

TopicMiner with Sentiment analysis



Manual

Version 96 (64 bits).

Saint-Petersburg

2018

Table of contents

Chapter 1. Preprocessing of documents.	4
1.1 Procedure for assembling and lemmatizing documents.	4
1.2. The second stage of preprocessing.	8
1.2.1. Creation of a list of stop words.	9
1.3. The third stage of preprocessing.	10
Chapter 2. View tmla format files.	10
Chapter 3. Topic modeling based on the Gibbs sampling model.	15
3.1. Interface of option 'Gibbs LDA sampling'.	15
3.2. Loading documents for topic modeling.	16
3.3. Topic modeling based on Gibbs sampling.	17
3.4. Visualization of the results of topic modeling.	19
3.4.1. Visualization of document distribution by words.	20
3.4.2. Visualization of word distributions by topics.	22
3.4.3. Visualization of distributions of sorted documents by topics.	23
3.4.2. Visualization of sorted word distributions by topics.	24
3.5. Saving the results of thematic modeling in the form of a project file.	26
3.6. Loading the results of topic modeling from the project file.	27
Chapter 4. Topic modeling by BigArtm models (multimodal topic modeling).	27
4.1. Parameter setting in multimodal TM models.	27
4.2. Visualization of the results of topic modeling.	28
4.3. Saving the results of topic modeling in the form of a project file.	29
4.4. Calculation of multimodal variant of TM.	29
Глава 5. Stability analysis of simulation results.	31
5.1. Download of topic solutions.	31
5.2. Comparison of topic solutions.	33
5.2.1. Matrix 'Kullback - Leibler distance'.	34
5.2.2. Matching topics from different solutions.	34
Глава 6. Visualization of the results of topic modeling on the map of the Russian Federation.	36
6.1. Calculation of the distribution of documents by regions.	36
6.2. Visualization of document distribution in Quantum GIS.	38
Chapter 7. Analysis of the tonality of texts.	43
7.1. Introduction.	43
7.2. Preparing the vocabulary for sentiment analysis.	43
7.2. Connecting the dictionary to the topic model.	44
7.3. Tonal calculation of the distribution of words by topics.	44
7.3.1. Unload matrix of words - topics with tonal estimates.	45

7.3.2. Prompt of topics.	46
7.4. Tonal calculation of the distribution of documents by topics.	46
7.4.1. Unloading matrix documents - topics with tonal estimates.....	47
7.5. Tonal calculation of the distribution of documents by topics for BigArtms.	48
Глава 8. Time trends in topic models.	48
8.1. Unification of time dates.	48
8.2. Construction of time trends in models based on multimodal thematic modeling.	50
Constructing a trend based on the distribution of documents by topic.	50
8.3. The construction of time trends in models based on Gibbs sampling.....	54
Conclusion.....	54

Introduction.

The TopicMiner program was developed in the Internet Research Laboratory (<http://linis.hse.ru/>) using external developments, including the BigARTM algorithms library, which is included in the program as a DLL. The program is designed for topic modeling of Russian-language and English-language documents. The program includes: 1. The option of preprocessing documents. 2. Option of topic modeling and visualization of calculation results. 3. Option to analyze the stability of the results of thematic modeling. When publishing scientific results based on the work of this program, it is necessary to refer to the Internet Research Laboratory, Higher School of Economics

Topic modeling is one of the modern machine learning applications for text analysis that has been actively developing since the late 1990s. The topic model of the collection of text documents determines which topics each document relates to and what words (terms) form each topic. Each text and word belong to a set of topics. More precisely, each text and word belong to each topic with different probability. The input data of the topic model is the matrix (table) of words and documents, where the elements (cells) are the frequencies of words in the documents. The output data are two matrices of smaller dimension (smaller size): words on topics and documents on topics, where elements are the probabilities of words or documents belonging to topics. The number of required topics is set by the user based on experience.

In machine learning problems, either the selection of characteristics leading to a reduction in the number of parameters is usually used to reduce the dimension of a matrix, or regularization by imposing additional constraints on the parameters. In particular, Bayesian regularization is based on the introduction of a priori probability distribution in the parameter space. This program uses two basic approaches to calculate distribution of words by topics and documents by topics.

In this version, the following topic models are implemented:

1. LDA (Gibbs sampling), GLDA (Gibbs sampling).
2. PLSA + регуляризаторы (E-M algorithm)
3. Multimodal topic modeling (E-M algorithm)
4. Variational LDA (E-M algorithm).

In addition, this version of the software has the procedure of sentiment analysis based on the dictionary approach. The dictionary obtained as a result of the project 'Development of a public database and crowdsourcing web resource for creating tools of sentiment analysis', No. 14-04-12031 is offered as a Russian-language dictionary.

Chapter 1. Preprocessing of documents.

Preprocessing of documents is an essential part of working with documents. Preprocessing consists of three stages: 1. The procedure for assembling a set of documents into one file and lemmatization. 2. Procedure of calculating word frequencies, selecting words from parentheses, and creating a list of stop words. 3. Removal of stop words from the lemmatized texts.

1.1 Procedure for assembling and lemmatizing documents.

The input data for the TopicMiner software is a directory with documents, in which each file contains one document in txt format. In addition, in this directory there may be a file with metadata describing each file. An example of such a file is shown below. Each column contains a separate metadata attribute.

	A	B	C	D	E
1	1	1	gutta_honey	http://gutta-honey.livejournal.com/298516.html	18.02.2012 5:49
2	2	2	gutta_honey	http://gutta-honey.livejournal.com/298998.html	20.02.2012 22:15
3	3	3	gutta_honey	http://gutta-honey.livejournal.com/299320.html	21.02.2012 23:40
4	4	4	gutta_honey	http://gutta-honey.livejournal.com/299748.html	22.02.2012 8:45
5	5	5	gutta_honey	http://gutta-honey.livejournal.com/300401.html	24.02.2012 15:44
6	6	6	gutta_honey	http://gutta-honey.livejournal.com/300630.html	25.02.2012 9:33
7	7	7	gutta_honey	http://gutta-honey.livejournal.com/301085.html	26.02.2012 12:35
8	8	8	gutta_honey	http://gutta-honey.livejournal.com/301440.html	27.02.2012 14:21
9	9	9	gutta_honey	http://gutta-honey.livejournal.com/301700.html	28.02.2012 22:21

In this file, each line contains a set of metadata. The maximum number of metadata can not exceed 20 (20 columns). The first column contains the file names containing the text. It is recommended to number the files and use their numbers as names.

The first stage of the preprocessing.

The general view of the preprocessing window is shown in Figure 1.1.

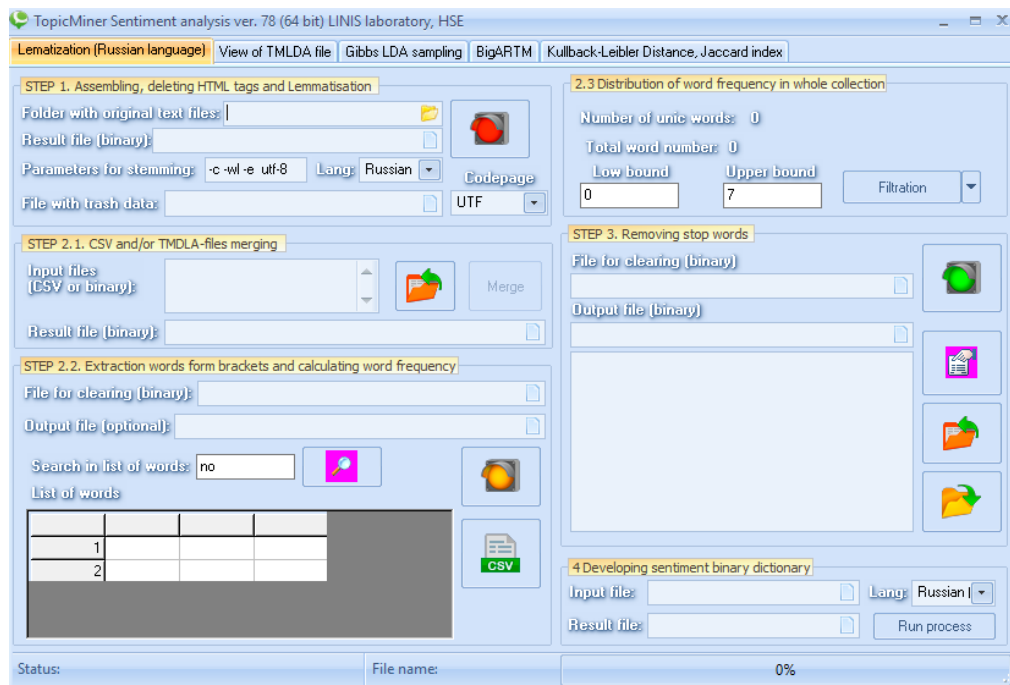


Fig. 1.1. General view of the window of the Russian preprocessing module.

Parameters of the first stage of preprocessing: 1. **The path to the directory with the initial data.** This path should be specified in the option:

Folder with original text files:

2. **The name of the file** where all original and lemmatized texts will be found. You can specify the file name in the following option:

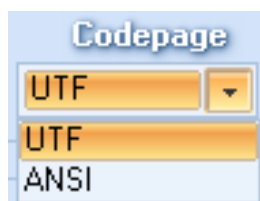
Result file (binary):

It is enough to specify only the file name. The program will automatically add an extension 'tmllda' (topic modeling LDA).

3. **The procedure of lemmatization** is based on the use of the lematizer 'mystem.exe' (development of the company 'Yandex', <https://tech.yandex.ru/mystem/>), which under the terms of the license can not be used for commercial purposes. To run the program 'mystem.exe' you must specify a set of parameters. In the TopicMiner program, these parameters are set automatically, based on encoding option selected by user. The list of parameters is specified in the line 'Parameters for stemming'.

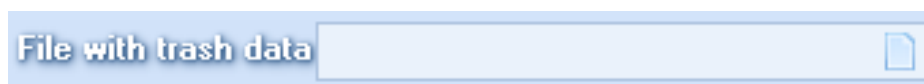


Selecting the encoding type for Russian texts. This program implements two types of encoding for source files.



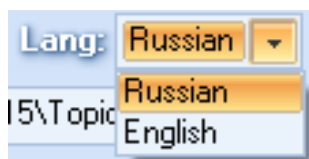
The user can select the encoding 'UTF' or 'ANSI'.

4. **File with a list of stop-symbols.** In original documents, symbols and groups of symbols may be present (for example, html markup, punctuation marks) that interfere with the analysis and should be removed from the texts. To perform the first stage of preprocessing, you must specify the name of the file in which such symbols are stored, and the path to it. This can be specified in the next option.



5. Language selection. In this version, two languages are supported, Russian and English.


The choice is made using the drop-down list:



The procedure of the lemmatization is carried out using the mystem and porter programs.

The completed parameter table for the first stage of the preprocessing can look like this (example):



After all the parameters are filled, in order to start the assembly and lemmatization process, you need to click on the button . The percentage of execution of the first stage - see Fig. 1.2.

Attention. Despite the fact that the process of lemmatization is parallel, the execution time of the first stage essentially depends on the number of source files and the total file size (kbytes ntrcnjd). For example, for 9 million short posts from a social network, the lemmatization time is approximately 13 days.

The result of preprocessing after the first stage.

The result of the preprocessing option after the first stage is a file with the extension tmla, which consistently contains pairs of texts in the original and lemmatized form. An example of the contents of such a file is shown in Figure 1.3. The program 'mystem.exe' converts each word in the documents to the initial form and puts each word in parentheses.

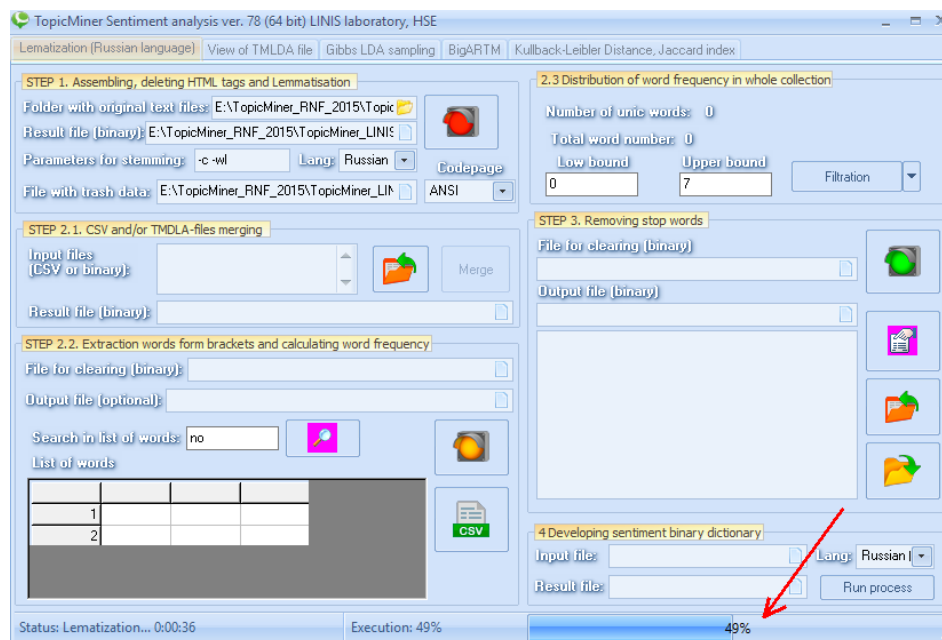


Fig. 1.2. An example of the process of lemmatization

ответов на вопрос, почему одни люди очень быстро спиваются, а другие могут годами пить потихонечку без особого вреда. Теперь решили исследовать, как конкретно алкоголь действует на мозг при отсутствии дофаминовых рецепторов данного типа. Как водится в ученых кругах вывели специальную линию мышек и стали их полгода пить раствором этилового спирта. Потом исследовали их мозг при помощи МРТ. Оказалось, что мыши без вышеназванного рецептора обнаруживали атрофию коры головного мозга и таламуса, в то время, как нормальные мыши не обнаруживали каких то заметных изменений. Людей совсем без этого рецептора, как утверждают опять же специалисты не встречается, но, а вот их сниженное количество в мозге может встречаться. Более того люди с низким количеством данного рецептора еще и быстрее развивают зависимость от алкоголя, по сравнению с другими.

<http://onlinelibrary.wiley.com/doi/10.1111/j.1530-0277.2011.01667.x/abstract.jsessionid=FB4EF53787D563FA8F1D6D2C3F205F0C.d01t01>{новость} {наука} {о} {зависимость}: {Чантиск??} {средство} {против} {курение}, {показывать} себе} {также} {положительно|положительный} {в} {отношение} {контроль} {над|нада} {прием} {алкоголь}. {тот}, кто} {принимать} {препарат??} {с} {цель} {бросать} {курить} {часто|частый} {сообщать}, {что} {у} {они} снижаться} {потребность} {в} {алкоголь}. {исследование} {показывать}, {что} {это|этот} действительно|действительный} {так}. {Чантиск??} {снижать} {ощущение} {удовольствие} {от} {прием} алкоголь} {и} {усиливать} {его|он|оно} {неприятный} {свойство}. {такой} {образ} {питие} {становиться} {совсем} безрадостный}. {исследование} {касаться} {только} {однократный} ({острый}) {прием} {препарат} {за} 3 {час} до} {прием} {алкоголь}. {длительный} {применение} {пок|пока} {не} {исследоваться}. {но} {тем|тема|то|тот} {не} мало|мнее|меней}, {предполагать}, {что} {препарат} {будет|быть} {снижать} {вероятность} {потерять} контроль} {на} {принимать} {алкоголь} {во} {время} {вечеринка}.

<http://www.uchospitals.edu/news/2012/20120215-alcoholism>{html} {еще} {один} {механизм}, который} {делать} {отказ} {от} {курение} {довольно|довольный} {трудный}. {в} {принцип}, {девать|дело} {вполне} ожидать}. {отказ} {от} {курение} {приводить} {к} {падение} {уровень} {дофамин} {в} {система} {вознаграждение}, что} {приводить} {к} {депрессия} {и} {к} {желание} {снова} {закуривать}. {подтверждение} {давать|данный} механизм} {делать} {применение} {дофаминергическх??} {препарат} {еще} {более|много}

Fig. 1.3. An example of the result of preprocessing after the first stage.

1.2. The second stage of preprocessing.


At the second stage of preprocessing, the words are extracted from brackets (see Figure 1.3) and the frequency of words across all documents is counted. The input data for the second stage is the file obtained after the first stage. You must specify the name and path to this file in the 'File for clearing (binary)' option (for example):

File for clearing (binary): D:\TopicMiner\polygon_RNF\data for orange\my_test1

In addition, you should specify the name of the file in which the results of the second stage of preprocessing will be stored. This should be done in the following 'Output file' option (for example):

Output file (optional): D:\TopicMiner\polygon_RNF\data for orange\my_test2.tmlc

The result of the second stage of preprocessing is the creation of a frequency dictionary of unique words and the conversion of lemmatized documents into a digital format. In this digital format, words in documents are replaced with numerical codes (IDs) of words from the list of unique

words. To start the second stage of preprocessing, you need to press the button . As a result of the work, new data (a frequency dictionary of unique words and digital documents) will be added to the file with the extension tmlcda. An example of the work is shown in Figure 1.4.

Attention. In this version, TF-IDF is implemented, but this option has not been fully tested yet.

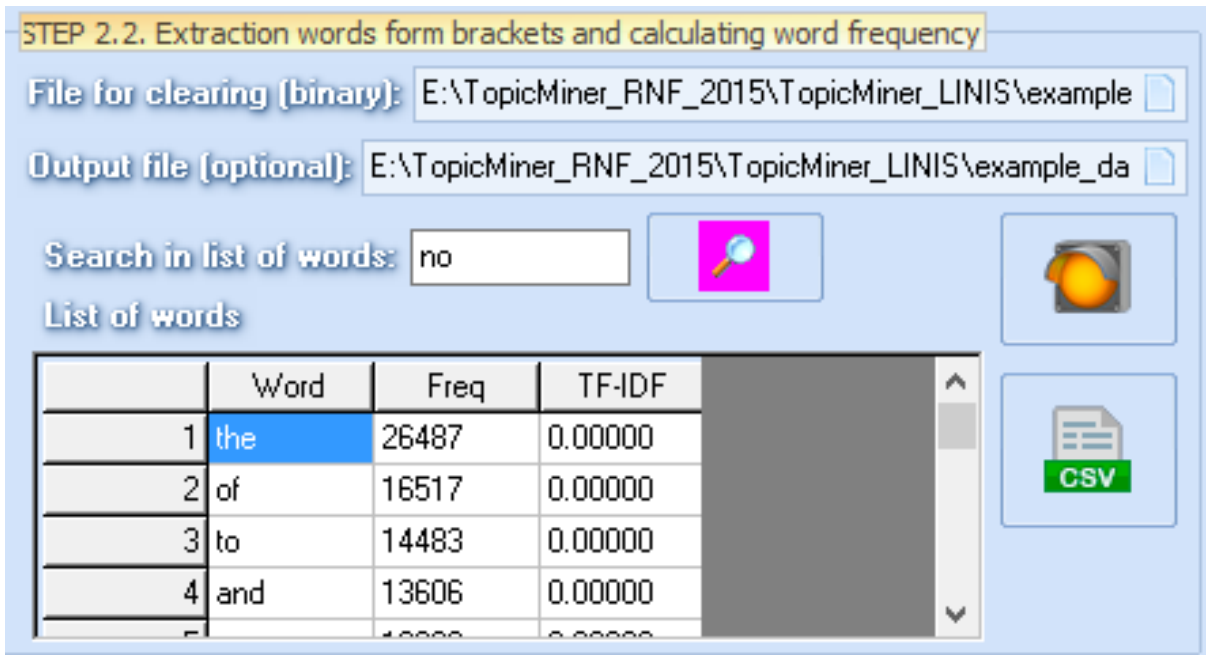




Fig. 1.4. An example of the result of preprocessing after the second stage.

The frequency dictionary can be downloaded in csv format to an external file. To do this, press the button  and specify the file name. If you need to find a word in the list of unique words, you need to specify it in the 'Search in list of words' window and click on the button . An example of the result is shown in Figure 1.5.

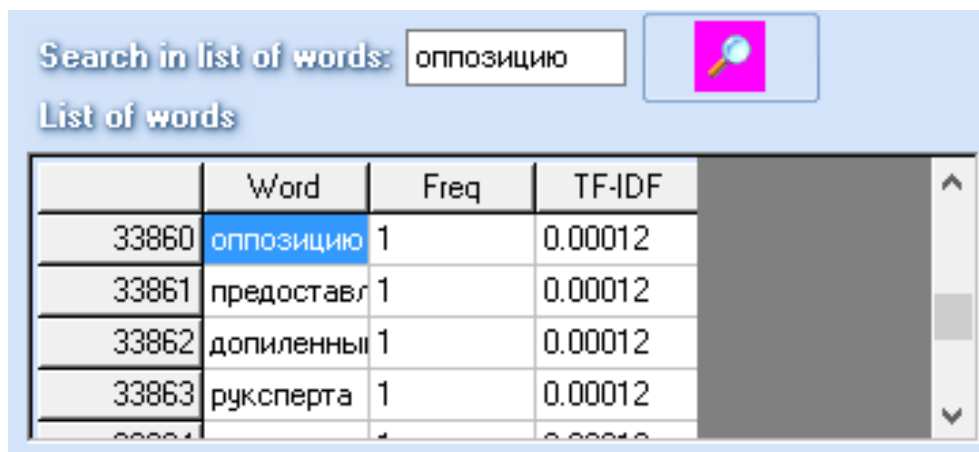


Fig. 1.5. An example of the result of preprocessing after the second stage.

1.2.1. Creation of a list of stop words.

In the second stage of preprocessing, you can create a list of stop words based on the list of frequencies of unique words. To do this, you must specify the upper and lower bounds for frequencies from the list of unique words in the 'Distribution of word frequency in whole collection' option:

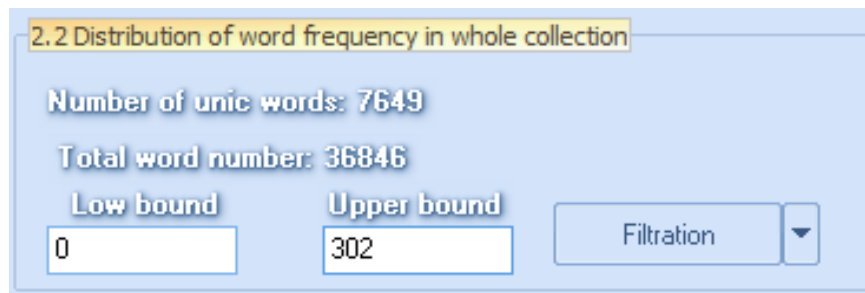


Fig. 1.6. Option to create a list of stop words.




After clicking the 'Filtration' button, a window will be opened where you need to specify the name of the file where the stop word list will be stored. The words whose frequencies are beyond specified limits will be saved there. In this example, the limits are the numbers '0' and '302'.

The result of preprocessing after the second stage is a file that contains original, lemmatized and digitized texts.

1.3. The third stage of preprocessing.

Here, stop words from digitized documents are deleted. The input data is the file that was output from the second stage; it must be specified. Then you need to specify the name of the output file, which will contain the original, lemmatized and digitized texts with deleted stop words. In addition, in this option you need to download a list of stop words from the text file. This can be a file created in the second stage, or an external file with any other list of words, or a file containing both.

This option contains the following buttons:

1. Button  : Clear the field for the stop word list.
2. Button  : Loading stop words from a text file.
3. Button  : Save a list of stop words to a text file.

The third button is needed if the user enters stop words in the TopicMiner field manually. The percentage of executed work concerning removing stop words is shown at the same place as the percentage of execution in the first stage of preprocessing.

Attention. It is necessary to go through all three stages of the preprocessing procedure

Chapter 2. View tmla format files.

The TopicMiner program provides the ability to view tmla format files, as well as the option to download texts (original and lemmatized) into a csv file. The view option is useful, since it allows you to see which stop words are not yet removed from the documents. Here, one can search for documents using the list of keywords and delete blank documents. This allows you to significantly reduce the size of the collection and, accordingly, increase the speed of topic modeling. The general view of the 'View of tmla files' option is shown in Figure 2.1.

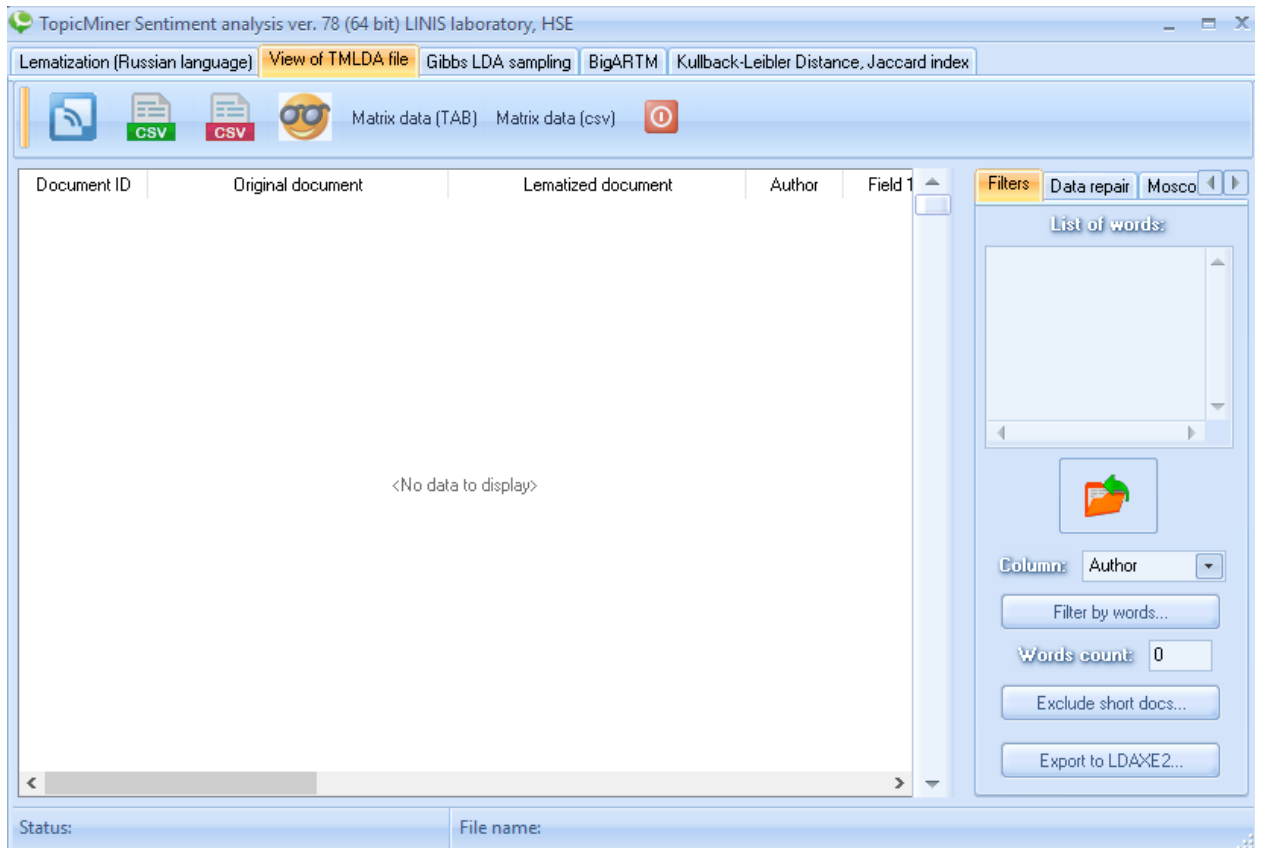






Fig. 2.1. The option to view tmla files.

