Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6745–6752
Marseille, 11–16 May 2020

TopicNet: Making Additive Regularisation for Topic Modelling Accessible

**Victor Bulatov, Evgeny Egorov, Eugenia Veselova,
Darya Polyudova, Vasiliy Alekseev,
Alexey Goncharov, Konstantin Vorontsov**
Moscow Institute of Physics and Technology (National Research University)
9 Institutskiy per., Dolgoprudny, Moscow Region, 141701, Russian Federation
viktor.bulatov@phystech.edu, egorov.eo@mipt.ru, veselova.er@phystech.edu,
wasya.alekseev@gmail.com, darya.polyudova@phystech.edu,
alex.goncharov@phystech.edu, k.v.vorontsov@phystech.edu

Abstract

This paper introduces TopicNet, a new Python module for topic modelling. This package, distributed under the MIT license, focuses on bringing additive regularization for topic modelling (ARTM) to non-specialists using a general-purpose high-level language. The module features include powerful model visualization techniques, various training strategies, semi-automated model selection, support for user-defined goal metrics, and a modular approach to topic model training.

Keywords: topic modelling, Python, supervised learning, unsupervised learning, latent Dirichlet allocation, additive regularization for topic modelling, BigARTM

1. Introduction

Topic modelling is a method to extract hidden word probability distributions called topics from text corpora. Being primarily introduced to find latent topics in text documents, topic models have proven to be relevant in a wide range of contexts (Boyd-Graber et al., 2017).

Over the time topic models have primarily been used for two purposes: either providing embeddings over collections of text for various recommendation and search purposes or giving a cohesive clustering of data for purposes of data mining and extracting information about the collection structure.

The first problem requires the researcher to tune their models to perform well on labelled datasets, thus, reducing the task to classification problem using obtained from topic model document embeddings as feature vectors. This makes topic models compete with various machine learning and deep learning algorithms in the field of this classical task.

Topic modelling does not necessarily show the best results in terms of classification metrics, but it has an edge when it comes to interpretability. Each topic can be interpreted as a probability distribution over words in that topic, and each coordinate in a document embedding has a meaning as a probability for that document to contain a certain topic in it. This property of topic models makes them better suited for the fields of AI where clarity of prediction matters or where one needs to correct undesirable biases introduced by the data into the model.

It should be noted that topic modelling is still a valuable tool in natural language processing (NLP) researcher arsenal. Currently, there is an ongoing work to extend state-of-the-art deep learning models to incorporate topical importance of words for text summarization (Narayan et al., 2018; Wang et al., 2018; Lebanoff et al., 2019) and song lyrics understanding (Choi et al., 2015; Fell et al., 2019; Watanabe et al., 2018). One can see that topic modelling and its numerous variations continues to be a workhorse in many NLP projects.

The second purpose — collection structure retrieval, however, is somehow unique to the field of topic modelling. Given that the researcher not necessarily wants to optimise the classification metric, but merely searches for answers about the collection structure and nature. This approach, for example, has been adopted in the fields of biology (Liu et al., 2016; Funnell et al., 2019) and humanities (Boyd-Graber et al., 2017; Antons et al., 2019). It provides an invaluable insight into big data that otherwise might have been missed by the researcher.

Jordan Boyd-Graber, Yuening Hu and David Mimno in their monograph (Boyd-Graber et al., 2017) show, that designing new topic model is a difficult enterprise and identify accessibility as the most important unsolved problem in topic modelling: "the primary research challenge of topic models is... to make them more accessible".

The same year another publication (Lee et al., 2017) argued, that existing topic models do not provide non-expert users with direct means to alter topic features perceived by these users to be bad. The study suggested several improvements for existing topic modelling user experience, concentrating on topic visualisation for the users.

The article from another group (Agrawal et al., 2018) demonstrate that existing popular topic models are failing to provide good results when used "off-the-shelf" with their default parameters.

To facilitate the research on mining latent topics from texts and text-like collections, we present a novel software package called TopicNet[1]. This software package contributes to the flexibility of topic model design and provides powerful out-of-the-box experience boosting the accessibility of topic modelling for the general audience.

[1]Source code can be downloaded from `github.com/machine-intelligence-laboratory/TopicNet`
The documentation is available at `machine-intelligence-laboratory.github.io/TopicNet`

2. Related Work

Existing topic modelling algorithms originate from latent semantic analysis approach — LSA (Deerwester et al., 1990), which is, in the essence, a singular value decomposition with the lowest possible rank for document×term matrix. Later, a probabilistic approach took the main place in the topic modelling field.

