

**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

<b>1</b>	<b>Abstract</b>	<b>1</b>
<b>2</b>	<b>Introduction</b>	<b>2</b>
<b>3</b>	<b>Related work</b>	<b>3</b>
<b>4</b>	<b>Latent Dirichlet Allocation (LDA)</b>	<b>6</b>
4.1	General description . . . . .	6
4.2	Examples and possible use cases . . . . .	10
4.3	Python library 'Gensim' . . . . .	11
<b>5</b>	<b>Acute Lymphoblastic Leukemia</b>	<b>12</b>
5.1	Types of Leukemia and its causes . . . . .	12
5.2	Examples for Genome Analysis: Next-Generation Sequencing (NGS) .	14
5.3	Data sources: National Center for Biotechnology Information (NCBI) and Ensembl genome browser 96 . . . . .	15
<b>6</b>	<b>Development of a solution for genetic analysis of Acute Lymphoblastic Leukemia (ALL) genomes by implementing LDA</b>	<b>18</b>
6.1	Problems and challenges of genetic analysis . . . . .	18
6.2	First steps: Draft of developed solution . . . . .	19
6.3	Proposed solution . . . . .	21
6.4	Results . . . . .	21
<b>7</b>	<b>Conclusion and Outlook</b>	<b>26</b>
7.1	Lessons learned . . . . .	26
7.2	Conclusion . . . . .	26
7.3	Outlook . . . . .	26

## List of Figures

5.1	NCBI query to find all ALL related gene sequences . . . . .	16
5.2	Detailed information about selected gene sequence . . . . .	16
5.3	Information about genome IKZF1, found in Ensembl genome browser 96	17
6.1	Extracted data (including genome's name and description), the green rectangle includes the descriptions . . . . .	19
6.2	Diagram of first draft of developed algorithm . . . . .	20
6.3	Diagram of developed algorithm to create basic topics among given genomes . . . . .	22
6.4	Diagram (created with PyLDAvis) which shows the distribution of the topics . . . . .	23
6.5	Diagram of LDA process . . . . .	24
6.6	Call of method LdaModel() from library Gensim and its output . . . . .	25

## List of Tables

# 1 Abstract

There exist many types of cancer. Every type of cancer is so individual and is treated particularly in practice. Experimental studies showed that cancer is the consequence of the incorrect mutation of genomes which makes it even harder to find out the appropriate solution and therapy for the patient (since every patient has its own individual set of genomes). Generally, our human genomes are nearly similar, except from small variations (due to ethnics, demography etc.). Besides, to find out which mutations can lead to a certain kind of cancer, data analysis algorithms can be a chance to find out or to 'predict' the appearing mutations.

On the other side, from a computer scientist's view, there exist many algorithms to analyze data. In this paper, the focus is on LDA, an unsupervised (autonomous) topic modelling technique. Given a certain dataset (including genomes which can cause ALL), topics and similarities between the genome data are figured out. The results are discussed in section 6.4.

The aim of this paper is to find out whether LDA is the appropriate solution to analyze genome data and to find out similarities between the genes. These similarities are expected as a solution approach to find a pattern among the disease-related genomes and to be able to predict these mutations. Furthermore, if some patient shows these genomes (or mutations of them), the geneticist or molecular biologist can inject the required sequence to the patient so that the cancer will not break out.

## 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 <sup>1</sup>.

In the year 2000, more than 209,000 children and adults died of leukemia globally. This number keeps growing year by year <sup>2</sup>.

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.

---

<sup>1</sup>cf.[Twinandilla et al. 2018]

<sup>2</sup>cf.[Tang, Cao, and Y. Wang 2011]

## 3 Related work

In this section, some examples of using text mining techniques in biology and medicine are described. Zhao et al.<sup>1</sup> 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.<sup>2</sup> 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. 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.<sup>3</sup>, micro Ribonucleic Acid (MIRNA)s build a class of 17-27 nucleotides single-stranded Ribonucleic Acid (RNA) molecules that regulate gene 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

---

<sup>1</sup>cf.[Zhao et al. 2016]

<sup>2</sup>cf.[Jurca et al. 2016]

<sup>3</sup>cf.[Xu et al. 2013]

improve diagnoses for stroke disease. Djatna et al.<sup>4</sup> 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.<sup>5</sup> built an automatic summarization algorithm for literature. It can include three steps: First, topic identification, second topic interpretation and third summary generation. While describing the process of textual analysis, 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<sup>6</sup>. Discourse-based approaches include Rhetorical Structure Theory (RST), Hidden-Markov-Models (HMM) or Bayesian models (BM).

Yang et al.<sup>7</sup> 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.<sup>8</sup> 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<sup>9</sup>. 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 step, Rela-

---

<sup>4</sup>cf.[Djatna, Hardhienata, and Masruriyah 2018]

<sup>5</sup>cf.[Lloret and Palomar 2012]

<sup>6</sup>cf.[Lloret and Palomar 2012]

<sup>7</sup>cf.[Yang et al. 2018]

<sup>8</sup>cf.[Yang et al. 2018]

<sup>9</sup>cf.[*Natural Language Toolkit — NLTK 3.4.1 documentation* 2019]

tion extraction (RE), the positions in sentences and entities are compared to generate instances that consist 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.<sup>10</sup> 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.

---

<sup>10</sup>cf.[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<sup>1</sup>. 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.<sup>2</sup> 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.<sup>3</sup> the text mining process can be divided into four steps: First, the information has to be retrieved by user queries (Information Retrieval (IR)). Sec-

---

<sup>1</sup>cf.[Zhao et al. 2016]

<sup>2</sup>cf.[L. A. F. Park and Ramamohanarao 2009]

<sup>3</sup>cf.[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-occurrence 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 clustering. 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.<sup>4</sup> 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.<sup>5</sup> define topic models as a text mining approach that assumes observed word co-occurrences 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<sup>6</sup>. The idea of LDA is to define a document as a collection of  $k$  topics. Each topic defines a multinomial 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.<sup>7</sup> mention three variables to be defined before the LDA process:  $\alpha$  (the diversity of sentence distribution),  $\beta$  (the diversity of topic distribution) and  $\gamma$  (the similarity between sentences and titles).