Downloading the tmla file. To download a file in tmla format, click the button . In the appeared window it is necessary to specify a file name. As a result, the specified file will be displayed in the table (the example is shown in Figure 2.2). It has the following columns: 1. Column with original documents. 2. Column with lemmatized documents. 3. A set of columns with metadata. The format for metadata is described in Chapter 1.

Uploading original documents in csv format. The csv format is supported by many external programs, in particular Excel (if the data is not very large). For upload in csv format, click on the button  and specify the file name.

Uploading lemmatized documents in csv format. Click the button  and specify the file name.

Uploading of lemmatized documents in TAB format. The TAB format is supported by a number of external software products, in particular, the statistical package Orange. For uploading in TAB format, you need to click on the button  and specify the file name.

Downloading a list of words for filtering documents. To download a list of words, click on the button  and specify the file name.

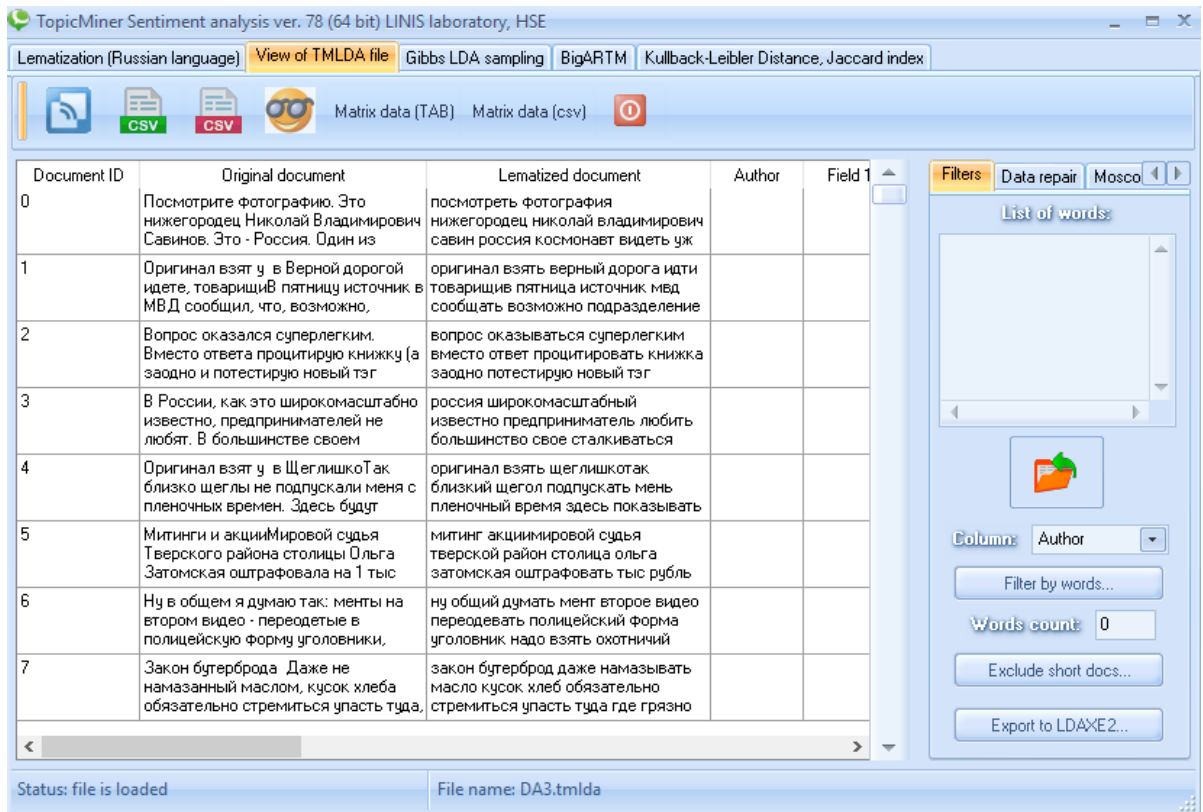


Fig. 2.2. An example of a downloaded file.

Note. The words in the text file must be presented in the following form: one word in a line. An example of a list of downloaded words is shown in Figure 2.3 on the right.

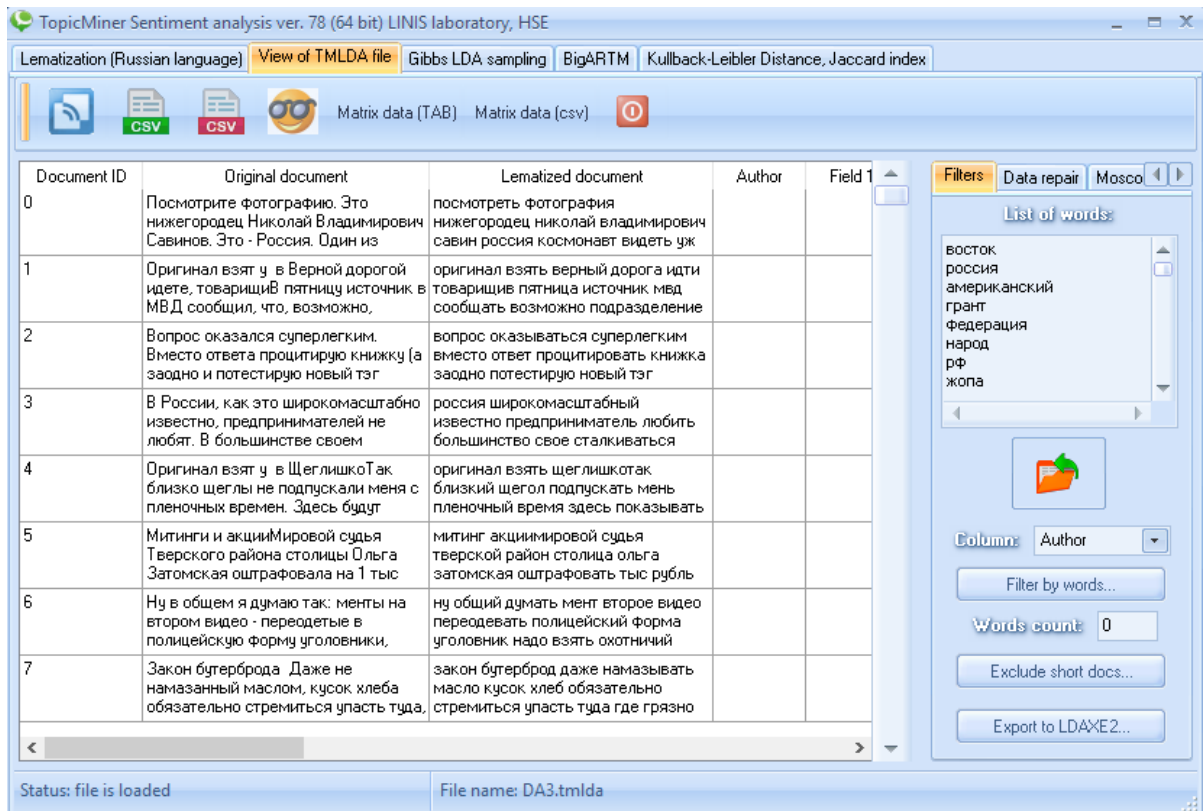






Fig. 2.3. Example of a loaded list of words

Uploading documents in tmla format by word list. To reduce the collection of documents according to the list of downloaded words, you need to click the button . The program will create a file in tmla format with the name of the originally downloaded file, however, the combination of the letters '_ww' will be added to the file name. For example, 'my_test2_ww.tmla'. The file will contain only those documents in which there is at least one word from the downloaded list.

Uploading documents in the 'tmla' format with deleted empty documents. Documents can be empty as a result of removing stop words, or initially - for example, these are social network records that contain only a photo. To reduce simulation time, it is recommended to delete such documents. To create a file in tmla format without any empty documents, you need to click on the button . The program will create a file in tmla format, with the name of the originally downloaded file, however, the combination of the letters '_we' will be added to the file name. For example, 'my_test2_we.tmla'.

Term-document matrix calculation (TAB format). When you click the button , the frequency of the list of words loaded into this option is calculated and the frequency matrix is downloaded for using this matrix in the statistical package 'Orange' (TAB separator). This matrix can be used to train classifiers such as 'Naïve Bayes', 'KNN', 'SVM'.

Calculation of the term-document matrix (CSV format). When you click the button , the frequencies of the list of words loaded in this option are calculated, and the frequency matrix is downloaded in CSV format. This matrix can be used to train classifiers such as 'Naïve Bayes', 'KNN', 'SVM'.

Forming 'tmla' files for multimodal models (BigARTM). The multimodal scheme of topic modeling includes the use of metadata fields (no more than 5 pieces of fields). As a result of the calculation of multimodal schemes, additional distribution matrices are formed for the selected metadata fields by topic. In order to generate data for the BigARTM model, select the option 'Dict for BigARTM' (see Figure 2.4)

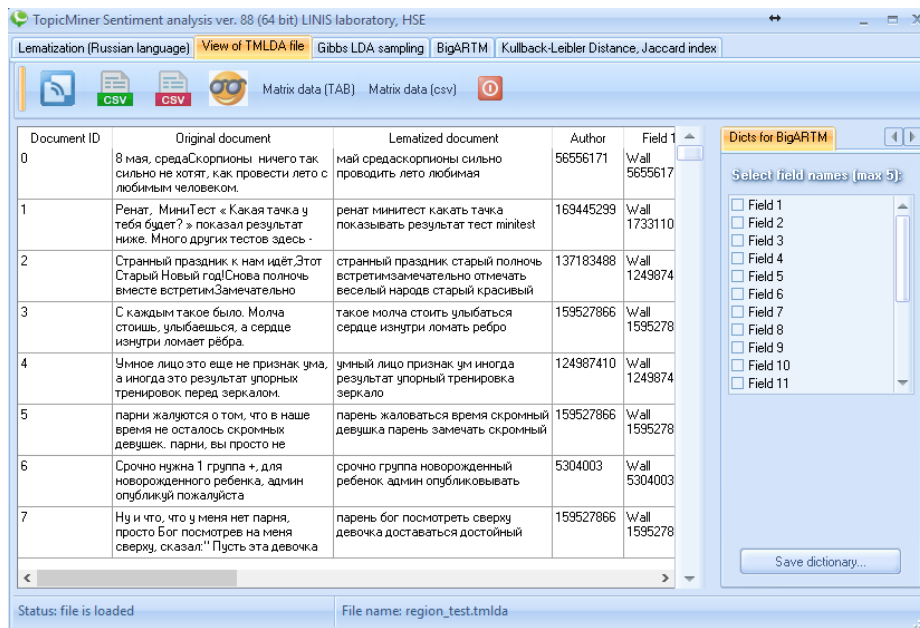



Fig. 2.4. Example of data generation option for BigArtm models.

In the field list, you must specify the required fields, for example, field 5 (the name of the author of the post) and field 7 (geotag). After that, press the button 

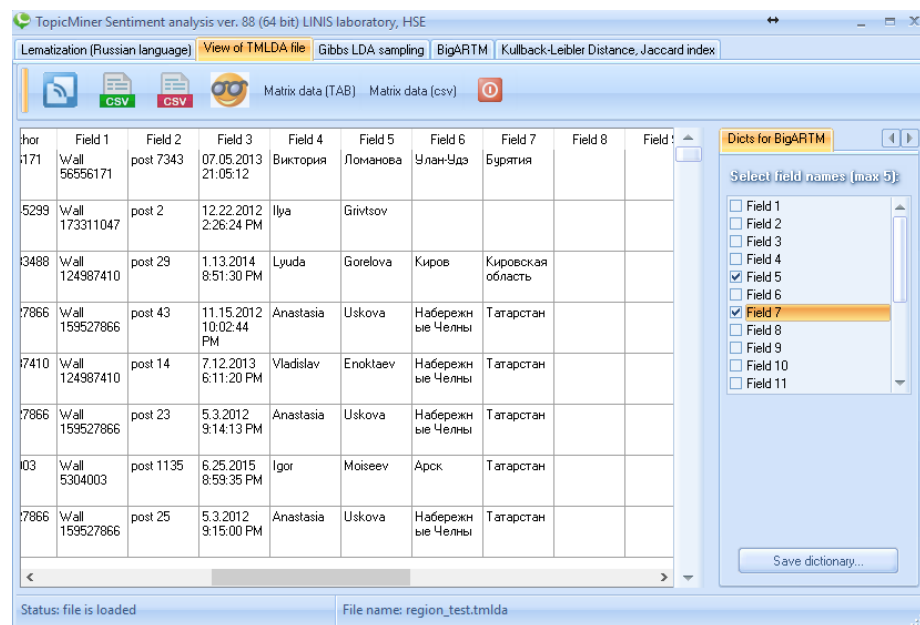


Fig. 2.5. Example of data generation option for BigArtm models.

As a result, the process of creating two files will start: 1. The Tmlda file, in which the selected metadata fields are formed. 2. The file with the metadata dictionaries. **Attention, this process takes a long time, as there is a procedure for the lematization of selected fields, converting the data to crc32 format and creating a list of unique words for the selected fields.**

Chapter 3. Topic modeling based on the Gibbs sampling model.

3.1. Interface of option 'Gibbs LDA sampling'.

The result of preprocessing is a file with the extension tmla. It contains lemmatized, original documents and documents in digital form. Each of the documents has its own ID (the ID of the lemmatized and original documents are the same). Lemmatized documents are used directly for thematic modeling, and original documents are easy to read.

The interface of the 'Gibbs LDA sampling' option looks like this (see Figure 3.1).



- data loading button for topic modeling.



- start button for topic modeling.



- stop button for topic modeling.



- button for viewing the matrix of the document distribution by topic (not sorted variant of the matrix).



- button for viewing the matrix of word distributions by topic (not sorted variant of the matrix).



- button for viewing the matrix of document distribution by topic (documents are sorted by probability in each topic in descending order).



- button for viewing the matrix of word distribution by topic (the words are sorted by probability in each topic in descending order).

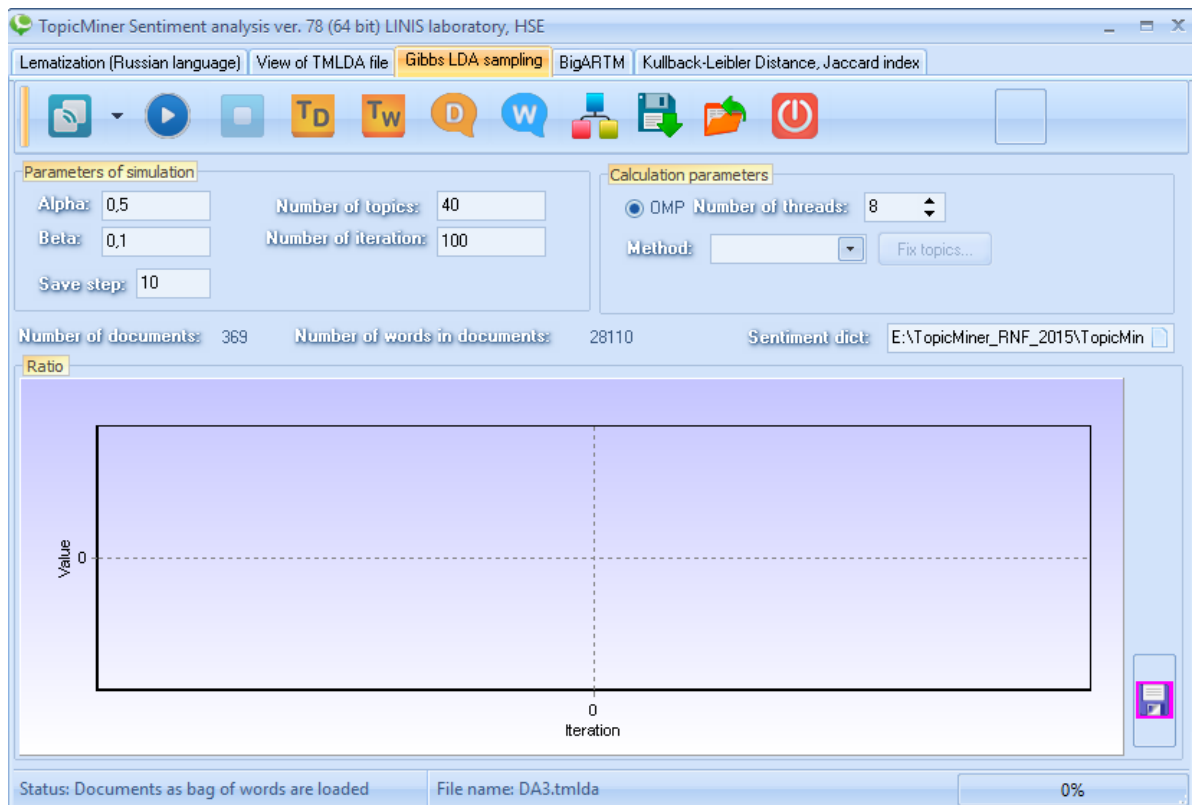



Fig. 3.1. Interface of the option 'Gibbs LDA sampling'

3.2. Loading documents for topic modeling.

To download documents to the program for models based on Gibbs sampling, you need to click on the button , and in the appeared window specify the file with the extension tmla (see Figure 3.2).

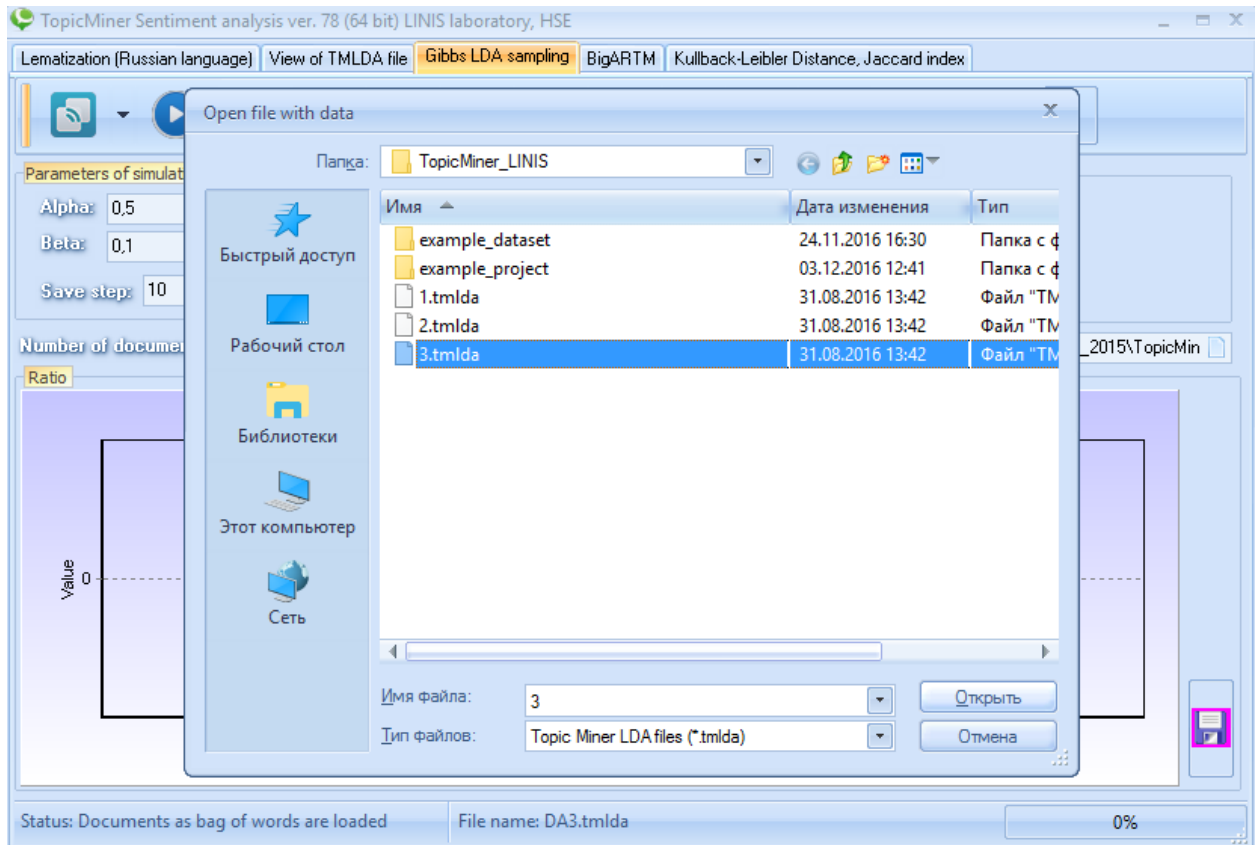


Fig. 3.2. An example of loading a data file.

An example of the data loading process is shown in Figure 3.3.

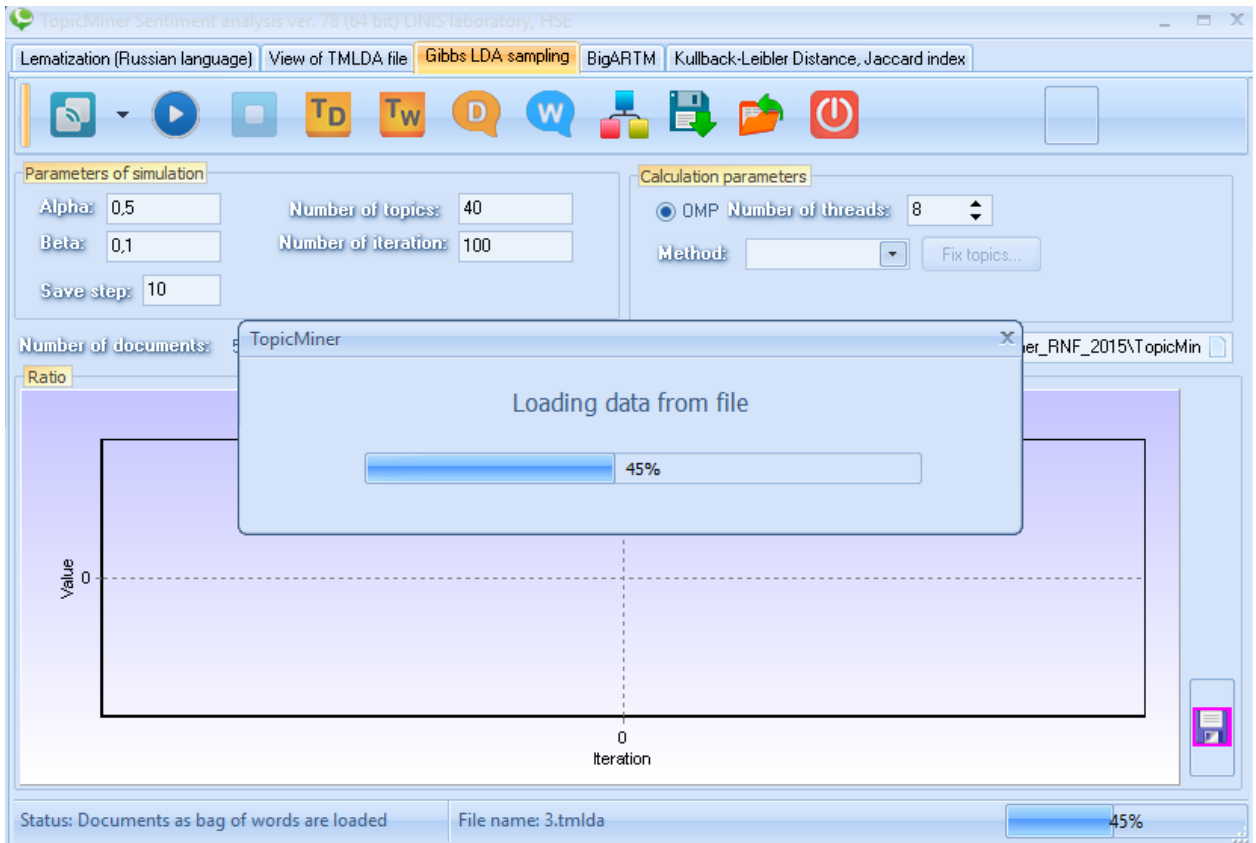


Fig. 3.3. Example of uploading a data file

After downloading the program will show statistics on documents and words (see Figure 3.4).

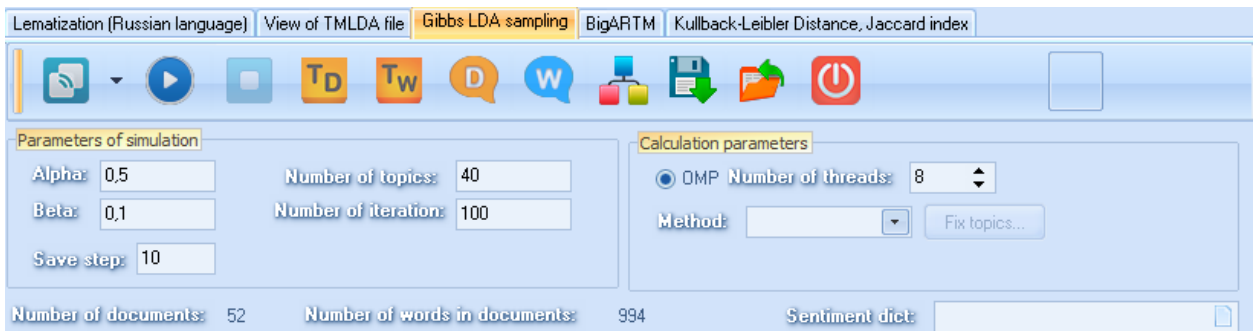


Fig. 3.4. Example of uploading a data file.

Number of documents is the number of documents in the collection (the number of documents in the tmla file).

Number of words in documents is the number of unique words in the collection.


A downloaded collection of documents can be used in topic modeling.


3.3. Topic modeling based on Gibbs sampling.

Before starting the simulation, you need to specify the following simulation parameters:

- 1) Coefficients α , β . Default values: $\alpha=0,5$, $\beta=1$. Beginners can use default values:

Alpha:	0,5
Beta:	0,1

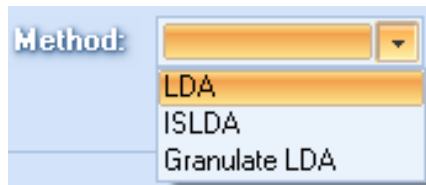
2) Number of topics. The number of topics can be set in the option: . The default value is 40 topics, however, all users are encouraged to experiment with the number of topics, usually, increasing from the default setting.