A probabilistic topical model can be considered as a black box that receives a text document collection at the input and produces two families of distributions at the output: the term probabilities for each topics $\phi_{wt} = p(w \mid t)$ and the topic probabilities for each document $\theta_{td} = p(t \mid d)$. The term×topic matrix Φ and topic×document matrix Θ are model parameters to be found during the training.

The development of probabilistic topic modelling started with two fundamental models: probabilistic Latent Semantic Analysis — pLSA (Hofmann, 1999) and Latent Dirichlet Allocation — LDA (Blei et al., 2003). LDA adds Dirichlet prior to pLSA: the generative LDA model draws term-topic and topic-document distributions from prior Dirichlet distributions, each with its own concentration parameter. While writing a document, authors usually use only handful of topics per each text. This sparsity property is reflected in concentration parameter. Varying value of this parameter, one can obtain distributions where only a few topics have high probability to be in a document with other topic probabilities being small. Thus, the sparsity of Dirichlet distributions is the probabilistic tool that encodes this intuition. However, according to the article (Wallach et al., 2009), LDA requires extensive hyperparameter optimization toward producing good results.

Over the past years, hundreds of pLSA and LDA model extensions have emerged, each taking into account the various problem-specific data features and providing desired solution properties. Starting with LDA model, Bayesian learning is the de facto standard in topic modelling. In this approach, one first describes the probabilistic generative model of data, specifies prior distributions of the model parameters, and then uses Bayesian inference to obtain the parameter posterior distributions. Bayesian inference method requires unique derivation and, consequently, unique implementation for each new model.

Additive regularization for topic modelling — ARTM (Vorontsov, 2014; Kochedykov et al., 2017) is a non-Bayesian multi-objective approach. It is based on the maximization of the log-likelihood together with a weighted sum of regularization criteria. Many of the well-known Bayesian topic models can be reformulated as a regularization over the PLSA model. After reformulation, they usually become much easier to understand and implement. The ARTM approach allows combining topic models simply by adding the regularizers. This gave rise to the modular technology for topic modelling implemented in the BigARTM[2] open-source software (Vorontsov et al., 2015c).

Given a generative model and data, inference must be executed for extracting probabilistic topic-depending distributions. There are many inference algorithms: expectation-maximization (EM) algorithm, Gibbs sampling, variational inference, gradient descent and message passing. In ARTM the regularized EM-algorithm is used to learn model parameters. The similarities between each of the algorithms were noted before in (Asuncion et al., 2009).

Gensim (Řehůřek and Sojka, 2010) is the most popular NLP framework for Topic Modeling. It implements several popular models such as LSA, pLSA, LDA, Hierarchical Latent Dirichlet Allocation (HLDA) and their derivatives. This framework also implements a coherence metric to improve topics from previously mentioned models as done in the article (Röder et al., 2015). This framework is written in Python and optimized for huge document corpora.

Stanford Topic Modelling Toolbox — TMT (Ramage et al., 2009), based on Scala, has LDA, Labelled LDA and PLDA models available for training. Stanford TMT also includes user-friendly interaction with Excel, allowing one to load data from Excel cells and generate rich output for tracking word usage across topics, time and other groupings of data.

MALLET (McCallum, 2002) is an essential framework for any researcher in Topic Modelling field, which is implemented in Java. This toolkit contains efficient, sampling-based implementations of LDA, Pachinko Allocation (Li and McCallum, 2006), and HLDA models. The framework provides advanced LDA models and often used for online services (Pol et al., 2017).

A Library of Short Text Topic Modelling — STTM (Qiang et al., 2018) is the Java framework that extends the range of available open-source modelling methods and integrates the state-of-the-art models of short text topic modelling algorithms. Primarily it aims at distilling meaningful topics from short texts, thus many high-performance models such as Dirichlet Multinomial Mixture DMM (Yin and Wang, 2014), Word Network Topic Model WNTM (Zuo et al., 2016b), Pseudo-Document-Based Topic Model PTM (Zuo et al., 2016a) and Self-Aggregation-Based Topic Model SATM (Quan et al., 2015) present in this framework. Some long-text topic models such as LDA and Latent Feature Model with LDA (Nguyen et al., 2015) are provided by STTM as well.

Familia (Jiang et al., 2018) is a framework that implements various topic models, including, but not limited to, chief LDA and Supervised LDA: Topics Over Time TOT (Wang and McCallum, 2006), Bilingual Topic Model (Gao et al., 2011), Location-Aware Topic Model LATM (Wang et al., 2007) and some more, using Gibbs sampling as a mathematical engine. To the best of our knowledge, despite authors claim that Familia provides the ability to "design their own topic models", we found no such evidence in the repository of the project (Lian, 2019).

As identified before, one of the main challenges in the field of topic modelling is the accessibility of the models to the general public. Each previously mentioned framework succeeded in closing that gap at the time of their release, but due to the nature of the Bayesian approach, the most popular frameworks (Gensim, MALLET) provide outdated models.

