Topic Modeling for Tatar language

Kamil Akhmetov
B17-DS-01
Data Science
Innopolis University
k.ahmetov@innopolis.ru

Ruslan Sabirov
B17-DS-02
Data Science
Innopolis University
r.sabirov@innopolis.ru

Abstract

The Tatar language is far from the most widespread in the world. As a result, natural language processing techniques and methods for it are not fully developed and implemented. However, there is a serious need in their applications in the Republic of Tatarstan and other places of Tatars' residence. Topic modelling of texts in Tatar language is one of the problems that arise. In our work we have designed and implemented the pipeline for Topic Modelling which runs on data, we have collected ourselves. We have composed a dataset of topic-tagged Tatar texts from Wikipedia. The model we have developed has shown strong discriminatory power (9285.568 perplexity and 0.607 Silhouette score).

1 Introduction

Minority languages usually are payed less attention, than others. This fact leads to the lack of explored techniques available for natural language processing. There is also lack of structured data ready for NLP. Tatar, being one of those languages, suffers the same problems. In our work we have developed a complete pipeline for Topic modelling of the Tatar texts from scratch. Starting from data collection and finishing by clusterization model. The rest of the paper is organized in the following way: Literature review, Data Collection, Preprocessing, Model description, Model selection, Model visualization, Conslusion.

2 Literature Review

Latent Dirichlet Allocation — one of the unsupervised topic modeling algorithms, that allows to describe a set of documents by unobserved groups (topics).

In the original paper (Pritchard et al., 2000) LDA was proposed in the context of population genetics. The authors have applied the algorithm to detect the presence of structured genetic variation in a group of individuals.

Its application in machine learning and topic modeling was found by later. The paper (Blei et al., 2003) proposed to use LDA to automatically classify documents and estimate their relevance to the set of topics.

3 Data Collection

There is lack of open-source quality data available for Tatar language. For the purpose of our work we have chosen to use articles from Tatar version of Wikipedia.

3.1 Topics Definition

First, we have parsed all main categories, resulting in a global list of about $25,000$ 'topics' sorted by declared amount of articles for each of them. Unfortunately, these raw categories are not ready to be used as topics, because of several reasons: internal categories (such as 'Wikipedia Common'), high imbalance of texts distribution, etc. Thus, we have chosen several topics ourselves and found all categories that contain a special keyword, defining that category. As a consequence we had a list of article titles and corresponding topics. These labels are taken to be the ground truth for further analysis.

3.2 Content Extraction

For each of the article titles we have extracted their contents (text) and finally constructed a set of data points with Tatar texts and topic labels, ready for further processing.

4 Pre-processing

There were several pitfalls regarding the data-set, thus in order to prepare the data for the further stage of our pipeline we have applied several pre-processing steps described below.

4.1 Normalization

The goal of this part is to transform the original text into the sequence of tokens. Normalization includes: lowering, removing URLs, stress and punctuation marks.

4.2 Transliteration

Tatar language has rich history of using different alphabet systems for writing. Before 1920s Tatar writing was based on Arabic writing system. Soviet Union government intended to transfer all Turkic languages to a unified newly developed system based on Latin alphabet - "Yana Alifba" (from tat. - New Alphabet.) It was used to be the official writing system for Tatar language for more than 10 years, and was substituted by Cyrillic version, which is currently in use. Thus we have had to be capable of working with 2 different writing systems.

Transliteration is one of the solutions of this issue. Fortunately, there exists almost-1-to-1 mapping between 2 systems. Rule based substitution helped as to transfer all Latin based texts to Cyrillic versions.

4.3 Stop-words removal

Some of the words are used in a large amount of different topics, so they suffer the lack of discriminatory power. Removing them, we do not lose any information concerning the particular topic of the text, but make the data more feasible. For this purpose we have used several sets of stop-words:

- Common stop-words for Tatar, Russian and English languages. They were found in the Internet.
- Domain specific stop-words (Wikipedia's keywords, years, toponyms, etc). They were collected manually by considering most popular words in dataset and words that were keywords for predicted topics.

4.4 Stemming

Tatar language is agglutinate. Any word consist of the root and different affixes. Each of the affixes adds exactly one meaning aspect to the word. This fact allows for design of accurate algorithm to remove those affixes and obtain a 'clean' form of the word. Thus, we have developed a rule-based approach to process all words as if they were nouns, adjectives, verbs and adverbs. We have also tried to remove last consonant vocalization, which happens when an affix starting with vowel is added. Now, the best results are achieved without using the last option.