3) The number of iterations. The number of iterations can be set in the option: . The default value is 100. Beginners can use this value.

4) Save step. This parameter shows the iteration step, which determines which step to visualize the calculation results. The default value is 10. You can change the value in the next option:




5) Type of model. In this version, three types of models are implemented (the standard LDA model, the ISLDA model and the granulated GLDA sampling method). Recommended for advanced users. The model is selected from the drop-down list.



6) The number of threads. This program implements parallelization of the thematic model based on Gibbs sampling using OpenMP technology. The number of threads can be specified

in the following option: 

After setting the parameters, you need to click on the button . The calculation process (iteration number) is shown in the lower left corner of the window (see Figure 3.5).

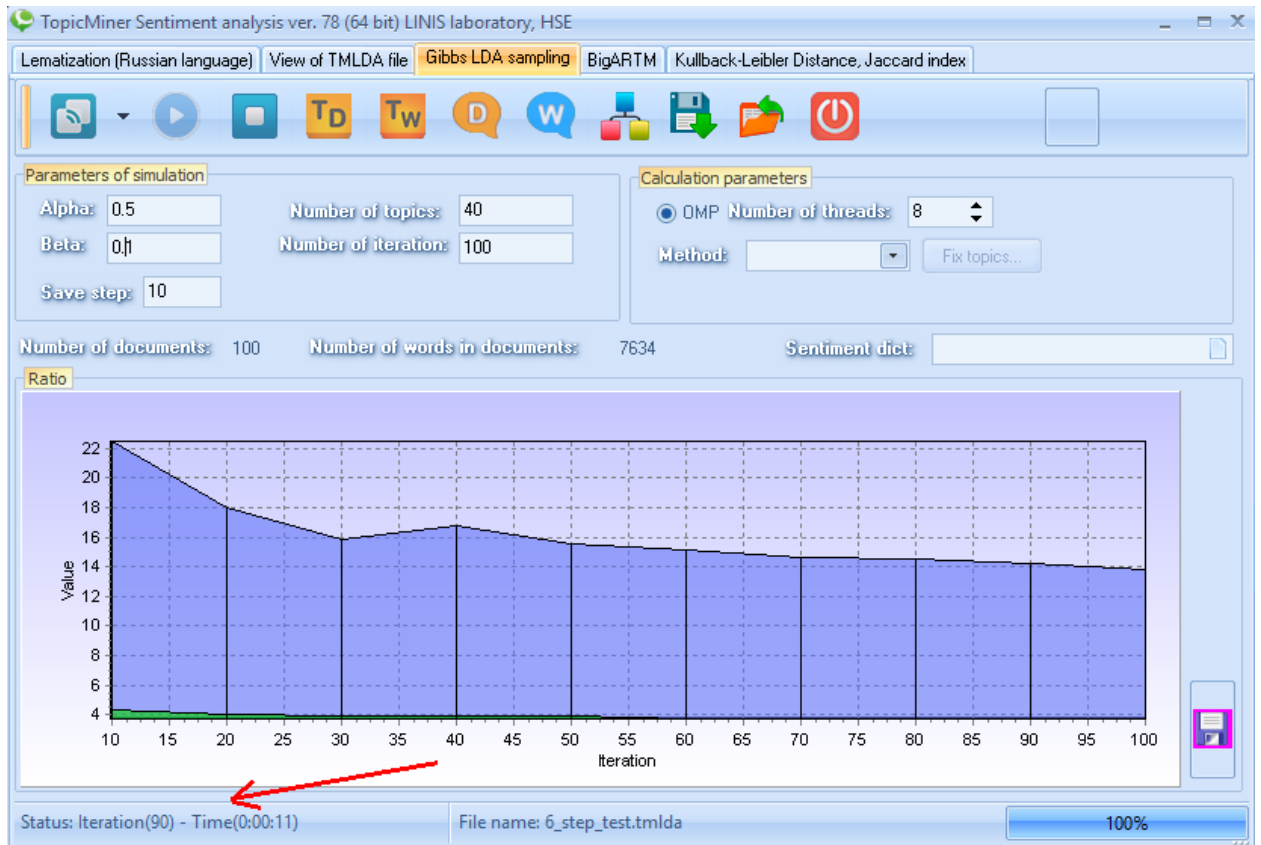


Fig. 3.5. The process of execution of the topic model.

During the process of execution of topic model, the program calculates the proportion of words and documents for which the probability is above average. The probability curves during the iterations are shown in the graph (see Figure 3.6).

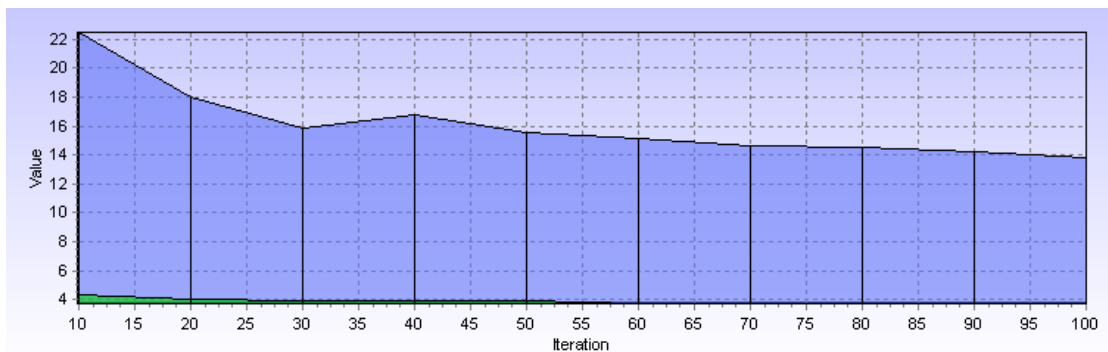


Fig. 3.6. The process of execution of the topic model.

The blue chart shows the proportion of documents, green shows the proportion of words. For example, for documents from the social network Live Journal, a typical number of documents with a probability above the average value of about 11%.

3.4. Visualization of the results of topic modeling

Visualization of topic modeling consists of the following items:

1. Visualization of document distribution by words.
2. Visualization of words by topics.
3. Visualization of sorted document distributions by topic.
4. Visualization of sorted word distributions by topic.

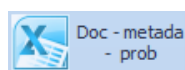


The visualization modules can be launched using the buttons

3.4.1. Visualization of document distribution by words.

To visualize the distribution of documents by topic, click on the button **TD**. A window will appear (see Figures 3.7 and 3.8). In the table, each line represents the text of the document (the column 'Orig text'), its metadata (starting with the column 'Nick' and ending with the 'Field 20' column) and the probability of belonging to the topics. Thus, TopicMiner allows you to use 21 columns for metadata (see Figure 3.7). The distribution of documents by topic is shown in the columns, starting with the column '1' and ending with the topic number, which is specified in the 'Number of topic' parameter.

In this window there are also several buttons that allow you to unload the results of thematic modeling into csv format files.



- Unload the results of topic modeling in the csv format in the form: original text - metadata - probabilities for all topics. An example of such unloading is shown in Figure 3.9.

	ID	Orig text	Nick	Field 1	Field 2
1					
2	1	Новости науки о зависимостях: Чантикс средство	1	gutta_honey	http://gutta-honey.livejournal.com/298516.html
3	2	Только сейчас и только для вас, настоящие идолы со	37	yuzilla	http://yuzilla.livejournal.com/926282.html
4	3	Этот пятилетний мальчик, британец Зак Эйвери, год	25	yuzilla	http://yuzilla.livejournal.com/923209.html
5	4	Никакого спора тут нет. Цивилизация в целом устр	61	alexlotov	http://alexlotov.livejournal.com/357885.html
6	5	Последние несколько лет Питер Липпманн (Peter Lipp	26	yuzilla	http://yuzilla.livejournal.com/923636.html
7	6	Проголосуем за Эюганова, чтобы он проконтролировал	85	alexlotov	http://alexlotov.livejournal.com/363880.html
8	7	Именно про него самого, ведь на нем родимом держат	38	yuzilla	http://yuzilla.livejournal.com/926693.html
9	8	Почему мы воюем" - Битва за Россию: The Battle of	62	alexlotov	http://alexlotov.livejournal.com/358134.html
10	9	Вот так представьте: взяли вы свою вторую половину	27	yuzilla	http://yuzilla.livejournal.com/923797.html
11	10	видео от grigoruk Оппозиция не против оккупации	63	alexlotov	http://alexlotov.livejournal.com/358294.html
12	11	Джен Старк художница из Майами.При помощи цветно	49	yuzilla	http://yuzilla.livejournal.com/929280.html
13	12	Явка обещает быть высокой, потому что нормальные з	86	alexlotov	http://alexlotov.livejournal.com/364282.html
14	13	Поставлена жирная точка в деле Юрия Луценко. Сегод	28	yuzilla	http://yuzilla.livejournal.com/924058.html
15	14	Есть такие европейские врачи-окулисты(чуть было не	39	yuzilla	http://yuzilla.livejournal.com/926959.html
16	15	Совершенно очевидно, что поднимать визг, вопли и в	64	alexlotov	http://alexlotov.livejournal.com/358436.html
17	16	Живем мы в такое время, что информация льется со	2	gutta_honey	http://gutta-honey.livejournal.com/298998.html
18	17	Представляю вам подборку изящных и медитативных ф	29	yuzilla	http://yuzilla.livejournal.com/924253.html
19	18	Меланизм преимущественное распространение тёмноо	50	yuzilla	http://yuzilla.livejournal.com/929771.html

Есть такие европейские врачи-окулисты(чуть было не написал окулисты :) утверждают, что вот длительный просмотр 3D-видео детьми является очень вредным для их зрения и его развития в целом. Да кто ж сомневался, оно и у взрослых вызывает головные боли даже. Вообще в первую очередь, это конкретно относится к такому 3D, для которого кстати не нужны те самые специальные стерео-скопические очки. Известный специалист Карен Спарроу окулист из официальной Европейской ассоциации окулистов четко поясняет, что для нормального здорового развития зрения у детей, им нужно исключительно чистое изображение, точнее сказать именно такое, которое спокойно воспринимается глазками детей. А 3D-видео уж очень губительно может повлиять на зрение у детей с возрастом до 6 лет и

Hide selected row | Hide by threshold: 0,5 | Reset hidden row | Doc - metadata - prob | Doc ID - prob | Number of documents for export: 100

Fig. 3.7. Visualization of document distribution by topics (first part).

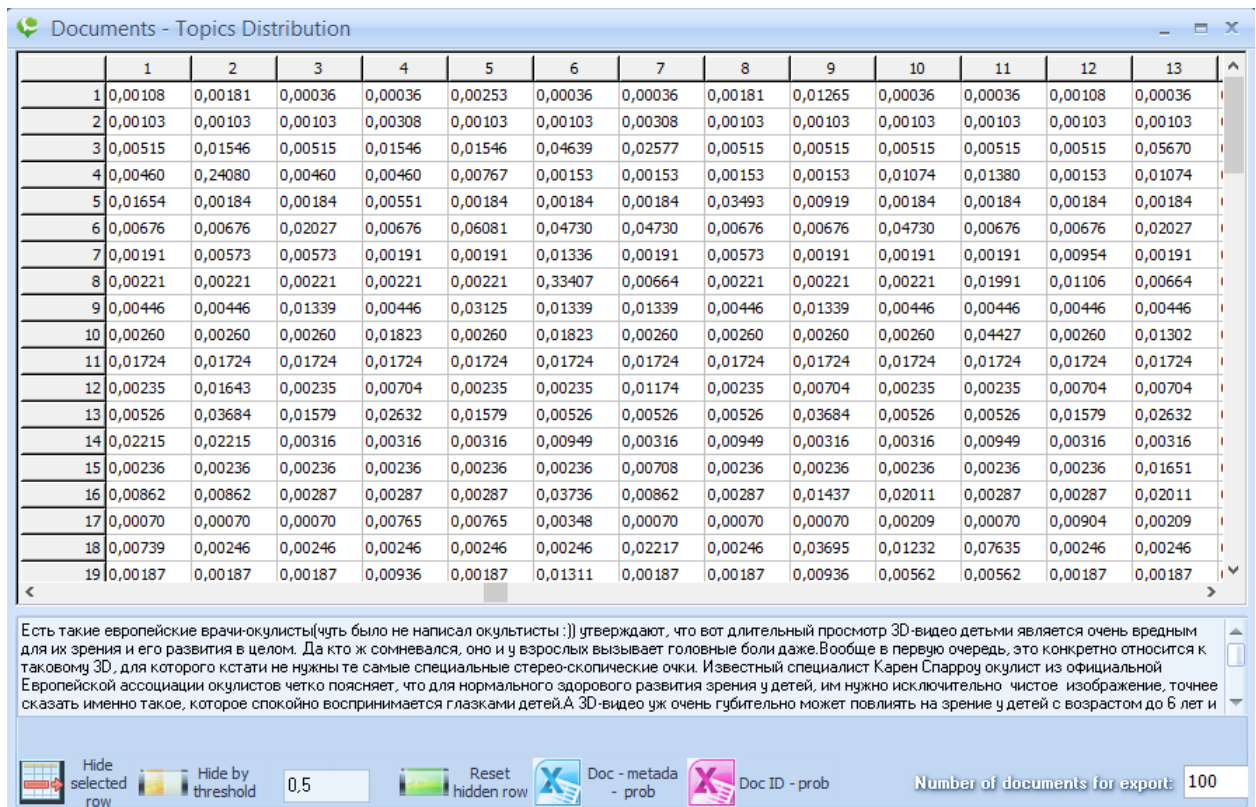
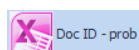


Fig. 3.8. Fig. 3.7. Visualization of document distribution by topic (second part).

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	
1	ID	Orig text	Nick	Field 1	Field 2	Field 3	Field 4	Field 5	Field 6	1	2	3	4	5	6	7	8	9
2										0,00108	0,00181	0,00036	0,00036	0,00253	0,00036	0,00036	0,00181	0,01265
3	1	Новости науки о зависимостях: Чанти	1	gutta_honey	http://gutta-honey.livejournal.com/298516.html					0,00103	0,00103	0,00103	0,00308	0,00103	0,00103	0,00308	0,00103	0,00103
4	2	Только сейчас и только для вас, настс	37	yuzilla	http://yuzilla.livejournal.com/926282.html					0,00515	0,01546	0,00515	0,01546	0,01546	0,04639	0,02577	0,00515	0,00515
5	3	Этот пятилетний мальчик, британец :	25	yuzilla	http://yuzilla.livejournal.com/923209.html					0,0046	0,2408	0,0046	0,0046	0,00767	0,00153	0,00153	0,00153	0,00153
6	4	Никакого спора тут нет. Цивилизаци	61	alexlotov	http://alexlotov.livejournal.com/357885.html					0,01654	0,00184	0,00184	0,00551	0,00184	0,00184	0,00184	0,03493	0,00919
7	5	Последние несколько лет Питер Лип	26	yuzilla	http://yuzilla.livejournal.com/923636.html					0,00676	0,00676	0,02027	0,00676	0,06081	0,04730	0,04730	0,00676	0,00676
8	6	Проголоусеум за Зоганова, чтобы он г	85	alexlotov	http://alexlotov.livejournal.com/363880.html					0,00191	0,00573	0,00573	0,00191	0,00191	0,01336	0,00191	0,00573	0,00191
9	7	Именно про него самого, ведь на нем	38	yuzilla	http://yuzilla.livejournal.com/926693.html					0,00221	0,00221	0,00221	0,00221	0,00221	0,33407	0,00664	0,00221	0,00221
10	8	Почему мы воюем - битва за Россию:	62	alexlotov	http://alexlotov.livejournal.com/358134.html					0,00446	0,00446	0,01339	0,00446	0,03125	0,01339	0,01339	0,00446	0,01339
11	9	Вот так представьте: взяли вы свою в'	27	yuzilla	http://yuzilla.livejournal.com/923797.html					0,0026	0,0026	0,0026	0,01823	0,0026	0,01823	0,0026	0,0026	0,0026
12	10	видео от gtrigotuk Опозиция не про	63	alexlotov	http://alexlotov.livejournal.com/358294.html					0,01724	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724
13	11	Джен Старк художница из Майами.Г	49	yuzilla	http://yuzilla.livejournal.com/929280.html					0,00235	0,01643	0,00235	0,00704	0,00235	0,00235	0,01174	0,00235	0,00704
14	12	Явкя обещает быть высокой, потому	86	alexlotov	http://alexlotov.livejournal.com/364282.html					0,00526	0,03684	0,01579	0,02632	0,01579	0,00526	0,00526	0,03684	
15	13	Поставлена жирная точка в деле Юри	28	yuzilla	http://yuzilla.livejournal.com/924058.html					0,02215	0,02215	0,00316	0,00316	0,00316	0,00949	0,00316	0,00949	0,00316
16	14	Есть такие европейские врачи-окули	39	yuzilla	http://yuzilla.livejournal.com/926959.html					0,00236	0,00236	0,00236	0,00236	0,00236	0,00236	0,00708	0,00236	0,00236
17	15	Совершенно очевидно, что поднимае'	64	alexlotov	http://alexlotov.livejournal.com/358436.html					0,00862	0,00862	0,00287	0,00287	0,00287	0,03736	0,00862	0,00287	0,01437
18	16	Живем мы в такое время, что инфор	2	gutta_honey	http://gutta-honey.livejournal.com/298998.html					0,0007	0,0007	0,0007	0,00765	0,00765	0,00348	0,0007	0,0007	0,0007
19	17	Представляю вам подборку изящных	29	yuzilla	http://yuzilla.livejournal.com/924253.html					0,00739	0,00246	0,00246	0,00246	0,00246	0,00246	0,02217	0,00246	0,03695
20	18	Меланизм преимущественное расп	50	yuzilla	http://yuzilla.livejournal.com/929771.html					0,00187	0,00187	0,00187	0,00936	0,00187	0,01311	0,00187	0,00187	0,00936
21	19	Самое дорогое путешествие на двои:	40	yuzilla	http://yuzilla.livejournal.com/927036.html					0,00211	0,00211	0,03165	0,00633	0,02321	0,27637	0,00211	0,00211	0,01899
22	20	Новая парадигма мировоззрения Ктс	73	alexlotov	http://alexlotov.livejournal.com/360899.html					0,00214	0,00071	0,00071	0,00071	0,00214	0,00071	0,00071	0,00071	0,00357
23	21	el_muridi: Арабская весна создала вес	87	alexlotov	http://alexlotov.livejournal.com/364440.html					0,05968	0,00161	0,00161	0,00161	0,00161	0,00484	0,00161	0,17258	0,00484
24	22	В случае конфликта , неудовлетвори	13	gutta_honey	http://gutta-honey.livejournal.com/303091.html					0,00038	0,00038	0,0019	0,00267	0,00038	0,00038	0,00343	0,00038	0,00038
25	23	Сегодня в Астрахани произошел взрь	30	yuzilla	http://yuzilla.livejournal.com/924445.html					0,00256	0,00256	0,00256	0,00256	0,01795	0,00769	0,01282	0,00256	0,00256
26	24	Проводимые разными компаниями и	41	yuzilla	http://yuzilla.livejournal.com/927292.html					0,01289	0,00773	0,00258	0,00258	0,00258	0,00258	0,00773	0,00258	0,01289

Fig. 3.9. Uploading the results of topic modeling in the format 'Original text - metadata - probabilities'.



- Uploading data in the form: document id - the number of words in the document - metadata - probabilities. The data is downloaded in csv format. An example of such an unloading is shown in Figure 3.10.


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	ID docum	Number o	Nick	Field1	Field2	Field3	Field4	Field5	Field6	Topic(1)	Topic(2)	Topic(3)	Topic(4)	Topic(5)	Topic(6)	Topic(7)	Topic(8)	Topic(9)	Topic(10)	Topic(11)
2	1	447	1	gutta_honey	http://gutta-honey.livejournal.com/298516.html	0,00103	0,00103	0,00103	0,00308	0,00103	0,00103	0,00308	0,00103	0,00103	0,00308	0,00103	0,00103	0,00103	0,00103	0,00103
3	2	74	37	yuzilla	http://yuzilla.livejournal.com/926282.html	0,00515	0,01546	0,00515	0,01546	0,00515	0,01546	0,00515	0,01546	0,00515	0,01546	0,00515	0,01546	0,00515	0,01546	0,00515
4	3	308	25	yuzilla	http://yuzilla.livejournal.com/923209.html	0,0046	0,2408	0,0046	0,0046	0,00767	0,00153	0,00153	0,00153	0,00153	0,00153	0,00153	0,00153	0,00153	0,00153	0,00153
5	4	253	61	alexlotov	http://alexlotov.livejournal.com/357885.html	0,01654	0,00184	0,00184	0,00184	0,00551	0,00184	0,00184	0,00184	0,00184	0,00184	0,00184	0,00184	0,00184	0,00184	0,00184
6	5	53	26	yuzilla	http://yuzilla.livejournal.com/923636.html	0,00676	0,00676	0,02027	0,00676	0,06081	0,0473	0,0473	0,00676	0,00676	0,0473	0,0473	0,00676	0,00676	0,0473	0,00676
7	6	243	85	alexlotov	http://alexlotov.livejournal.com/363880.html	0,00191	0,00573	0,00573	0,00191	0,00191	0,01336	0,00191	0,00191	0,01336	0,00191	0,00573	0,00191	0,00191	0,00191	0,00191
8	7	211	38	yuzilla	http://yuzilla.livejournal.com/926693.html	0,00221	0,00221	0,00221	0,00221	0,00221	0,33407	0,00664	0,00221	0,33407	0,00664	0,00221	0,00221	0,00221	0,00221	0,01991
9	8	102	62	alexlotov	http://alexlotov.livejournal.com/358134.html	0,00446	0,00446	0,01339	0,00446	0,03125	0,01339	0,01339	0,00446	0,03125	0,01339	0,01339	0,00446	0,01339	0,00446	0,00446
10	9	172	27	yuzilla	http://yuzilla.livejournal.com/923797.html	0,0026	0,0026	0,0026	0,01823	0,0026	0,01823	0,0026	0,01823	0,0026	0,01823	0,0026	0,0026	0,0026	0,0026	0,04427
11	10	9	63	alexlotov	http://alexlotov.livejournal.com/358294.html	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724
12	11	247	49	yuzilla	http://yuzilla.livejournal.com/929280.html	0,00235	0,01643	0,00235	0,00704	0,00235	0,00235	0,01174	0,00235	0,00235	0,01174	0,00235	0,00704	0,00235	0,00235	0,00235
13	12	78	86	alexlotov	http://alexlotov.livejournal.com/364282.html	0,00526	0,03684	0,01579	0,02632	0,01579	0,00526	0,00526	0,00526	0,01579	0,00526	0,00526	0,00526	0,03684	0,00526	0,00526
14	13	138	28	yuzilla	http://yuzilla.livejournal.com/924058.html	0,02215	0,02215	0,00316	0,00316	0,00316	0,00949	0,00316	0,00949	0,00316	0,00949	0,00316	0,00949	0,00316	0,00316	0,00949
15	14	194	39	yuzilla	http://yuzilla.livejournal.com/926959.html	0,00236	0,00236	0,00236	0,00236	0,00236	0,00236	0,00236	0,00236	0,00236	0,00708	0,00236	0,00236	0,00236	0,00236	0,00236
16	15	154	64	alexlotov	http://alexlotov.livejournal.com/358436.html	0,00862	0,00862	0,00287	0,00287	0,00287	0,03736	0,00862	0,00287	0,03736	0,00862	0,00287	0,01437	0,02011	0,00287	0,00287
17	16	714	2	gutta_honey	http://gutta-honey.livejournal.com/298998.html	0,0007	0,0007	0,0007	0,00765	0,00765	0,00348	0,0007	0,0007	0,0007	0,0007	0,0007	0,0007	0,0007	0,00209	0,0007
18	17	190	29	yuzilla	http://yuzilla.livejournal.com/924253.html	0,00739	0,00246	0,00246	0,00246	0,00246	0,00246	0,00246	0,00246	0,00246	0,00246	0,00246	0,00246	0,03695	0,01232	0,07635
19	18	247	50	yuzilla	http://yuzilla.livejournal.com/929771.html	0,00187	0,00187	0,00187	0,00936	0,00187	0,01311	0,00187	0,00187	0,01311	0,00187	0,00936	0,00187	0,00936	0,00562	0,00562
20	19	216	40	yuzilla	http://yuzilla.livejournal.com/927036.html	0,00211	0,00211	0,03165	0,00633	0,02321	0,27637	0,00211	0,00211	0,01899	0,02321	0,00211	0,01899	0,02321	0,01899	0,01899
21	20	688	73	alexlotov	http://alexlotov.livejournal.com/360899.html	0,00214	0,00071	0,00071	0,00071	0,00214	0,00071	0,00071	0,00071	0,00214	0,00071	0,00071	0,00071	0,00357	0,00071	0,00071
22	21	294	87	alexlotov	http://alexlotov.livejournal.com/364440.html	0,05968	0,00161	0,00161	0,00161	0,00161	0,00484	0,00161	0,00161	0,00484	0,00161	0,17258	0,00484	0,00806	0,00161	0,00161
23	22	1315	13	gutta_honey	http://gutta-honey.livejournal.com/303091.html	0,00038	0,00038	0,0019	0,00267	0,00038	0,00038	0,00343	0,00038	0,00038	0,00343	0,00038	0,00038	0,00495	0,00038	0,00038
24	23	175	30	yuzilla	http://yuzilla.livejournal.com/924445.html	0,00256	0,00256	0,00256	0,00256	0,01795	0,00769	0,01282	0,00256	0,00256	0,00769	0,01282	0,00256	0,00256	0,00256	0,00256
25	24	172	41	yuzilla	http://yuzilla.livejournal.com/927292.html	0,01289	0,00773	0,00258	0,00258	0,00258	0,00258	0,00773	0,00258	0,00258	0,00773	0,00258	0,01289	0,00773	0,01804	0,01804
26	25	370	65	alexlotov	http://alexlotov.livejournal.com/358710.html	0,0013	0,0039	0,00649	0,0013	0,0013	0,0013	0,0013	0,0013	0,0013	0,0013	0,0013	0,0013	0,0013	0,0039	0,0013

Fig. 3.10. Uploading the results of thematic modeling in format 'Document Id - metadata - probabilities'.

In these uploads, the number of uploaded documents can be specified in the options:

Number of documents for export:

3.4.2. Visualization of word distributions by topics.

To visualize the distribution of words by topic, you need to click on the button . When you click this button, a window will appear (see Figure 3.11).