The other challenge is building complex, tractable and multi-objective topic models from scratch. While the basic model can be implemented in a reasonable amount of time, improving it remains both a time-consuming and an

[2] bigartm.org

error-prone task (Jiang et al., 2018)

With our work, we aim not only to close the gap between novel models and popular models as our predecessors do but also provide a tool that will allow anyone to construct new types of topic models. Due to ARTM formalism, TopicNet offers natural language processing community access to Python-based multimodal topic modelling that supports large documents and huge corpora.

3. Underlying Technologies

Having carefully considered all mentioned approaches to topic modelling, we decided to build a framework that will allow a better user experience with the BigARTM library. Below, we discuss in greater detail pluses of the BigARTM library python API and its minuses that we wanted to deal with in the TopicNet.

3.1. BigARTM Strengths

BigARTM is a fast and flexible library for topic modeling (Frei and Apishev, 2016), based on Additive Regularization of Topic Models (ARTM) formalism (Vorontsov, 2014). The idea behind ARTM is to replace likelihood with regularized likelihood and optimize this functional using modified EM-algorithm. The regularization serves two purposes. First, it ensures robustness and limits solution area, reducing the instability of inference. Second, each regularization term is used to pursue different solution characteristics such as sparsity, diversity, coherence or to take extra information into account. As an example of dealing with auxiliary information, BigARTM makes it very easy to include document metadata (e.g. authors, timestamps, tags, and n-grams) in a single model. This is because the likelihood of each additional modality could be considered as a regularizer applied to the topic model over words.

Works that take advantage of ARTM's and BigARTM's flexibility include: exploratory search quality improvement (Yanina et al., 2018), learning interpretable topical word embeddings through word network topic model (Potapenko et al., 2017), hierarchical topic modelling (Chirkova and Vorontsov, 2016), multi-label text categorization (Vorontsov et al., 2015b), improving topics for document vector representation in text regression problems (Sokolov and Bogolubsky, 2015), finding rare ethnically relevant topics in social media (Apishev et al., 2016a; Apishev et al., 2016b), incorporating language features (Popov et al., 2019), topic selection through entropy regularization (Vorontsov et al., 2015a), improving topics through text segmentation (Skachkov and Vorontsov, 2018), directly improving topic coherence (Mavrin et al., 2018), surpassing the bag-of-words hypothesis by using a new intra-text coherence measure (Alekseev et al., 2018). The review (Kochedykov et al., 2017) shows how Bayesian topic models can be re-formulated in a much simpler way from the ARTM point of view, including multimodal, multilingual, temporal, hierarchical, graph-based, and short-text topic models.

The development of both ARTM theory and BigARTM software is still ongoing. Many of the existing widely used regularizers were contributed by community.

To the best of our knowledge, among other topic modelling frameworks, only Familia offers comparable flexibility. However, the ability to construct a custom topic model or even to train one on a given corpus is absent from the open-source release of Familia (Lian, 2019).

3.2. BigARTM Flaws

Given a precise specification of a regularized model, BigARTM can infer it's parameters in a very fast, scalable and efficient way. However, it is unclear where such specification comes from. While the number of topics is an unresolved question in many topic modelling frameworks, the issue is further complicated by the ability to combine many different regularizers, each having an unknown individualised regularization coefficient. BigARTM offers virtually no guidance regarding the selection of regularizers and their structural parameters.

An additional factor contributing to the high entry barrier is somewhat inconvenient and inconsistent API in BigARTM library. This is a natural result of implementing new functionality before "best practices" of its usage are established and inability to change API afterwards due to backwards compatibility concerns. Therefore, as applications of BigARTM were becoming more diverse and the algorithms were gradually refined, the high-level interface of BigARTM was getting less well-suited for "best practices". Another shortcoming of BigARTM is the difficulty in extending it. From a technical point of view, BigARTM library consists of a core written in C++ and several Python wrapper classes. Low-level C++ routines are multithreaded and highly optimized, giving BigARTM an edge in performance. At the same time, it makes any modification of low-level functionality challenging. Meanwhile, the high-level API does not always offer enough flexibility to, for example, experiment with new custom regularizers.