4.5 Documents removal

During the processing of the documents, we have noticed some articles being just a template for a future article to be written. Such documents were useless for us and were successfully removed from the database.

5 Model description

The model uses LDA and Agglomerative clusterization components. LDA component takes a matrix of words occurrences in documents of size $|D| \times |W|$ as an input and produces the output matrix of size $|D| \times |T|$ – topics probability distribution for each document ($|D|$ – number of documents, $|W|$ – number of unique words, $|T|$ – number of predefined topics).

Further we name output of LDA for a particular document as LDA-embedding of that document.

6 Model selection

One of the most important hyper-parameters is the number of topics.

Even though it is known from the data-set, we have decided to put an experiment and see if changing it would give us better results, as soon as a topic may be divided into several subtopics or, oppositely, be a combining part of some bigger topic. It was stated previously that we have taken several Wikipedia categories that match some predefined keyword to contribute in a single topic.

Grounds for choosing the number of topics is reasoned by the scores of different clustering metrics, including supervised (Fowlkes-Mallows, Normalized Mutual Info and Adjusted Rand) and unsupervised (Silhouette, Calinski-Harabasz and Davies Boulduin) metric scores.

For each number of topics a metric is computed using either the most probable topic predictions (LDA labels) or Agglomerative Clustering predictions (clustering runs on LDA-embeddings as input).

The second hyper-parameter is the clustering linkage method (ward, complete, average or none, if we use LDA labels).

Plots of Silhouette and Fowlkes-Mallows metric scores are presented in the Figures 1 and 2 below. Applying the elbow method we have heuristically determined the optimal number of topis to be 13.

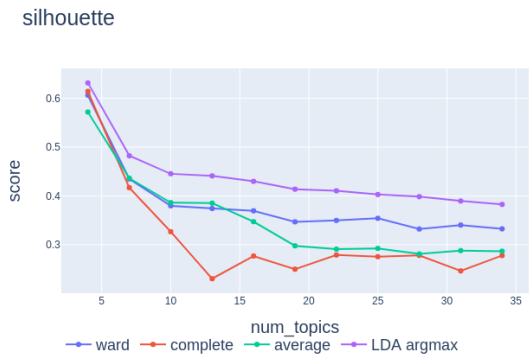

Figure 1: Silhouette score

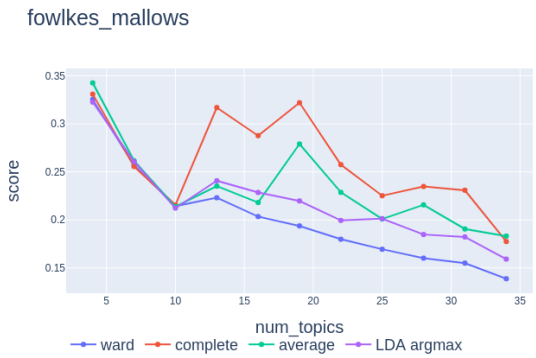

Figure 2: Fowlkes-Mallows score

There are several more hyper-parameters including distance metrics (cosine, euclidean, manhattan), vectorization model (count matrix, TF-IDF matrix) and labels source (LDA labels or Agglomerative clustering labels). To fine tune them, the data is divided into two parts: train and test split. LDA is trained on train split with number of topics equal to 13 and metrics are computed on the test set. Results are shown in the table below.

	Count Matrix		**TF-IDF matrix**	
	LDA embeddings argmax	Agglomerative Clust on LDA embeddings	LDA embeddings argmax	Agglomerative Clust on LDA embeddings
cosine	0.607	0.562	0.475	0.431
euclidean	0.387	0.354	0.307	0.271
manhattan	0.346	0.326	0.253	0.258

Table 1: Silhouette Score on the test dataset depending on used matrix and distance metric

We have also computed perplexity for both vectorization models: perplexity for **Count matrix** is 9285.568 and **TF-IDF** – 861043.810.

Finally we have chosen to use LDA without Agglomerative clustering and have chose the following hyper-parameters as the optimal ones.

