TALLINN UNIVERSITY OF TECHNOLOGY

Faculty of Information Technology

Department of Computer Systems

Aleksandr Kuhhar 104516

Topic modeling system based on additive regularization methods for online classification of song lyrics.

Bachelor's Thesis

Supervisor: Eduard Petlenkov

Prof.

Tallinn 2019

Declaration: I hereby declare that this Bachelor's thesis, my original investigation and achievement, submitted for the Bachelor's degree at Tallinn University of Technology, has not been submitted for any degree or examination.

Deklareerin, et käesolev bakalaureusetöö, mis on minu iseseisva töö tulemus, on esitatud Tallinna Tehnikaülikooli bakalaureusekraadi taotlemiseks ja selle alusel ei ole varem taotletud akadeemilist kraadi.

Aleksandr Kuhhar

Date: .....

Signature: .....

Contents

Abstract

Topic modeling system based on additive regularization methods for online classification of song lyrics.

Topic modeling algorithms family is an essential component in the natural language processing science field and plays a key role in extracting hidden topics from text documents collection. An algorithm analyzes documents and computes probabilities of the relation of words to topics and topics to document.

On the base of these algorithms was tested a hypothesis of possibility to analyze lyrics documents of all music genres with the model which was learned only on one music genre. Usage of the proposed music genre explained by the usage of much larger average vocabulary in documents and word count used to generate a song text.

Discussed thesis conducts a comparison results of classic algorithms such as *LDA*, *PLSA*, and proposed novel method of Additive Regularization of Topic Modeling, which shows the superiority of the last approach. These findings were proven by third-party assessors, which ranked topics quality by subjectively feeling of relation set of words to the abstract topic by their opinion. *ARTM* methodology suggests an opportunity for creating special restrictions depend on tasks.

With the usage of the *ARTM* approach was created an online system for classification a topic relation of previously unseen lyrics text.

This thesis is in English and is 24 pages of text, 3 chapters and 3 figures.

Kokkuvõte

Temaatilise modelleerimise süsteem, mis põhineb additiivse regulariseerimise meetoditel lugude klassifitseerimiseks internetis

Temaatilise modelleerimise algoritmide perekond on oluline osa loomuliku keele töötlemise teadusvaldkonnast ja mängib võtmerolli peidetud teemade ekstraheerimisel tekstdokumentide kogumitest. Algoritm analüüsib dokumente ja arvutab seose tõenäosust sõnade ja teemade vahel ning teemade ja dokumentide vahel.

Nende algoritmide alusel oli testitud hüpotees võimalusest analüüsida dokumente kõikide muusikažanride laulusõnadega mudeli abil, mis oli õpetatud vaid ühe muusikažanriga. Kavandatud muusikažanri kasutamine on seletatav sellega, et laulusõnade genereerimiseks on keskmiselt kasutatud oluliselt suurem sõnavara ja sõnade arv.

Arutatud lõputöö viib läbi klassikaliste algoritmide, näiteks LDA, PLSA ja välja pakutud uudse meetodi – Teematilise modelleerimise additiivne regulariseerimine (ARTM) - tulemuste võrdlust, mis näitas uue meetodi paremust. Tulemused olid kinnitatud kolmandate osapoolte poolt, kes reastasid teemade kvaliteedi, subjetiivselt hinnates sõnade seost abstraktse teemaga. ARTM metodoloogia pakub võimalust luua erilisi restriktsioone, sõltuvalt ülesannetest.

ARTM meetodi abil oli loodud veebisüsteem varem nägemata lugude temaatiliseks klassifitseerimiseks laulusõnade järgi.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 24 leheküljel, 3 peatükki, 3 Figuret.

List of Figures

List of abbreviations and terms

API Application programming interface

ARTM Additive regularization of topic models

EM Expectation-Maximization Algorithm

HTTP Hypertext Transfer Protocol

JSON JavaScript Object Notation

KL Kullback–Leibler divergence

LDA Latent Dirichlet allocation

PLSA Probabilistic latent semantic analysis

POS Part-of-speech

Python Programming language

REST Representational state transfer

Chapter 1

Introduction

1.1 Objectives and Contributions

The original idea of this work is creating a stable system for online classification topics of song lyrics, which was trained on a base of documents related to the hip-hop music genre. Such a specific choice of music genre can be explained with the help of descriptive analyses of corpora with song documents, which includes different music genres [1, 2, 3]. These analyses have shown that hip-hop music genre differs significantly from all other songs genres by an average number of used words in song and usage of unique words. This effect is fair due to specifics of creation song lyrics for hip–hop music genre. We will omit most primitive types of hip-hop songs and concentrate on complex examples of lyrics because they can include themes of primitive types due to usage of more abstract word plays, paraphrases, and other techniques to interconnect word and meaning.

To find hidden themes in the text collection will be used a machine learning method called Topic Modeling, which broadly used for the analysis of natural language processing tasks. Most nowadays used probability-based models will be described in the chapter about topic modeling methods, additionally will be demonstrated a new approach that outperforms classic models. The problem of Topic Modeling is that the model must satisfy different requirements which depend from task and data. The reviewed approach aims directly to increase the stability of the model and to form efficiently such topic models that perform to the task and can be calibrated directly for data used. On the base of this new approach will be created several topic modeling models and compared with each other. The most

suitable of them will be used as a classifier for new data. To make stable models, we need to solve several tasks, like data collecting, selecting optimal corpus of data on which models will be trained, calibrate hyperparameters of the models, optimize a number of topics on which corpus will be divided and analyze the inner structure of these topics. With the usage of the best type of topic model will be created a web classifier to detect topics on unseen data over the Internet via REST API HTTP requests.

1.2 Thesis Outline

Thesis contains of 3 chapters: introduction, theoretical description of topic modeling algorithms and chapter with practical implementation of topic modeling systems,

Chapter 2 present a brief introduction to the specifics of the Topic Modeling problem and a most used algorithms to solve these types of problems. The main focus will be given to methods based on additive regularization and description of their inner structure, learning process and the difference between model types with the usage of math equations. Overview of methods for automatic validation of performance and quality of topic models will be given at the end of the chapter.

In Chapter 3 will be described phases in development process of topic modeling system which further will be used for classification task. This process includes data collection, data preparation, selecting optimal parameters for each model and comparing models between each other.