---

<sup>4</sup>cf.[Djatna, Hardhienata, and Masruriyah 2018]

<sup>5</sup>cf.[Lu, Wei, and Hsiao 2016]

<sup>6</sup>cf.[Hoffman, Bach, and Blei 2010]

<sup>7</sup>cf.[Twinandilla et al. 2018]

In their article 'Multi-document summarization using k-means and LDA-significant sentences, Twinandilla et al. <sup>8</sup> 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 eliminating punctuation), stopword removal (deleting words that appear often without a particular meaning) and stemming (changing words in a document that appear often without 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.

---

<sup>8</sup>cf.[Twinandilla et al. 2018]

As reported by Blei et al.<sup>9</sup>, 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 occurrences for each word) is formed<sup>10</sup>. There are three terms to be distinguished: First, a word is a basic unit of discrete data, defined to be an item from the vocabulary indexed by  $1 \dots V$ . 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 occurrences 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<sup>11</sup>: 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-occurring 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.

---

<sup>9</sup>cf.[Blei, Ng, and Jordan 2003]

<sup>10</sup>cf.[Blei, Ng, and Jordan 2003]

<sup>11</sup>cf.[X. 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) strains. 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  $\alpha$  and  $\beta$ . 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'<sup>12</sup>. By using the strain-topic matrix derived from topic modelling, relationships or similarities between the strains serotypes can be found out.

Hoffman et al.<sup>13</sup> 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.<sup>14</sup> 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-

---

<sup>12</sup>cf.[Zhao et al. 2016]

<sup>13</sup>cf.[Hoffman, Bach, and Blei 2010]

<sup>14</sup>cf.[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.<sup>15</sup> 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.

Lee et al.<sup>16</sup> analyzed the direct and indirect interactions between DNA copy number and gene expression changes. 'DNA copy number aberrations and gene expression changes provide valuable information for studying chromosomal instability and its consequences in cancer.' Three types of cancer have been analyzed: brain, bladder and breast cancer. The results of the analysis lead to a significant association between DNA copy number and its gene expression. By monitoring gene expression data, different classes of tumors can be distinguished successfully.

### 4.3 Python library 'Gensim'

'Gensim' stands for 'generate similar' and is a python library, created by Radim Rehurek and Petr Sojka<sup>17</sup>. Firstly created in 2008, it was a collection of various Python scripts to generate a short list of the most similar articles to a given article. Gensim supports scalable statistical semantics to analyze plain-text documents for semantic structure and to retrieve semantically, similar documents.

---

<sup>15</sup>cf.[Lu, Wei, and Hsiao 2016]

<sup>16</sup>cf.[Lee, Kong, and P. J. Park 2008]

<sup>17</sup>cf.[Řehůřek and Sojka 2010]

## 5 Acute Lymphoblastic Leukemia

### 5.1 Types of Leukemia and its causes

According to Jurca et al.<sup>1</sup>, 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 functions are impaired. This is also why malignant 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 Montaña et al., ALL is a malignant disorder originating from hematopoietic B-/T-cell precursors which are characterized by marked heterogeneity at molecular and clinical levels<sup>2</sup>. 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.

Jagadev et al.<sup>3</sup> define leukemia as blood cancer which begins in the bone marrow where it causes formation of a large number of abnormal cells. According to Jagadev et al. there are four common types of leukemia: ALL, Acute Myeloid Leukemia (AML), Chronic Lymphocytic Leukemia (CLL), Chronic Myeloid Leukemia (CML). Furthermore, the normal and leukemic lymphocytes differ significantly.

---

<sup>1</sup>cf.[Jurca et al. 2016]

<sup>2</sup>cf.[Montaña et al. 2018]

<sup>3</sup>cf.[Jagadev and Virani 2017]

As explained by Tang et al.<sup>4</sup>, every subtype of leukemia can have individualized treatments and therapies.

According to Andersson et al., ALL is caused the expansion of immature hematopoietic cells (blasts) in the bone marrow and peripheral blood<sup>5</sup>. Furthermore, ALL counts to the most common childhood malignancies, appearing between 2 to 5 years of age. ALL has two subtypes: T-Acute Lymphoblastic Leukemia (T-ALL) and B-cell precursor Acute Lymphoblastic Leukemia (BCP-ALL). The new cases of T-ALL are 15% childhood and 25% adult ALL. Compared to BCP-ALL, T-ALL is characterized by older ages of onset, male sex predominance, inferior outcome and chromosomal abnormalities. In 1960, the 'Philadelphia chromosome' in CML was discovered which groups it to 'genetic disease'. The result of the exchange of genetic material between two chromosomes is called chromosomal translocations. These serve as hallmarks of leukemia and are intimately associated with a specific leukemia type and its prognosis. Moreover, the different types of ALL are commonly associated with genetic events, such as DNA copy-number alterations (CNA), a sub-microscopic deletions or gains, amplifications and sequence mutations.

**Therapies against ALL** The Childhood Acute Lymphoblastic Leukemia Collaborative Group (CALLCG) first met in 1994 to find out a specific treatment for childhood ALL. In 2009, CALLCG describe a therapy using anthracyclines to treat childhood ALL<sup>6</sup>. In fact, some studies suggest that anthracyclines might cause cardiotoxicity. In most cases of the therapeutic approach, anthracyclines significantly reduced the bone marrow relapse when added to the standard therapy. But it did not significantly increase an event-free survival. One year after that, CALLCG reviewed the addition of vincristine plus steroid pulses (prednisone and prednisolone pulses) to the treatment of childhood ALL<sup>7</sup>. As a result, the combination of vincristine pulses significantly reduced the overall event rate by around 30% to 10% improvement in event-free survival (EFS) by five years.