Number of topics	Distance Metric	Vectorization Model	Clustering
13	Cosine	Count Matrix	None (LDA labels)

Table 2: Final hyper-parameters configuration

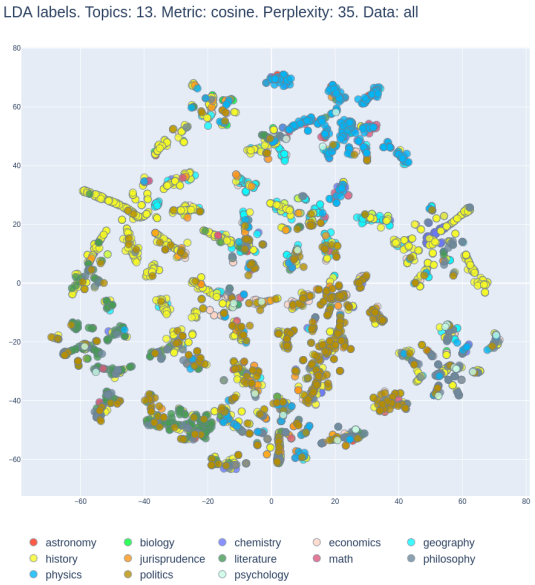

Figure 3: All data labeled

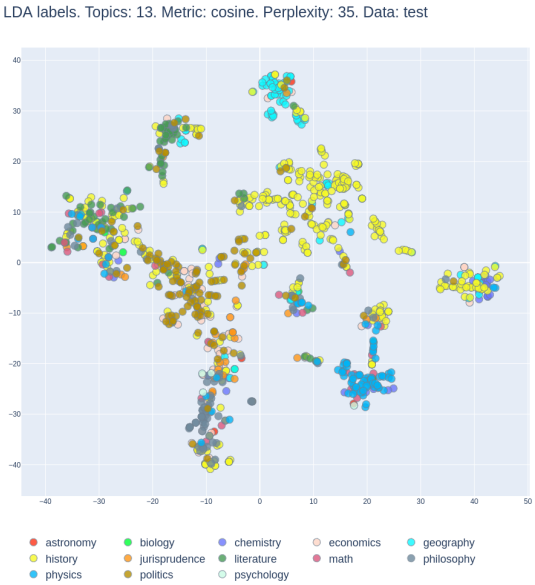

Figure 4: Test data labeled

7 Model visualization

We visualize the model by plotting document LDA-embeggins (t-SNE dimensionality reduction technique is used) and their labels (ground truth from the dataset). See figures 3 and 4.

Our observations:

- History and Politics topics are spread across the entire space. One of the reasons is time and location difference of data points (e.g., lexicons of World War II and Medieval plague would differ significantly).
- Literature is also spread across space. We suppose that it is because of the different genres and written dates.
- Physics and Chemistry are highly mixed, because they are highly interconnected and often contain similar subjects of interest.
- Clusters of Geography, Math, and Philosophy are easily distinguishable, due to very specific languages each of these have.

8 Conclusion

To conclude, the model has shown strong discriminatory power for texts in Tatar language. We have achieved 9285.568 perplexity and 0.607 Silhouette score. Original paper on LDA shows around 2 times better results for English. This shows us that Tatar language highly suffers lack of research in the field of NLP. By performing this project we have:

- Achieved desirable results, concluded that topic modeling can be applied even for small data-sets and observed interesting properties of the topics;
- Contributed to NLP of Tatar language: developed our own transliteration and lemming techniques and implemented the full pipeline for topic modeling;
- Practiced in NLP techniques (normalization, transliteration, lemmatization, LDA) and writing reports in LaTeX.

References

Tatar Wikipedia. Wiki Communtity. 2003 – present day. In *https://tt.wikipedia.org/wiki/*.

Stopwords Russian (RU). GitHub: @stopwords-iso. 2016. In *https://github.com/stopwords-iso/stopwords-ru/*.

Stopwords English (EN). GitHub: @stopwords-iso. 2016. In *https://github.com/stopwords-iso/stopwords-en/*.

Stopwords Tatar (TT). GitHub: @aliiae. 2018. In *https://github.com/aliiae/stopwords-tt*.

Pritchard J. K.; Stephens M.; Donnelly, P. 2000. Genetics. *Inference of population structure using multilocus genotype data*, 155(2):945–959.

Blei, David M.; Ng, Andrew Y.; Jordan, Michael I. 2003. "Latent Dirichlet Allocation". *Journal of Machine Learning Research*, 4(4-5):993–1022.