possible use cases

Zhao et al. describe the process of analyzing genomes as follows: First, each document corresponds to one of the total number of Desoxyribonucleic acid (DNA) straints. Second, all documents had the same number of words. Third, the distribution of words for topics as well as the distribution of topics in documents were described by random variables obeying Dirichlet distributions with parameters α and β. After that, nucleotides and their orders in NGS sequences could be treated as words and the genetic information in sequences was translated and exhibited as a 'bag of words'[12]. By using the strain-topic matrix derived from topic modelling, relationships or similarities between the strains serotypes can be found out.

Hoffman et al.[13] describe the development of an online variational Bayes algorithm for LDA which is based on stochastic optimization with a natural gradient step. This step converges to a local optimum of the variational Bayes objective function. To be more precise, Bayesian models provide a natural way to encode assumptions about observed data. There can be distinguished between two approaches: First, sampling approaches are based on Markov Chain Monte Carlo (MCMC) sampling. Here, a Markov chain defines the stationary distribution. Second, there are optimization approaches which are usually based on variational inference MCMC. In this case, variational Bayes optimizes the simplified parametric distribution.

Twinandilla et al.[14] developed a 'multi-document summarization using k-means and LDA-significant sentences' on yellow journalism. The term 'yellow journalism' stands for 'redundant news documents' which makes it difficult to distinguish documents containing fact or opinionated information. After defining the corpus, Twinandilla et al. describe two different summarization processes: Abstractive as well as extractive summarization. Abstractive summarization means summarizing documents by creat-

[12] [Zhao et al. 2016]
[13] [Hoffman, Bach, and Blei 2010]
[14] [Twinandilla et al. 2018]

ing new sentences (with the same information as the original document). Extractive summarization suggests summarizing a document by selecting a part of a sentence in that document.

As described in chapter 3 on page 3, Lu et al.[15] used Multiple-channel Latent Dirichlet Allocation (MCLDA) to estimate latent health status groups. MCLDA constructs latent relations among diagnoses, medications, contextual variables in different status groups. To be more precise, it is a dimensional reduction method that summarizes each record using a probability vector over a latent health status group. The prediction tasks were performed by using a Collapsed Gibbs model (CGS) based inference model and inferred methods. During the topic modelling process, Lu et al. refer to two associations among the data: diagnosis-medication associations to identify the clinical use of medications and diagnosis-diagnosis associations to create a network structure among the diseases.

4.3 Python package 'Gensim'

[15] [Lu, Wei, and Hsiao 2016]

5 Acute Lymphoblastic Leukemia

5.1 Types of Leukemia and its causes

According to Jurca et al. [1], cancer is the result of damage, especially of mutations to cell's DNA which leads to a cell losing its normal functionality and gains the ability to indefinitely multiply until normal tissue funtions are impaired. This is also why malitious cancer is distributing so fast. Besides, each patient develops a different set of cancerous mutations in various genes which lead to multiple subtypes of cancer. Furthermore, some genes can be up-regulated (which means that they are transcribed and expressed more), down-regulated (which means that they are not expressed) or can be co-expressed (which means that they are expressed at the same time).
As stated by Montano et al.[2], ALL is a malignant disorder originating from hematopoietic B-/T-cell precursors which are characterized by marked heterogenity at molecular and clinical levels. There are many approaches to analyze these precursors, such as analyzing targeting of transcriptional factors (PAX5) which are involved in the pathogenesis of B-ALL. Other therapeutic and clinical approaches are genome editing techniques, i.e. the design of new therapies (Chimeric Antigen Receptors (CAR)s) and the study of genes involved in the evolution of pathogenesis.