---

<sup>4</sup>cf.[Tang, Cao, and Y. Wang 2011]

<sup>5</sup>cf.[Andersson et al. 2016]

<sup>6</sup>cf.[Group (CALLCG) 2009]

<sup>7</sup>cf.[“Systematic review of the addition of vincristine plus steroid pulses in maintenance treatment for childhood acute lymphoblastic leukaemia – an individual patient data meta-analysis involving 5659 children” 2010]



## 5.2 Examples for Genome Analysis: NGS

NGS refers to post-Sanger sequencing methods<sup>8</sup>. 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 experiments. According to Gasperskaja et al.<sup>9</sup>, 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 aberration. 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 Montaña et al.<sup>10</sup>, 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

---

<sup>8</sup>cf.[Zhao et al. 2016]

<sup>9</sup>cf.[Gasperskaja and Kučinskas 2017]

<sup>10</sup>cf.[Montaña et al. 2018]

are able to manipulate virtually any gene in a diverse range of cell types and organisms. To give an example, Gasperskaja et al. and Montaña 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 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 Montaña et al.<sup>11</sup>, 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

**NCBI** NCBI was established in Nov. 4, 1988 as a national resource for molecular biology information<sup>12</sup>. Its main purpose was to create automated systems for storing and analyzing knowledge about molecular biology as well as biochemistry and genetics. The community of NCBI conducts research on fundamental biomedical problems at a molecular level using mathematical and computational methods. Moreover NCBI collaborates with several National Institutes of Health (NIH) and develops, distributes, supports, and grants access to various databases and software for scientific and medical communities. NCBI also works as an organization for standardization of databases, data deposition and data exchange and biological nomenclature. As a challenge, NCBI needs to find new approaches to deal with the volume and complexity of data. Another complex issue NCBI is facing is how to provide researchers better access to analysis and computing tools. These tools will be useful to understand the genetic legacy and its role in health and disease.

In this article, NCBI was used to query all genomes related to the disease ALL which can be seen in figure 5.1. After selecting the first gene 'IKZF1', the web interface

---

<sup>11</sup>cf.[Montaña et al. 2018]

<sup>12</sup>cf.[Biotechnology et al. 2019]

shows details about the given genome sequence (see figure 5.2). Below the given point 'See related', a link to the related source from Ensembl genome browser 96 can be clicked. The next paragraph describes the information from the Ensembl genome browser 96.

NCBI Resources How To Sign in to NCBI

Gene (acute lymphoblastic leukemia) AND "Homo sapiens" Search

Gene sources: Genomic, Mitochondria, Organelles

Categories: Alternatively spliced, Annotated genes, Non-coding, Protein-coding

Sequence content: CCDS, Ensembl, RefSeq, RefSeqGene

Status: Current

Search results: Items: 1 to 20 of 588

Name/Gene ID	Description	Location	Aliases	MIM
IKZF1 ID: 10320	IKAROS family zinc finger 1 [Homo sapiens (human)]	Chromosome 7, NC_000007.14 (50303453..50405101)	CVID13, Hs.54452, IK1, IKAROS, LYF1, LyF-1, PPP1R92, PRO0758, ZNFN1A1	603023
ARID5B ID: 84159	AT-rich interaction domain 5B [Homo sapiens (human)]	Chromosome 10, NC_000010.11 (61901254..62096948)	DESRT, MRF-2, MRF2	608538
CEBPE ID: 1053	CCAAT enhancer binding protein epsilon [Homo sapiens (human)]	Chromosome 14, NC_000014.9 (23117306..23119611, complement)	C/EBP-epsilon, CRP1	600749
DDC ID: 1644	dopa decarboxylase [Homo sapiens (human)]	Chromosome 7, NC_000007.14 (50458436..50565460, complement)	AADC	107930
TSPAN7 ID: 7102	tetraspanin 7 [Homo sapiens (human)]	Chromosome X, NC_000023.11 (38561478..38688918)	A15, CCG-B7, CD231, DXS1692E, MRX58, MXS1, TALLA-1, TM4SF2, TM4SF2L	300096

Filters: Manage Filters

Results by taxon: Top Organisms [Tree] Homo sapiens (587) Human mastadenovirus A (1)

Find related data: Database: [Select] Find items

Search details: acute lymphoblastic leukemia[All Fields] AND "Homo sapiens"[All Fields] AND alive[prop] Search See more...

Recent activity: Turn Off Clear

Figure 5.1: NCBI query to find all ALL related gene sequences

IKZF1 IKAROS family zinc finger 1 [Homo sapiens (human)]

Gene ID: 10320, updated on 4-Jun-2019

Summary

Official Symbol: IKZF1 provided by HGNC

Official Full Name: IKAROS family zinc finger 1 provided by HGNC

Primary source: HGNC:HGNC:13176

See related: Ensembl:ENSG00000185811 MIM:603023

Gene type: protein coding

RefSeq status: REVIEWED

Organism: Homo sapiens

Lineage: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo

Also known as: IK1; LYF1; LyF-1; CVID13; IKAROS; PPP1R92; PRO0758; ZNFN1A1; Hs.54452

Summary: This gene encodes a transcription factor that belongs to the family of zinc-finger DNA-binding proteins associated with chromatin remodeling. The expression of this protein is restricted to the fetal and adult hemo-lymphopoietic system, and it functions as a regulator of lymphocyte differentiation. Several alternatively spliced transcript variants encoding different isoforms have been described for this gene. Most isoforms share a common C-terminal domain, which contains two zinc finger motifs that are required for hetero- or homo-dimerization, and for interactions with other proteins. The isoforms, however, differ in the number of N-terminal zinc finger motifs that bind DNA and in nuclear localization signal presence, resulting in members with and without DNA-binding properties. Only a few isoforms contain the requisite three or more N-terminal zinc motifs that confer high affinity binding to a specific core DNA sequence element in the promoters of target genes. The non-DNA-binding isoforms are largely found in the cytoplasm, and are thought to function as dominant-negative factors. Overexpression of some dominant-negative isoforms have been associated with B-cell malignancies, such as acute lymphoblastic leukemia (ALL). [provided by RefSeq, May 2014]