	Word	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	в	0,02077	0,00006	0,00012	0,00011	0,00008	0,00009	0,00005	0,00009	0,00011	0,00011	0,02139	0,00008	0,00007	0,00010
2	и	0,00008	0,00006	0,00012	0,00011	0,00008	0,00009	0,00005	0,00009	0,00011	0,00011	0,00006	0,00008	0,00010	0,00010
3	не	0,00008	0,00006	0,00012	0,00011	0,00008	0,00009	0,00005	0,00009	0,00011	0,00011	0,00006	0,00008	0,00007	0,00010
4	на	0,00008	0,00006	0,00012	0,00011	0,02029	0,00009	0,00005	0,00009	0,00011	0,00011	0,00006	0,00008	0,00007	0,00010
5	что	0,00008	0,00006	0,00012	0,00011	0,00008	0,00009	0,00005	0,00009	0,00011	0,00011	0,00006	0,00008	0,00007	0,00010
6	это	0,00008	0,00006	0,00012	0,00011	0,00008	0,00009	0,00005	0,00009	0,00011	0,00011	0,00006	0,00008	0,00007	0,00010
7	с	0,00008	0,00006	0,00012	0,00011	0,00251	0,00009	0,00106	0,00009	0,00011	0,00011	0,00006	0,00008	0,00007	0,00010
8	весь	0,00008	0,00065	0,00012	0,00011	0,00008	0,00009	0,00005	0,00009	0,00011	0,00011	0,00006	0,00008	0,00007	0,00010
9	то	0,00008	0,00006	0,00012	0,00011	0,00008	0,00009	0,00106	0,00009	0,00011	0,00011	0,00006	0,00008	0,00007	0,00010
10	быть	0,00008	0,00006	0,00012	0,00011	0,00008	0,00009	0,00005	0,00009	0,00011	0,00011	0,00006	0,00008	0,00007	0,00010
11	он	0,00008	0,00006	0,00012	0,00011	0,00008	0,00009	0,00005	0,00009	0,00011	0,00011	0,00006	0,00260	0,00007	0,00010
12	как	0,00008	0,00006	0,00012	0,00011	0,00089	0,00009	0,00005	0,00009	0,00011	0,00011	0,00006	0,00008	0,00007	0,00010

Hide selected row Reset hidden rows Export to Excel... Number of words for export: Boundary for probability:

Fig. 3.11. Example of visualization of word distributions by topic

The size of the upload (number of documents) is regulated by two parameters: 1. 'Number of words for export'. 2. 'Boundary for probability' (see Figure 3.11). The first parameter regulates the number of words for export in the csv format, the second parameter specifies the probability of a word in the subject, the minimum required to get the word into the download. Words with lower probabilities are not unloaded.

The button allows you to hide the selected row in the table. The latent word does not participate in the unloading in the csv format.

The button  allows you to restore all previously hidden words.


To unload the distribution of words by topic in a csv file, click on the button



and in the appeared window specify the file name.

Attention: this unloading is useful in researching the stability of thematic modeling or when comparing the performance of several models with each other. A similar comparison is discussed in Chapter 5.

3.4.3. Visualization of distributions of sorted documents by topics.

To open a window in which document distribution by topic is presented in descending order of probability, you need to click on the button . As a result, a window will appear; sorting in it is done by probabilities, so that at the top (in each topic) there is a document with the greatest probability of belonging to this topic. An example of such sorting is shown in Figure 3.12. In each cell of this table is the document number and its probability. If you click on the selected cell, the original text appears at the bottom of the screen.

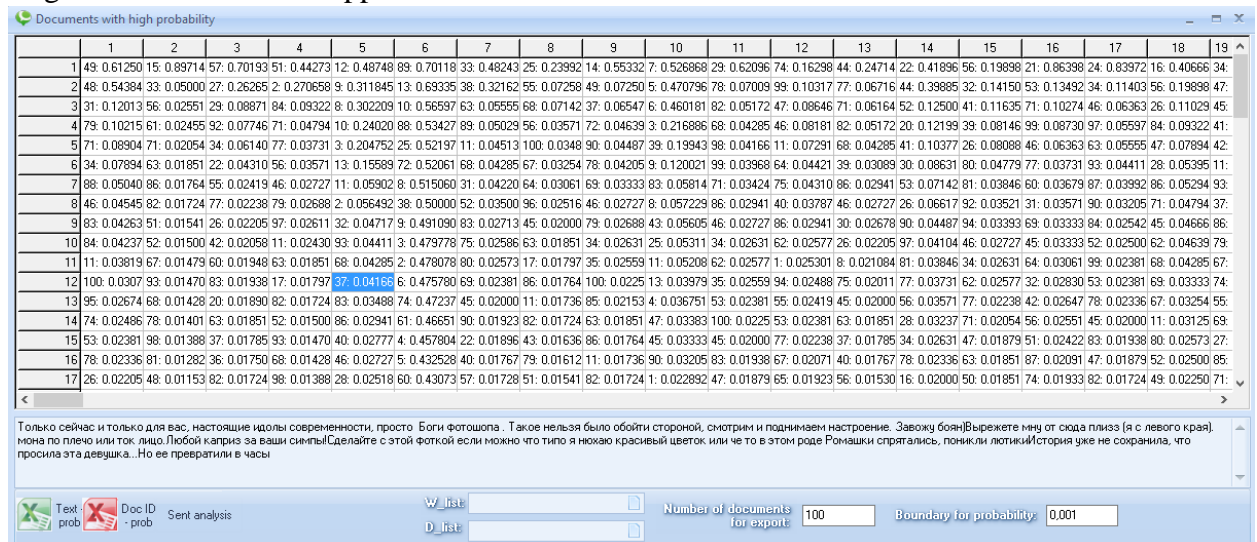


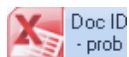
Fig. 3.12. An example of visualization of document distribution by topic.

Unload the sorted results.

In this window, several options are available for uploading the sorted data into a csv file.



- texts of documents and their probabilities are unloaded.




- unload documents id and their probabilities. An example of such unloading is shown in Figure 3.13.

	A	B	C	D	E	F	G	H
1	ID doc(Topic 1)	Prob. of Doc.(Topic 1)	ID doc(Top	Prob. of D	ID doc(Top	Prob. of D	ID doc(Top	Prob. of D
2	94	0,788783	53	0,39734	59	0,106481	86	0,134868
3	65	0,084615	62	0,307404	19	0,031646	31	0,048673
4	32	0,071782	3	0,240798	31	0,022124	32	0,042079
5	21	0,059677	32	0,14604	42	0,021186	57	0,041667
6	38	0,054124	12	0,036842	5	0,02027	39	0,038462
7	58	0,052326	78	0,032468	49	0,01875	78	0,032468
8	44	0,04321	40	0,032297	44	0,018519	12	0,026316
9	36	0,038889	13	0,022152	83	0,018293	67	0,023214
10	79	0,038182	99	0,020833	58	0,017442	65	0,023077
11	67	0,030357	81	0,020147	10	0,017241	42	0,021186
12	99	0,026786	33	0,019481	61	0,016129	48	0,019084
13	84	0,025862	10	0,017241	12	0,015789	49	0,01875

Fig. 3.13. An example of unloading document distribution by topic (id-probability).

Attention, the description of the sentiment analysis is given in Chapter 7.

3.4.2. Visualization of sorted word distributions by topics.

To open a window in which word distributions by topic are presented in order of decreasing probabilities, you need to click on the button . As a result, a window will appear in which the probability sorting is done in such a way that at the top (in each topic) there is a word with the highest probability of belonging to each topic. An example of such sorting is shown in Figure 3.15. This window also makes it possible to unload the sort results into a csv file. The size of the discharge is regulated by two parameters: 1. The number of words for unloading. 2. Boundary in probability.

Number of words for export: Boundary for probability:


The first parameter specifies the maximum number of words to be uploaded. The second parameter defines the boundary. Words with probabilities below the specified boundary will not be unloaded.

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Field1	Field2	Field3	Field4	Field5	Field6	Number words in doc(1)	Probability(1)	Document Nick	Field1	Field2	Field3	Field4	Field5	Field6	
2	yuzilla	http://yuzilla.livejournal.com/922914.html					395	0,788783	53	54	yuzilla	http://yuzilla.livejournal.com/930651.html				
3	alexlotov	http://alexlotov.livejournal.com/363587.html					46	0,084615	62	43	yuzilla	http://yuzilla.livejournal.com/923209.html				
4	yuzilla	http://yuzilla.livejournal.com/927501.html					174	0,071782	3	25	yuzilla	http://yuzilla.livejournal.com/927501.html				
5	alexlotov	http://alexlotov.livejournal.com/364440.html					294	0,059677	32	42	yuzilla	http://yuzilla.livejournal.com/927501.html				
6	alexlotov	http://alexlotov.livejournal.com/359355.html					176	0,054124	12	86	alexlotov	http://alexlotov.livejournal.com/364282.html				
7	alexlotov	http://alexlotov.livejournal.com/360194.html					66	0,052326	78	99	alexlotov	http://alexlotov.livejournal.com/367512.html				
8	alexlotov	http://alexlotov.livejournal.com/361912.html					59	0,04321	40	33	yuzilla	http://yuzilla.livejournal.com/925211.html				
9	alexlotov	http://alexlotov.livejournal.com/365095.html					75	0,038889	13	28	yuzilla	http://yuzilla.livejournal.com/924058.html				
10	alexlotov	http://alexlotov.livejournal.com/357587.html					258	0,038182	99	11	gutta_hor	http://gutta-honey.livejournal.com/302369.html				
11	alexlotov	http://alexlotov.livejournal.com/366522.html					267	0,030357	81	100	alexlotov	http://alexlotov.livejournal.com/367668.html				
12	gutta_hor	http://gutta-honey.livejournal.com/302369.html					148	0,026786	33	53	yuzilla	http://yuzilla.livejournal.com/930544.html				
13	yuzilla	http://yuzilla.livejournal.com/928575.html					35	0,025862	10	63	alexlotov	http://alexlotov.livejournal.com/358294.html				
14	yuzilla	http://yuzilla.livejournal.com/922308.html					311	0,025	63	83	alexlotov	http://alexlotov.livejournal.com/363461.html				
15	yuzilla	http://yuzilla.livejournal.com/925448.html					43	0,02459	36	90	alexlotov	http://alexlotov.livejournal.com/365095.html				
16	alexlotov	http://alexlotov.livejournal.com/363461.html					129	0,02349	11	49	yuzilla	http://yuzilla.livejournal.com/929280.html				
17	yuzilla	http://yuzilla.livejournal.com/931203.html					88	0,023148	46	3	gutta_hor	http://gutta-honey.livejournal.com/299320.html				
18	yuzilla	http://yuzilla.livejournal.com/930817.html					43	0,022727	61	82	alexlotov	http://alexlotov.livejournal.com/363073.html				
19	yuzilla	http://yuzilla.livejournal.com/924058.html					138	0,022152	2	37	yuzilla	http://yuzilla.livejournal.com/926282.html				
20	alexlotov	http://alexlotov.livejournal.com/364963.html					187	0,021845	77	98	alexlotov	http://alexlotov.livejournal.com/367193.html				
21	yuzilla	http://yuzilla.livejournal.com/926122.html					218	0,018519	52	93	alexlotov	http://alexlotov.livejournal.com/365861.html				
22	alexlotov	http://alexlotov.livejournal.com/358294.html					9	0,017241	39	68	alexlotov	http://alexlotov.livejournal.com/359528.html				
23	alexlotov	http://alexlotov.livejournal.com/357885.html					253	0,016544	48	78	alexlotov	http://alexlotov.livejournal.com/362198.html				
24	yuzilla	http://yuzilla.livejournal.com/928978.html					143	0,016447	60	81	alexlotov	http://alexlotov.livejournal.com/362988.html				
25	alexlotov	http://alexlotov.livejournal.com/363073.html					11	0,016129	70	58	yuzilla	http://yuzilla.livejournal.com/931721.html				
26	alexlotov	http://alexlotov.livejournal.com/367193.html					14	0,013889	30	32	yuzilla	http://yuzilla.livejournal.com/925144.html				

Fig. 3.14. An example of unloading document distribution by topic (metadata - the number of words in the document - the probability of the document).

Fig. 3.15. An example of visualization of word distributions by topics.

3.4.2.1. Export sort the results in csv file format.

To unload the results of sorting into a csv file, click the button . In the appeared window it is necessary to specify a file name. An example of such an unloading is shown in Figure 3.16.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		1 prob		2 prob		3 prob		4 prob		5 prob		6 prob	7
2	в	0,020765	кукла	0,020673	р	0,002468	северный	0,004602	на	0,020293	интернет	0,01176	просить
3	февраль	0,017583	семья	0,014194	ji	0,002468	стрип	0,002357	инвалид	0,014633	хостинг	0,009067	желание
4	кабул	0,015992	невеста	0,011249	машин	0,002468	немало	0,002357	февраль	0,013825	вы	0,008169	где
5	afp	0,012014	друг	0,009482	иран	0,002468	американ	0,002357	павлин	0,013016	ваш	0,007272	мост
6	photo	0,010422	женех	0,008893	нестабил	0,001293	также	0,002357	ла-пас	0,011399	отдых	0,005476	место
7	shah	0,009627	свадьба	0,008893	альфред	0,001293	смеяться	0,002357	полицейс	0,011399	сайт	0,005476	из
8	демонстр	0,008036	молодая	0,007715	прислуга	0,001293	два	0,002357	боливия	0,009783	качество	0,004579	башня
9	город	0,008036	девушка	0,007126	посидеть	0,001293	дурак	0,002357	david	0,008974	дорогой	0,004579	улица
10	полицейс	0,006444	родители	0,005949	квинсе	0,001293	жертва	0,002357	mercado	0,008166	по	0,003681	загадыват
11	тысяча	0,006444	младенец	0,00536	динамиче	0,001293	быстрый	0,002357	reuters	0,008166	надежны	0,003681	житель
12	афганский	0,004853	религиозн	0,00536	агонизире	0,001293	rasquini	0,001235	автомоби	0,006549	магазин	0,003681	исполнен
13	демонстр	0,004853	дитя	0,004771	замешате	0,001293	пинобрест	0,001235	фотограф	0,006549	мобильны	0,003681	считаться
14	запад	0,004853	реборн	0,004771	хлопья	0,001293	leon	0,001235	путь	0,00574	машин	0,003681	построит
15	сша	0,004853	ортодокс	0,004771	отличите	0,001293	жанейро	0,001235	набрасыв	0,004932	мегафон	0,002783	удача
16	военный	0,004853	живой	0,004771	мстить	0,001293	jesseглаз	0,001235	костыль	0,004123	страна	0,002783	пора
17	оскорбля	0,004058	женщина	0,004182	уоллес	0,001293	насквозь	0,001235	район	0,004123	турист	0,002783	верона
18	ар	0,004058	зак	0,004182	актерский	0,001293	росиии	0,001235	из	0,003315	любой	0,002783	город
19	километр	0,004058	мальчик	0,003593	определе	0,001293	жужноаме	0,001235	перегора	0,003315	любая	0,002783	один
20	мусульма	0,004058	также	0,003593	победите	0,001293	устремле	0,001235	плаз	0,003315	холод	0,002783	находит
21	нато	0,004058	мама	0,003593	джа	0,001293	источники	0,001235	численно	0,002506	бесплатн	0,002783	колодец

Fig. 3.16. An example of unloading word distributions by topics in the csv format.

3.4.2.2. Visualization of the distribution by weight of the topic.

Attention, this option is temporarily disabled, as the option is supposed to be upgraded. It is supposed to add visualization of distributions by sentiment weight.

Usually, it is important to quickly estimate the sum of weights of all probabilities in a given topic (within the given number of words) and sort all topics by weight. This can be done by clicking on

the button Topic Distribution. As a result, a window will appear in which the sorted distribution of the topics on the scales is visualized. An example of such a distribution is shown in Figure 3.17. The graph also shows the 6 most probable words in each topic.

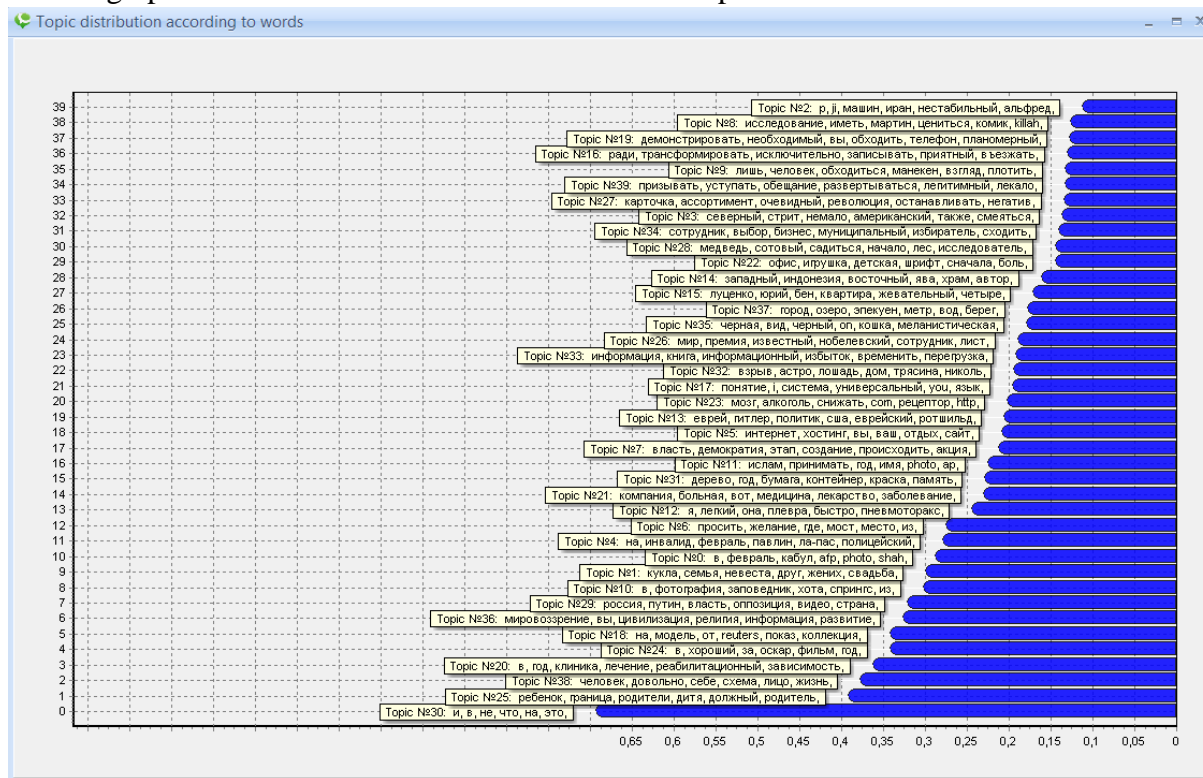



Fig. 3.17. An example of visualizing the distribution of topics by weight of the topic.

3.5. Saving the results of thematic modeling in the form of a project file.


Topic model is based on data downloaded from a file with the extension tmla (e.g., 2_step_test.tmla). As a result of thematic modeling, two matrices are created: 1. The matrix for the distribution of documents by topic (matrix phi). 2. The matrix of the distribution of words by topic (theta matrix). That is, two additional files appear in the directory: 2_step_test_phi.bin and 2_step_test_theta.bin. Thus, it is necessary to always store a combination: the original data plus the simulation results. This can be done by clicking on the button . In the window that appears, you must specify a file name. The program will create a project file (for example, my_test.tproj), which will contain the paths to the source data (tmla) and the results of thematic modeling, that is, the matrices _phi.bin and _theta.bin.

An example of such a file is shown below:

```
<?xml version="1.0" encoding="UTF-8"?>
<TopicMinerProject><LDAFileName>D:\TopicMiner\poligon_RNF\data for
orange\2_step_test_we.tmla</LDAFileName><PhiFileName>D:\TopicMiner\poligon_RNF\da
ta for
orange\2_step_test_we_phi.bin</PhiFileName><ThetaFileName>D:\TopicMiner\poligon_RNF
\data for orange\2_step_test_we_theta.bin</ThetaFileName></TopicMinerProject>
```

This will allow you to download only one project file for later analysis, and the program will automatically load all other files. A project file is a text file that can be easily changed when the project is transferred to another computer or to another directory.

3.6. Loading the results of topic modeling from the project file.

To download previously obtained results of thematic modeling, you need to click on the button . In the appeared window you need to specify the name of the project file. The program will automatically load all the necessary files based on the paths specified in the project file.

Chapter 4. Topic modeling by BigArtm models (multimodal topic modeling).

4.1. Parameter setting in multimodal TM models.

The multimodal version of the topic modeling is based on the BigARTM regularization procedure, which involves metadata, for example, the date of the post or the geotag of the post. In addition, version 88 implements the ability to specify a range of topics, which allows you to calculate the optimal number of topics. Topic modeling based on additive regularization and multimodal schemes are implemented on the 'BigArtm' tab. An example of the 'BigArtm' interface is shown in Figure 4.1.

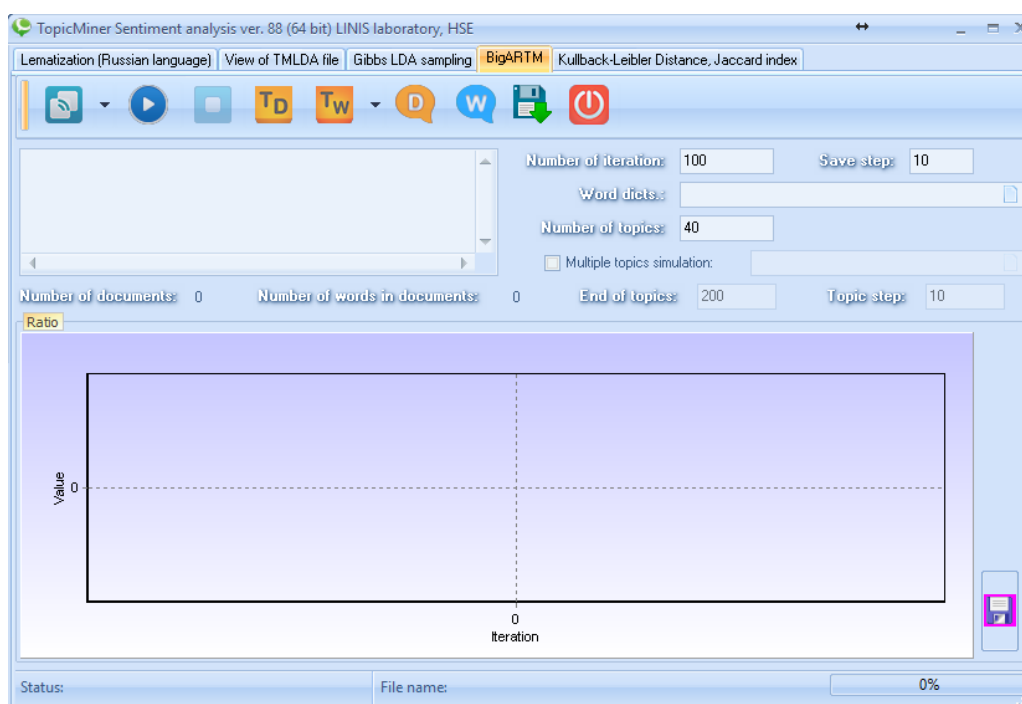


Fig. 4.1. Example of the interface 'BigArtm'.

Models 'BigArtm' are characterized by the following parameters (similar to models based on Gibbs sampling):

1. Number of topics.

Number of topics:

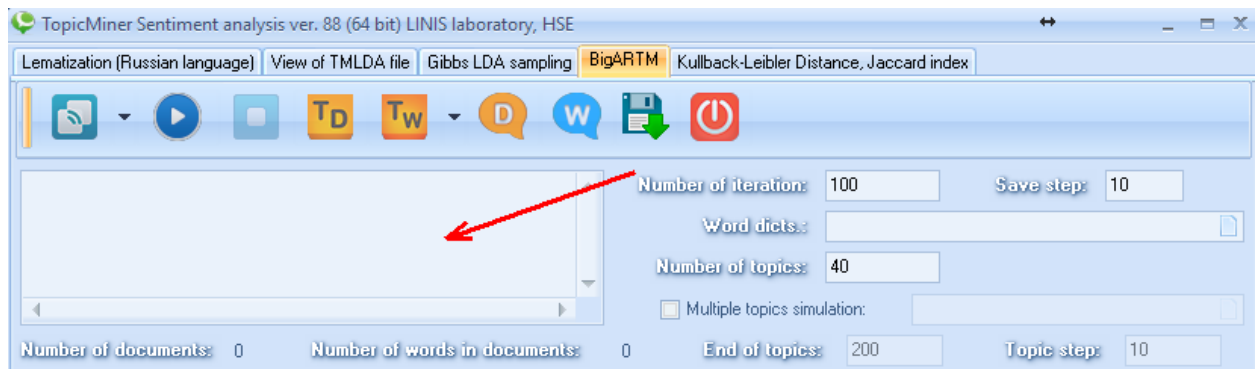
This parameter is used in calculations with a fixed number of topics.

2. Number of iterations.

Number of iteration:

3. **Save step:** . Step: the number of iterations, after which the results of the calculation are visualized.
4. **Word dict:** In this option, a dictionary (in bin format) is specified that contains lists of unique words for the selected metadata fields (see Chapter 2 of the User's Guide).
5. **Multiple topic simulation:** this option enables the ability to set the range of the number of topic
6. **End of topics:** a final number of topics. Attention, the initial number of topics is determined by the parameter 'Number of topics'.
7. **Topic step:** parameter that determines the step by topic.

A significant difference from other models is the way to specify regularizers. Regularizers are set as text in the following window:



In this version of the software, the following options for specifying regularizers are:

1. Model pLSA (do not enter any parameters).
2. Model with very sparse matrix Theta (Td) and dense matrix Phi (Tw).
Example of a regularizer job: `--regularizer "0.2 SparseTheta"`
The value of the regularizer (0.2) can be varied.
3. A model with a very sparse matrix Phi (Tw) and a dense matrix Theta (Td)
Example of the regularizer: `--regularizer "0.5 SparsePhi"`
4. The value of the regularizer (0.5) can be varied.
5. A model in which regularizers are applied to fixed columns.
Example of the regularizer task: `--topics obj: 35, back: 5 --regularizer "0.2 SmoothTheta #back" --regularizer "0.5 SparseTheta #obj" =>` the first 35 columns of the Td matrix are sparse, the remaining five columns are dense.
6. Model of topic decorrelation. Example of a regularizer job: `--regularizer "1000 decorrelation"`. The value of the regularizer (1000) can be changed.

A detailed description of models and regularizers can be found at: <http://bigartm.org/>

Attention, visualization of word_ratio and doc_ratio for BIGARTM is not implemented in this version.

4.2. Visualization of the results of topic modeling.

Visualization of topic modeling consists of the following items:

1. Visualization of the distribution of documents by topic.
2. Visualization of the distribution of words by topic.

3. Visualization of sorted document distributions by topic
4. Visualization of sorted word distributions by topic.




The visualization modules can be launched using the buttons

The action of buttons is similar to the action of buttons in models based on Gibbs sampling.