5.2 Examples for Genome Analysis: NGS

NGS refers to post-Sanger sequencing methods[3]. Since NGS produces large volumes of sequence data it might be very useful implementing topic modelling techniques in order to maintain the flexibility for the level of resolution required for given

[1] [Jurca et al. 2016]
[2] [Montaño et al. 2018]
[3] [Zhao et al. 2016]

experiments. According to Gasperskaja et al.[4], NGS does not require a priori knowledge about genomic feature, it only requires a low amount of DNA or RNA as input.
The step before analyzing two or more (multiple) genomes is called alignment which includes a comparison of two genomes. There are many different types of alignments, but Zhao et al. refer to the Multiple Sequence Alignment (MSA) by describing Multiple Sequence Comparison by Log- Expectation (MUSCLE) and CLUSTAL.
Gasperskaja et al. mention an important question which should be asked before every genome analysis: 'Is the variance pathogenic?' and whether there is any relationship between genotype and phenotype which means that it can lead to a disease or can cause a number of disorders. Moreover, there can be distinguished between beneficial (Single Nucleotide Polymorphism (SNP)) and pathogenic (nonsense variant) single nucleotide changes, large microscopically visible or chromosomal aberation. To find out whether a genome mutation is pathogenic, Gasperskaja et al. explain that substantial information about functional genomics can be found through the analysis of messenger RNA (MRNA) or complementary Ribonucleic Acid (CRNA) (which is a copy from MRNA by reverse transcription Polymerase Chain Reaction (PCR). Methods to measure RNA expression are the following: Serial Analysis of Gene Expression (SAGE) or Quantitative real-time Polymerase Chain Reaction (QPCR). By using complementary Desoxyribonucleic acid (CDNA) microarray assays important genome-wide information about changes of gene expression in various cell lines can be found out.
As claimed by Montano et al.[5], the development of NGS techniques implicates vast amount of data which need to be translated. One important question is to find out how the genotype (the genetic expression of a biological attribute) influences the phenotype. By integrating genome editing systems into their research process, investigators are able to manipulate virtually any gene in a diverse range of cell types and organisms. To give an example, Gasperskaja et al. and Montano et al. describe Clustered Regularly Interspaced Short Palindromic Repeats Cas-9 (CRISPR-CAS9). In fact, this genome analysis includes generating a direct cut in the double strand of DNA by Cas9 nuclease. Cas9 is driven by a single 20-nucleotide RNA strand which marks the direct breakpoint. After cutting the DNA, the repair machinery of the host cell

[4][Gasperskaja and Kučinskas 2017]
[5][Montaño et al. 2018]

repairs errors and promotes a modification of the original sequence by a mutation (e.g. insertion, deletion, inversion). But why should be changed the shape of DNA? In the opinion of Montano et al.[6], the use of genetically modified cell lines and animal models help us to better understand the functions of genes and their pathogenesis in diseases, such as cancer.

5.3 Data sources: NCBI and Ensembl genome browser 96

[6] [Montaño et al. 2018]

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

6.1 Problems and challenges of genetic analysis

Quality of data Is the data complete, which includes that it contains all required genomes which can cause ALL.
our assumptions can lead to false content or solutions

6.2 First steps: Draft of developed solution

To get useful data, the NCBI [1] was used to get all currently detected mutations of genomes which may cause LDA.
The first idea was to build a parsing application, which iterates over the found 582 genomes. After the iteration, it compares the oncogenes with the healthy genomes and to figure out where the differences are. The results might be displayed in a diagram. It might be possible to create clusters from the differences between the two groups or practice LDA on the differences.

6.3 Proposed solution

6.4 Results

[1] [Biotechnology et al. 2019]