Chapter 2

Methods and criteria of Topic Modeling

In this work, we are interested in findings of hidden relations of text documents by their inner text structure and joining them together by features that will be found. Speaking differently, we are looking for topics of interest in the unlabeled corpus of documents. Problem with analyzing song lyrics can be reduced to analyzing of plain text documents. It is good that the fundamental approaches to solving these types of problem have long been developed and already successfully used in all kinds of structures where is needed to work with this type of data.

General machine learning methods for natural language processing, which currently exist and widely used to analyze discrete text data like large text collections, can be usually divided into vector-based and probability-based approaches. In probabilistic based approaches, every topic usually described by the discrete distribution of word probabilities and every document by a discrete distribution of topics probabilities. These methods usually using matrix factorization of distribution of tokens to documents matrix to compute stochastic token-topic and topic-document matrices.

This means that we can describe each document as a mixture of topics and each topic as a mix of words. The same idea of describing a relation between objects to some sets of data is used in soft-clustering algorithms family. Onward we will use the term “Topic Modeling” to refer to probability-based approaches of analyzing text collection. Further will be briefly described about all most used topic modeling algorithms and how they will be used in this project.

When we are talking about the vector-based approach, we refer to word embedding mod-

els, where is semantically similar vectors of words will be situated closer between each other in this vector space. In vector-based models, each word is represented by vector space with usually more than 50 dimensions. This giving possibility to compare not only a distance but also compare entities by degrees of similarity of their vectors, nevertheless, coordinates of these vectors is impossible to interpret to human. Similarly, of words comparing, we can compare vectors of sentences or vectors of documents. The most used and most intuitive metric of comparing vectors is *cosine similarity.* At the moment most popular word embedding models is WORD2VEC [4] and FASTTEXT [5]. Word embedding or vector-based methods is beyond of scope of this work.

Topic modeling algorithms can be interpreted as soft-clustering algorithms which relating tokens to topics-clusters with a some probability. Currently the most popular and widely used method for creating such models is a *Latent Dirichlet Allocation* [6] (LDA), which was proposed as a solution of problem with overfitting of early suggested algorithm - PLSA [7]. In this work will be mainly used a new approach - *Additive Regularization of Topic Models* [8] (ARTM), which was used idea of decomposition as a PLSA algorithm, but with added combination of penalties for estimation log-likehood of distributions in Expectation-Maximization phase.

To move further, we need to formulate three hypotheses for the probability-based approach in the context of text documents analyze.

Bag-of-words hypothesis: We treat, that order of words in documents is not important for selecting a topic. We can mix words in document in any way and this will have no effect on the topic to which this document belongs, also this principle is true for the order of documents in documents corpus.

Hypothesis of existence of topics: Every occurrence of term $w \in W$ in songs lyrics document $d \in D$ is related to topic *t* from a set of specific *T.*

Hypothesis of conditional independence of topics: The appearance of term $w \in W$ in document $d \in D$ by topic $t \in T$ is independent of document and depends only from topic and can be described as a unified probabilities distribution:

$$p\left(w|d\right) = \sum_{t \in T} p\left(w|d,t\right) p\left(t|d\right) = \sum_{t \in T} p\left(w|t\right) p\left(t|d\right) = \sum_{t \in T} \varphi_{wt} \theta_{td} \tag{2.1}$$

where $\varphi_{wt} = p(w|t)$ - is a probability for token w in topic $t \in T$ and $\theta_{td} = p(t|d)$ is a probability of occurring of topic t in document $d \in D$.

However if chosen method based on stochastic matrix factorization it can lead to unsustainable solutions because we have infinite variants how we can factorize this type of matrix, due to this problem we need to use additional parameters which will affect on factorization process and lead to more controlled solutions. Generally this restrictions called as *regularization* and parameters of this regularization is selected and calibrated depends of task and data structure.

2.1 Topic modeling algorithms

2.1.1 LDA

LDA algorithm relates to generative probabilistic algorithms family where data is originated from the *generative process* which includes hidden variables. One of the relations between this hidden and observed variables is named a *joint distribution*, which is created by generative process. On the base of this joint distribution produced a *conditional distribution* of the hidden variables given the observed variables.

The idea of this model can be freely described in the context of the generative process: documents contain several topics and the proportion of these topics is different. Every token in each document is selected from one of the topics in the topics set, where selected topics is chosen from the per-document distribution over topics.

We can say more formally that each topic t in $t \in T$ is selected over distribution of tokens $w \in W$ and denoted as ϕ_t. Proportions of t for *d*th document in $d \in D$ is a θ_d. Both of ϕ_t and θ_d are drawn from *Dirichlet distribution* using hyper parameters $\beta = (\beta_w)_{w \in W}$ and and $\alpha = (\alpha_t)_{t \in T}$ respectively. Dirichlet distribution can create sparse and dense vectors of discrete distributions.

All computations of probabilities are based on *Bayesian inference* which used to compute *posterior probabilities* of distribution. The algorithm to approximate posterior inference usually based on a *Gibbs sampling* algorithm or more newer *Collapsed Gibbs Sampler* [9]. Techniques of approximate Bayesian inference not give a possibility for an easy

combination of models and formalize special requirements for them without needing to recalculate math equations and rewriting a programming base.

At the moment models based on LDA algorithm and its modifications are the most used approaches to find hidden topics of interest in text documents. Currently exist a lot of modifications of LDA algorithm [10] and successfully implemented in the production environment.

Further will be shown how to derive an LDA model in the context of additive regularization methodology.

2.1.2 PLSA and ARTM

Probabilistic latent semantic analysis (PLSA) was the first probability-based topic model. Equation (2) shown how collection of documents *D* generated from known distributions of $p(w|t)$ and $p(t|d)$ matrices.

Learning process of a topic model is the opposite problem and consider as a factorization of word-to-document probabilities matrix to two new stochastic matrices - documents-topics and words-topics.

Model learning is organized around a EM algorithm the task of which is to maximize a log-likehood with some linear restrictions. As known EM algorithm converges weakly without these constraints. [11].

Optimization problem can be written as follows [8]:

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} ln \sum_{t \in T} \phi_{wt} \theta_{td} \to \max_{\Phi, \Theta} \tag{2.2}$$

On each iteration EM algorithm trying to solve a system of non-linear equations:

$$\begin{aligned} p_{tdw} &= norm_{t \in T}(\phi_{wt} \theta_{td}); \\ \phi_{wt} &= norm_{w \in W}(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}}); \\ \theta_{td} &= norm_{w \in W}\left(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}}\right); \end{aligned}$$

The system of non linear equations used to approximate values in matrices ϕ_{wt} and θ_{td} on base of these matrices computed p_{tdw}variables, which used to approximate next values of matrices $\phi_{wt}$and θ_{td}. This process will continue until convergence or the limit of iterations.