4.3. Saving the results of topic modeling in the form of a project file.

Topic model is based on data downloaded from a file with the extension tmla (e.g., 2_step_test.tmla). As a result of thematic modeling, two matrices are created: 1. The matrix for the distribution of documents by topic (matrix phi). 2. The matrix of the distribution of words by topic (theta matrix). Thus, two additional files appear in the directory: 2_step_test_phi.bin and 2_step_test_theta.bin. You must always store a combination: the original data plus the simulation

results. This can be done by clicking on the button . In the window that appears, you must specify a file name. The program will create a project file (for example, my_test.tproj), which will contain the paths to the source data (tmla) and the results of thematic modeling, that is, the matrices _phi.bin and _theta.bin.

Attention: download the calculation results for the BigArtm model on the 'Gibbs LDA sampling' tab, because Gibbs sampling models and additive regularization models are similar in structure. In other words, in both cases the results are matrices _phi.bin and _theta.bin.

4.4. Calculation of multimodal variant of TM.

In order to start the calculation of the multimodal thematic model, you first need to download the 'tmla' file for BigARTM and a dictionary containing unique words for the selected metadata. After loading the data, you need to set the parameters and start the calculation. As an example, you can use the data from the 'test_bigartm' directory. An example of calculation is shown in Figure 4.2.

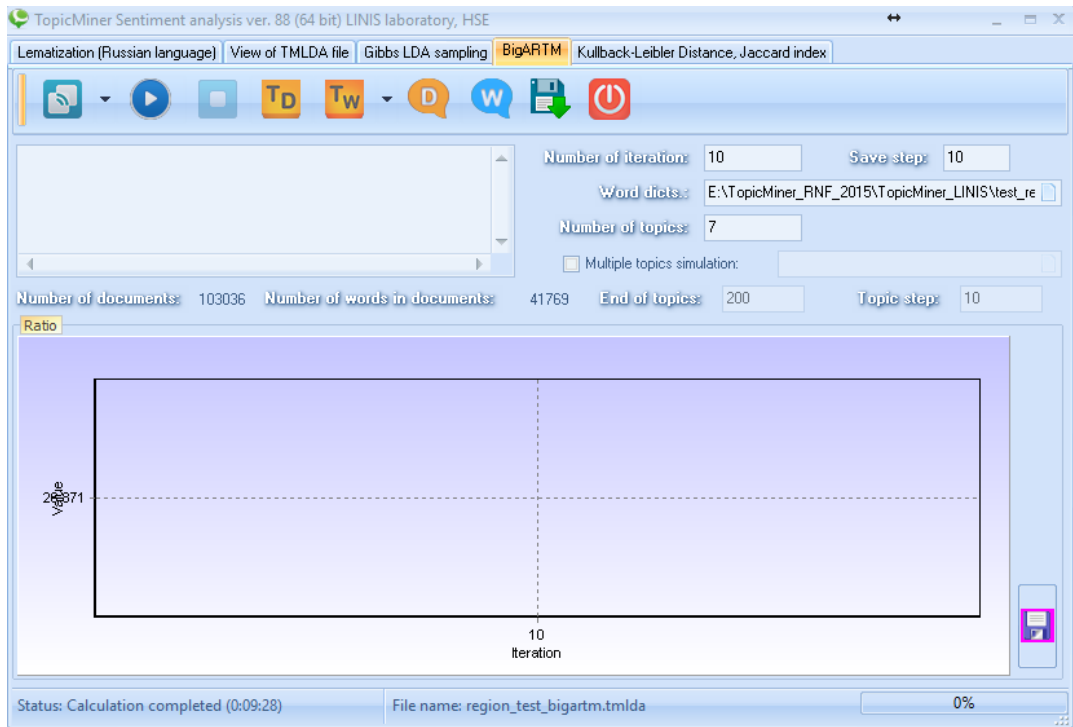


Fig. 4.2. Example of the interface 'BigArtm'.

As a result, multimodal topic model will have additional matrix words on topics of distribution. The matrix data is the distribution of words from the selected fields by topic. In this example, two fields were used: 1. The field 'the author's name of the post', 2. The field 'the geotag of the author of the post'. Figure 4.3 shows an example of the matrix of the author's surnames by topics.

Word	1	2	3	4	5	6	7
1 Ломанова	0.00000	0.00000	0.00053	0.00000	0.05934	0.00000	0.00000
2 Gritsov	0.00000	0.00000	0.00000	0.00015	0.00000	0.00000	0.00000
3 Gorelova	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00005
4 Uskova	0.00000	0.00000	0.00032	0.00000	0.00000	0.00000	0.00000
5 Enoktaev	0.00000	0.00002	0.00000	0.00000	0.00019	0.00000	0.00000
6 Moiseev	0.00071	0.00008	0.00011	0.00007	0.00184	0.00000	0.00063
7 Petrov	0.00048	0.00000	0.00909	0.00125	0.00157	0.00000	0.00020
8 Левинцева	0.00000	0.00000	0.00049	0.00096	0.00980	0.00000	0.00000
9 Novikov	0.00000	0.00000	0.00892	0.00000	0.00229	0.00000	0.00003
10 михайл	0.00028	0.00000	0.00002	0.00006	0.00729	0.00000	0.00000
11 буян	0.00002	0.00000	0.00000	0.00025	0.01072	0.00000	0.00002
12 иван	0.00091	0.00588	0.00001	0.00002	0.01872	0.00001	0.00155
13 Воу	0.00170	0.00004	0.00059	0.00113	0.07370	0.00000	0.00681

Fig. 4.3. An example of visualizing the distribution of names by topics.

An example of geotagging by topics is shown in Figure 4.4.

Word	1	2	3	4	5	6	7
1 бурятия	0.12427	0.00812	0.02494	0.07089	0.63573	0.00001	0.00061
2 Кировский	0.00000	0.00011	0.00000	0.00000	0.00000	0.00000	0.00028
3 область	0.21300	0.47643	0.22329	0.33429	0.00951	0.04825	0.41700
4 Татарстан	0.43156	0.00493	0.48459	0.21358	0.33516	0.84205	0.14492
5 Тверская	0.22207	0.41873	0.17816	0.31097	0.01030	0.03703	0.40958
6 Санкт-Петербург	0.00000	0.00535	0.01193	0.00023	0.00562	0.00009	0.00599
7 Крым	0.00000	0.00015	0.00043	0.00000	0.00000	0.00078	0.00000
8 Архангельский	0.00000	0.00045	0.00072	0.00000	0.00000	0.00041	0.00112
9 Московская	0.00000	0.00699	0.00000	0.00000	0.00000	0.00429	0.00002
10 Ташкентский	0.00000	0.00000	0.00000	0.00000	0.00000	0.00164	0.00000
11 Краснодарский	0.00000	0.00294	0.00002	0.00000	0.00000	0.00001	0.00001
12 край	0.00001	0.00835	0.00000	0.00000	0.00000	0.00740	0.00007
13 Запорожский	0.00000	0.00018	0.00007	0.00000	0.00000	0.00000	0.00065


Рис. 4.4. An example of geotagging by topics.

The obtained data for different authors can be exported in the 'csv' format. To do this, you need to use the "Export to Excel" button.

Глава 5. Stability analysis of simulation results.

When studying the topic structure by different models, and also when analyzing the stability of thematic models, it is necessary to compare thematic solutions with each other. The software has implemented the option of comparing two solutions based on two measures: 1. Kulbak-Leibler measure. 2. The measure of Jacquard. The general view of this option is shown in Figure 5.1.

5.1. Download of topic solutions.

To compare the two solutions, you must first download them. As a solution, we use the unloading of words distribution by topic (see the paragraph '3.4.2. Visualizing word distributions by topic'). To download the first thematic solution, you need to click on the button . The window that appears should contain the file name. An example of the loaded first solution is shown in Figure 5.2.

The result of loading is the matrix, in which the first column contains the word codes in the format crc32, and in the second - words. The subsequent columns contain the probabilities of words belonging to the topics. Download the second solution using the button



. An example of downloading two solutions is shown in Figure 5.3.

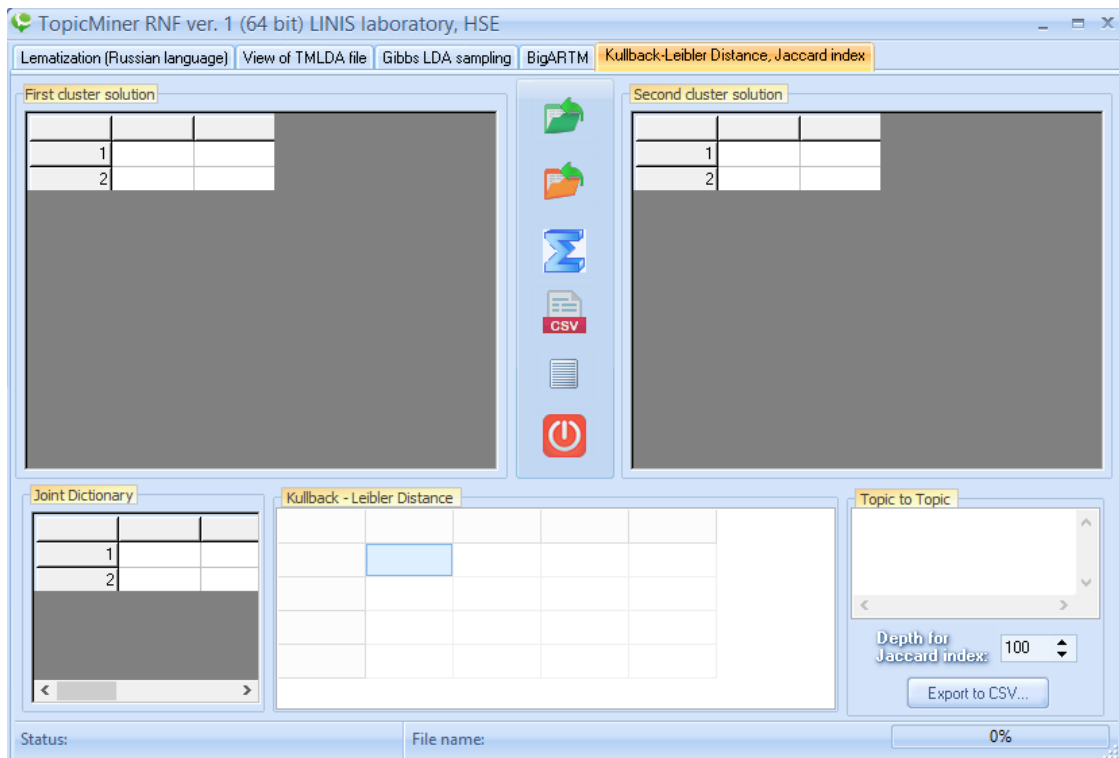


Fig. 5.1. An example of an interface for comparing two topic solutions.

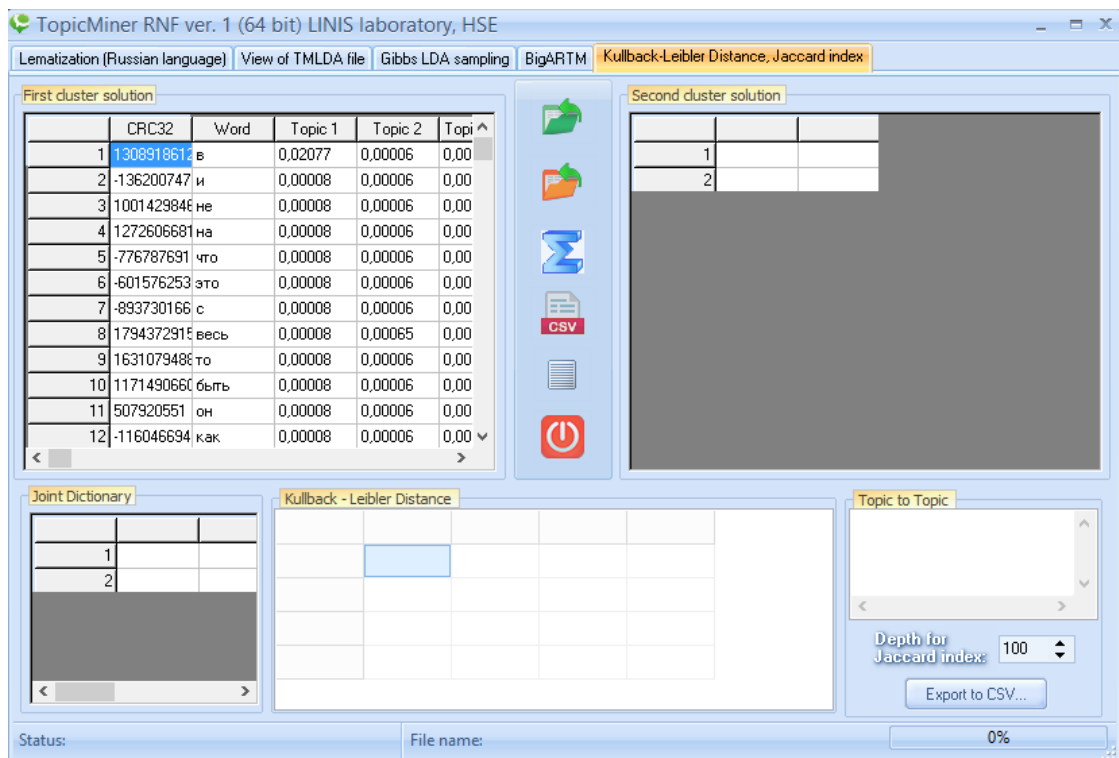


Fig. 5.2. An example of downloading the first topic solution.

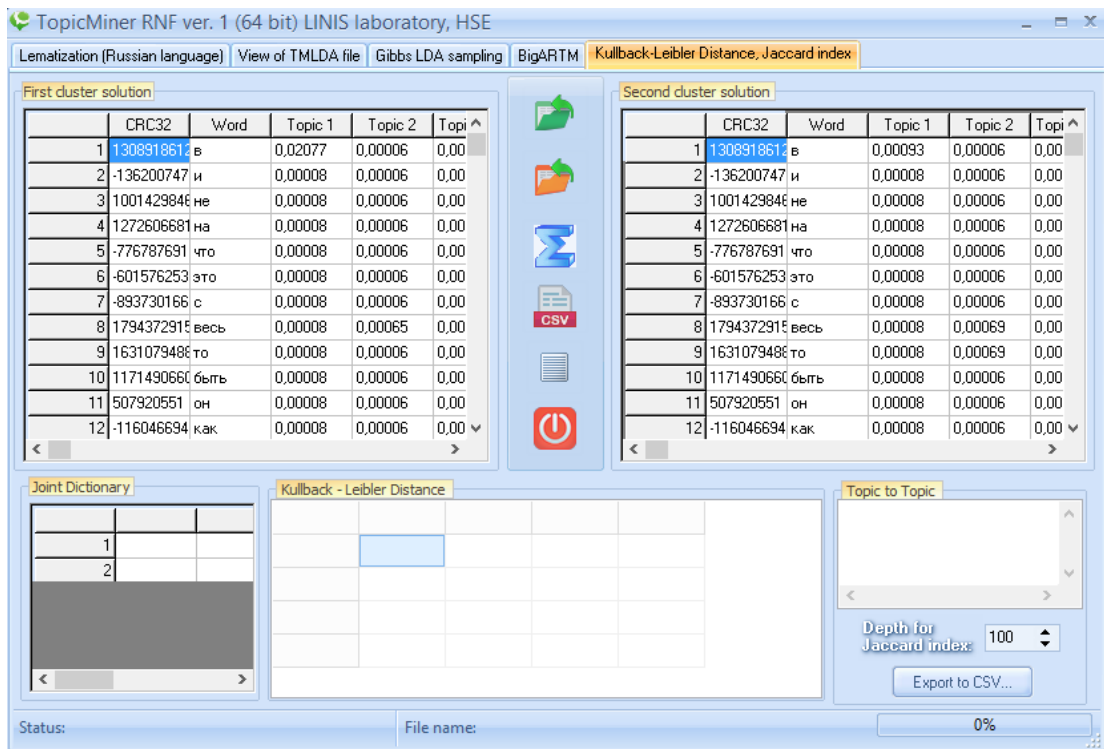



Fig. 5.3. An example of downloading two topic solutions.

5.2. Comparison of topic solutions.

To start the procedure of pairwise comparison (topic1 vs topic2) of two topic solutions, you need to click on the button . After that, a comparison procedure will start, in which each topic from the first solution will be compared with each topic from the second solution. An example is shown in Figure 5.4.

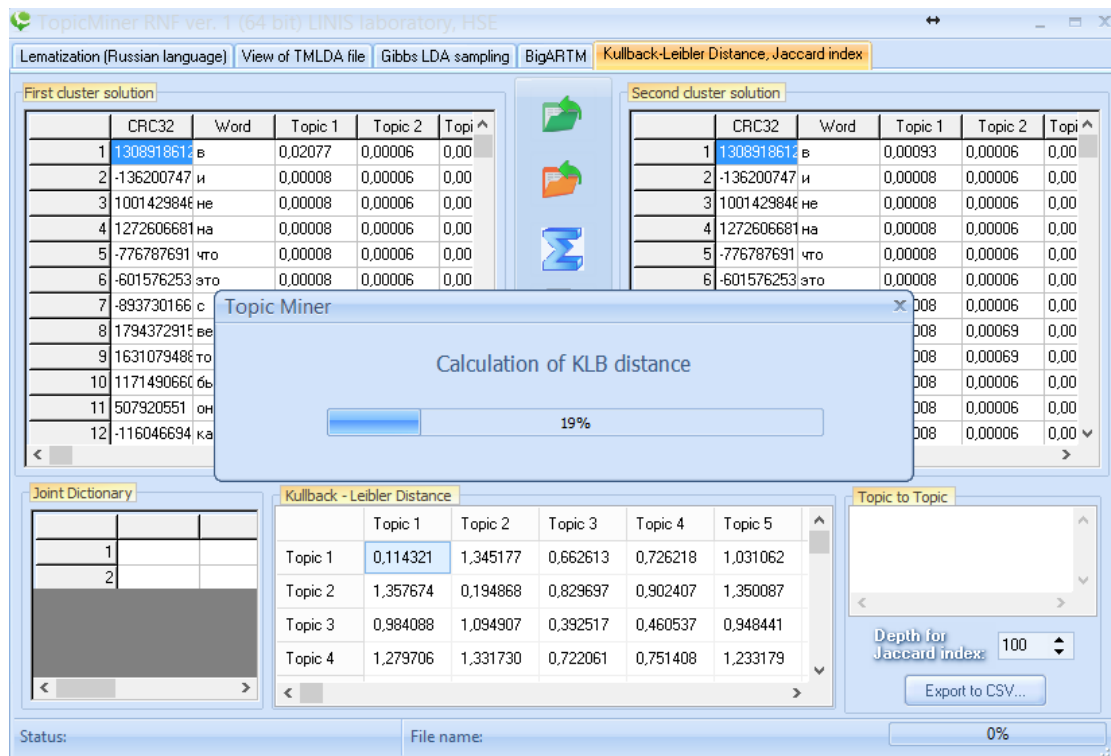


Fig. 5.4. An example of downloading two topic solutions.

As a result, the matrices 'Joint Dictionary', 'Kullback - Leibler distance' will be filled.

The screenshot shows a software interface with three main panels. The 'Joint Dictionary' panel on the left contains a table with columns 'CRC32' and 'Word'. The 'Kullback - Leibler Distance' panel in the center shows a matrix of similarity percentages between seven topics. The 'Topic to Topic' panel on the right displays a list of topic mappings and includes a 'Depth for Jaccard index' dropdown set to 100 and an 'Export to CSV...' button.

'Joint Dictionary' is a list of unique words collected from two topic solutions. 'Kullback - Leibler distance' - matrix, where in each cell is the percentage of similarity between the two topics. 100% corresponds to the maximum similarity.

5.2.1. Matrix 'Kullback - Leibler distance'.

The 'Kullback - Leibler distance' matrix can be downloaded in csv format by pressing the button




. In the appeared window it is necessary to specify a file name. An example of unloading is shown in Figure 5.5.

	A	B	C	D	E	F	G	H	I	J
1		Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
2	Topic 1	95,942	51,802	65,064	54,57	70,552	68,352	54,147	72,448	52,919
3	Topic 2	52,245	93,082	61,13	52,723	66,454	63,804	50,307	69,044	49,678
4	Topic 3	76,477	70,545	86,065	74,366	89,375	86,847	73,731	91,222	72,46
5	Topic 4	74,219	67,964	83,651	73,325	88,417	86,703	71,778	88,533	70,828
6	Topic 5	63,397	52,071	66,33	56,221	73,721	69,768	57,237	73,683	62,038
7	Topic 6	63,414	58,78	73,477	62,524	78,561	76,246	62,459	80,618	61,407
8	Topic 7	55,234	49,725	61,583	51,369	67,449	65,285	54,652	70,081	52,401
9	Topic 8	63,943	58,83	71,638	61,338	78,11	75,811	61,405	80,23	59,857
10	Topic 9	73,901	69,117	81,72	72,245	86,856	85,536	74,147	89,405	70,562

Fig. 5.5. An example of unloading the comparison results for 'Kullback - Leibler distance'.

Comparison of the maximum values of 'Kullback-Leibler distance' on all topics are displayed in the window:


The 'Topic to Topic' window displays a list of topic mappings. The first line shows '38 89,4487772173344 28', indicating that topic 38 of the first solution is 89.44% similar to topic 28 of the second solution. The second line shows '39 68,988183622686 19' and the third line shows '40 95,1161384148887 40'.


In this example, the topic number 38 of the first solution is similar to topic 28 of the second decision, at 89.44%. The results of the mapping are unloaded using the button .

5.2.2. Matching topics from different solutions.

The program can compare (place side by side) the most similar topics from two different thematic decisions, and also calculate the measure of Jacquard. Unlike the 'Kullback-Leibler distance',

which is considered throughout the list of unique words, for Zhakar's measure it is necessary to specify the depth by words, that is, the number of words by which the measure can be calculated.

This depth can be specified in the next option: . A typical value of 100 most probabilistic words. To unload the table of matching similar topics, in the form of a collection

of words, you need to click on the button . In the appeared window it is necessary to specify a file name. An example of such an unloading is shown in Figure 5.6.

	A	B	C	D	E	F	G	H
1	1 - 95,941	1 - 0,4599	2 - 93,082	2 - 0,4706	3 - 91,222	8 - 0,0363	4 - 88,533	8 - 0,0000
2	в	февраль	кукла	кукла	машин	великобр	северный	великобр
3	февраль	кабул	семья	семья	ji	почувствс	немало	почувствс
4	кабул	аfr	невеста	невеста	р	населени	также	населени
5	аfr	photo	друг	друг	иран	сообщест	американ	сообщест
6	photo	shah	жених	жених	сажать	без	жертва	без
7	shah	демонстра	свадьба	свадьба	потребов	конфликт	два	конфликт
8	город	город	молодая	молодая	текущий	отчетливи	быстрый	отчетливи
9	демонстр	тысяча	девушка	девушка	интересо	религия	смеяться	религия
10	полицейс	полицейск	родители	религиоз	род	такать	стрит	такать
11	тысяча	демонстра	религиоз	младенец	сильный	источникт	дурак	источникт
12	сша	афганский	младенец	ортодокс	замешате	действие	создавать	действие
13	демонстр	getty	дитя	живой	сантьяго	погон	источники	погон
14	афганский	нато	реборн	реборн	креативн	празднов	интерпре	празднов
15	военный	ар	ортодокс	сначала	канители	сеть	адекватн	сеть
16	запад	километр	живой	свадебны	отводить	будущее	прекраща	будущее
17	getty	мусульман	женщина	этап	волокно	хранение	лишь	хранение
18	мусульма	авиабаза	зак	встреча	губить	блог	спецслуж	блог
19	нато	баграм	нея	еврей	предназн	закрывает	объезжат	закрывает
20	оскорбля	тагаі	также	будущее	диктатор	логотип	подписые	логотип
21	ар	атаковать	мама	малыш	альфред	купиться	sařina	купиться
22	километр	сша	мальчик	родители	тема	интуиция	ужасный	интуиция
23	провинци	провинция	свадебны	договор	лет	ложиться	гои	ложиться
24	баграм	военный	этап	знакомств	отличите	пингвин	туризм	пингвин

Fig. 5.6. An example of unloading the comparison results for the 'Kullback-Leibler distance' and the measure of Jacquard and the unloading by words.

Let's see what this example shows. In the first pair of columns, 'A' and 'B', two topics are presented from two different solutions, which turned out to be the most similar. In this case, this is the theme №1 of the first decision and the topic № 1 of the second decision; the coincidence of their numbers is random. This is indicated in the headers of the two columns. In the heading of the column 'A' the value 'Kullback - Leibler distance' is given, in this case it is 95.941%, and in the header of the column 'B' the measure of Jaccard is given, and in this case it is 0.4599. In cells of columns 'A' and 'B' the most probabilistic words in these two topics are given. In the following pairs of

columns, for example 'C' and 'D', the following pair of the most similar themes are given. The number of column pairs is equal to the number of topics in the solution.

Глава 6. Visualization of the results of topic modeling on the map of the Russian Federation.

6.1. Calculation of the distribution of documents by regions.

Visualization of the results of thematic modeling is realized with the help of the free map system Quantum GIS (download the map system at: <http://www.qgis.org/ru/site/forusers/download.html>).

Attention: in this project the regions "Crimea" and "Sevastopol" are not yet represented. These regions will be added in the next version. Part of this project is a file with the dfb extension, which contains a list of regions in the cartographic project and a column 'Topic', which is automatically populated in the 'TopicMiner' program. The cartographic project is in the directory 'RNF_RF_visualisation'. The main project file is 'full_project.ggs'.

Before you visualize the results of topic modeling in the cartographic system, you need to calculate the amount of probabilities for a given topic across all regions. To do this, you need to download a project in TopicMiner with the done topic modeling or conduct topic modeling. For example, open the finished project from the 'Vk_data_example' directory.

ID	Orig text	Nick	Field 1	Field 2	Field 3	Field 4	Field 5	Field 6	Field 7	Field 8	Field 9	Field 10	Field 11	Field 12	Field 13
1															
2	люблю	244159479	Wall 244159	post 2	22.02.2014	Олеся	Малодзінэ	Улан-Удэ	Бурятия						
3	С днем рож	15538973	Wall 124987	post 33	4.1.2014 9:	Alsu	Asarova	Набережны	Татарстан						
4	Новый год?	150682985	Wall 124987	post 27	12.31.2013	Lyudmila	Trofimova	Набережны	Татарстан						
5	?? Ван откр	137183488	Wall 124987	post 36	4.20.2014 1	Lyuda	Gorelova	Киров	Кировская						
6	?? Ван откр	137183488	Wall 124987	post 35	4.1.2014 6:	Lyuda	Gorelova	Киров	Кировская						
7	лишь 2% лн	56556171	Wall 565561	post 7348	08.05.2013	Виктория	Ломанова	Улан-Удэ	Бурятия						
8	что или ктс	0	Wall 124987	comments of	7.22.2013 8	Vladislav	Enoktaev	Набережны	Татарстан						
9	Отправлен:	237671145	Wall 124987	post 34	4.1.2014 2:	Ruslan	Enoktaev	Москва							
10	?Отправлен	97170510	Wall 124987	post 8	5.8.2013 11	Ruslan	Enoktaev								
11	С НАСТУПА	22306292	Wall 124987	post 28	12.31.2013	Vera	Yasnikovska	Набережны	Татарстан						
12	С НАСТУПА	63412409	Wall 124987	post 26	12.30.2013	Irinochka	Aldemirova	Уржум	Кировская						
13	не приказы	124987410	Wall 124987	post 22	7.25.2013 2	Vladislav	Enoktaev	Набережны	Татарстан						
14	?Отправлен:	137183488	Wall 124987	post 31	2.14.2014 8	Lyuda	Gorelova	Киров	Кировская						
15	Братишка п	200339270	Wall 124987	post 2	3.18.2013 9	Alexander	Makarov	Тюмень	Тюменская						
16	?Отправлен:	97170510	Wall 124987	post 7	5.5.2013 1:	Ruslan	Enoktaev								
17	мы жден др	56556171	Wall 565561	post 7345	07.05.2013	Виктория	Ломанова	Улан-Удэ	Бурятия						
18	?Отправлен:	97170510	Wall 124987	post 6	5.4.2013 10	Ruslan	Enoktaev								
19	Лови позит	171916464	Wall 173311	post 4	5.18.2013 1	Azalia	Giniatullina	Бавлы	Татарстан						

Fig. 6.1. An example of the visualization of the distribution of documents by topic with metadata.

