



TopicMiner with Sentiment analysis



Manual Version 96 (64 bits).

Saint-Petersburg

2018

Table of contents Chapter 1 Preprocessing of documents	Л
1.1 Procedure for assembling and lammatizing documents	4 Л
1.2 The second store of proprocessing	4
1.2.1 Creation of a list of stop words	ە م
1.2. The third store of preprocessing	9
Chapter 2. View tolde formet files	10
Chapter 2. View tinua format mes.	10
3.1. Interface of option 'Gibbs I DA sampling'	15
3.2. Loading documents for tonic modeling	15
3.3 Topic modeling based on Gibbs sampling	10
3.4. Visualization of the results of tonic modeling	10
3.4.1 Visualization of document distribution by words	20
3.4.2 Visualization of word distributions by topics	20
3.4.3 Visualization of distributions of sorted documents by topics	22
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	24
3.6 Loading the results of tonic modeling from the project file	20
Chapter 4. Tonic modeling by BigArtm models (multimodal tonic modeling).	27
4.1 Parameter setting in multimodal TM models	27
4.2. Visualization of the results of topic modeling.	
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
Γ_{Λ} ава 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 lematized 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.

	А	В	С	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.

오 TopicMiner Sentiment analysis ver. 78 (64 bit) LINIS	laboratory, HSE	_ = ×
Lematization (Russian language) View of TMLDA file Git	obs LDA sampling BigARTM Ku	ullback-Leibler Distance, Jaccard index
STEP 1. Assembling, deleting HTML tags and Lemmatisation Folder with original text files Result file (binary): Parameters for stemming: -c-wl-e utf-8 Lange File with trash data: STEP 2.1. CSV and/or TMDLA-files merging Input files [ESV or binary]:	n Russian V Doideparge UTF V	2.3 Distribution of word frequency in whole collection Number of unite words: U Total word number: U Low bound Upper bound 0 7 Filtration STEP 3. Removing stop words File for elearing (binary) Uutput file (binary)
Result file (binary): STEP 2.2. Extraction words form brackets and calculating File for elearing (binary): Uutput file (optional): Search in fist of words: no List of words 1 2	word frequency	ADeveloping sentiment binary dictionary ADeveloping sentiment binary dictionary ADeveloping sentiment binary dictionary Russian [] Rus
Status:	File name:	0%

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:



It is enough to specify only the file name. The program will automatically add an extension 'tmlda' (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'.

Parameters for stemming -c -wl -e utf-8

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

Codepage				
UTF	-			
UTF				
ANSI				

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):

STEP 1. Assembling, deleting HTML tags and Lemmatisation	
Folder with original text files: E:\TopicMiner_RNF_2015\Topic 📂	
Result file (binary): E:\TopicMiner_RNF_2015\TopicMiner_LINIS	
Parameters for stemming: -c -wl -e utf-8 Lang: Russian 💌	Codepage
File with trash data: E:\TopicMiner_RNF_2015\TopicMiner_LIN	UTF 💌

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 (lkbys 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 tmlda, which consistently contains pairs of texts in the original and lematized 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.

😍 TopicMiner Sentiment analysis ver. 78 (64 bit) LINIS laboratory, HSE 👘	_ = ×
Lematization (Russian language) View of TMLDA file Gibbs LDA sampling Big	gARTM Kullback-Leibler Distance, Jaccard index
STEP 1. Assembling, deleting HTML tags and Lemmatisation Folder with original task files: E:\TopicMiner_RNF_2015\TopicMiner_LINIS Parameters for stemming: -c-wl Lang: Russian File with trash data; E:\TopicMiner_RNF_2015\TopicMiner_LIN ANSI STEP 2.1. CSV and/or TMDLA-files merging Input files (SSV or binary):	2.3 Distribution of word frequency in whole collection Number of unterwords: U Total word number: U Low bound Upper bound 0 7 Filtration STEP 3. Removing stop words File for clearing (binary) Merge Dumut file (binary)
Hesult file (binary): STEP 2.2. Extraction words form brackets and calculating word frequency File for alearing (binary): Dutput file (optional): Search in fist of words: List of words 1 2	A Developing sentiment binary dictionary Hasult files Runsprocess
Status: Lematization 0:00:36 Execution: 49%	49%