Additive Regularization of Topic Models approach is based on idea of maximization of linear combination of log-likehood and regularizers with coefficients of regularization τ_i

.

$$L(\Phi, \Theta) + R(\Phi, \Theta) \rightarrow \underset{\Phi,\Theta}{max}$$

where $R(\Phi, \Theta) = \sum_{i=1} \tau_i R_i(\Phi, \Theta)$ and $L(\Phi, \Theta)$ is formed in (2.1.2)

2.2 Regularization

The main task of Topic Modeling is a selection of *good* topics, where good can be defined by a set of requirements. Good topics should be easily interpreted and be decorrelated from other topics. One of the conditions of interpretability is a sparseness of the data. Each topic should be described by a small set of tokens and document relates to small amount of topics, that should be expressed in a large number of zero columns in $\phi_{wt}$and $\grave{\theta_{dt}}$matrices. Topics should be decorrelated, which means that they should differ as much as possible from another topic. To form such restrictions in the initial work of ***ARTM*** was proposed to use a combination of regularizers. The process of regularization is used to restrain an EM algorithm in a way that is needed to achieve suitable results for a specific task on our data.

We can define each of the previously submitted probabilistic topic models in the context of *Additive Regularization of Topic Models*.

LDA model can be rewritten in terms of $R(\Phi, \Theta)$ as follows: $R(\Phi, \Theta) = \sum_{t,w} \beta_w ln \phi_{wt} + \sum_{t,d} \alpha_d ln \theta_{td}$, where variables α and β is a parameters of Dirichlet distributions.

PLSA model can be interpreted as a topic model with zero as regularization $R(\Phi, \Theta) = 0$ because PLSA model cannot contain degenerated topics or documents otherwise this will lead to zeroes in columns of matrices, what is unfavorable to process of maximization of log-likelihood.

In ARTM approach we treat $R(\Phi, \Theta)$ as a combination of constraints $R(\Phi, \Theta) = C(\tau, \Phi, \Theta) = \sum_i \tau_i R_i(\Phi, \Theta)$, where $\tau_i$is a coefficient of regularization for each regularizer. This leads to the possibility to fine-tune all restrictions specially for the task which is needed to be solved. It is possible to combine them freely between each other and enable or disable in any moment between learning processes. Regularization criteria can be expressed not only in the language of probability but also as a numeric restrain.

To compare the difference between multinominal distributions in the learning process is used a KL divergence, KL is a non-symmetric function which is used to compute a measure of inclusiveness of model distribution q into empirical distributions p and infers as:

$$KL(p||q) = \sum_{i}^{n} ln(\frac{p_i}{q_i})p_i$$

Minimizing of KL divergence leads to increasing log-likehood between distributions.

In this work will be used several regularizers, which characteristics will be described below.

ϕ_{wt} *sparseness regularization* - τ can be as negative and as positive number. Decreasing values of τ leads to increasing number of zero elements in ϕ_{wt}matrix, which minimizing KL divergence and increasing values of τ is maximizing KL divergence between distributions.

θ_{td} *smoothiness regularization* - Same idea as ϕ_{wt}but for topic-documents matrix

Topic decorrelation regularization - idea behind this regularization is to increase a probabilities of more important topics t in token w and decrease non important topics probabilities in ϕ_{wt} for this token.

However process of selection best values of τ_i for regularization is a heuristic process which depends from many factors as a size of collection, volume of vocabulary and other parameters of data. To optimize the process of finding suitable values of τ_i in this work is proposed a usage of *Distributed Asynchronous Hyperparameter Optimization [12]* (Hyperopt) approach.

2.3 Multi-modal and hierarchy model

2.3.1 Multi-modal ARTM model

We can define a document as a container with several modalities that can be used to describe it. In this work we will create a models which is primarily based on two modalities: song text and refrain. Text in them will be used as a bag of words, which means that we don't care about the order of tokens in them. Author name and other meta-data will be omitted because we are not interested in building a model for finding relations of authors to topics or building some kind of temporal model if we are talking about time-related modalities. *ARTM* extension [13] to handle modalities can be denoted and inferred as:

Let M is a set of modalities, every $m \in M$ have own vocabulary W_m which is non-intersected pairwise. Multi-modal model is created by maximization of weighted sum of log-likehood of modalities and regulizers.

$$\sum_{m \in M} \mu_m \sum_{d \in D} \sum_{w \in W} n_{dw} ln \sum_{t \in T} \phi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

$$\sum_{w \in W_m} \phi_{wt} = 1;\ \phi_{wt} \geqslant 0;\ \sum_{t \in T} \theta_{td} = 1;\ \theta_{td} \geqslant 0;$$

where $\mu_m$is a coefficients of modalities which can be adjusted accordingly of their importance or influence rate on model. Distribution of topics in each document is a same for all of modalities.

2.3.2 Hierarchy ARTM model

Idea of dividing topics recursively for sub-topics is widely and successfully used in LDA models, In related work [14] idea of hierarchies was adopted to ARTM approach and based on the idea of hierarchical divisive clustering. Zero level of this model will be a whole collection, first level will contain a small number of ancestor topics and all other levels will have a greater amount of topics and be build iteratively based on parent-child topics coefficients of relation. Each level can be described in the context of *flat ARTM model*. Each child topic will have at least one parent topic as an ancestor, number of

parent topics usually is a relatively small 1-5. On each iteration is used Φ or Θ interlevel regularizer to compute parents for topics:

$$\sum_{t\in T} n_t KL_w \left((p(w|t)|| \sum_{s\in S} p(w|s)p(s|t) \right) = \sum_{t\in T} n_t KL_w \left(\frac{n_{wt}}{n_t} || \sum_{s\in S} \phi_{ws}\psi_{st} \right) \to \min_{\Phi,\Psi}$$

where $\Psi = (\psi_{st})_{SxT}$ is a link matrix which is used as additional matrix of parameters for child level model. Topics of all levels represented by their distributions over words.

By last studies most powerful model in context of analyze of huge text collections and generating most valuable topics is build on base of a hARTM approach. In this work hARTM idea used to build simple hierarchy model just to try this approach among other ARTM based models.