After downloading the project, you need to click on the button . In the window that appears (see Figure 6.1), you need to click on the button . As a result, the following window will appear (see Figure 6.2).

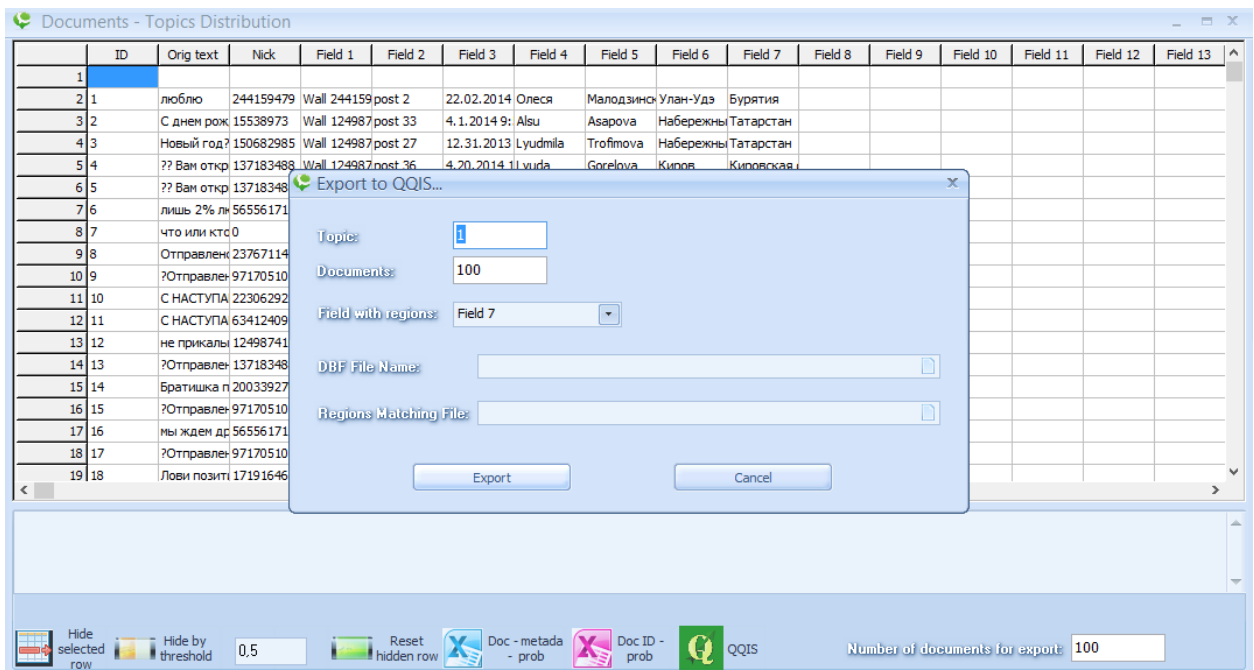


Fig. 6.2. Example of data export to the Quantum GIS map system.

To calculate a topic by region, you need to set the following parameters:

1. **'Topic'**. Number of the topic. For example, set the number of topic number 1, as shown in Figure 6.2
2. **'Documents'**. Number of documents whose geotags will be used in the calculation. For example, specify 100 documents, as shown in Figure 6.2. The program will select 100 most likely documents for a given topic and calculate for each region the sum of the probabilities of all documents belonging to the given region. Belonging to the region is determined by geotagging its author.
3. **'Field with regions'**. In this option, you need to specify the column number in which the names of regions will be located. For example, in the test collection from VKontakte, the names of regions are in column No. 7 (see Figure 6.2)
4. **'DBF file name'**. In this option, you must specify the file name from the map project. For example, the file 'regions2010_sib_5.dbf'. This file contains the names of the regions chosen for visualization, and the corresponding sums of probabilities for the selected topic. In this column, each region is assigned a color according to the severity of the selected topic in the region. This color Quantum GIS paints this region on the map of the Russian Federation.
5. **'Regions Matching File'**. Since the names of regions in the cartographic project and in the metadata from different social networks can differ, it is necessary to create a file that maps these names. In this option, you must specify the name of this file. **Attention: in this version of the monitoring system, a file is created in which the names of the regions from the cartographic project are compared with the names from the social network VKontakte. The name of this file is: vktm.dbf.**

After all the fields are filled (see Figure 6.3), you need to click on the 'Export'.

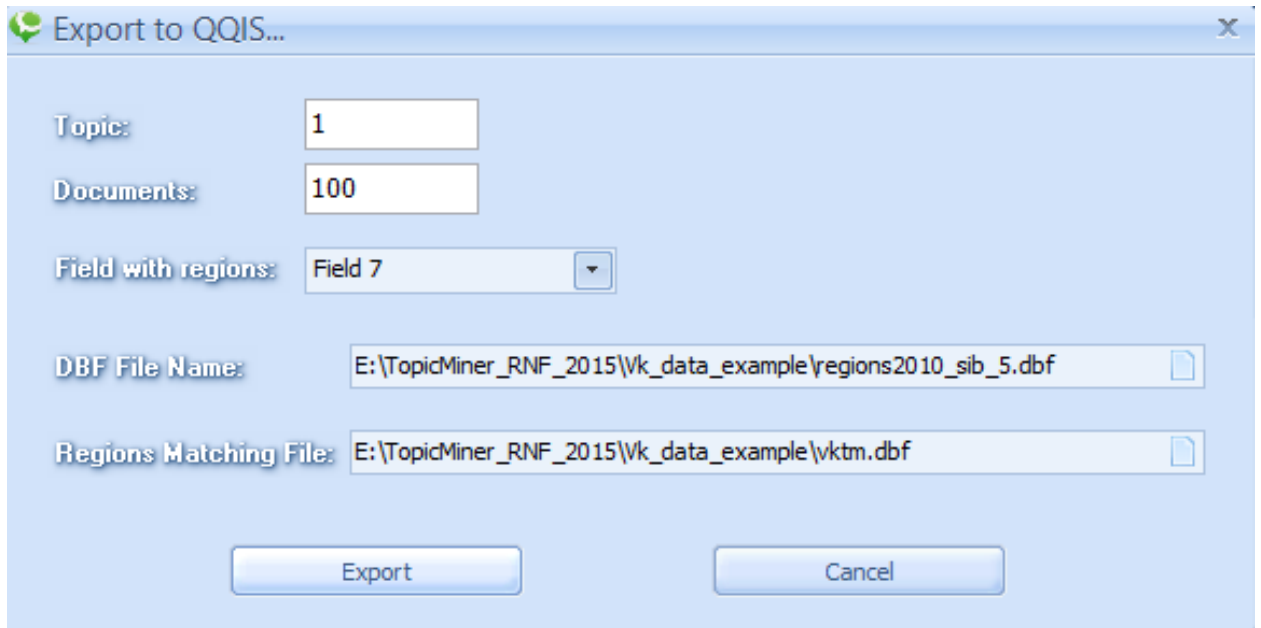
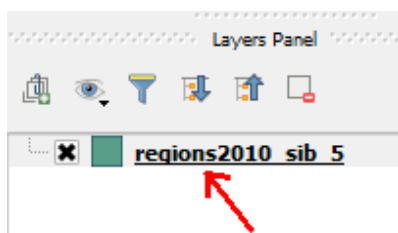


Fig. 6.3. Example of data export to the Quantum GIS map system.

During the calculation, the program will perform the following actions. 1. Define the list of regions that are present in the specified number of sorted documents (based on the mapping file). 2. Calculates the sum of the probabilities of documents for each region. 3. Save the calculated sums of probabilities in the file 'regions2010_sib_5.dbf' (see Figure 6.3).

6.2. Visualization of document distribution in Quantum GIS.

The ready project with a set of maps of the regions of the Russian Federation is located in the directory 'RNF_RF_visualisation'. To visualize the received data, you need to copy the 'regions2010_sib_5.dbf' file to this directory, that is, replace the old file with the same name with the new file. After that, click (twice) on the file 'full_project.qgs'. Attention: the 'Quantum GIS' map system must already be installed. As a result, the 'Quantum GIS' will start and the project 'full_project.qgs' will load (see Figure 6.4). First, all regions will be highlighted in one color. To color the regions in colors according to the sum of the probabilities, you need to change the drawing style. To change the style, double click on the project name, as shown by the red arrow in the picture below.:



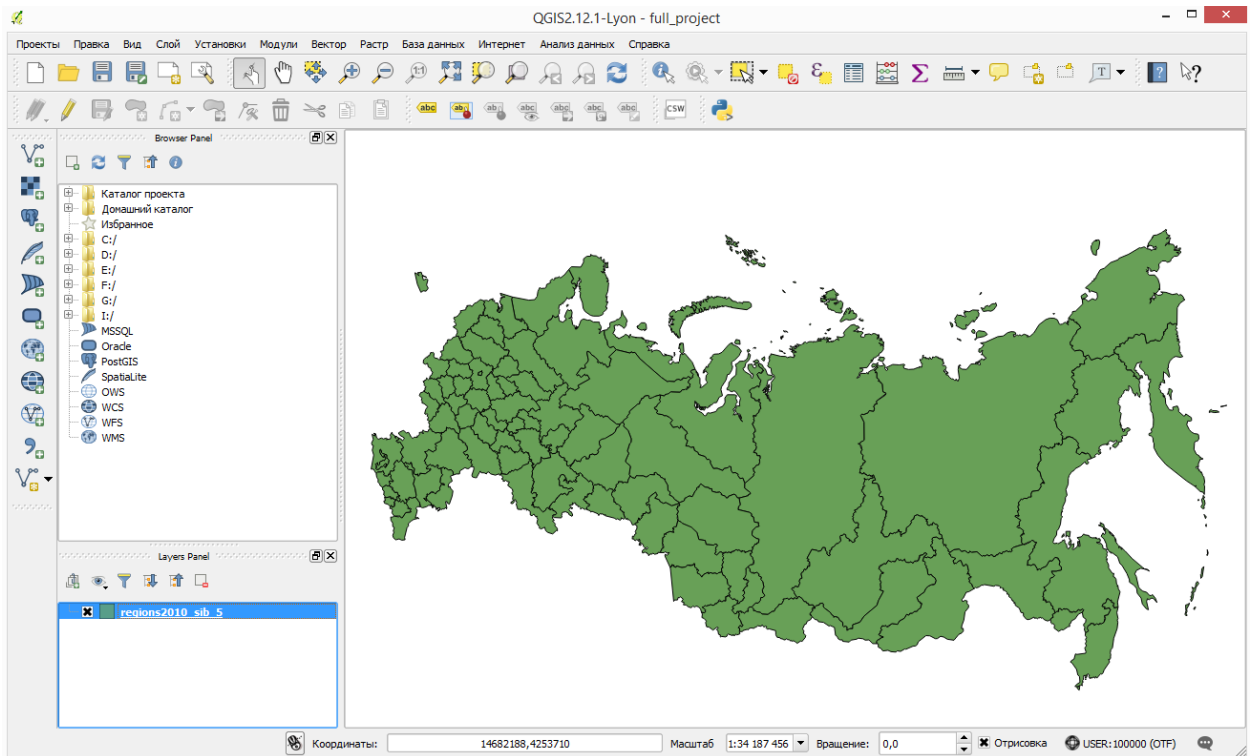


Fig. 6.4. Example of data export to the Quantum GIS map system.

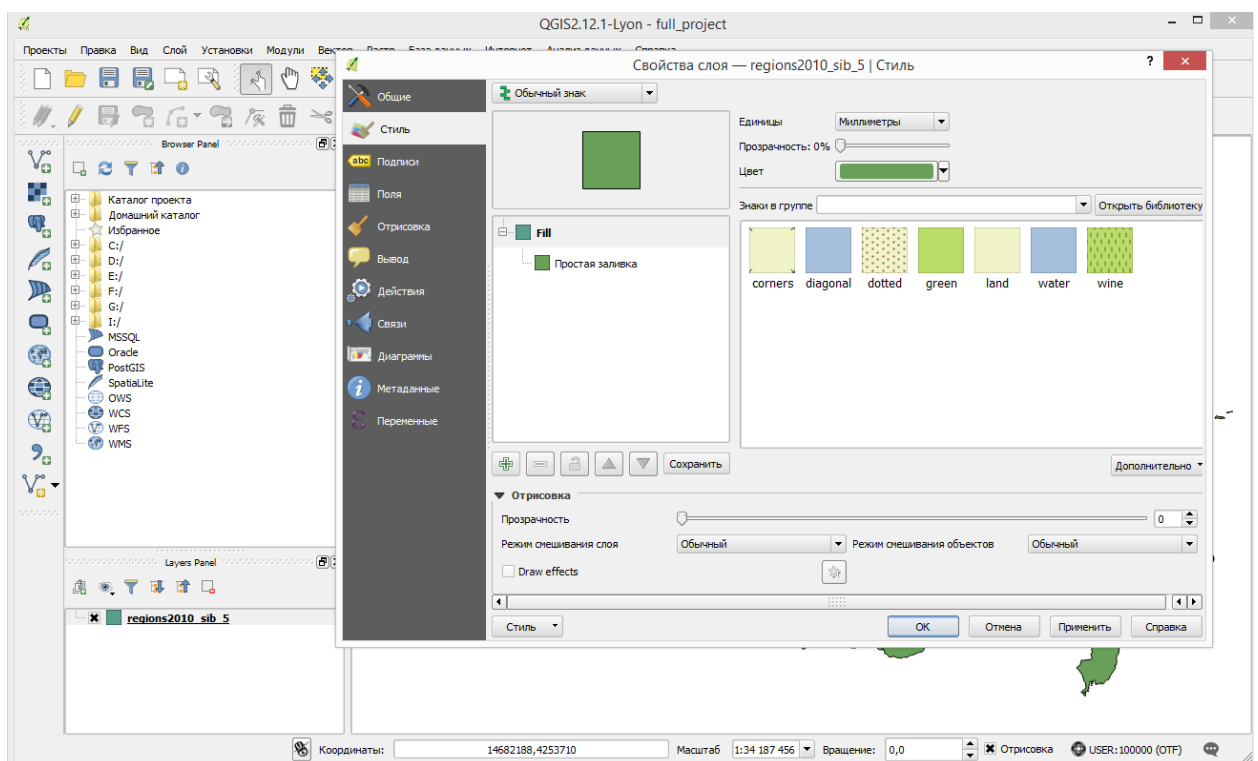


Fig. 6.5. Example of changing the style in Quantum GIS.

As a result, a window opens in which you can change the drawing style (see Figure 6.5). To do this, select the "Unique values" in the drop-down menu where the "Normal character" is by default, as shown in Figure 6.6.

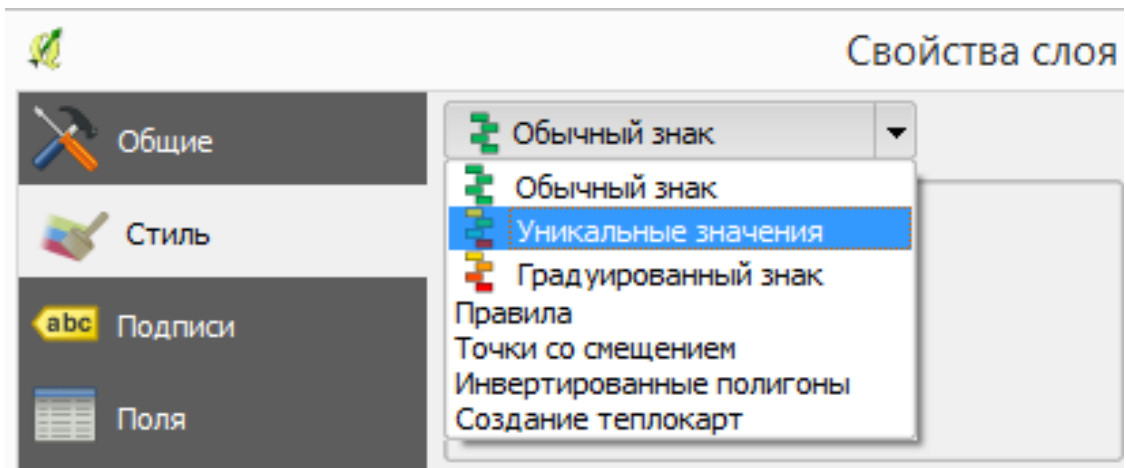


Fig. 6.6. Example of changing the style in Quantum GIS.

Then select the field on which you want to calculate the unique values. In our case, this is the 'Topic' field, which contains the sums of probabilities for each region. This data is taken from the file 'regions2010_sib_5.dbf'. After that, click on the 'classify' button (see Figure 6.7).

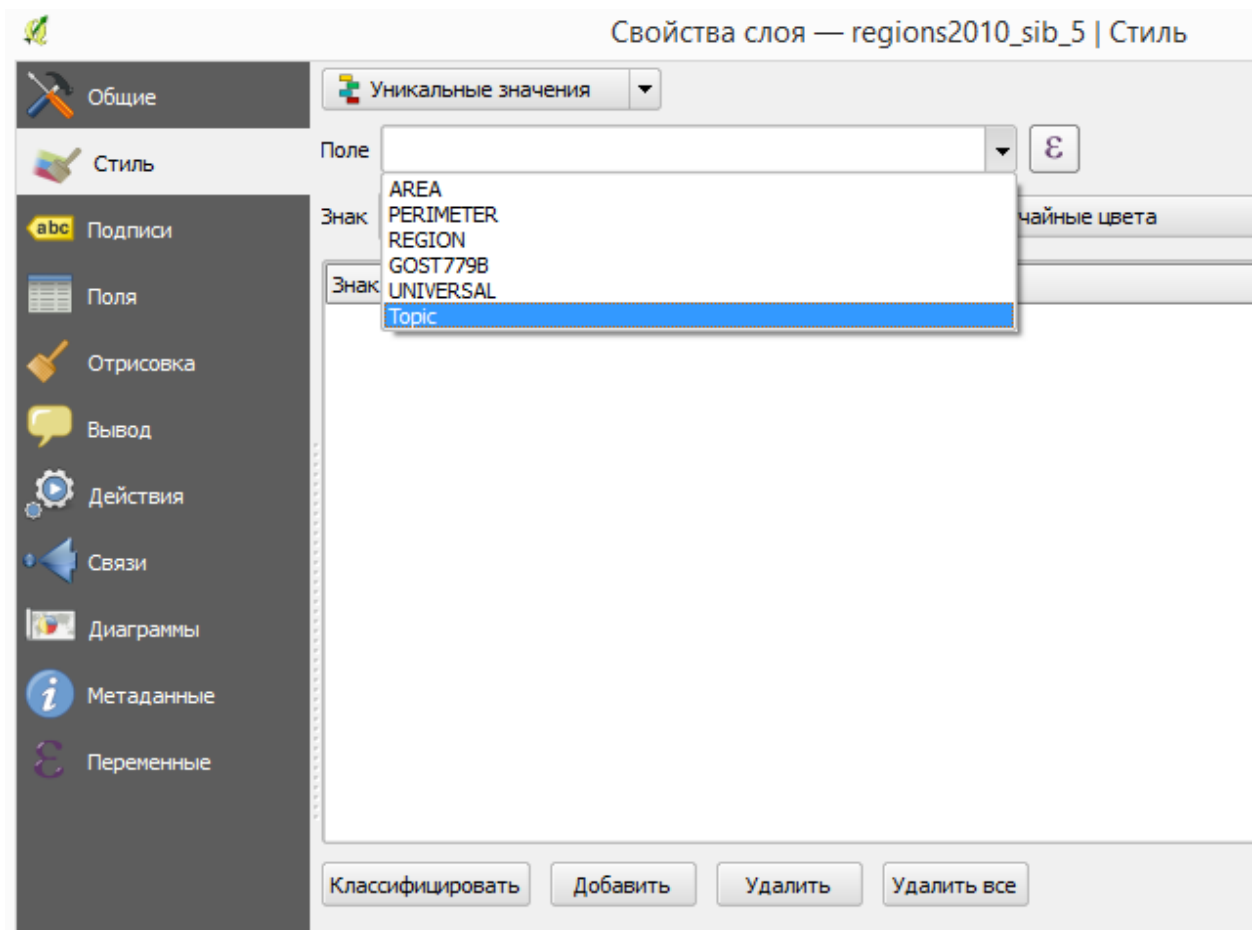


Fig. 6.7. Example of changing the style in Quantum GIS.

As a result of the Quantum classification, GIS will determine all unique values (see Figure 6.8 for an example). Now you need to specify the type of coloring for the found values. This can be done in the 'Gradient' option (see Figure 6.9).

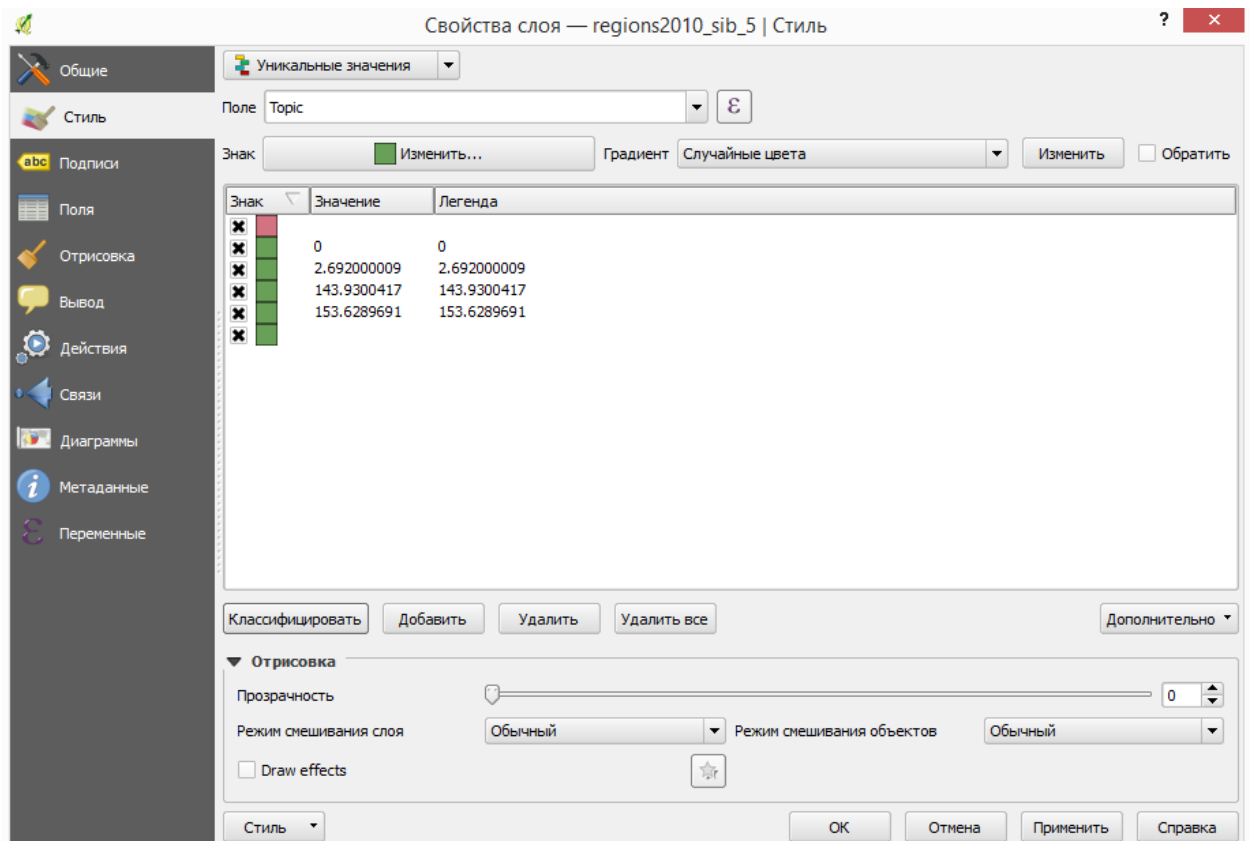


Fig. 6.8. Example of changing the style in Quantum GIS.

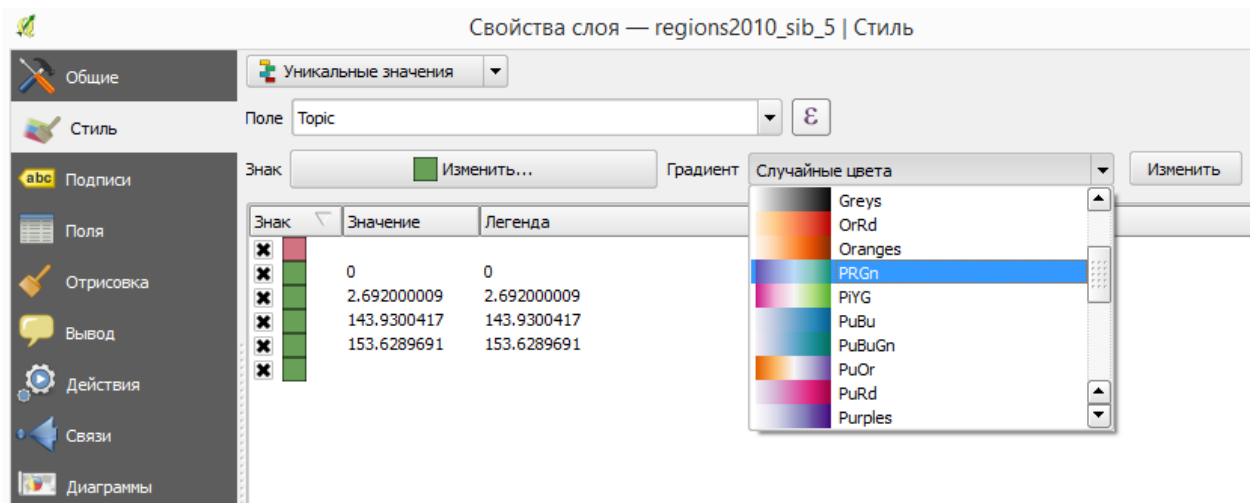


Fig. 6.9. Example of changing the style in Quantum GIS.