4. Project Vision

The main motivation of the TopicNet is to close the gap between non-experts and power users. It does not mean that both groups will use the library in the same way; rather, it means that both groups should be able to communicate with each other. We formulated the following requirements necessary to achieve these aims:

- Modularity: it should be possible to use just a small part of TopicNet functionality as "plugin" inside an independently existing project. The reason behind this was twofold. The first point was that it should make the adoption easier: power users are not forced to dramatically change their existing projects to start getting benefit from TopicNet. The second point is related to the building an open-source community around TopicNet: modular open source projects are more welcoming to the contributors.
- Visualisation tools: the library should provide ready-to-use powerful visualisation tools. Such tools play an important role in error analysis and might be helpful for downstream tasks (e.g. exploratory search). Following our previous requirement, this module should be as stand-alone as possible; ideally, it would allow

the community to incorporate best practices from literature into TopicNet.

- Concision: the library should remove the low-level details, freeing more time for substantial problems. The reason behind this was threefold. The first point was the need to improve the artistic process of constructing a target function for the task at hand. The second point was the adherence to the "convention over configuration" philosophy: by reducing the amount of explicitly declared things and providing sensible defaults, we can enforce "best practices" such as automatically saving trained models or storing data batches separately for different datasets. The final point is readability: when working with concise and uniform code, the user can see more code lines on the screen which aids better reading and understanding the experiment. As a result, it is much simpler to share, review and debug finished experiments.

- Work out of the box: the library should include some pre-coded training pipelines which are ready-to-use and produce good results. Moreover, these pipelines should absorb the best-known approaches for as many modelling tasks as possible. The positive "out of the box" experience is vital for involving novice users, whereas power users may deliver their experience through pre-coded pipelines.

5. Code Design

The TopicNet consists of two large modules: `viewers` and `cooking machine`.

The purpose of `Viewers` module is to provide powerful visualisation tools. The design adheres to the Unix philosophy: each viewer has a limited area of responsibility and returns the result as a JSON-convertible object by default. Consequently, the module has a high degree of composability and modularity while still having convenience methods returning `pandas.DataFrame` or rendered HTML. Examples of the viewers output, like the TopTokensViewer and the DocumentClusterViewer, can be seen in Figures 1, 2 respectively.

The `Cooking Machine` module contains all modelling toolbox, embodied in the semi-hierarchical structure of the main modelling classes. These classes are responsible for building and training model of the given structure, for selecting models according to various constraints and for saving, loading and logging through the modelling process.

Following our "convention over configuration" principle, we enforce some assumptions on which kind of experiments TopicNet supports. We hold that model training pipeline could be represented as a tree. Each node is a topic model and directed edges represent parent-children relationship, such as "model Y was obtained from model X with the transformation T_{XY}". We restrict allowed transformations by linking them to their depth in the experiment tree: we require that each edge of the same level describe the same transformation aside from a set of individualised parameters.

A non-exhaustive list of such transformations:

- Applying a regularizer with arbitrary regularization coefficient or changing parameters of the existing regularizer

- Training a model for several iterations

- Adding topics inferred on a different corpus to the model

The `Experiment` class is responsible for storing, logging and maintaining this structure.

The transformations are connected to the instances of `Cube` class. Each `Cube` acts as a blueprint for all model transformations performed on the current stage of the experiment. In a way, a training pipeline is several cubes stacked together sequentially.

Cube is responsible for two essential functions. The first use is *specification*: during the initialization stage, cube converts user-defined parameters into a multidimensional search space. The second use is *alteration*: given a point in search space and a topic model, cube alters one or many model hyperparameters. That way it acts as an incubator for models, which is reflected in the class name. Scheme of the training process with two cubes, applied to the model, can be seen in Figure 3.

Taken together, `Experiment` and `Cube` classes make logging and complex training pipelines more concise and accessible. To capitalise on this decision, we implemented a `config_parser` module which allows specifying complex training pipelines as a plaintext config file in YAML format.

Another key area described as very verbose and confusing is model selection. In real experiments, not every model has descendants; most models are rejected based on some criteria. Aside from expensive manual inspection of top words and top documents, other conventional criteria include perplexity and coherence. BigARTM library adds several other metrics such as sparsity, purity and contrast (Vorontsov and Potapenko, 2015).

To reduce the burden of manual inspection, we implemented a simple domain-specific language for model selection (see Figure 4 for example). This language simplifies the selection task by conveniently leveraging various metrics.