Expression: Biased expression in lymph node (RPKM 13.3), appendix (RPKM 8.6) and 10 other tissues See more

Orthologs: mouse all

Table of contents: Summary, Genomic context, Genomic regions, transcripts, and products, Expression, Bibliography, Phenotypes, Variation, HIV-1 interactions, Pathways from BioSystems, Interactions, General gene information, Markers, Homology, Gene Ontology, General protein information, NCBI Reference Sequences (RefSeq), Related sequences, Additional links

Genome Browsers: Genome Data Viewer, Variation Viewer (GRCh37.p13), Variation Viewer (GRCh38)

Figure 5.2: Detailed information about selected gene sequence

**Ensembl genome browser 96** The second genome database used in this article, was the Ensembl genome browser 96<sup>13</sup>. It is a special genome browser for vertebrate genomes. In comparison to NCBI, Ensembl annotates genes and computes multiple alignments and predicts regulatory function. Besides, Ensembl collects disease data and supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. In April 2019, Ensembl published a new release which included a Representational State Transfer (REST) and File Transfer Protocol (FTP) Application Programming Interface (API). Apart from new genomes, Ensembl Release 96 offers a new interface to configure the regulation tracks.

Figure 5.3 shows the Ensembl result of the selected gene from NCBI database.

The screenshot displays the Ensembl genome browser interface for the gene **IKZF1** (ENSG00000185811). The interface is organized into a left-hand navigation menu and a main content area. The navigation menu includes sections such as 'Gene-based displays' (with 'Summary' selected), 'Sequence', 'Comparative Genomics', 'Ontologies', 'Phenotypes', 'Genetic Variation', 'Gene expression', 'Pathway', 'Regulation', 'External references', 'Supporting evidence', 'ID History', and 'Gene history'. The main content area is titled 'Gene: IKZF1' and provides detailed information about the gene, including its description, location, synonyms, and various identifiers. The 'Summary' section highlights that the gene has 23 transcripts, 182 orthologues, 11 paralogues, and is a member of 1 Ensembl protein family. It also mentions that the gene is associated with 68 phenotypes. The 'Transcripts' section includes a 'Show transcript table' button. The 'Summary' section further details the gene's name (IKZF1), its CCDS set (CCDS59055.1, CCDS69299.1, CCDS75596.1, CCDS75597.1, CCDS78232.1, CCDS78233.1, CCDS87502.1), UniProtKB identifiers (Q13422), and RefSeq information. The 'Ensembl version' is listed as ENSG00000185811.18. The 'Other assemblies' section notes that there is no ungapped mapping of this gene onto the GRCh37 assembly. The 'Gene type' is identified as 'Protein coding'. The 'Annotation method' section states that the annotation for this gene includes both automatic annotation from Ensembl and Havana manual curation.

Figure 5.3: Information about genome IKZF1, found in Ensembl genome browser 96

<sup>13</sup>[Ensembl genome browser 96 2019]

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

### 6.1 Problems and challenges of genetic analysis

First of all, importing all the needed data from NCBI was very easy. The NCBI website supports SQL-similar syntaxes so that the query could be as easy as follows: 'acute lymphoblastic leukemia 'AND' HOMO SAPIENS' (to only get the human genomes). By the way, the user of NCBI can choose between many different types of animal genomes.

After importing the data by using pandas, a dataframe object was created and all the relevant data (columns) were filtered. To be more precise, the columns 'symbol' and 'other<sub>designations</sub>' were very interesting since they gave insights about the 'core characteristics' of each gene. A value map was created automatically. During the process of data retrieval, there came up the problem of 'binding protein'. So, they had to be added by using a custom stopword list.

Furthermore, the most important words that should be clustered or put into topics, where compound expressions, such as 'enhancer binding protein'. One idea was to create bigrams or trigrams within the dictionary and to get all the appearances of these special expressions.

Is the data complete, which includes that it contains all required genomes which can cause ALL.

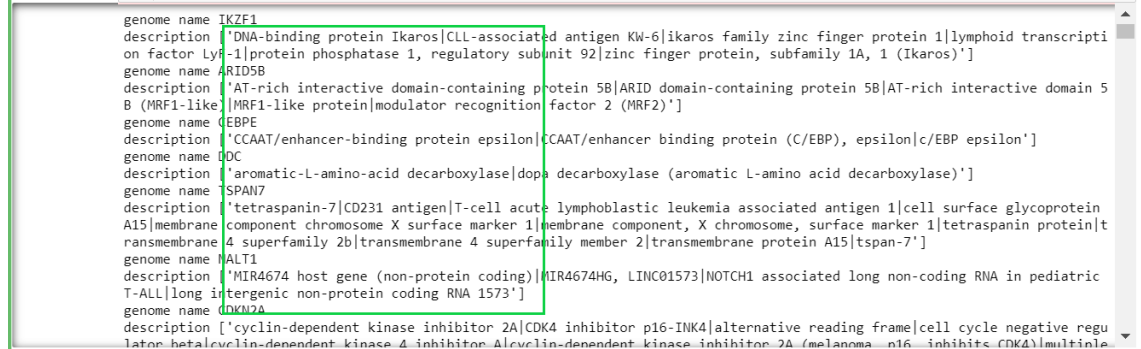
Another problem that came up with the time was the quality of the data. During the analyzation process, the data was extracted from the whole dataset. As can be seen in figure 6.1, every genome was described differently. Some are described by a

combination of numbers and letters (e.g. 'tetraspanin-7', others are described by their functionality (e.g. 'DNA-binding protein Ikaros'). These differences made the analysis and data preparation very challenging, since every description should be compared to the others.

```

17 for keys, values in df2_map.items():
18     print('genome name', keys)
19     print('description', values)