To apply the color gamut, you need to click on the 'Apply' button. The result for the three unique values found (that is, for the three regions found) is shown in Figure 6.10. Note: only those regions are highlighted for which documents with high probabilities for the selected topic were found in the data. Figure 6.11 shows an example of visualizing a topic by region based on 222546 documents in topic № 1..

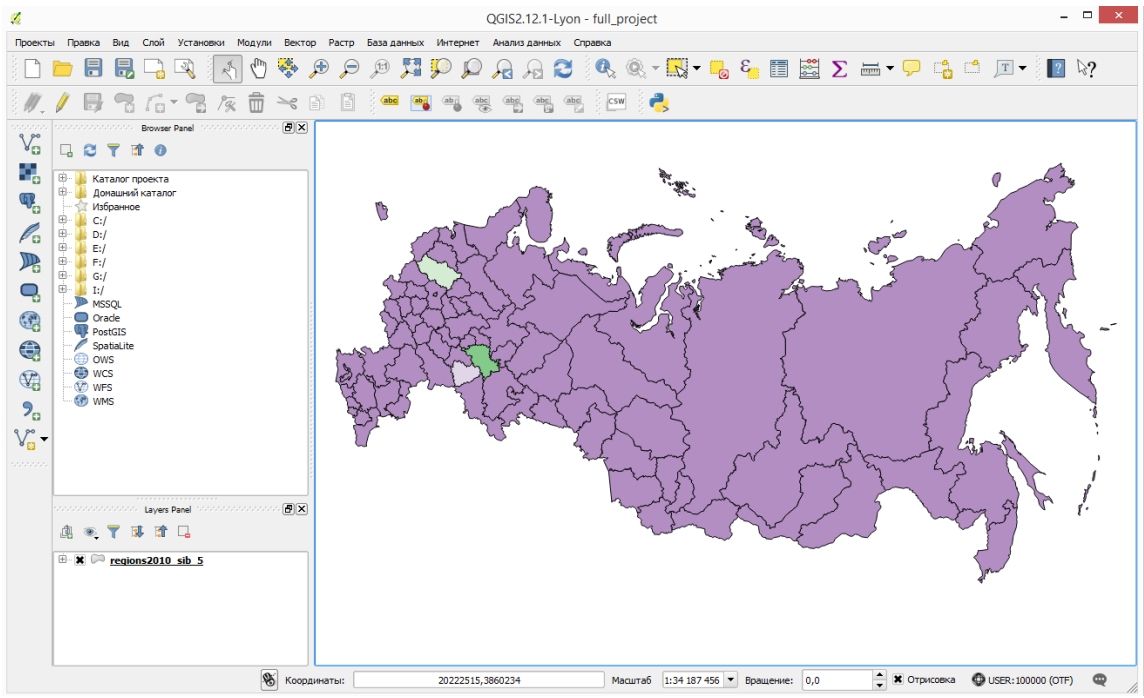


Fig. 6.10. Example of visualizing a topic in Quantum GIS in three regions.

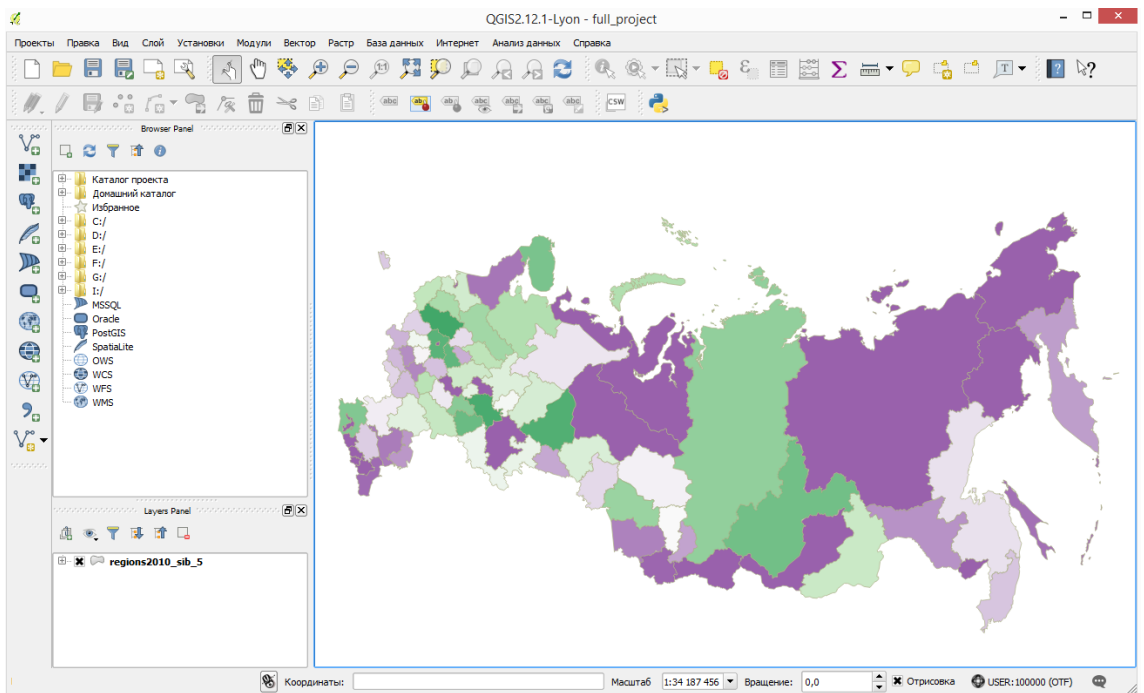


Fig. 6.11. Example of visualizing a topic in Quantum GIS for a set of regions.

Chapter 7. Analysis of the tonality of texts.

7.1. Introduction.

Tonal analysis (Sentiment analysis) or automated analysis of the emotional coloring of texts (bad / good, like / dislike, etc.) can be attributed to the field of computer linguistics, however, the tasks of its application, basically, lie outside the actual linguistics. They can be divided into two broad areas: marketing (primarily as an analysis of feedback on goods and services) and sociology / political science. The latter includes, first, an analysis of the media texts to identify how certain socially important questions are presented to the audience and, accordingly, what response can be expected from them to the public. Secondly, this study of the texts of blogs, social networks, forums and other user content in order to identify public opinion - more precisely, the views of the Internet-active part of the population. Within the framework of this software, a dictionary-based approach is implemented. As the initial dictionary, a set of words was used, which was obtained in the framework of the RGNF project 'Development of a public database and crowdsourcing web resource for creating sentimental analysis tools'. Application number: 14-04-12031. However, this software implements a common technology, the connection of any dictionary, which includes ethnicity, ethnophilism and so on.

7.2. Preparing the vocabulary for sentiment analysis.

Due to the fact that the result of thematic modeling is the matrix of the distribution of the lemmatized words on the themes, accordingly the sentiment evaluation should be conducted on the basis of the lemmatized dictionary. Preparing the dictionary is as follows. The 'Lematization' tab provides the following option (see Figure 7.1)

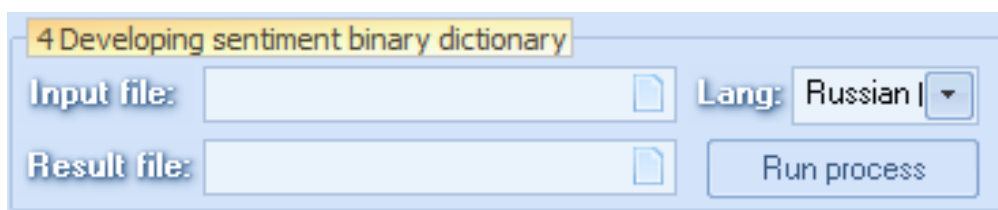


Fig. 7.1. Option to prepare a tonal dictionary.

The input file must be a file of the following type:

Words; average rate

alcoholic; -2

apathetic -2

distressing; -2

mediocrity; -2

ruthlessly; -2

mindless; -2

unpunished; -2

ugly; -2

unanswered; -2

An example of such a dictionary is included in the current version. For the list of words, you must specify the language (Russian, English) and the encoding type (UTF, Ansi).

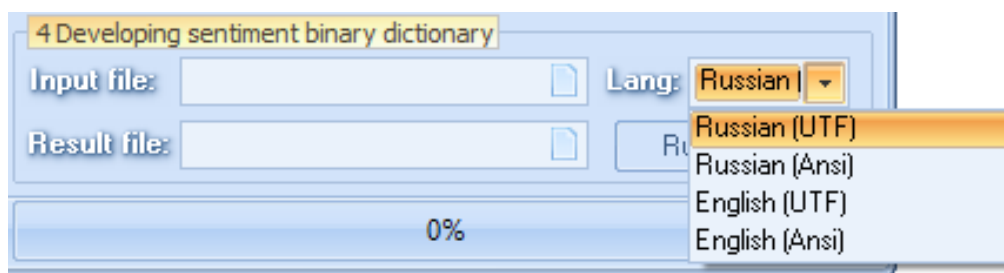



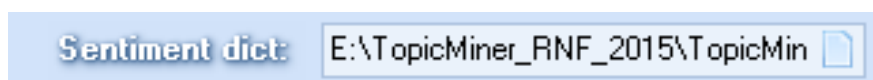
Fig. 7.2. Option to prepare a tonal dictionary.

The language and encoding options are shown in Fig. 7.2.

The result of preprocessing is a binary file containing words and estimates in a binary format. Keeping a dictionary in binary form allows you to significantly speed up the calculation of the tonality of the distribution of words and documents on topics. To convert a dictionary from text to binary format, you need to click on the button .


7.2. Connecting the dictionary to the topic model.


Calculation of tonal estimates is made for a ready-made thematic solution. The connection option is located on the 'Gibbs LDA sampling' tab. The connection consists in specifying the path and name of the dictionary in binary format, see the following figure:



7.3. Tonal calculation of the distribution of words by topics.

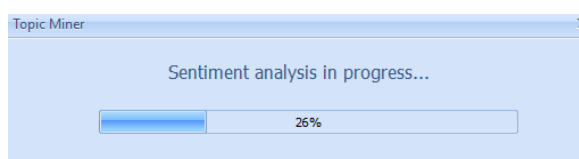
Calculation of tonal estimates is made for a ready-made thematic solution. This means that either the thematic modeling should be done, or the project made earlier should be loaded into the program. Calculation of tonal estimates for the distribution of words by topic is implemented in the window 'words

with high probability'. In order for this window to appear, you need to click on the button . As a result, the following window will appear (see Figure 7.3). In this window, the matrix of the distribution of words by themes is visualized. At the moment, each cell contains the word and the probability of the word

in the subject. In order to add tonal estimates, you need to click on the button . As a result, the tonality calculation starts. However, in the case of an unconnected dictionary, then a warning window will appear (see below):



If the dictionary is connected, then the calculation process is shown (see below).



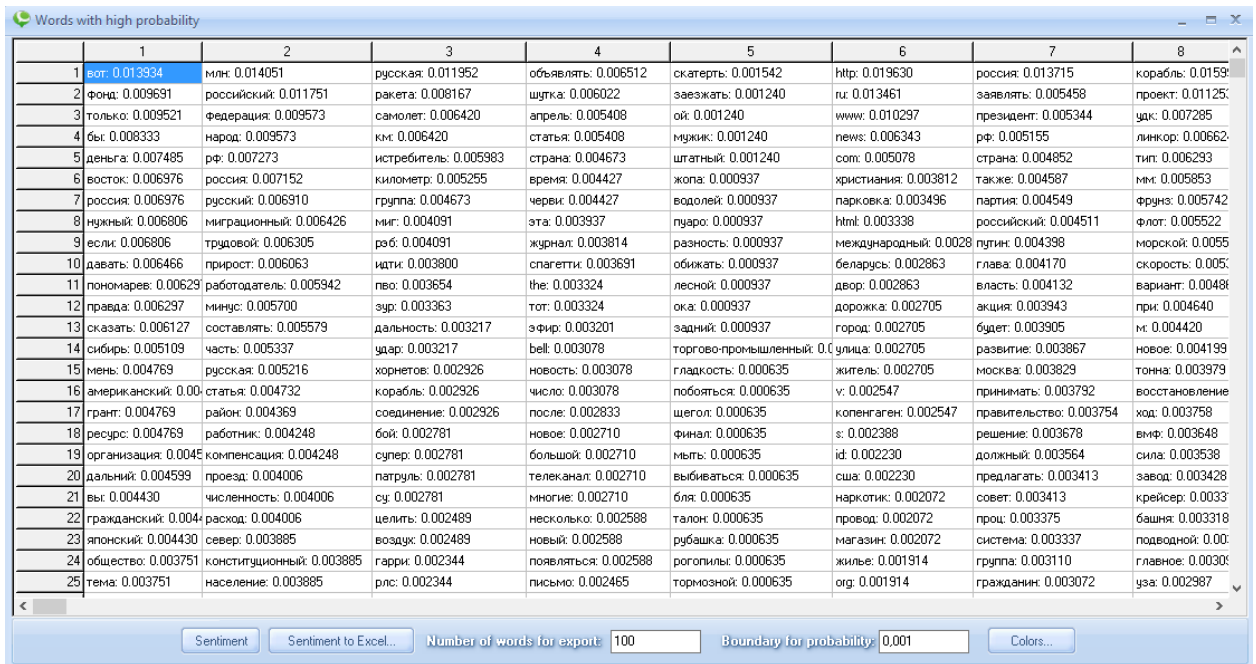


Fig. 7.3. Matrix of the distribution of words by topics before calculating the tonality.

As a result of calculation, in each cell, in addition to the probability, there will be an integer that characterizes the tonality of the word (see Figure 7.4)..

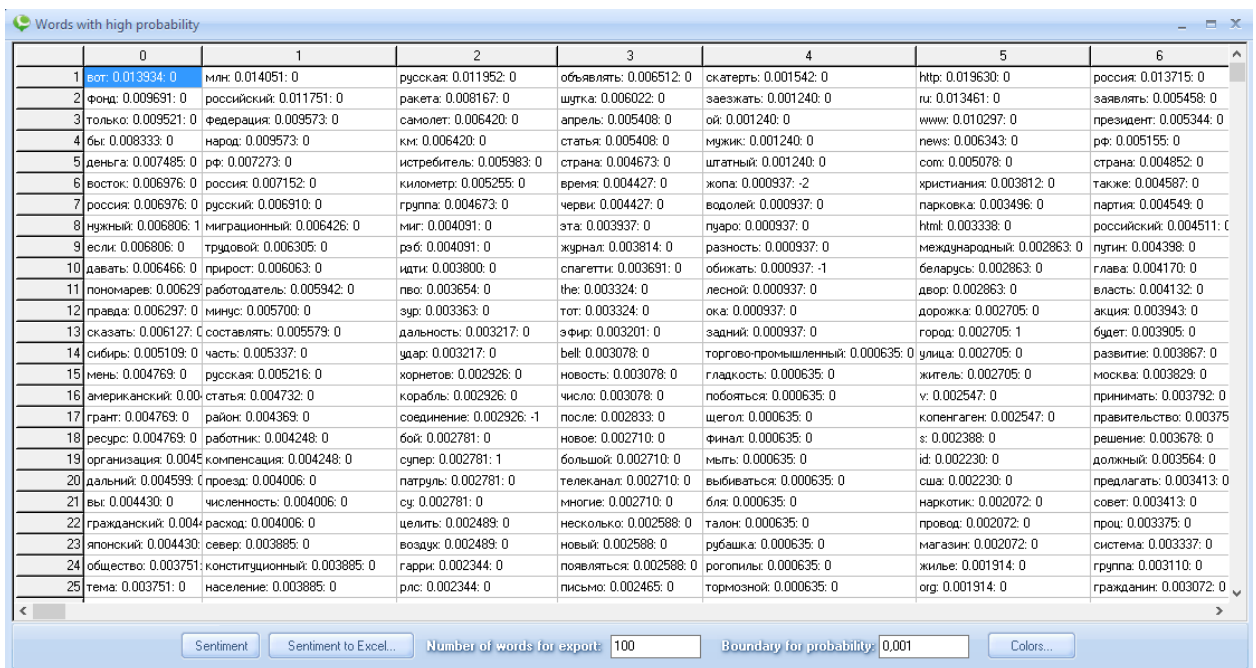


Fig. 7.4. The word distribution matrix for topics and sentiment analysis.

Attention, the tonality calculation is made for a fixed number of most probabilistic words. The number of words in each topic for which the calculation is made is determined in the next option:

Number of words for export:

. The default is 100 words.

7.3.1. Unload matrix of words - topics with tonal estimates.

The unloading of the matrix of word distributions by themes together with tonal estimates is performed by pressing the button (you must specify a file name). As a

result, all topics are unloaded. The depth of word unloading is determined by the parameters **Number of words for export:** and **Boundary for probability:** . As a result of the upload, a file with the given name appears and contains besides the probabilities, also tonal estimates of the words.

7.3.2. Prompt of topics.

In real calculations, the number of topics can range from several tens to several hundred. The search for necessary topics can take a long time. In order to optimize the work of the information system user, color highlighting of the necessary topics is realized. The highlight of the theme is implemented as follows. The user must click on the button **Colors...**. The user specifies the file, which contains a list of words on which to search for calculated topics. After that, the program reads the file, and calculates how many words from the list are present in among the top words in each topic. All numbers are divided into 4 categories, topics with maximum numbers (the maximum number of words from the list) is colored in red. Topics with minimal numbers are tinted green. An example of such a hint is shown in Figure 7.5.

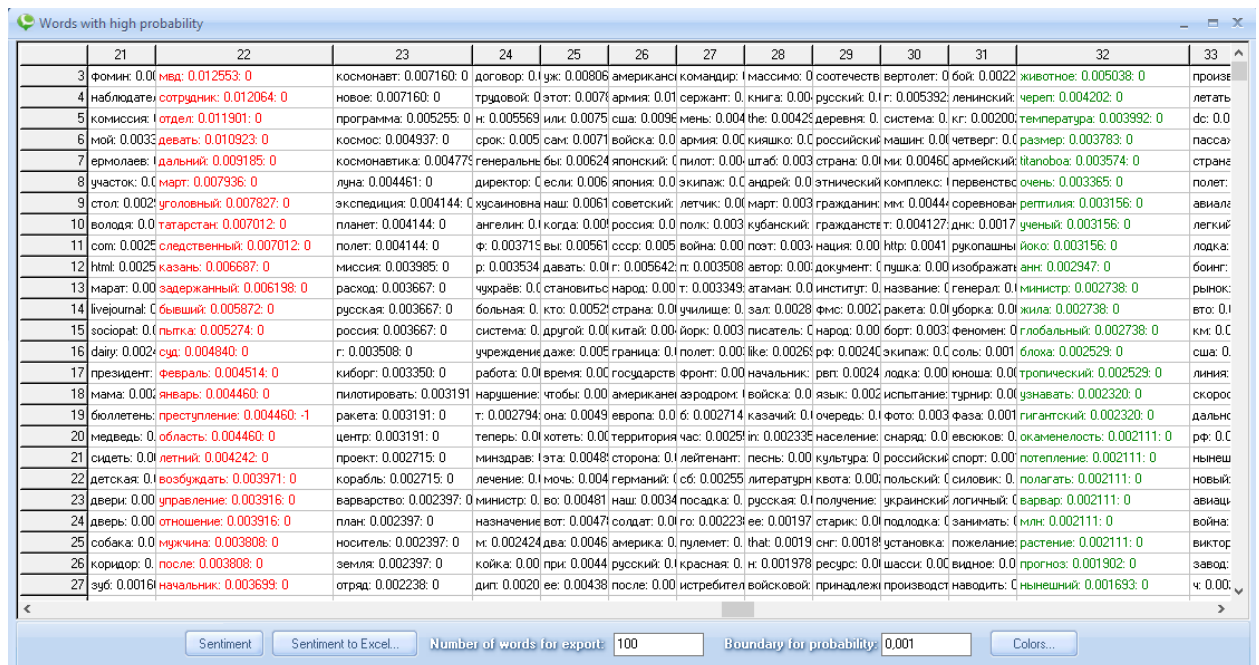


Fig. 7.5. Example of color hint in the word distribution matrix by topics.

As practice has shown, this color differentiation is extremely convenient.

7.4. Tonal calculation of the distribution of documents by topics.

Calculation of tonal estimates for documents is made for a ready-made topic solution. In order to go to the sentiment calculation of documents you need to click on the button **D**. As a result, the following window will appear (see Figure 7.6)

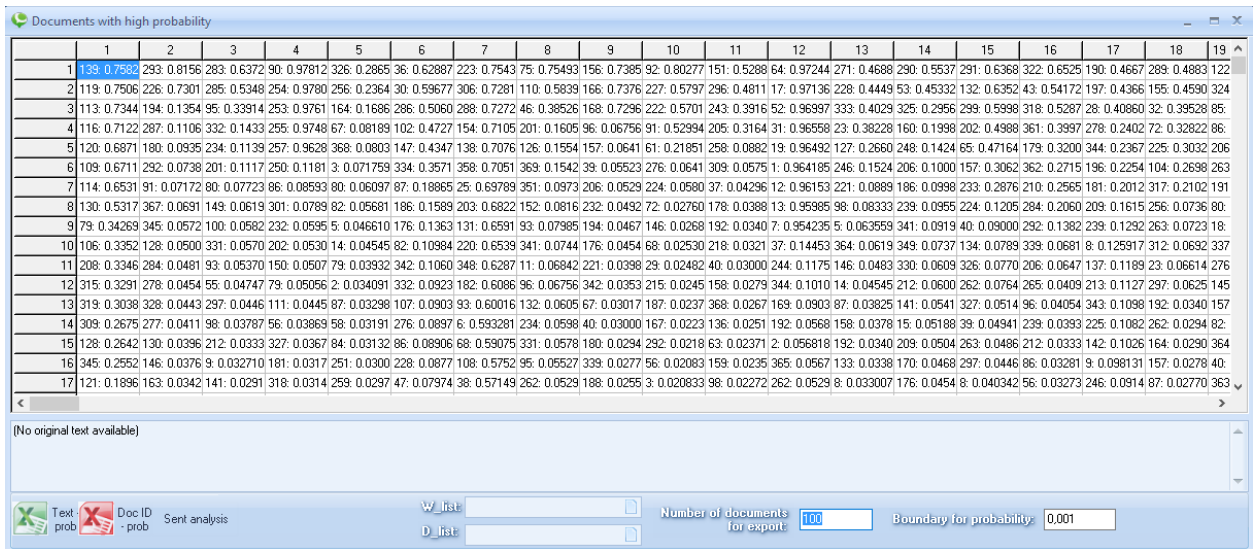
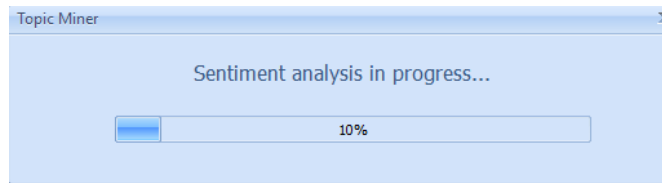


Fig. 7.6. Matrix of the distribution of documents on the themes before calculating the tonality.

In order to calculate the tonality of documents by topic, you need to click on the button Sent analysis. As a result (if a dictionary is connected), a window will appear (see below), which reflects the calculation process.



As a result of the calculation, in each cell containing the document number, the probability of the document is added the document evaluation sentiment (see figure 7.7).

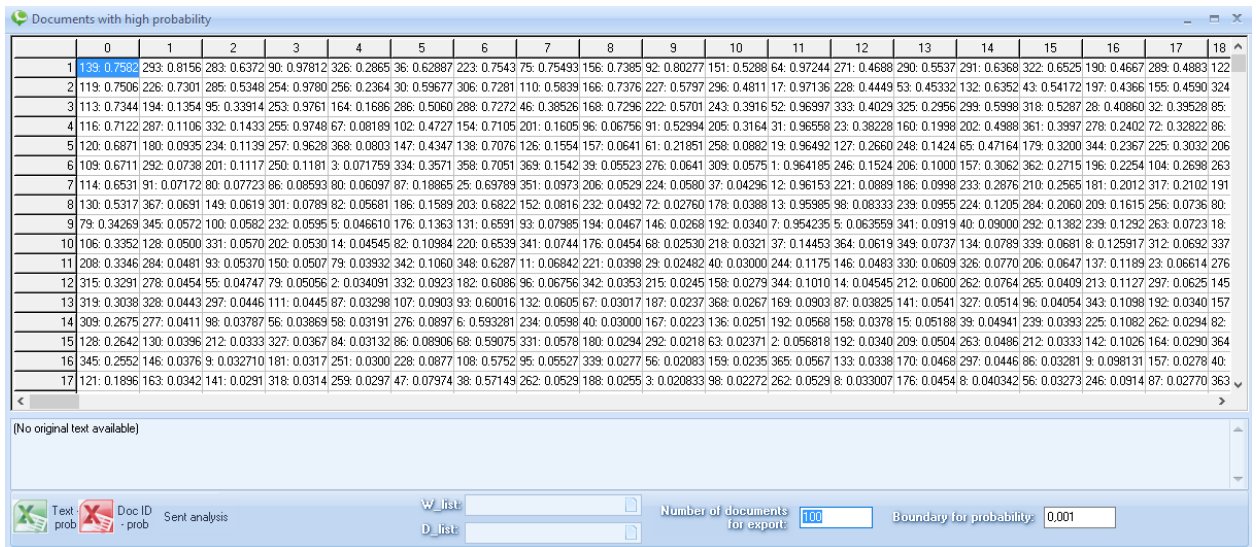

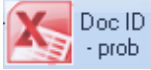


Fig. 7.7. The matrix of document distribution on topics after the calculation of the tonality.

7.4.1. Unloading matrix documents - topics with tonal estimates.

The unloading of the matrix, in this version, is implemented in the form of two functions (two buttons). The first unloading is realized in the form of uploading texts + probability + sentiment

of evaluation. In order to get such an unloading you need to click on the button . In the appeared window it is necessary to specify a file name. The data will be saved in csv format. The second upload is implemented as a combination of document IDs + probabilities + evaluation

sentiment. In order to unload the matrix in this form it is necessary to press the button . In the window that appears, you must specify a file name. The data will be saved in csv format.

The number of uploaded documents is determined by two parameters: 1. Number of documents for export (as shown in the figure below):



2. Boundary for probability (as shown in the figure below).



As a result, only those documents are uploaded (on all topics) that satisfy the above conditions.

7.5. Tonal calculation of the distribution of documents by topics for BigArtms.

Calculation of the tonality of the model, calculated within the BGARTM approach, can be performed as follows. The model calculated in the BIGARTM option should be saved as a project. Then open this project on the tab 'Gibbs sampling'. Next, calculate the tonality as described in the above. This approach is due to the fact that the data format calculated by the BigArtm and Gibbs sampling models are identical.

Глава 8. Time trends in topic models.