6. Comparison with Related Work/Benchmarks

To compare the performance of the TopicNet with other frameworks, we chose a few essential areas. First of all, we wanted to check that we do not loose much time on training a single model or spend too many resources on it compared to the other frameworks. The other test would be measuring topic interpretability and diversity for each model. To measure interpretability we are using Umass coherence provided by Palmetto web service (Röder et al., 2015) as it is shown to correlate with the desired property of the topics (Mimno et al., 2011). The Jaccard metric is used to estimate topic diversity across the model topics.

topic_10		topic_10		topic_11		topic_11		topic_12		topic_12	
token @bigram		token @lemmatized		token @bigram		token @lemmatized		token @bigram		token @lemmatized	
male_female	0.0693	population	0.0086	jesus_christ	0.0202	church	0.0111	hall_fame	0.0318	game	0.0214
north_carolina	0.0225	female	0.0063	eastern_orthodox	0.0194	language	0.0094	black_hole	0.0289	team	0.0141
life_expectancy	0.0191	male	0.0061	holy_spirit	0.0178	christian	0.0058	league_baseball	0.0249	player	0.0139
hiv_aid	0.0122	specie	0.0055	bipolar_disorder	0.0097	word	0.0052	super_bowl	0.0217	season	0.0104
mortality_rate	0.012	year	0.0043	singular_plural	0.0085	god	0.0041	grand_prix	0.0211	play	0.0085
demographic_statistic	0.0106	human	0.0038	anti_semitism	0.0078	english	0.0037	red_sox	0.0163	win	0.0083
golden_ratio	0.0106	rate	0.0032	martin_luther	0.0071	century	0.0036	white_sox	0.0157	league	0.0081
infant_mortality	0.0101	est	0.0031	ecumenical_council	0.0061	jesus	0.0033	internet_explorer	0.0125	football	0.0047
religious_affiliation	0.0092	birth	0.0031	apostolic_succession	0.0059	catholic	0.003	san_francisco	0.0111	first	0.0044
fertility_rate	0.0084	animal	0.003	nicene_creed	0.005	use	0.003	abu_bakr	0.0109	ball	0.0043

Figure 1: Output of the TopTokensViewer. Token score in the topic is calculated for every token, score function can be specified at the stage of a viewer initialization.

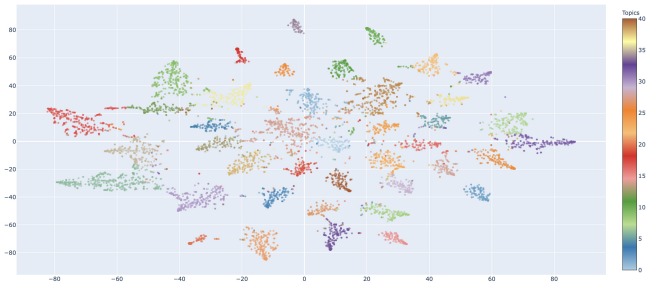

Figure 2: Visualisation of reduced document embeddings colored according to their topic made by DocumentClusterViewer.

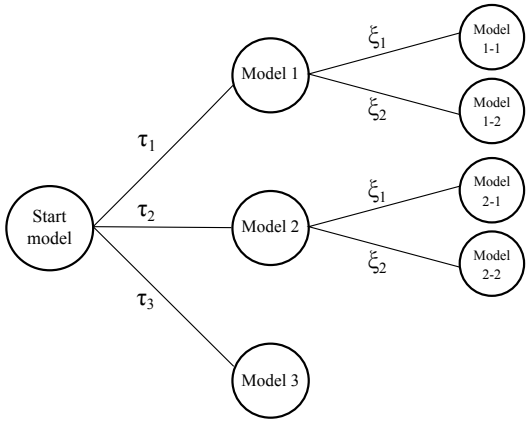

Figure 3: Example of the two-stage experiment scheme. At the first stage, regularizer with parameter τ taking values in some range $\{\tau_1, \tau_2, \tau_3\}$ is applied. Best models after the first stage are *Model 1* and *Model 2* — so *Model 3* is not taking part in the training process anymore. The second stage is connected with another regularizer with parameter ξ taking values in range $\{\xi_1, \xi_2\}$. As a result of this stage, two descendant models of *Model 1* and two descendant models of *Model 2* are obtained.

6.1. Resources Usage

Important library feature, from the scope of accessibility, is the usage of computer resources during the model training. If we want to engage a broader community such as students or aim at production applications of our models the framework should be as fast and lightweight as possible. We chose the NIPS dataset to run a performance test (McCallum, 1996). From the table 1 one can see that

	RAM, MB	Training time, s
TopicNet	991	222.6 (15)
TopicNet multiprocess	1084	51.4 (15)
Gensim LDA	3559	282.3 (3)
STTM DMM	3202	52997 (1)
STTM PTM	1604	84677 (1)
STTM WNTM	18663	157819 (1)

Table 1: In the first column we consider all the memory taken by the process during the training. The second column represents time needed to complete the training and number of models trained during the session.

even single processor version beats other frameworks in terms of performance. Meaning that TopicNet overhead over BigARTM did not decrease its performance too drastically.