2.4 Criteria of model quality

Topic modeling can be characterized as semi-supervisor learning problem what mean, that model can be learned by itself, but control of quality of this model is often lying on the shoulders of human. Of course, currently developed a lot of good metrics which is correlated with human decisions in question 'what is a good topic?', but often after topic extraction is needed to resort to human assessment at least at the stage of experiments because not all tasks can be reduced only to better metric values. In this work we will use a model metrics which is *Perplexity*, *Sparseness ratio, Coherence* and several topics metrics, such as a *Purity, Contrast, Size.*

Perplexity can be described as a measure of surprise of occurring tokens in document and inferred as

$$perp(D;p) = exp\left(-\frac{1}{n_d} \sum_{d\in D} \sum_{w\in W} n_{dw} \ln p\left(w|d\right) \right)$$

With smaller values model prediction of sample is greater. This is the most used metric to compare topic model quality. One of indicator of converging of model parameters is a decreasing of difference of perplexity between iterations.

Bad side of perplexity that metrics will vary from vocabulary on which model is built

and this means that we can't compare by this metric same models which were trained on different vocabulary

Sparseness ratio - Topic should be described by a small set of words as we mentioned before. To achieve higher values of sparseness ratio we need to increase count of zeroes in ϕ_{wt}or θ_{td}matrices can accordingly. It can be done with usage of certain regularization for sparse or smooth of matrices. In previous works it was noticed that model with larger number of good and interpreted topics showing a high sparseness ratio.

One of the methods that allow to rank topics independently and increase their interpretability consists in the introduction of the *topic kernel* [15], which is based on the idea that topic should consist of tokens which are used more frequently in certain topic and less inside other topics. The kernel defined as a $W_t = \{w \in W | p\,(t|w) > 0.25\}$, where t is a set of tokens which probabilities is higher than a conditional probability $p\,(p|w) = \phi_{wt}\frac{n_t}{n_w}$ of this topic. Conditional quality of topic kernel can be measured by *purity*, *contrast* and *size* of the topic.

Topic purity infer as $\sum_{w \in W_t} p\,(w|t)$ and topic contrast as $\frac{1}{|W_m|}\sum_{w \in W_t} p\,(t|w)$ where bigger values is usually is better for intrepretability of topic. Kernel size is a count of tokens that belong to this topic and can be useful if we have a division between subject and background topics where background topics should have most frequent words across all topics. Model quality can be also ranked by average value of each of kernel characteristics.

Further, as a framework for topic modeling will be used a brilliant open-source library BigARTM [16] which is based on ARTM methodology. In this framework is already implemented all necessary types of regularization and metrics. Framework is very fast, supports multithreading by default and extensible because users can implement their own regularization rules and metrics. It also support learning in offline or online mode. Learning in offline mode usually used to build a model when it is possible due to size of data. If data is so big that can't be stored and flow of this data is fast can be used a online learning method. In offline mode algorithm updating weights after each iteration by whole document collection and in online mode weights in matrices updated on fly after every document or batches of documents. In this work be used offline learning method due to we have a limited size of documents and collection of data can be stored locally.

Chapter 3

System development

The creation of a sustainable model is a difficult task because all requirements should be meet and phases should be done. The development process can be separated into several phases, such as data preparation, model prototyping, model validation and building a web system. In data preparation phase will be described steps of data collection, data preprocessing and statistic of the data. Several topic models will be built with the help of BigARTM framework in model prototyping and validating phases. In the phase of development a web system will be described the process of creation a working solution for online classification of song lyrics.

3.1 Data preparation

One of the most important tasks of the development process of a sustainable machine learning model is the creation of an appropriate pool of data on the base of which model can be learned. A phase of data preparation consists of data collection, data preprocessing, descriptive analyze and creation pipeline for the model. Python programming language and several third party libraries were used to create a web-scraper script that parsed data from several websites with Russian hip-hop songs lyrics. The total was collected 13118 documents with author name, song lyrics and song title attributes. Usually song lyric formatted as a text document where every new sentence is placed on the next line and consists of several text blocks of couplets of the song and refrain which is alternate each other. Often in the hip-hop genre, there are no refrains and whole text consist of one big

couplet.

Each document has a song title, artist name and song text modalities. For every document was generated a refrain modality that was extracted from song text of this document and related to most frequent repeated sentences in this song text. Sentences of refrain modality was stored as a unique sentence, which appears more than one time in song text. 33% of songs doesn't contain any refrains and the average inclusiveness of refrains to song text is 15% because of natural aspects. Words in the documents can be described in various forms but can have the same meaning; this relates to the conjugation of words, the difference between endings of words and grammar mistakes. To improve overall model performance and return words to its base form is used *lemmatization* technique which uses the morphological analysis of the words to process. Task of lemmatization was done with help of the library MyStem [17], which was created specially for analysis of Russian language. To increase the diversity of initial data also was decided to form two additional corpora of documents with POS tagged words and one with *bigrams* of words lemmas.

POS tagging can be described as a process of marking up words to it part of speech where selected part of speech depends of the context. For our task was selected only two speech parts such as verbs and nouns and all other words was omitted. Tagging process was performed by RusVectōrēs [18] framework on base of "Taiga" corpus [19] which have almost 5 billion of words and weight 331 megabytes unpacked. Tagging process was finished in about 13 hours. This approach can be unacceptable for classification of song lyrics on fly because of slow speed of words tagging and necessity to store the whole tagging model in memory.

At the end of the data preparation stage was selected 2 corpora of documents: single modality with words lemma, text song and refrains modalities based on words lemmas. The main focus of the work will concentrate on most simple corpora with single modality of text lyrics lemmas. These corpora were selected by the author as *optimal* for initial model testing. Of course, data can be more various and contain more modalities such as author name, published date or any other feature which will be valuable to describe data in a more complex way and to increase descriptive power of model. All corpora were translated to VowpalWabbit text document format which saves each corpus to a single text file and each line contains song data accordingly requirements of BigARTM. The initial songs lemmas vocabulary contains of 83212 unique tokens. To reduce vocabulary

on which model operates is used additional filtering of tokens by their frequencies of occurrence in the collection and frequency of documents which contain this token. Minimal term frequency was set to 100 and maximal to 15000 where documents frequencies were set to 40 documents for minimal occurrence of rare words and for upper threshold was selected value of 2600 documents. This procedure is required because helps to filter tokens which occur very often and unique tokens such as a rare words or typos. After all filtering procedures remains 3515 tokens.

3.2 Model prototyping

Whole process of building and learning models for topic modeling was accomplished by usage Python and BigARTM topic modeling framework. At the early stages of model building was used a web application for rapid Python code prototyping and data presentation called *Jupyter Notebook* nonetheless development platform was changed due to the lack of the ability to properly organize the project, size of the project and mismatch of file format which is usually used to launch the Python code.