In the version of 2017, the possibility of plotting time trends is realized. The possibility of building is realized on the basis of two things. First, documents must have timestamps. Secondly, it is possible to build time charts as a basis for labels for sorted documents in a given topic, and in multimodal thematic models, where an additional matrix for the distribution of dates by themes.

8.1. Unification of time dates.

Due to the fact that in the collections of documents can meet different formats for presenting dates, the module 'view tmllda' implements the option of unifying dates. To start the unification process, open the 'View of tmllda' tab, load the data containing the timestamps (see Figure 8.1).

Next, you need to open the 'Data / Time repair' option (see figure 8.2), specify the field number that contains the timestamps. In this example, this field is № 2. Then you need to click on the 'Save as TMLDA' button and specify the name of the new file..

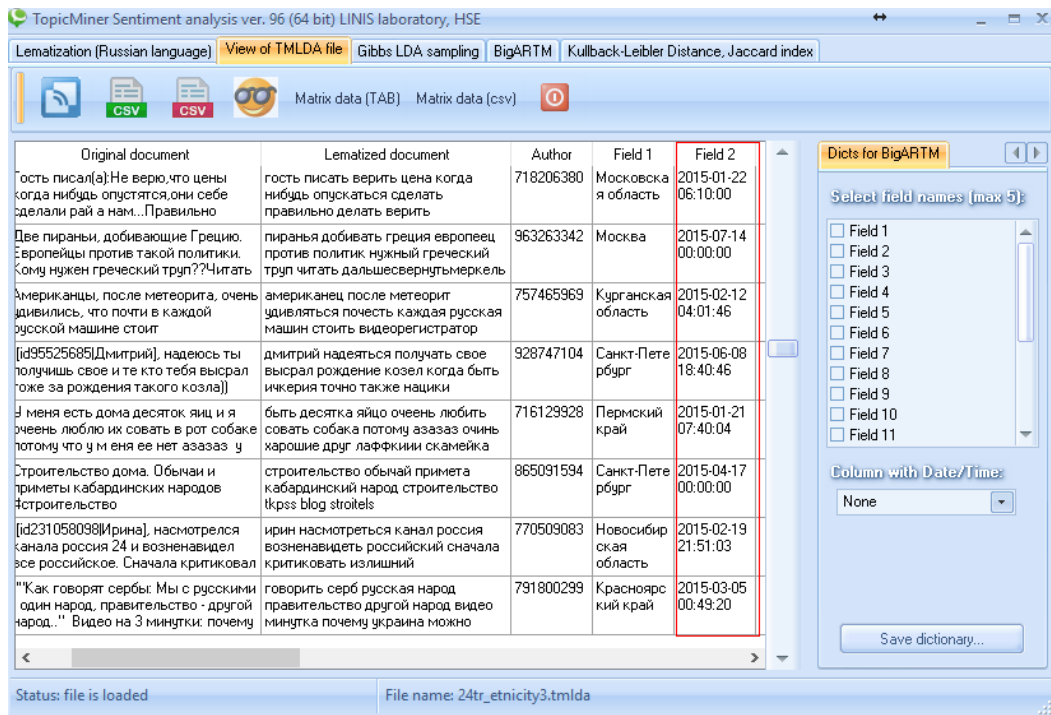


Fig. 8.1. Example collection with dates

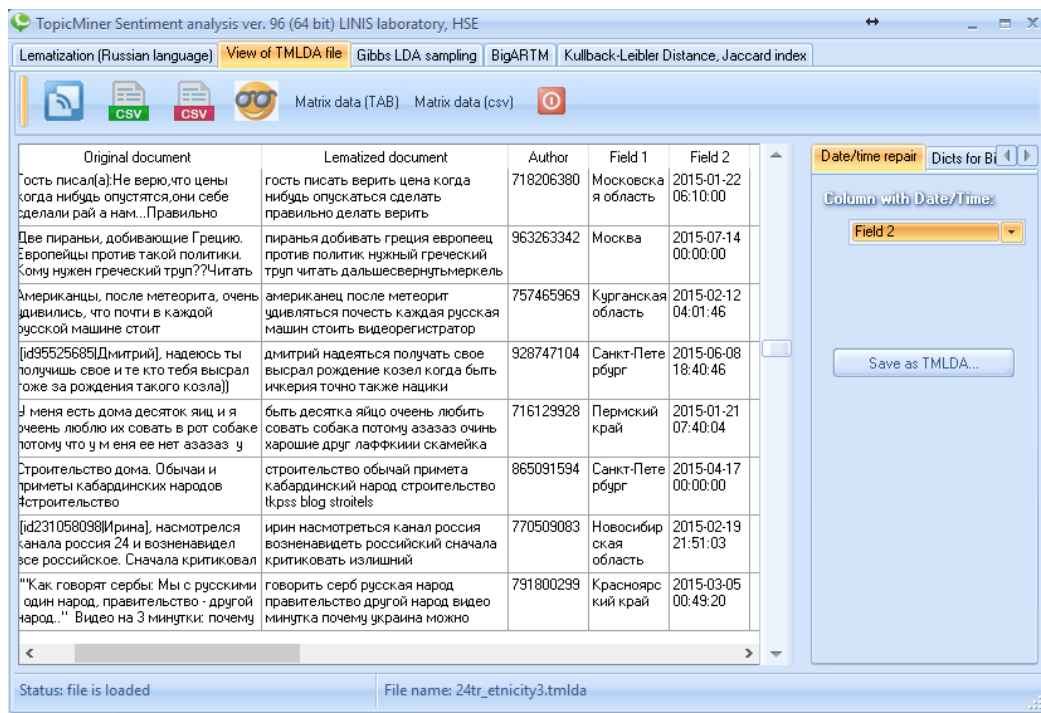


Fig. 8.2. Example of 'Data / Time repair' option.

After that, the process of unifying the temporary data will start, and the results will be written as a new tmla file. To make sure that all timestamps are unified enough to load a new file.

Next, you need to create a special tmla file and a dictionary for multimodal thematic modeling. To create such files, go to the 'Dics for BigARTM' option (see figure 8.2.1). Next, you need to specify the fields by which additional matrices will be built when calculating the multimodal models.

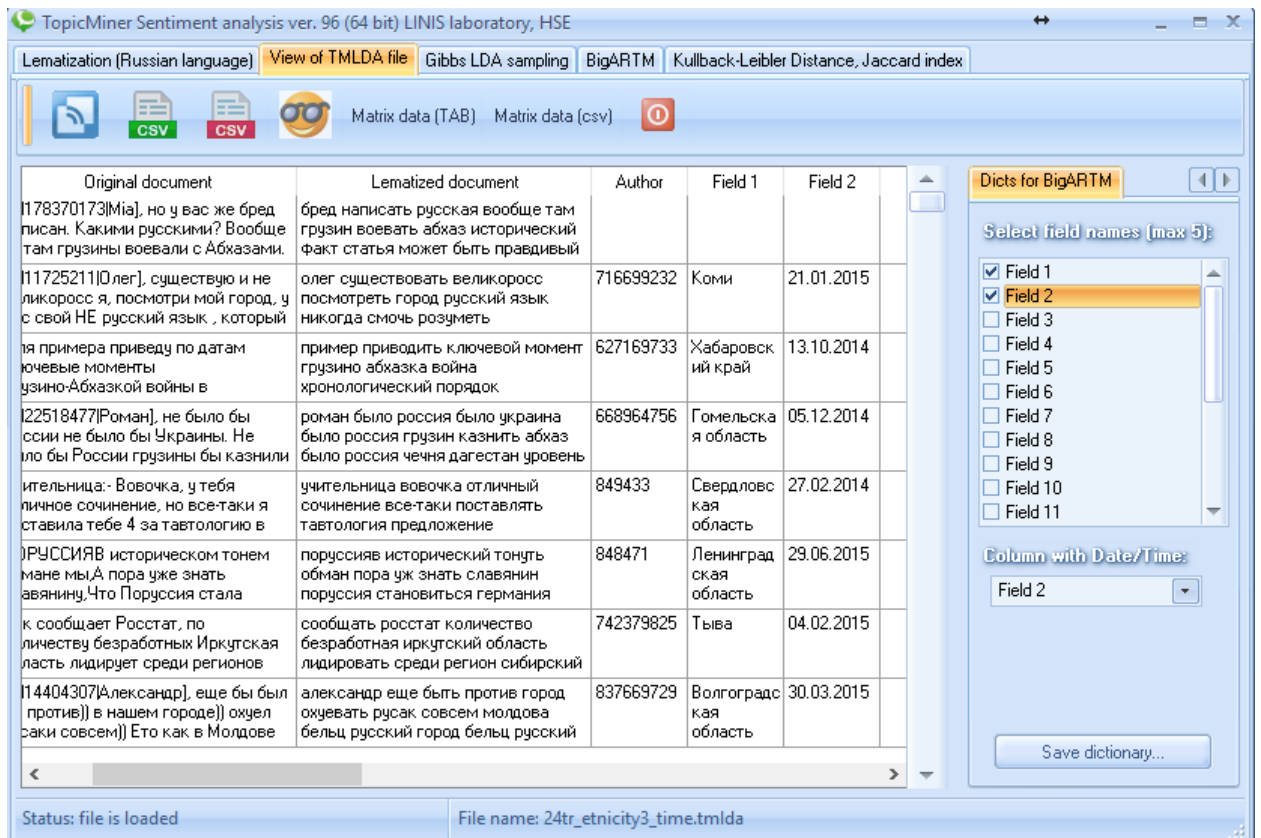



Fig. 8.2.1. An example of creating a dictionary with metadata for BigARTM.

Attention, the process of creating a dictionary can take a long time.

8.2. Construction of time trends in models based on multimodal thematic modeling.

Constructing a trend based on the distribution of documents by topic.

The construction of time trends for models on multimodal TM is possible only for sorted matrices of document distribution by topic. In order to build such a trend, you need either make a thematic calculation or upload the calculations already made. As a result, you need to open the distribution

matrix of sorted documents (button ) on the 'BigARTM' tab. The result is a matrix (see Figure 8.3).

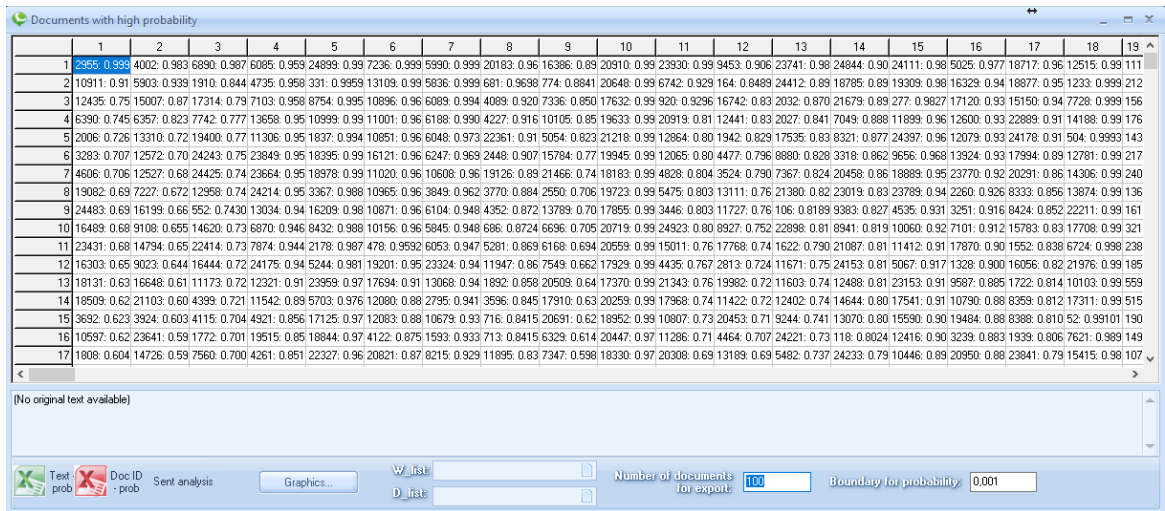


Fig. 8.3. An example of constructing time trends in the distribution matrix of documents by topics.

In this window there is a button 'Graphics'. When you click this button, the chart window appears.

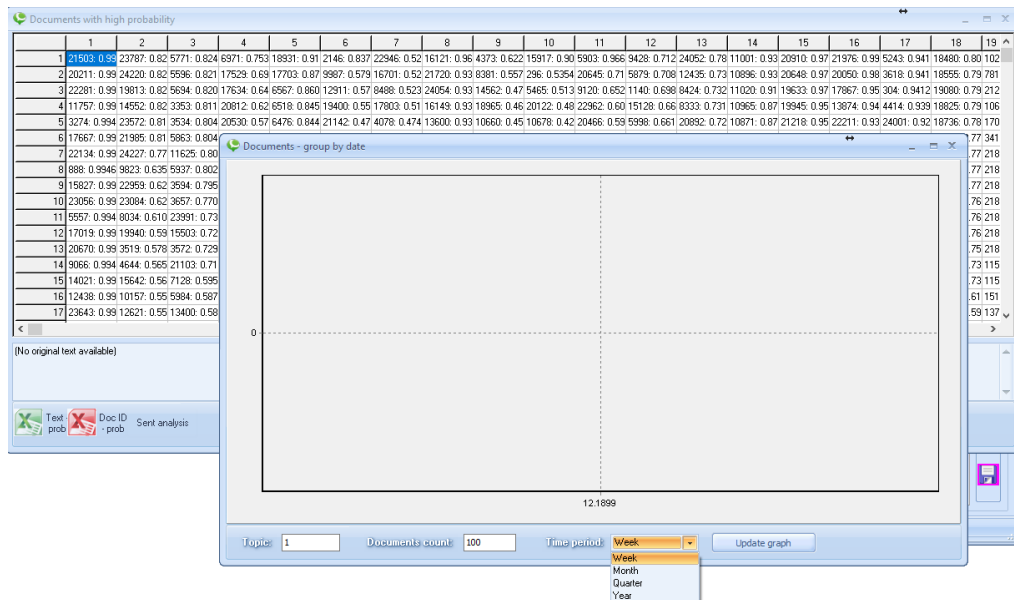


Fig. 8.4. An example of constructing time trends in the distribution matrix of documents by topics.

In this window, you first need to specify the topic number, the number of documents whose tags will be used to plot the graph and the aggregation period (see Figure 8.4). In order to update the contents of the graph, you need to click on the 'update graph' button. An example of such a graph on the topic № 11 is shown in Fig. 8.5

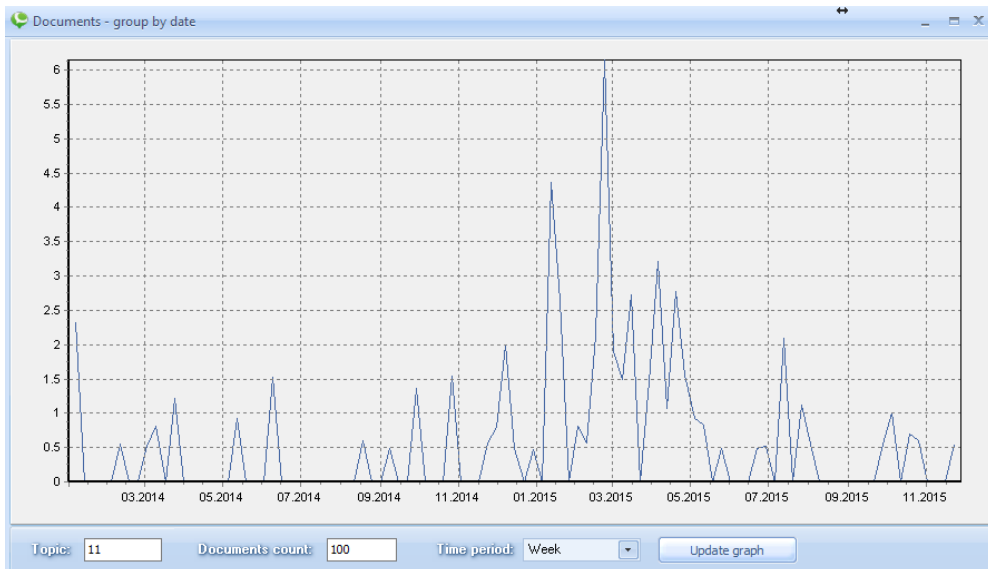



Fig. 8.5. An example of constructing time trends by weeks.

Distribution of time labels by topic.

In multimodal calculations, the field with timestamps can be used directly in the calculation. The result of the calculation is an additional matrix of time-stamp distribution by topic. For example, consider the '24tr_eticity3_time_bigartm.tmla' dataset. This data set is part of the information system. In this dataset, the timestamps are in field # 2, respectively, the additional matrix also has

the name 'field2'. To open the date distribution matrix by topic, select the button . An example of selecting a matrix is shown in Figure 8.6.

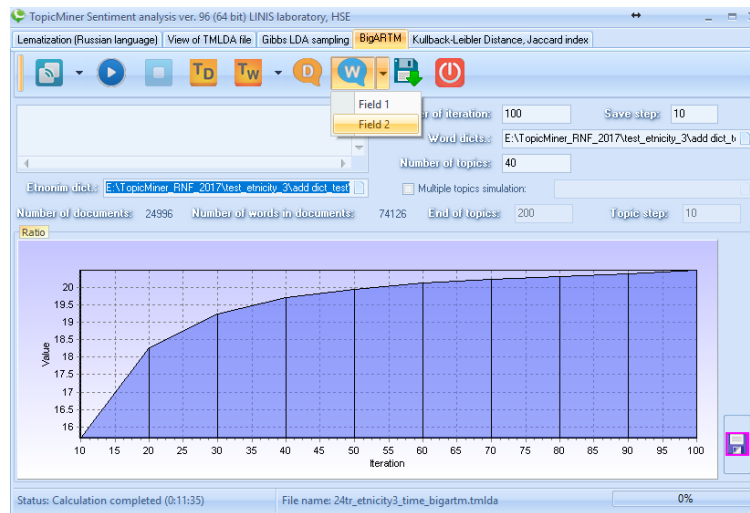


Fig. 8.6. Example of opening a matrix of time-stamp distribution by topics.

As a result, we get this matrix (see Figure 8.7).

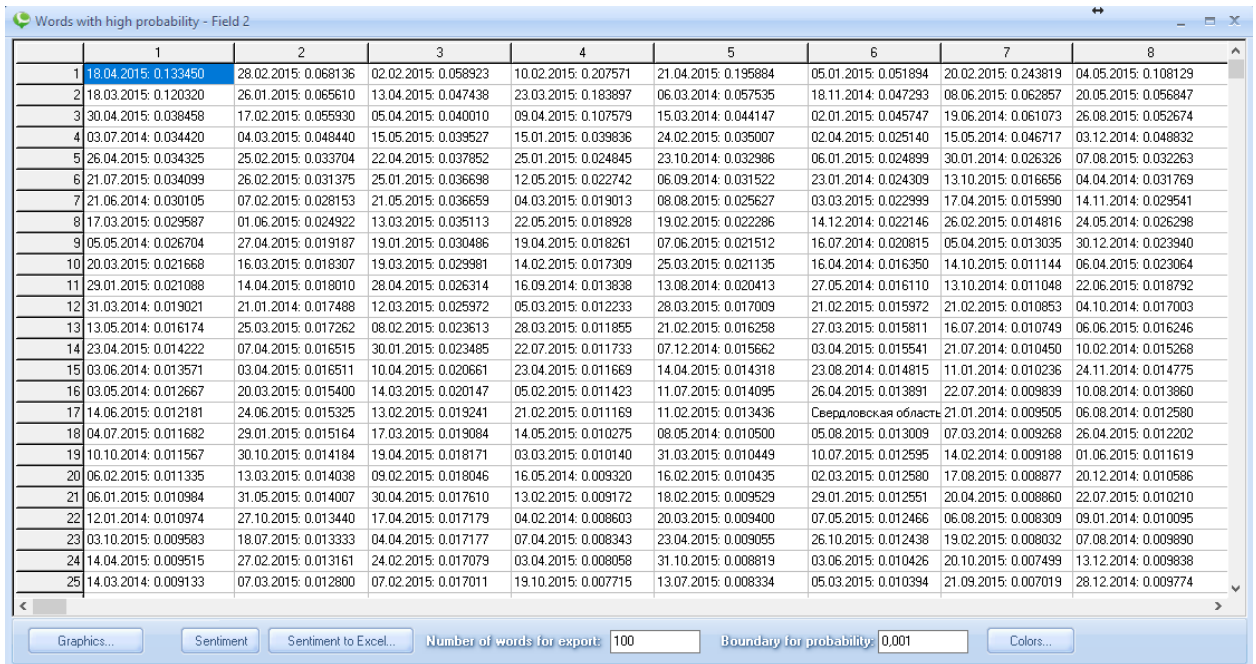


Fig. 8.7. Example of a matrix of time-stamp distribution by topics.

Each cell in this matrix contains the date and probability of the date in the corresponding topic.

In order to construct a timeline, you need to click on the 'graphics' button and in the window that appears, you need to specify the topic number, the number of documents whose tags will be used to plot the graph and the aggregation period. An example of plotting the graph is shown in Figure 8.8.

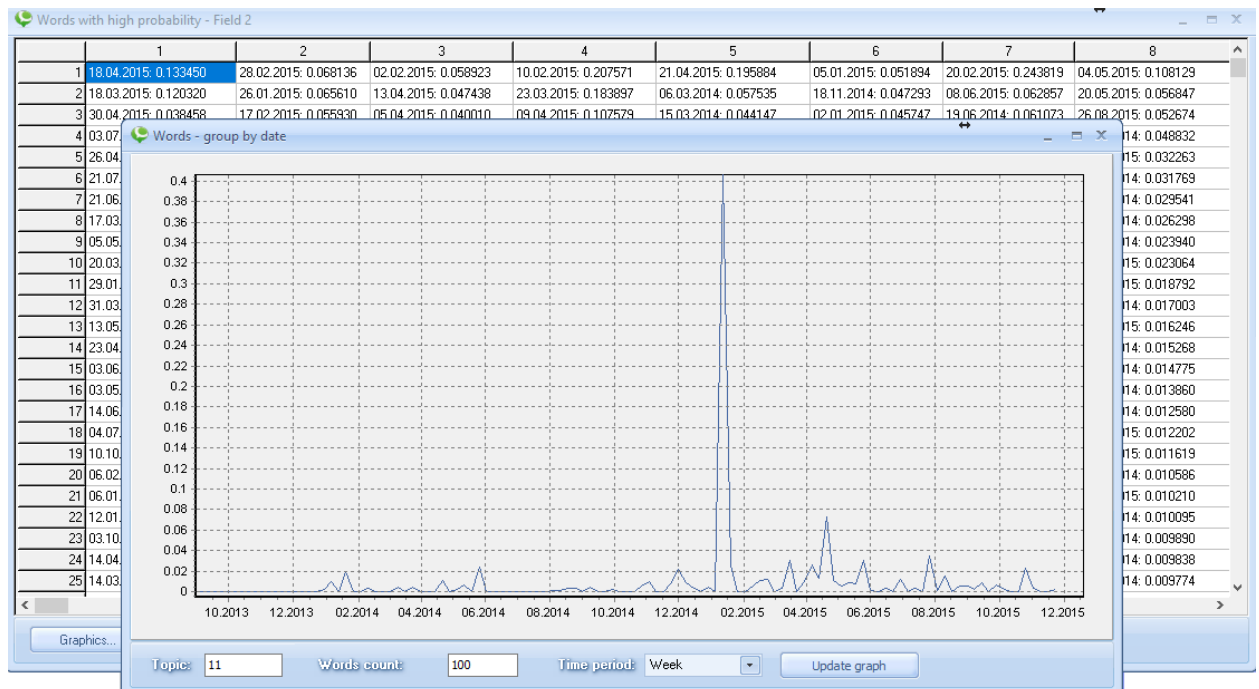



Fig. 8.8. Example of a matrix of time-stamp distribution by topics.

8.3. The construction of time trends in models based on Gibbs sampling.

Models based on the Gibbs sampling frame allow you to build a time trend only for the document

distribution matrix for topics. To do this, press the button  on the tab 'Gibbs LDA sampling'. In the window that appears, press the 'Graphics' button. Next, in the new window, specify the following parameters: 1. The topic number (for which you want to build a trend). 2. The number of documents that will be used to create the trend. 3. Aggregation period (week, month ..). 4. The number of the field that contains the date of the particular document. After setting the parameters, you need to click on the 'update graph' button. An example of such a graph is shown in Figure 8.9.

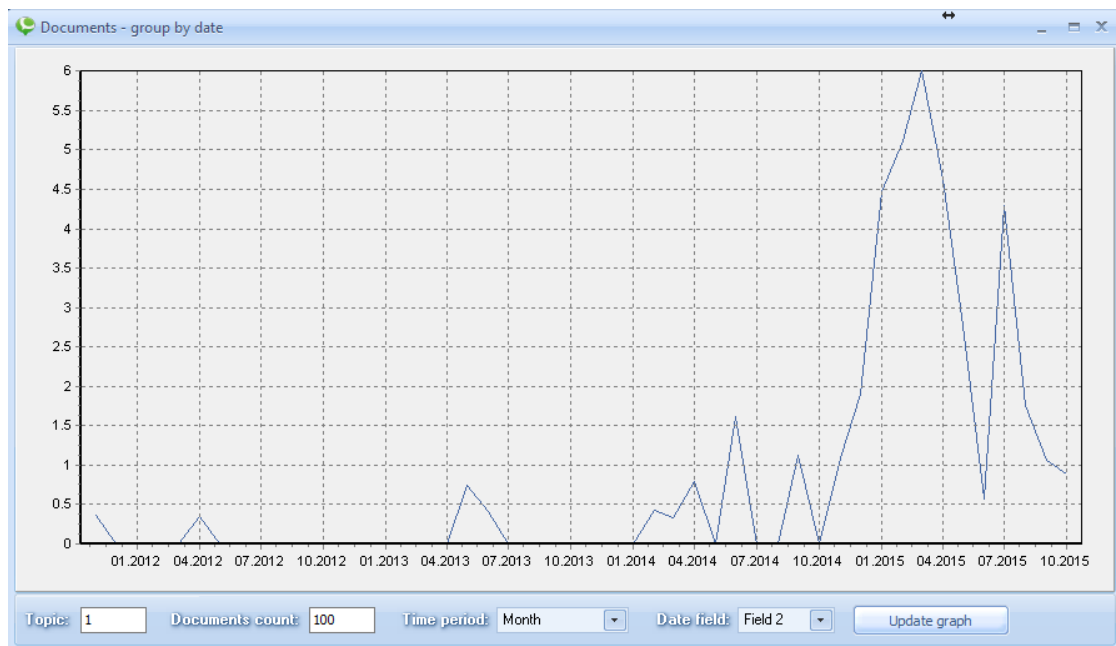


Fig. 8.8. Example of a time trend for Gibbs sampling.

Conclusion.

All questions on the application of the monitoring system 'TopicMiner' and 'Web TopicMiner', please send to the laboratory of Internet research, to Koltsov SN (skoltsov@hse.ru)