6.2. Model Quality

Following our library developing aims, we look for a good out of the box experience for a new user. To simulate such conditions we train TopicNet model from `ARTM baseline` recipe and compare it with models from other frameworks. For each framework in our comparative experiment, one model is trained with default or built-in parameters. LDA and all STTM models had the number of topics equal 20, for TopicNet model we set 19 topics and one "background" topic, which has a special set of regularizers to collect polythematic documents. TopicNet model was constructed from the baseline training recipe without any parameter tuning. As an example of how to use TopicNet we provide the code needed to train baseline topic model discussed above 5. As a demonstration dataset, we chose a well known 20 newsgroups dataset. To make a fair model comparison we used Umass coherence score provided on Palmetto Demo website [3]. To assess the diversity of the topics provided by each model we used the Jaccard similarity coefficient. Both measures were calculated on the top ten tokens for each topic and every model.

[3] `palmetto.aksw.org/palmetto-webapp`

```
TopicKernel@word.average_contrast > 0.95 * MAXIMUM(TopicKernel@word.average_contrast)
   and PerplexityScore@all < 1.1 * MINIMUM(PerplexityScore@all)
   and SparsityPhiScore@word -> max
   COLLECT 3
```

Figure 4: This expression returns three models which are in the top 5% according to contrast, has acceptable perplexity and as sparse as possible. `SparsityPhiScore` stands for the fraction of zeros in $\phi_{wt} = p(w \mid t)$ distribution.

```python
from topicnet.cooking_machine.recipes import (
    ARTM_baseline as config_string
)

dataset_path = '/data/datasets/NIPS/dataset.csv'

specific_topics   = [f'spc_topic_{i}' for i in range(19)]
background_topics = [f'bcg_topic_{i}' for i in range( 1)]

config_string = config_string.format(
    dataset_path=dataset_path,
    modality_list=['@word'],
    main_modality='@word',
    specific_topics=specific_topics,
    background_topics=background_topics
)
experiment, dataset = (
    build_experiment_environment_from_yaml_config(
        yaml_string=config_string,
        experiment_id='sample_config',
        save_path='sample_save_path'
    )
)
experiment.run(dataset)
```

Figure 5: Example of the TopicNet baseline experiment.

	Jaccard measure of topic dissimilarity	Average topic coherence
TopicNet	**0.00169**	-2.551
Gensim LDA	0.01374	-2.747
STTM DMM	0.37541	-2.726
STTM PTM	0.02485	**-2.510**
STTM WNTM	0.01997	-3.572

Table 2: Topic quality comparison

As one can see from the above table 2, TopicNet model was a second-best model in terms of interpretability and the best model in terms of topic diversity. Which can be accounted to a Decorrelation regularization implemented in the `ARTM baseline` recipe. Together with a custom score provided by the TopicNet library, Decorrelation regularizer allows to strike a balance between perplexity minimization and topic interpretability. Further information on a topic coherence can be seen in Figure 6.

7. Conclusion

In this paper, we propose a configurable and fast framework for topic modelling and demonstrate its advantages over its competitors. TopicNet provides extensive functionalities such as building models from scratch, rich model customization and a possibility to fine-tune any previously constructed models.

The library provides modelling recipes capturing best-known practices of building an ARTM model for a certain task. As demonstrated in the previous chapter, an ARTM model is capable of beating most popular models in terms of providing cohesive and diverse topics. With TopicNet, engineers can customize it further by applying custom regularizers and finding best hyperparameter values for multi-criteria training scenario. Taken together, this allows the user to build their own type of topic model.

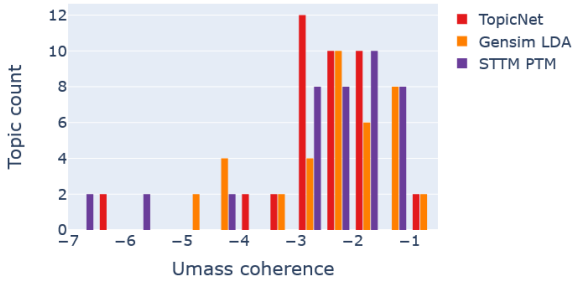

Figure 6: Topic coherence distribution.

Aside from tools for flexible and quick model development, we enable users to control their model quality. TopicNet provides model evaluation through a variety of built-in scores and the ability to add custom scores, which are used to track model characteristics during training. The library supports a variety of visualisation tools both conventional, such as top tokens and top documents viewers, and experimental ones.

Moreover, our framework is delivered as an open-source project, with potential for further extension. The design philosophy behind `viewers` and `cooking machine` modules allow open source community to incorporate new developments and discoveries into TopicNet library.

We hope that our framework will be equally valuable to software engineers and digital humanities researchers.

8. Acknowledgments

We are thankful to our colleagues Rosa Aisina, Murat Apishev, Ilya Zharikov, Filipp Nikitin, Artem Popov, Vera Shishkina, and Oleksandr Frei who provided the expertise that greatly assisted the project, although they may not agree with all of the project implementations presented in this paper.