```



The screenshot shows a Jupyter Notebook with a dictionary of genome names and descriptions. A green rectangle highlights the 'description' values for several genomes, including IKZF1, ARID5B, CEBPE, DDC, SPAN7, and TAL1. The descriptions are as follows:

- IKZF1: DNA-binding protein Ikaros|CLL-associated antigen KW-6|ikaros family zinc finger protein 1|lymphoid transcription factor 1|protein phosphatase 1, regulatory subunit 92|zinc finger protein, subfamily 1A, 1 (Ikaros)'
- ARID5B: AT-rich interactive domain-containing protein 5B|ARID domain-containing protein 5B|AT-rich interactive domain 5B (MRF1-like)|MRF1-like protein|modulator recognition factor 2 (MRF2)'
- CEBPE: CCAAT/enhancer-binding protein epsilon|CCAAT/enhancer binding protein (C/EBP), epsilon|c/EBP epsilon'
- DDC: aromatic-L-amino-acid decarboxylase|dopa decarboxylase (aromatic L-amino acid decarboxylase)'
- SPAN7: tetraspanin-7|CD231 antigen|T-cell acute lymphoblastic leukemia associated antigen 1|cell surface glycoprotein A15|membrane component chromosome X surface marker 1|membrane component, X chromosome, surface marker 1|tetraspanin protein|transmembrane 4 superfamily 2b|transmembrane 4 superfamily member 2|transmembrane protein A15|tspan-7'
- TAL1: MIR4674 host gene (non-protein coding)|MIR4674HG, LINC01573|NOTCH1 associated long non-coding RNA in pediatric T-ALL|long intergenic non-protein coding RNA 1573'

Figure 6.1: Extracted data (including genome's name and description), the green rectangle includes the descriptions

## 6.2 First steps: Draft of developed solution

To get the disease related data, NCBI <sup>1</sup> was used to get all currently detected mutations of genomes which can cause LDA.

The first idea was to build a parsing application, which iterates over the found 582 genomes. During the iteration, four pre-defined topics are built. The topics included 'the most common nucleotide in the sequence', 'the nucleotide followed by its complementary nucleotide', 'the nucleotide followed by other nucleotides (excepts its complementary)' and 'the double distribution of the same nucleotide' (see figure 6.2).

After the iteration, it compares the oncogenes with the healthy genomes and to figure out where the differences are. The results could be displayed in a diagram. After that, the idea was to create clusters among the two groups or run the LDA algorithm.

After research, this approach was very naive and supervised because it does not give relevant information about the genomes' mutations. Moreover, there exist many sequencing techniques like Basic Local Alignment Search Tool (BLAST) to parse genome sequences and to find out alignments (similar sequences). For that reason,

<sup>1</sup>cf.[Biotechnology et al. 2019]