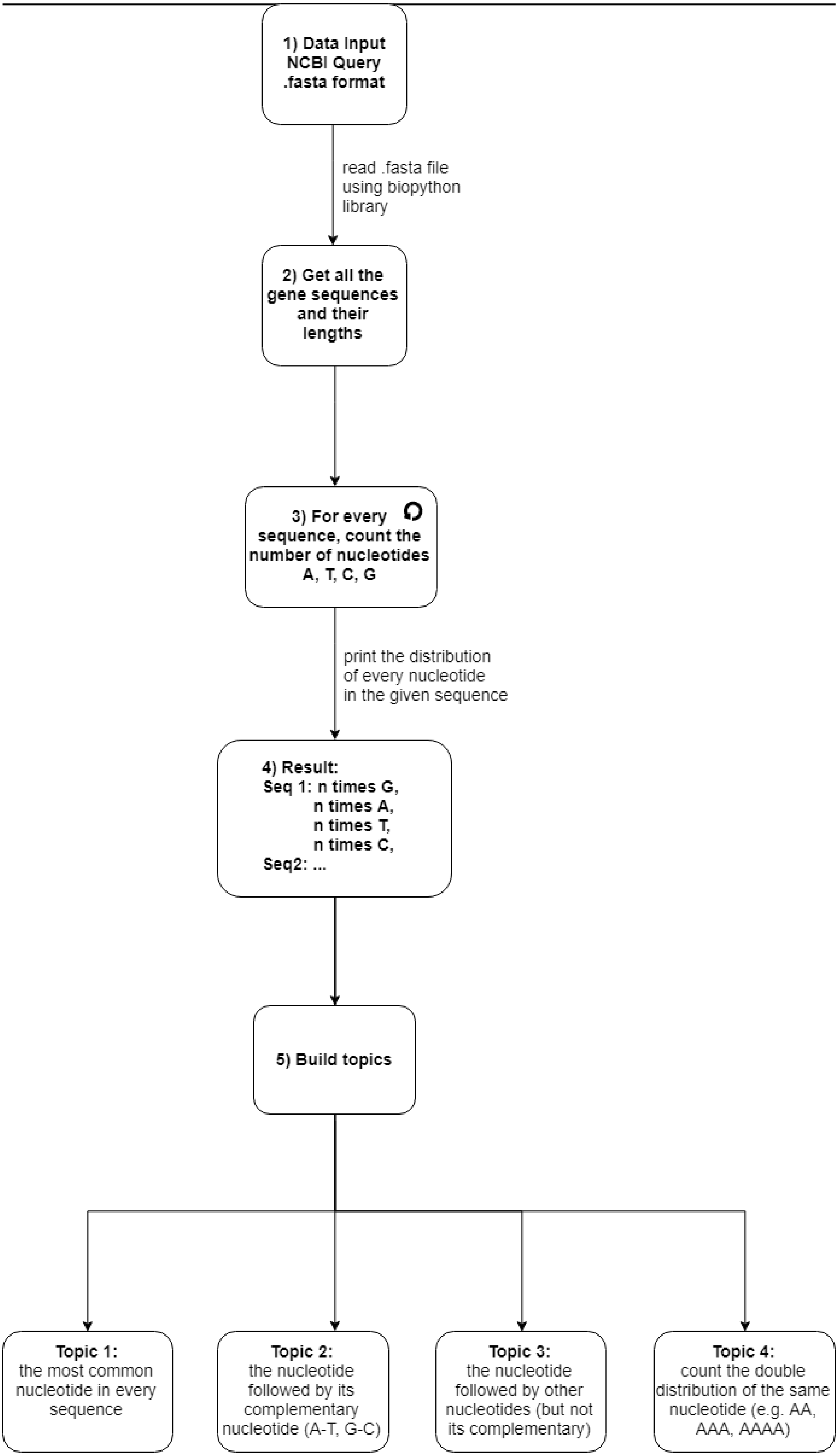

Figure 6.1: Diagram of developed algorithm to create basic topics among given gene sequences

7 Conclusion and Outlook

7.1 Lessons learned

7.2 Conclusion

7.3 Outlook

> *Ein Zitat*

[]

Literaturverzeichnis

Biotechnology, National Center for et al. (2019). *National Center for Biotechnology Information*. en. URL: https://www.ncbi.nlm.nih.gov/ (visited on 05/09/2019).

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3.Jan, pp. 993–1022. ISSN: ISSN 1533-7928. URL: http://www.jmlr.org/papers/v3/blei03a.html (visited on 04/29/2019).

Djatna, Taufik, Medria Kusuma Dewi Hardhienata, and Anis Fitri Nur Masruriyah (Dec. 2018). "An intuitionistic fuzzy diagnosis analytics for stroke disease". en. In: *Journal of Big Data* 5.1, p. 35. ISSN: 2196-1115. DOI: 10.1186/s40537-018-0142-7. URL: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-018-0142-7 (visited on 04/07/2019).

Gasperskaja, Evelina and Vaidutis Kučinskas (2017). "The most common technologies and tools for functional genome analysis". In: *Acta Medica Lituanica* 24.1, pp. 1–11. ISSN: 1392-0138. DOI: 10.6001/actamedica.v24i1.3457. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5467957/ (visited on 05/09/2019).