Models which were built: flat LDA model, flat PLSA, flat ARTM model with single text modality, flat multi-modal ARTM model with song text and refrain modalities , hierarchy ARTM which is based on song text modality. LDA and PLSA based models will be used as a baseline which scores is need to beat.

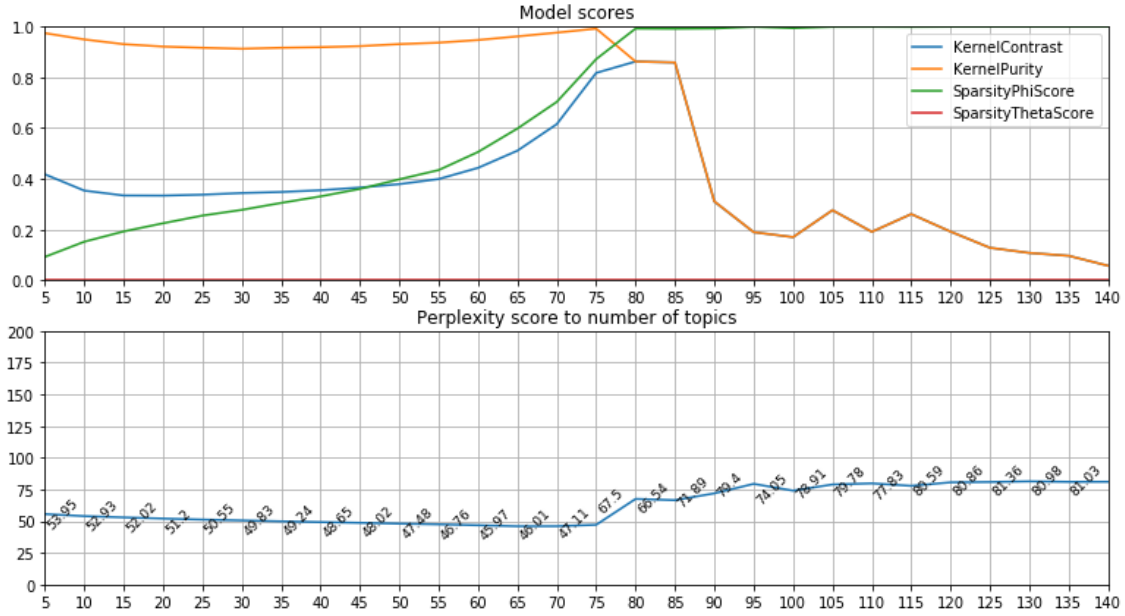
Flat and hierarchy ARTM models with single modality were learned on all corpora of data and flat multi-modal only on a suitable corpus of data with two modalities. Special restrictions of ARTM models will be formed by combination of regularization. All ARTM based models will use a same combination of regularization such as *SmoothSparsePhiRegularizer, SmoothSparseThetaRegularizer, DecorrelatorPhiRegularizer*. Besides coefficients of regularization all models had additional hyper-parameters such as a number of topics and a number of iterations. Hierarchy ARTM model also require to configure inter-level coefficients and multi-modal requires to configure a coefficients how modality impact to topic. The impact of the modality of refrains to song text modality will be 1:4 because of refrain modality had much less size in the context of proportion of raw song text. All regularization will be added to model from start of the learning process and will be active from the start. All other strategies with adding and enabling regularization at a certain point of

the learning process were unsuccessful. They shows good metrics, but topics were degenerated and contain elementary words that barely can describe a topic. Additionally was tested several other strategies of regularization such as: enabling *SmoothPhiRegularizer* for background topics at the start and adding *SparsePhiRegularizer* to all subject topics after, enabling all base combination of regularization at start and increase their strength every several iterations over the corpus. To obtain a optimal strategy should be evaluated and tested all other strategies of learning a model. BigARTM is allowing to easily enable, remove and change values of regularization when learning process is not active.

Selecting a proper number of topics is very important for the quality of topics because model with a lower number of topics will have vague topics, that will contain other topics. A high number, on the contrary, leads to the fact that model will select too specific topics. Optimal number of topics is depended of corpus size and requires a heuristic approach that requires iteratively build a several models with specified and same coefficients of regularization but a various number of topics.

To form a dependency plot of number of topics and model scores was used a flat ARTM model with a single modality, coefficients of this model was selected by Hyperopt algorithm, which tries to minimize the loss function of the model by selecting and evaluating parameters from restricted hyperparameters space. In our case the Hyperopt algorithm tried to select models with smaller values of perplexity score. As already known, matrix factorization problem can be solved in infinite ways so it is very common to have almost the identical perplexity scores but radically different coefficients of regularization. Manually was selected top result of this optimal parameters searching process and evaluated on topic number range space from 5 to 140 with step of five.

As we can see in Fig 3.1 after seventy-five topics model metrics starts to drastically change their behavior. Kernel purity and kernel contrast metrics started to going down and sparsity of ϕ_{wt}drastically increased to the value close of 1 what can sign to model degrading due to only biggest topics will be filled with a minimal count of tokens, all other topics will be empty or contain minimal amount of unrelated words. Very sparse ϕ_{wt}matrix can be useful where we have huge amount of data due to we can create many versatile clusters. The sparseness of $\theta_{td}$is close to zero because HyperOpt algorithm selected strategy to make θ_{td}more smoother instead of sparse. Lowest perplexity score was achieved on 60 topics and after 75 topics perplexity is started to raise. We will select 65 topics as optimal



This plot describes the relation between tuned ARTM model parameters to number of topics.

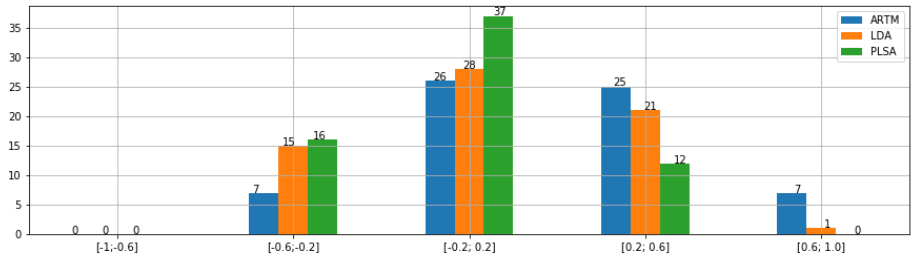
Figure 3.1: Topic number selection

because a difference of perplexity scores between 60 and 65 is almost the same but model with 65 topics have better metrics of all other scores.