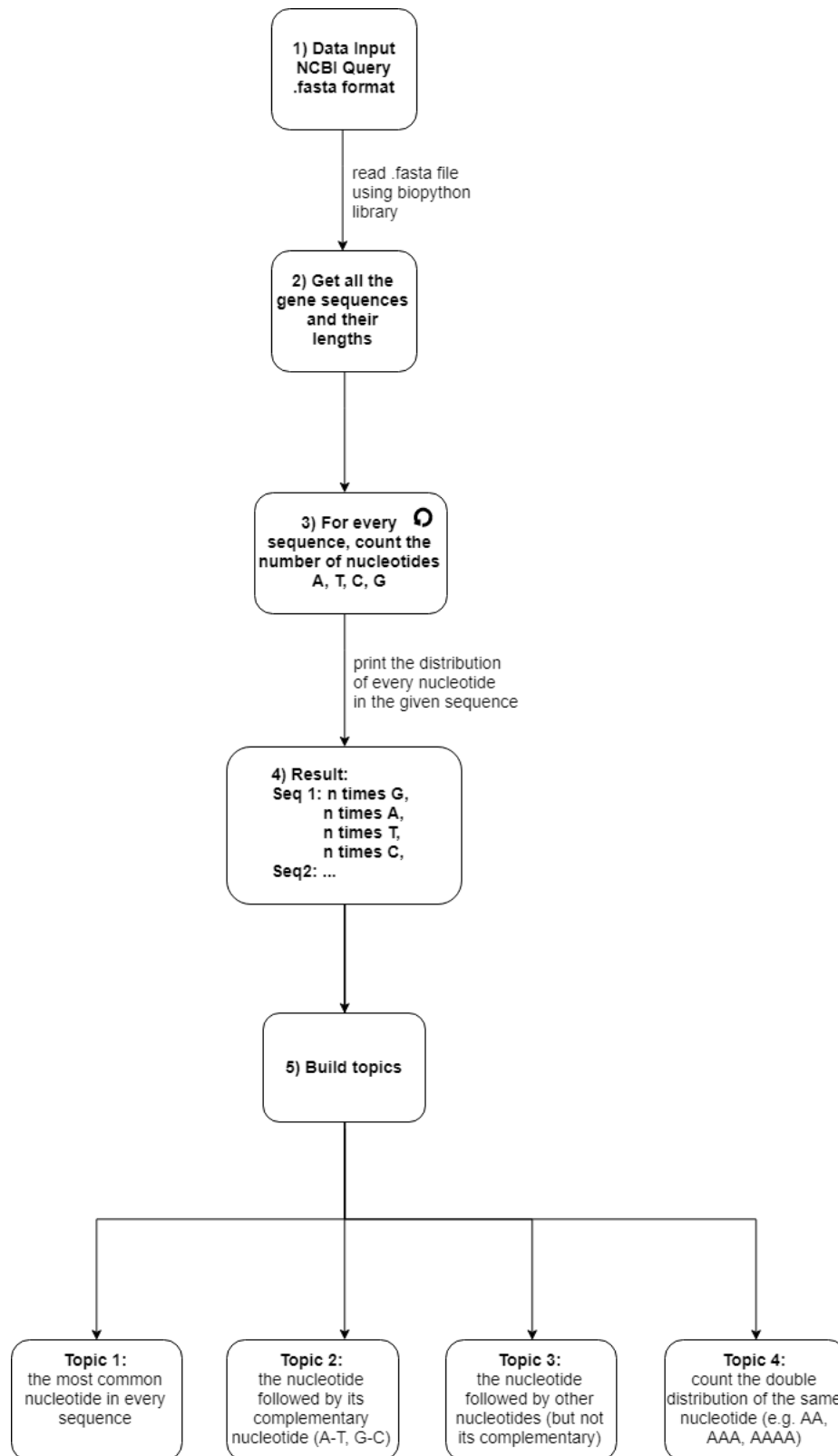


Figure 6.2: Diagram of first draft of developed algorithm

the first algorithm 6.2 was not important for the research question.

## 6.3 Proposed solution

The proposed solution can be divided into two processes: The preprocessing and the central LDA process. In the following both processes are described in detail showing diagrams.

Figure 6.3 shows the implemented solution. First of all, the required libraries and python packages have to be imported. After that, the dataset (genome data from NCBI) is imported and converted into a Pandas dataframe (table-like data structure). In the next step, only the relevant data is extracted from the dataset. Step five summarizes the LDA process (which is described in figure 6.5). In the last step, the topics are drawn in a topic distribution diagram using the library 'PyLDAvis'. An example of the digram can be seen in figure 6.4. PyLDAvis gives an overview of all large topics and draws them as large circles. Each circle can be clicked and shows the related words of the selected topic.

After the 'Preprocessing' step, the second process includes several NLP techniques, such as lemmatization, defining stopwords and removing them (see figure 6.5). Since the dataset differentiates from natural language, there had to be added genome-specific stopwords (see step four). With the rest of the data, a dictionary was created. This is relevant for the next step: define the number of topics and call the `LdaModel()` method from the gensim library. There are several parameters, such as the number of topics (which should be adjusted to the size of the dataset) or the number of passes (which means how often the dataset is being parsed) which can be configured. The result of this operation includes the most common words per topic multiplied with their distribution factor (dirichlet factor, see figure 6.6).

## 6.4 Results



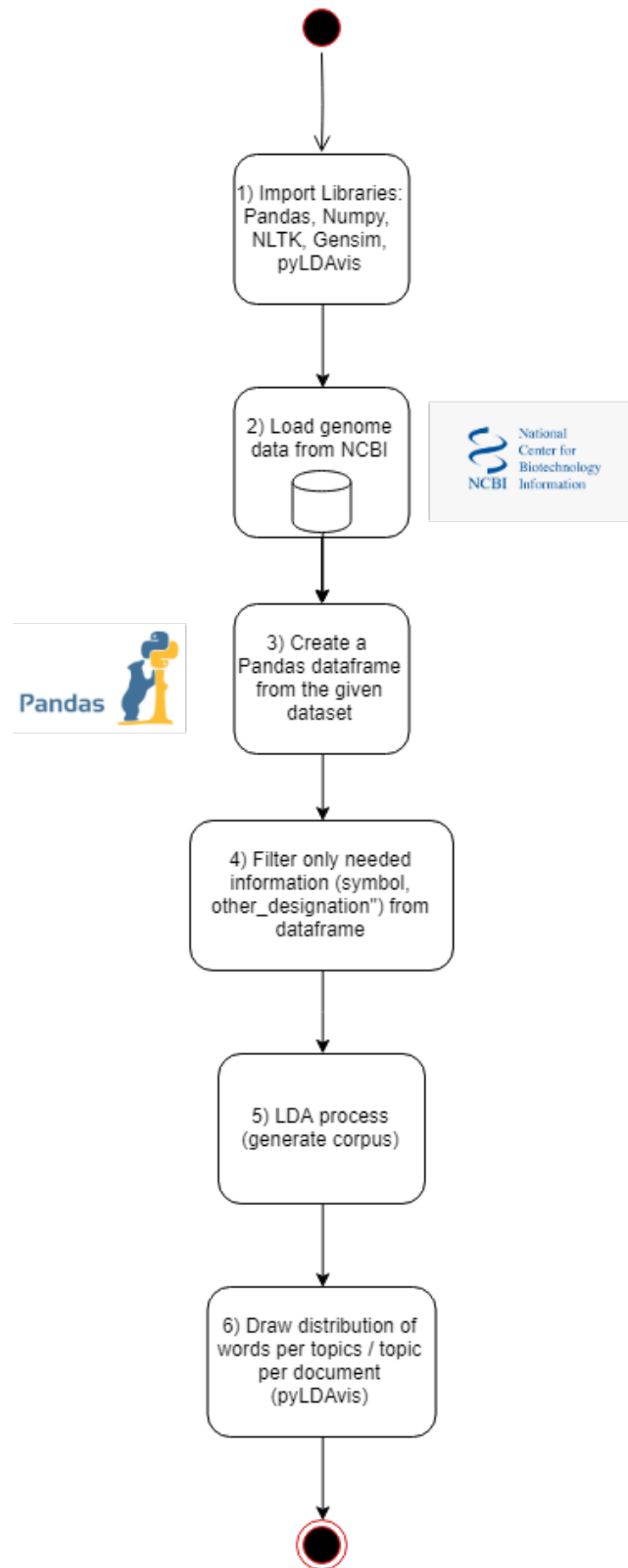


Figure 6.3: Diagram of developed algorithm to create basic topics among given genomes