Hoffman, Matthew, Francis R. Bach, and David M. Blei (2010). "Online Learning for Latent Dirichlet Allocation". In: *Advances in Neural Information Processing Systems 23*. Ed. by J. D. Lafferty et al. Curran Associates, Inc., pp. 856–864. URL: http://papers.nips.cc/paper/3902-online-learning-for-latent-dirichlet-allocation.pdf (visited on 04/19/2019).

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://bmcresnotes.biomedcentral.com/articles/10.1186/s13104-016-2023-5 (visited on 04/06/2019).

Lloret, Elena and Manuel Palomar (Jan. 2012). “Text summarisation in progress: a literature review”. In: *Artificial Intelligence Review* 37.1, pp. 1–41. ISSN: 1573-7462. DOI: 10.1007/s10462-011-9216-z. URL: https://doi.org/10.1007/s10462-011-9216-z.

Lu, Hsin-Min, Chih-Ping Wei, and Fei-Yuan Hsiao (Apr. 2016). “Modeling healthcare data using multiple-channel latent Dirichlet allocation”. In: *Journal of Biomedical Informatics* 60, pp. 210–223. ISSN: 1532-0464. DOI: 10.1016/j.jbi.2016.02.003. URL: http://www.sciencedirect.com/science/article/pii/S1532046416000253 (visited on 04/29/2019).

Montaño, Adrián et al. (July 2018). “Targeted genome editing in acute lymphoblastic leukemia: a review”. In: *BMC Biotechnology* 18.1, p. 45. ISSN: 1472-6750. DOI: 10.1186/s12896-018-0455-9. URL: https://doi.org/10.1186/s12896-018-0455-9 (visited on 04/19/2019).

Natural Language Toolkit — NLTK 3.4.1 documentation (2019). URL: https://www.nltk.org/ (visited on 05/09/2019).

Park, Laurence A. F. and Kotagiri Ramamohanarao (2009). “The Sensitivity of Latent Dirichlet Allocation for Information Retrieval”. en. In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by Wray Buntine et al. Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 176–188. ISBN: 978-3-642-04174-7.

Twinandilla, Shiva et al. (Jan. 2018). “Multi-Document Summarization Using K-Means and Latent Dirichlet Allocation (LDA) – Significance Sentences”. In: *Procedia Computer Science*. The 3rd International Conference on Computer Science and Computational Intelligence (ICCSCI 2018) : Empowering Smart Technology in Digital Era for a Better Life 135, pp. 663–670. ISSN: 1877-0509. DOI: 10.1016/j.procs.2018.08.220. URL: http://www.sciencedirect.com/science/article/pii/S1877050918315138 (visited on 04/29/2019).

Wang, Xiaogang and Eric Grimson (2008). “Spatial Latent Dirichlet Allocation”. In: *Advances in Neural Information Processing Systems 20*. Ed. by J. C. Platt et al. Curran Associates, Inc., pp. 1577–1584. URL: http://papers.nips.cc/paper/3278-spatial-latent-dirichlet-allocation.pdf (visited on 04/19/2019).

Xu, Yahong et al. (Sept. 2013). “RETRACTED ARTICLE: Integrated gene network analysis and text mining revealing PIK3R1 regulated by miR-127 in human bladder cancer”. In: *European Journal of Medical Research* 18.1, p. 29. ISSN: 2047-783X.

DOI: 10.1186/2047-783X-18-29. URL: https://doi.org/10.1186/2047-783X-18-29.

Yang, Xi et al. (Dec. 2018). “Constructing a database for the relations between CNV and human genetic diseases via systematic text mining”. en. In: *BMC Bioinformatics* 19.S19, p. 528. ISSN: 1471-2105. DOI: 10.1186/s12859-018-2526-2. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2526-2 (visited on 04/11/2019).

Zhao, Weizhong et al. (May 2016). “A novel procedure on next generation sequencing data analysis using text mining algorithm”. In: *BMC Bioinformatics* 17.1, p. 213. ISSN: 1471-2105. DOI: 10.1186/s12859-016-1075-9. URL: https://doi.org/10.1186/s12859-016-1075-9.

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)