Author understands that selecting optimal coefficients of regularization which is based only on one metrics such as perplexity is not the best method because this type of optimization task have only a local minimum and depends from initial state of regularization. One of the solutions of findings optimal minimum can be in a combination of proposed metrics [20] and Hyperopt algorithm or approach [21] with comparison of the intersection of kernel coherence and kernel purity but a realization of this ideas is out of the scope of current work.

After evaluating of proposed ARTM model was found that inverting of the regularization coefficient of θ_{td} tends to generate much more interpretable topics but with lower overall metrics, instead of idea to increase smoothness as was selected by HyperOpt algorithm. Further this model will be used as optimal ARTM model.

LDA model coefficients was selected as equal to $\alpha = \frac{1}{n_t}$ and $\beta = \frac{1}{n_t}$ as proposed as optimal in related work [22]. Regularization parameters of PLSA model were set up to zero. All models were learned on the same corpus of data and passed the whole collection for 75 times, the trend to increase values of collection passes didn't bring desired results.



The plot describes the relation between mean assessors choice to the interpretability of the topic.

Figure 3.2: Topics quality by assessors choice

Topic quality of all models was evaluated by third-party assessors with the usage of *Yandex Toloka* platform. Each topic of each model was evaluated by 75 random assessors by a mix of top ten words related to topic. Assessors had 3 options to vote: 'Words forms the topic', 'Hard to select one topic, possible that topic set describes more than one topic', 'Words can't be bounded into topics e.g random words, noise words, bad correlation between words and their meaning' and bounded to number values 1, 0, -1 accordingly. Interval of $[-1; 1]$ was divided to five equal parts e.g $intervals = [-1; -0.6) \cup [-0.6; -0.2) \cup [-0.2; 0.2) \cup [0.2; 0.6) \cup [0.6; 1]$. Answers was processed and mean values was formed by the answers for every of topic. On base of this mean values of topics was selected alignment of topic to specified intervals.

We can see on Fig3.2 that topics formed by ARTM approach are showing better results than LDA and PLSA based models in context of interpretability for users. This data can be used as necessary evidence in favor of using the ARTM based model as a optimal approach for *Topic modeling* task on this data corpus. Similar to user decision of tokens-to-topics correlation can be used a metric named *coherence* which is highly correlated with human decision [23]. Unfortunately, implementation of the *coherence* metric for this work was failed due to topics that were selected on base of *coherence* score was too 'data specific'. The coherence ratio was compared with same corpus, comparison on the external corpus was didn't performed.

Other than flat ARTM model was also evaluated a ARTM model with two modalities of song text and refrains, and Hierarchy ARTM model, but metrics will be not shown due to reasons which will be described further. Idea of learning ARTM model with two modalities seems optimistic at first, but evaluating on practice leads to bad topics selection due

to refrain model is formed with deleting refrains from song text. This extraction creates a refrain modality which is very limited by words and because of that vocabulary of song text modality is greatly reduced what have an adverse effect on the model. This leads to harder process of validation especially when refrain meaning is highly decorrelated with overall meaning.

To test the hierarchy ARTM approach was created three flat ARTM models and stacked together to generate a hierarchy architecture. First level extracts 16 topics, second - 32 topics and third 64 topics. As was mention before it is needed to increase the number of topics on each next level due to generating a sub-topics on the base of existing parent topic. Last level generates topics which can be described as a child topic but diffidently not always. Interesting that lower level started to select very strict topics such as 'about colors', 'about numbers' and etc. This behavior is definitely needed to be investigated deeper in future. Usually hierarchy ARTM model showing much better results on bigger and more cohesive data corpus and this is true in view of the fact that with hierarchy ARTM approach is it possible to divide any topic as much as it needed to liquidate moment when topic is so big that contain other sub-topics. But this versatile leads to problems with tuning of hyper-parameters because count of them is increasing greatly for each new added level and you need to tune not only hyper-parameters on each level and also a coefficient of inter-level regularization. Overall hierarchy ARTM approach shows a great versatile of how model can be configured and which strategies can be used to teach a model, but this level of complexity is not really needed for our task.

Figure 3.3 reveals inner parameters and scores of the tuned flat ARTM model. The sparseness of ϕ_{wt} matrix is close to 75% percent and this result is acceptable in the context of our data due to increase of sparseness will lead to degrading a model. Usually sparseness of ϕ_{wt} matrix is should be close to 90% or more, but this result is unobtainable without a degrading a other model parameters. Our result can be connected to fact about the data that tokens inside document is not so cohesive as a traditional text documents where topics can be selected much easily and text of this documents contain more strict relation to logic than in our corpus data due to poetic nature of our data. Sparseness of θ_{td} matrix is close to 48% percent but is not that important for our task because model will be focused to classify new documents on base of ϕ_{wt} matrix tokens. As we can see perplexity score was greatly reduced from the previously score of trained model with values of ~46 of