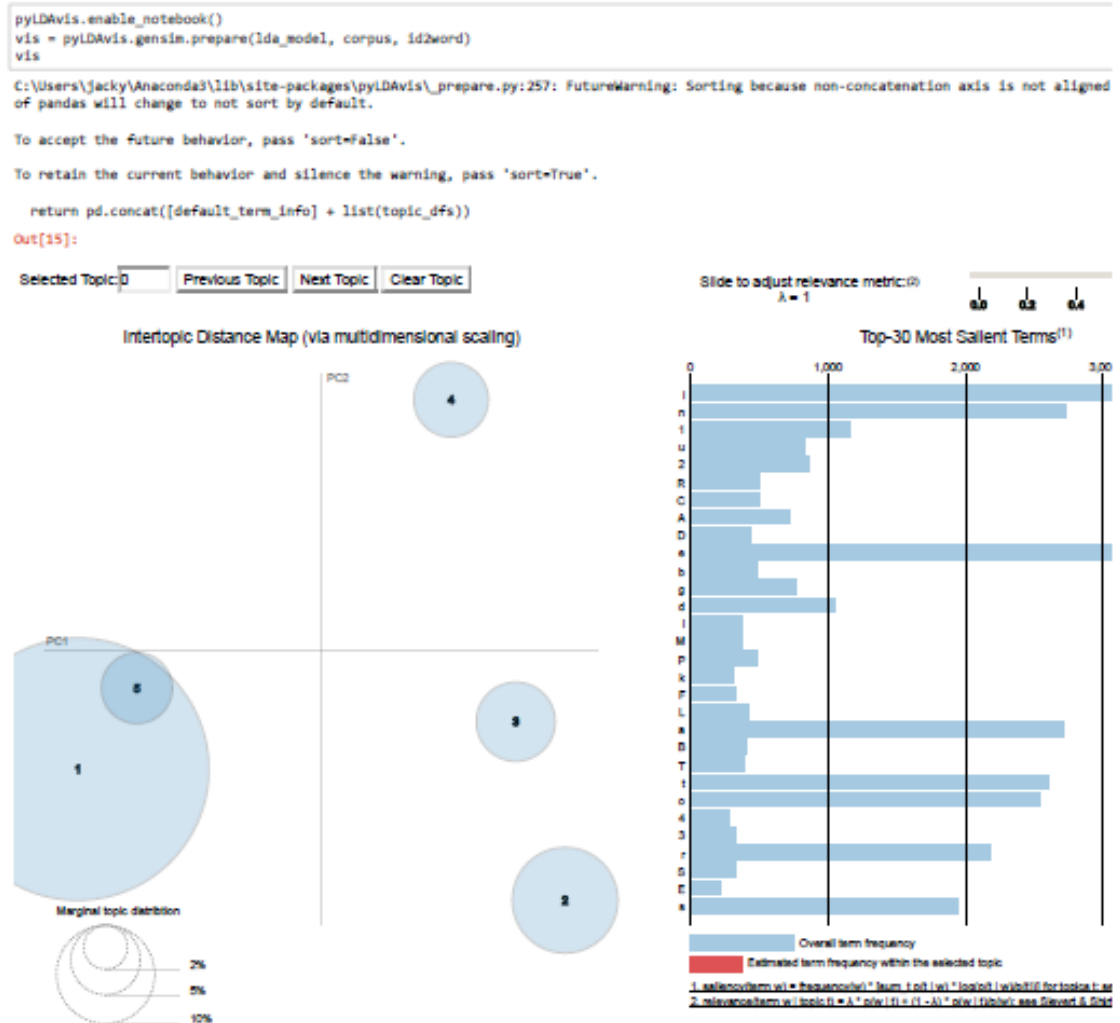


Figure 6.4: Diagram (created with PyLDAvis) which shows the distribution of the topics

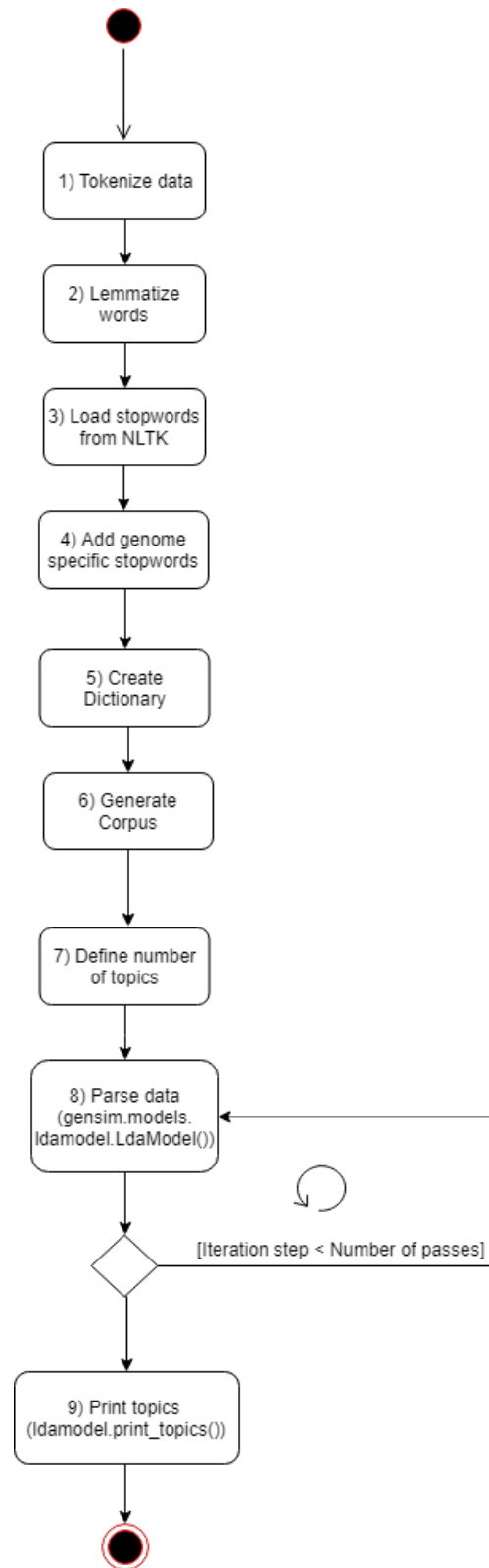


Figure 6.5: Diagram of LDA process

#### Build LDA Model (using corpus and dictionary)

- 5 topics
- 10 passes

In [12]:

```
# Build LDA model
lda_model = gensim.models.ldamodel.LdaModel(corpus=corpus,
                                             id2word=id2word,
                                             num_topics=5,
                                             random_state=100,
                                             update_every=1,
                                             chunksize=100,
                                             passes=10,
                                             alpha='auto',
                                             per_word_topics=True)
```