9. References

Agrawal, A., Fu, W., and Menzies, T. (2018). What is wrong with topic modeling? and how to fix it using search-based software engineering. *Information and Software Technology*, 98:74–88.

Alekseev, V., Bulatov, V., and Vorontsov, K. (2018). Intratext coherence as a measure of topic models' interpretability. In *Computational Linguistics and Intellec-*

tual Technologies: Papers from the Annual International Conference Dialogue, pages 1–13.

Antons, D., Joshi, A. M., and Salge, T. O. (2019). Content, contribution, and knowledge consumption: Uncovering hidden topic structure and rhetorical signals in scientific texts. *Journal of Management*, 45(7):3035–3076.

Apishev, M., Koltcov, S., Koltsova, O., Nikolenko, S., and Vorontsov, K. (2016a). Additive regularization for topic modeling in sociological studies of user-generated text content. In *MICAI 2016, 15th Mexican International Conference on Artificial Intelligence*, volume 10061, pages 166–181. Springer, Lecture Notes in Artificial Intelligence.

Apishev, M., Koltcov, S., Koltsova, O., Nikolenko, S., and Vorontsov, K. (2016b). Mining ethnic content online with additively regularized topic models. *Computacion y Sistemas*, 20(3):387–403.

Asuncion, A., Welling, M., Smyth, P., and Teh, Y. W. (2009). On smoothing and inference for topic models. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*, pages 27–34.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Boyd-Graber, J., Hu, Y., Mimno, D., et al. (2017). Applications of topic models. *Foundations and Trends® in Information Retrieval*, 11(2-3):143–296.

Chirkova, N. A. and Vorontsov, K. V. (2016). Additive regularization for hierarchical multimodal topic modeling. *Journal Machine Learning and Data Analysis*, 2(2):187–200.

Choi, K., Lee, J. H., Willis, C., and Downie, J. S. (2015). Topic modeling users' interpretations of songs to inform subject access in music digital libraries. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 183–186. ACM.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Fell, M., Cabrio, E., Gandon, F., and Giboin, A. (2019). Song lyrics summarization inspired by audio thumbnailing. In *RANLP*, pages 328–337.

Frei, O. and Apishev, M. (2016). Parallel non-blocking deterministic algorithm for online topic modeling. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 132–144. Springer.

Funnell, T., Zhang, A. W., Grewal, D., McKinney, S., Bashashati, A., Wang, Y. K., and Shah, S. P. (2019). Integrated structural variation and point mutation signatures in cancer genomes using correlated topic models. *PLoS computational biology*, 15(2):e1006799.

Gao, J., Toutanova, K., and Yih, W.-t. (2011). Clickthrough-based latent semantic models for web search. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 675–684. ACM.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc.

Jiang, D., Song, Y., Lian, R., Bao, S., Peng, J., He, H., and Wu, H. (2018). Familia: A configurable topic modeling framework for industrial text engineering. *arXiv preprint arXiv:1808.03733*.

Kochedykov, D., Apishev, M., Golitsyn, L., and Vorontsov, K. (2017). Fast and modular regularized topic modelling. In *2017 21st Conference of Open Innovations Association (FRUCT)*, pages 182–193. IEEE.

Lebanoff, L., Song, K., Dernoncourt, F., Kim, D. S., Kim, S., Chang, W., and Liu, F. (2019). Scoring sentence singletons and pairs for abstractive summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2175–2189, Florence, Italy, July. Association for Computational Linguistics.

Lee, T. Y., Smith, A., Seppi, K., Elmqvist, N., Boyd-Graber, J., and Findlater, L. (2017). The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies*, 105:28–42.

Li, W. and McCallum, A. (2006). Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, pages 577–584. ACM.

Lian, R. (2019). Project title. `https://github.com/baidu/Familia/issues/81`.

Liu, L., Tang, L., Dong, W., Yao, S., and Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5(1):1608, Sep.

Mavrin, A., Filchenkov, A., and Koltcov, S. (2018). Four keys to topic interpretability in topic modeling. In *Conference on Artificial Intelligence and Natural Language*, pages 117–129. Springer.

McCallum, A. K. (1996). Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. *http://www. cs. cmu. edu/mccallum/bow/*.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. *http://mallet. cs. umass. edu*.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 262–272, Stroudsburg, PA, USA. Association for Computational Linguistics.

Narayan, S., Cohen, S. B., and Lapata, M. (2018). Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807.

Nguyen, D. Q., Billingsley, R., Du, L., and Johnson, M. (2015). Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, 3:299–313.