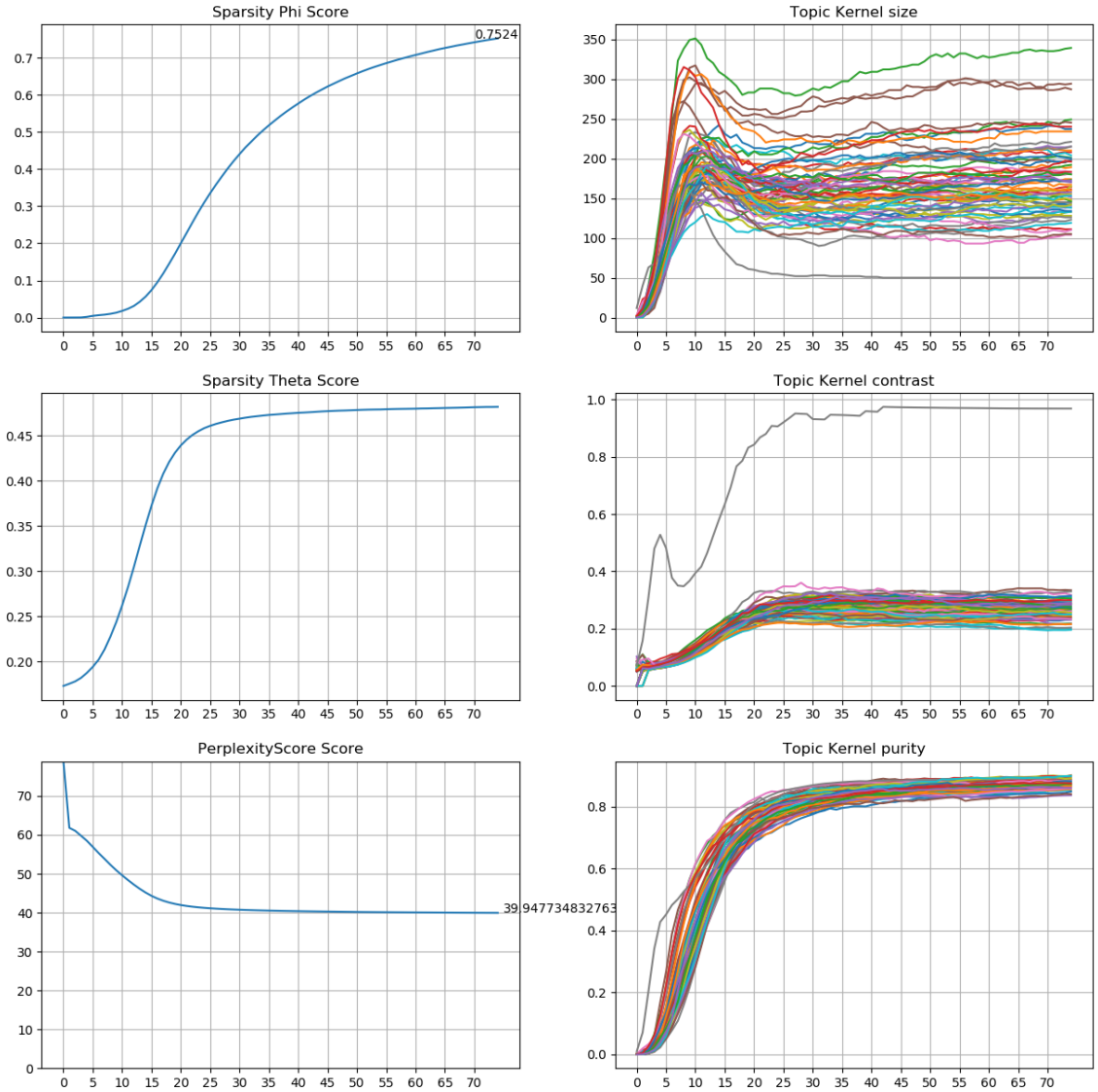


Figure 3.3: Evolution of tuned flat ARTM model in relation to iterations over data corpus

perplexity to the final value of 39.94 perplexity score. Each topic in average have around 150 words and topic purity score is around 87% for all topics. We can observe an interesting moment with topic contrast score. One of the topics achieved a much higher score in kernel purity metric. This is a topic that contains only English words and this is true due to the fact that all other tokens of the corpus are related to the Russian language. But average kernel contrast is close to only 28%. This final state of the model will be used further to classify new text documents.

On provided link[1] you can find web directories that contain a HTML files with final models metrics and visualization of each model. One HTML document contains a descriptive analysis of the model such as coefficients of regularization, list of tokens which relate to every topic, model final scores and plot of model scores which depends on iteration count over the collection. Other HTML files contains an interactive visualization of multidimensional scaling of topics which generated by LDAvis algorithm [24]. Interactive visualization is available in two variants: t-Distributed Stochastic Neighbor Embedding [25] (t-SNE) and Multidimensional scaling (MDS). These visualization algorithms is trying to reduce initial dimension of n-dimension space data corpus into lower dimension space. These visualization algorithms is widely used to get a understanding how clusters of the data is relation to each other. Clusters with intersections commonly have same words in vocabulary. LDA model directory contain only visualization file due to LDA model was created as a separated class of the BigARTM framework.

3.3 Building online classification system.

The classification task on new unseen data will be performed by our tuned ARTM model. Currently, topics don't have any names and it is needed to give names to them. The process of naming is accomplished by the author and based on the subjective feeling of relating overall meaning of topic tokens to some topic name. In the situation when topics contain more than one meaning will be used an abstract unification to single topic and named accordingly to contained mean of this topic. Topics that contains of non related words will be omitted.

The topic classification web-based system was created with help of FASTAPI Python

[1] `https://deepfatsnail.eu/topic_modeling_models/` (available till 10.11.2020)

framework and backed by Nginx server. FASTAPI framework is developed especially for rapid API prototyping and can work in asynchronous mode when results of sent user query will be served by the server when they will be ready without blocking all further user requests. Communication between client and server followed REST principles by using a uniform and predefined set of the stateless operations.

As a document, the server awaits a JSON file with three attributes: mandatory *doc attribute* and non required attributes *with_tokens and top_k_topics.* *doc* attribute should contain a song lyrics text and be a type of *string*. It is recommended to separate sentences by new line sign due to preprocessing of the documents, but it is not a strict guideline and whole song text can be stored without these special characters. Attributes *with_tokens* and *top_k_topics* defines a fullness of response from the server.

The server response will include also a tokens that model relate to topic if *with_tokens* is set to *true.* Amount of top returned topics defined by *top_k_topics* and should be *integer*.

Documentation and guidelines with usage a classification system can be found on related site[2].

The classification system shows satisfactory results, but not as good as it was expected at first. The model can predict topics with high probability if song text meaning describes a feeling, feature or another process which present in topics context of our model. Model will skip a general theme of documents if such theme is not presented in the initial model. For example, song text from rock music genre where signs about process of 'love to rock', part about 'rock music' will be ignored due to this topic is not present in initial corpus in decent size, but model will select topic about love. All texts which are semantically related to most major topics of our model, such as *relation with police, street life, love, family* is predicted with much higher probability of success.

[2] `http://deepfatsnail.eu:7650/docs` in future system can migrate to the subdomain `https://lt.deepfatsnail.eu/docs` (available till 10.11.2020)

Conclusions

In present graduation thesis was investigated a possibility of creating a stable web-based theme extractor and classifier of song lyrics documents. The main peculiarity of the proposed probability-based topic modeling task was a creation such topic model which was learned on documents corpus related to the single music genre with possibility of classification all other music genres. This model should meet other requirements to achieve acceptable results.

Initial corpus of data was collected by scraping sites with rap song lyrics and contains of 13118 documents.

Evaluation models performance on the data reveals a hidden question about an optimal number of topic for corpus of data. Further studies about optimal number of topics show that this specific corpus of data can be maximally divided to 75 sustainable topics but best results of the model evaluation were achieved at 65 topics.