In [13]:

```
# Print the Keyword in the 10 topics
print(lda_model.print_topics())
doc_lda = lda_model[corpus]

[(0, "0.248*I" + 0.165*n" + 0.137*k" + 0.097*b" + 0.095*2" + 0.083*u" + 0.055*v" + 0.048*g" + 0.036*d" + 0.029*q"), (1, "0.226*L" + 0.076*I" + 0.065*3" + 0.063*5" + 0.048*H" + 0.044*5" + 0.038*0"), (2, "0.117*e" + 0.091*I" + 0.086*a" + 0.082*t" + 0.088*2" + 0.061*s" + 0.052*c"), (3, "0.177*c" + 0.155*0" + 0.148*2" + 0.098*4" + 0.098*8" + 0.069*9" + 0.068*A" + 0.064*K" + 0.0 + 0.148*I" + 0.146*M" + 0.128*f" + 0.085*E" + 0.085*2" + 0.084*6" + 0.065*7" + 0.029*h" + 0.019*y")]
```

Figure 6.6: Call of method LdaModel() from library Gensim and its output

## 7 Conclusion and Outlook

### 7.1 Lessons learned

There are many genome analysis tools in the internet. NIH provides several interfaces for genome alignment methods to find out the similarity between two given genome.<sup>1</sup>. But in the given case, all given gene sequences should be compared with each other in order to find appropriate clusters.

Another point is the algorithm of LDA. Since it is an unsupervised algorithm and is more convenient for natural language than for genome-specific data, it is difficult to build topics among the genome specific data and measure their similarities.

### 7.2 Conclusion

### 7.3 Outlook

As described in 7.1, the problem of the unsupervised clustering algorithm and the genome-sepcific data came up. As a proposal for future analysis, a supervised technique like K-Means-Clustering could be used to identify similar genomes and to build topics.

---

<sup>1</sup>cf.[NCBI BLAST ; Sequence Similarity Searching ; EMBL-EBI 2019]

## Bibliography

- Andersson, Anna et al. (Mar. 2016). "Acute lymphoblastic leukaemia". en. In: *The Genetic Basis of Haematological Cancers*. Ed. by Sabrina Tosi and Alistair G. Reid. Chichester, UK: John Wiley & Sons, Ltd, pp. 223–264. ISBN: 978-1-118-52794-8. DOI: 10.1002/9781118527948.ch5. URL: <http://doi.wiley.com/10.1002/9781118527948.ch5> (visited on 05/31/2019).
- 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).
- Ensembl genome browser 96 (2019). URL: <https://www.ensembl.org/index.html> (visited on 06/02/2019).
- 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).
- Group (CALLCG), Childhood Acute Lymphoblastic Leukaemia Collaborative (2009). "Beneficial and harmful effects of anthracyclines in the treatment of childhood acute lymphoblastic leukaemia: a systematic review and meta-analysis". en. In: *British Journal of Haematology* 145.3, pp. 376–388. ISSN: 1365-2141. DOI: 10.1111/j.

- 1365–2141.2009.07624.x. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2141.2009.07624.x> (visited on 05/31/2019).
- 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).
- Jagadev, P. and H. G. Virani (May 2017). “Detection of leukemia and its types using image processing and machine learning”. In: *2017 International Conference on Trends in Electronics and Informatics (ICEI)*, pp. 522–526. doi: 10.1109/IC0EI.2017.8300983.
- 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://bmresnotes.biomedcentral.com/articles/10.1186/s13104-016-2023-5> (visited on 04/06/2019).
- Lee, Hyunju, Sek Won Kong, and Peter J. Park (Apr. 2008). “Integrative analysis reveals the direct and indirect interactions between DNA copy number aberrations and gene expression changes”. eng. In: *Bioinformatics (Oxford, England)* 24.7, pp. 889–896. ISSN: 1367-4811. doi: 10.1093/bioinformatics/btn034.
- 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).
- NCBI BLAST ; Sequence Similarity Searching ; EMBL-EBI* (2019). URL: <https://www.ebi.ac.uk/Tools/sss/ncbiblast/nucleotide.html> (visited on 06/06/2019).
- 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.
- Řehůřek, Radim and Petr Sojka (May 2010). "Software Framework for Topic Modelling with Large Corpora". English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. <http://is.muni.cz/publication/884893/en>. Valletta, Malta: ELRA, pp. 45–50.
- "Systematic review of the addition of vincristine plus steroid pulses in maintenance treatment for childhood acute lymphoblastic leukaemia – an individual patient data meta-analysis involving 5659 children" (2010). en. In: *British Journal of Haematology* 149.5, pp. 722–733. ISSN: 1365-2141. DOI: 10.1111/j.1365-2141.2010.08148.x. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2141.2010.08148.x> (visited on 05/31/2019).
- Tang, W., H. Cao, and Y. Wang (July 2011). "Subtyping of Leukemia with Gene Expression Analysis Using Compressive Sensing Method". In: *2011 IEEE First International Conference on Healthcare Informatics, Imaging and Systems Biology*, pp. 76–80. DOI: 10.1109/HISB.2011.60.
- 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)