Pol, M., Walkowiak, T., and Piasecki, M. (2017). Towards clarin-pl ltc digital research platform for: Depositing, processing, analyzing and visualizing language data.

In *International Conference on Reliability and Statistics in Transportation and Communication*, pages 485–494. Springer.

Popov, A., Bulatov, V., Polyudova, D., and Veselova, E. (2019). Unsupervised dialogue intent detection via hierarchical topic model. In *RANLP*, pages 932–938.

Potapenko, A., Popov, A., and Vorontsov, K. (2017). Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. In *Communications in Computer and Information Science, vol 789. AINL-6: Artificial Intelligence and Natural Language Conference, St. Petersburg, Russia, September 20-23, 2017*, pages 167–180. Springer, Cham.

Qiang, J., Li, Y., Yuan, Y., Liu, W., and Wu, X. (2018). Sttm: A tool for short text topic modeling. *arXiv preprint arXiv:1808.02215*.

Quan, X., Kit, C., Ge, Y., and Pan, S. J. (2015). Short and sparse text topic modeling via self-aggregation. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Ramage, D., Rosen, E., Chuang, J., Manning, C. D., and McFarland, D. A. (2009). Topic modeling for the social sciences. In *Neural Information Processing Systems (NIPS) Workshop on Applications for Topic Models: Text and Beyond*, Whistler, Canada, December.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. `http://is.muni.cz/publication/884893/en`.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408. ACM.

Skachkov, N. and Vorontsov, K. (2018). Improving topic models with segmental structure of texts. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference Dialogue*, pages 652–661.

Sokolov, E. and Bogolubsky, L. (2015). Topic models regularization and initialization for regression problems. In *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications*, pages 21–27, New York, NY, USA. ACM.

Vorontsov, K. V. and Potapenko, A. A. (2015). Additive regularization of topic models. *Machine Learning, Special Issue on Data Analysis and Intelligent Optimization with Applications*, 101(1):303–323.

Vorontsov, K. V., Potapenko, A. A., and Plavin, A. V. (2015a). Additive regularization of topic models for topic selection and sparse factorization. In A. Gammerman et al., editor, *The Third International Symposium On Learning And Data Sciences (SLDS 2015). April 20-22, 2015. Royal Holloway, University of London, UK.*, pages 193–202. Springer International Publishing Switzerland 2015.

Vorontsov, K., Frei, O., Apishev, M., Romov, P., Suvorova, M., and Yanina, A. (2015b). Non-bayesian additive regularization for multimodal topic modeling of large collections. In *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications*, pages 29–37, New York, NY, USA. ACM.

Vorontsov, K., Frei, O., Apishev, M., Romov, P., and Suvorova, M. (2015c). BigARTM: Open source library for regularized multimodal topic modeling of large collections. In *AIST'2015, Analysis of Images, Social networks and Texts*, pages 370–384. Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS).

Vorontsov, K. V. (2014). Additive regularization for topic models of text collections. *Doklady Mathematics*, 89(3):301–304.

Wallach, H. M., Mimno, D. M., and McCallum, A. (2009). Rethinking lda: Why priors matter. In *Advances in neural information processing systems*, pages 1973–1981.

Wang, X. and McCallum, A. (2006). Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433. ACM.

Wang, C., Wang, J., Xie, X., and Ma, W.-Y. (2007). Mining geographic knowledge using location aware topic model. In *Proceedings of the 4th ACM workshop on Geographical information retrieval*, pages 65–70. ACM.

Wang, L., Yao, J., Tao, Y., Zhong, L., Liu, W., and Du, Q. (2018). A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization. In *International Joint Conference on Artificial Intelligence*.

Watanabe, K., Matsubayashi, Y., Inui, K., Fukayama, S., Nakano, T., and Goto, M. (2018). Modeling storylines in lyrics. *IEICE TRANSACTIONS on Information and Systems*, 101(4):1167–1179.

Yanina, A., Golitsyn, L., and Vorontsov, K. (2018). Multi-objective topic modeling for exploratory search in tech news. In Andrey Filchenkov, et al., editors, *Communications in Computer and Information Science, vol 789. AINL-6: Artificial Intelligence and Natural Language Conference, St. Petersburg, Russia, September 20-23, 2017*, pages 181–193. Springer International Publishing, Cham.

Yin, J. and Wang, J. (2014). A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 233–242. ACM.

Zuo, Y., Wu, J., Zhang, H., Lin, H., Wang, F., Xu, K., and Xiong, H. (2016a). Topic modeling of short texts: A pseudo-document view. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2105–2114.

Zuo, Y., Zhao, J., and Xu, K. (2016b). Word network topic model: a simple but general solution for short and imbalanced texts. *Knowledge and Information Systems*, 48(2):379–398.