On the base of knowledge about an optimal number of topics was performed test and comparison across popular topic modeling algorithms, which help select the most suitable topic modeling approach. Hyperparameters was tuned with the help of HyperOpt algorithm which use Bayesian Optimization for finding suitable values of parameters to minimize the loss function. To simulate loss function was used a popular perplexity metric. However, process of model optimization which based only on one metric cannot be considered sufficient and question about optimization of the model parameters is remains open.

Third-party assessors ranked all topics across all models and achieved distribution show superiority of ARTM over classic algorithms. Surprisingly, that flat ARTM based model show better results than multi-modal ARTM model which based on song text and refrain modality. However was evaluated the hierarchical approach of ARTM which show better

results in most recent studies, but due to complex strategy of regularization process is not suites well for thesis purpose.

The process of labeling topics to its meaning was tedious and from author opinion requires further investigation. The proposed idea can be in using word embedding models to mark-up a topics automatically by their top tokens.

After all evaluating phases was created a REST principles based web system which contains of tuned flat ARTM model to perform classification task on unseen data.

Additional validating of the model shows that corpus of data should be much bigger to handle all document specific topics, but overall quality of predictions was acceptable if

This was an interesting topic for study in connection with which formed a solid understanding of probability based Topic Modeling algorithms, math logic behind them and behavior on the data. To accomplish data preparation, models creation and further evaluating of models was written helper library on Python language contains more than 2500+ lines of code, which can be shown by request. This helper library in development phase and not ready for public release due to handles only flat models, but planned to add a support for hierarchical models too.

Further investigation should be done in direction of determination optimal number of topics more precisely, automatic topics labeling and development of optimal strategies of regularization for ARTM based approaches. Topic modeling algorithms based on additive regularization can be successfully combined with vectors of word embedding models as proposed in recent studies [26]. This open a new possibilities of text data exploration.

Bibliography

[1] M. Fell and C. Sporleder, "Lyrics-based analysis and classification of music," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, Aug. 2014, pp. 620–631. [Online]. Available: https://www.aclweb.org/anthology/C14-1059

[2] "The Largest Vocabulary in Music," Jul 2015, [Online; accessed 8. Dec. 2019]. [Online]. Available: https://lab.musixmatch.com/largest_vocabulary

[3] M. Antoniou, "Text analytics & topic modelling on music genres song lyrics," *Medium*, Jun 2018. [Online]. Available: https://towardsdatascience.com/text-analytics-topic-modelling-on-music-genres-song-lyrics-deb82c86caa2

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.

[5] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *arXiv preprint arXiv:1607.04606*, 2016.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[7] T. Hofmann, "Probabilistic latent semantic analysis," in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, ser. UAI'99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, pp. 289–296. [Online]. Available: http://dl.acm.org/citation.cfm?id=2073796.2073829

[8] K. Vorontsov and A. Potapenko, "Additive regularization of topic models," *Mach. Learn.*, vol. 101, no. 1, pp. 303–323, Oct 2015.

[9] J. S. Liu, "The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem," *Journal of the American Statistical Association*, vol. 89, no. 427, pp. 958–966, 1994. [Online]. Available: https://doi.org/10.1080/01621459.1994.10476829

[10] A. Daud, J. Li, L. Zhou, and F. Muhammad, "Probabilistic topic models survey," *Journal of Frontiers of Computer Science in China (FCS), Volume 4(2), pp. 280-301, June, 2010. SCIE*, vol. 4, 06 2010.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, vol. 39, no. 1, pp. 1–38, 1977.

[12] J. Bergstra, D. Yamins, and D. D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ser. ICML'13. JMLR.org, 2013, pp. I–115–I–123. [Online]. Available: http://dl.acm.org/citation.cfm?id=3042817.3042832

[13] K. Vorontsov, O. Frei, M. Apishev, P. Romov, M. Suvorova, and A. Yanina, "Non-bayesian additive regularization for multimodal topic modeling of large collections," in *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications*, ser. TM '15. New York, NY, USA: ACM, 2015, pp. 29–37. [Online]. Available: http://doi.acm.org/10.1145/2809936.2809943

[14] N. Chirkova and K. Vorontsov, "Additive regularization for hierarchical multimodal topic modeling," *Journal Machine Learning and Data Analysis*, vol. 2, no. 2, pp. 187–200, 2016.

[15] K. Vorontsov and A. Potapenko, "Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization," in *Analysis of Images, Social Networks and Texts*, D. I. Ignatov, M. Y. Khachay, A. Panchenko, N. Konstantinova, and R. E. Yavorsky, Eds. Cham: Springer International Publishing, 2014, pp. 29–46.

[16] K. Vorontsov, O. Frei, M. Apishev, P. Romov, and M. Dudarenko, "BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections," *SpringerLink*, pp. 370–381, Apr 2015.

[17] “Mystem,” Nov 2019, [Online; accessed 8. Dec. 2019]. [Online]. Available: https://nlpub.mipt.ru/Mystem

[18] A. Kutuzov and E. Kuzmenko, *WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models*. Cham: Springer International Publishing, 2017, pp. 155–161. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-52920-2_15

[19] S. O. A. Shavrina T., “To the methodology of corpus construction for machine learning: "taiga" syntax tree corpus and parser.” *Proceedings of the international conference "CORPUS LINGUISTICS - 2017"*, pp. 78–84, 2017.

[20] A. Mavrin, A. Filchenkov, and S. Koltsov, *Four Keys to Topic Interpretability in Topic Modeling: 7th International Conference, AINL 2018, St. Petersburg, Russia, October 17-19, 2018, Proceedings*, 01 2018, pp. 117–129.

[21] F. Krasnov and A. Sen, “The number of topics optimization: Clustering approach,” *Machine Learning and Knowledge Extraction*, vol. 1, pp. 416–426, 01 2019.

[22] M. Hoffman, F. R. Bach, and D. M. Blei, “Online learning for latent dirichlet allocation,” in *advances in neural information processing systems*, 2010, pp. 856–864.

[23] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, “Automatic evaluation of topic coherence,” in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, ser. HLT ’10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 100–108. [Online]. Available: http://dl.acm.org/citation.cfm?id=1857999.1858011

[24] C. Sievert and K. Shirley, “Ldavis: A method for visualizing and interpreting topics,” in *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, 2014, pp. 63–70.

[25] L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008. [Online]. Available: http://www.jmlr.org/papers/v9/vandermaaten08a.html

[26] A. Potapenko, A. Popov, and K. Vorontsov, “Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks,” *arXiv*, Nov 2017. [Online]. Available: https://arxiv.org/abs/1711.04154