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 поить 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=FB4EF53787D563FA8F1D6D2C3 F205F0C.d01t01{новость} {наука} {o} {зависимость}: {Чантикс??} {средство} {против} {курение}, {показывать} {себе} {также} {положительно|положительный} {в} {отношение} {контроль} {над|нада} {прием} {алкоголь}. {тот}, {кто} {принимать} {прапарат??} {c} {цель} {бросать} {курить} {часто|частый} {сообщать}, {что} {y} {они} {снижаться} {потребность} {в} {алкоголь}. {исследование} {показывать}, {что} {это|этот} {действительно|действительный} {так}. {Чантикс??} {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|менее|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??} {eщe} {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} {дофаминергических??} {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\poligon_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\poligon_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 tmlda. 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.

STEP 2.2. Ext	raction words	form bracke	ts and calcul	ating word frequenc	у	
File for clea	aring (binar,)): E:\Topic	Miner_RNF_	2015\TopicMiner_LI	NIS\example 📄	
Output file (optional): E:\TopicMiner_RNF_2015\TopicMiner_LINIS\example_da						
Search in List of wor	list of word ds	s: no		<u>/</u>		
	Word	Freq	TF-IDF	^		
1	the	26487	0.00000			
2	of	16517	0.00000		CSV	
3	to	14483	0.00000			
4	and	13606	0.00000			
-	1	40000	0.00000	*		

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.

Search in List of wor	list of word ds	s: оппозиці	ию	<u> </u>	
	Word	Freq	TF-IDF	· · · · · · · · · · · · · · · · · · ·	~
33860	оппозицию	1	0.00012		
33861	предоставл	1	0.00012		
33862	допиленны	1	0.00012		
33863	руксперта	1	0.00012		
00001		-	0.0004.0		۳.

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:

2.2 Distribution of word frequency in whole collection					
Number of unit	Number of unic words: 7649				
Total word number: 36846					
Low bound	Upper bound				
0	302	Filtration			

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, lematized 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 Error 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 tmlda format files.

The TopicMiner program provides the ability to view tmlda 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 tmlda files' option is shown in Figure 2.1.

📮 TopicMiner Sentiment analysis ver. 78 (64 bit) LINIS laboratory, HSE 📃 🗖 🗶						
Lematization (Russian language)	View of TMLDA file Gib	bs LDA sampling BigARTM Kullbac	k-Leibler Distance, Jaccard i	ndex		
	Matrix data (T	AB) Matrix data (csv) 🧿				
Document ID Drig	jinal document ⊲No dat	Lematized document	Author Field 1	Filters Data repair Mosco		
Status:		File name:				

Fig. 2.1. The option to view tmlda files.

Downloading the tmlda file. To download a file in tmlda 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 **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 99 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.

🜻 TopicMiner S	entiment analysis ver. 78 (64 bit) LINIS	laboratory, HSE			_ = ×
Lematization (Rus	ssian language) View of TMLDA file Gib	obs LDA sampling BigARTM Kullback-I	Leibler Distance, Jac	card index	
	sv Csv 🧭 Matrix data (T	(AB) Matrix data (csv) 🧿			
Document ID	Original document	Lematized document	Author Fie	d 1 📥	Filters Data repair Mosco
0	Посмотрите фотографию. Это нижегородец Николай Владимирович Савинов. Это - Россия. Один из	посмотреть фотография нижегородец николай владимирович савин россия космонавт видеть уж			List of words:
1	Оригинал взят у в Верной дорогой идете, товарищиВ пятницу источник в МВД сообщил, что, возможно,	оригинал взять верный дорога идти товарищив пятница источник мвд сообщать возможно подразделение			
2	Вопрос оказался суперлегким. Вместо ответа процитирую книжку (а заодно и потестирую новый тэг	вопрос оказываться суперлегким вместо ответ процитировать книжка заодно потестирую новый тэг			
3	В России, как это широкомасштабно известно, предпринимателей не любят. В большинстве своем	россия широкомасштабный известно предприниматель любить большинство свое сталкиваться			
4	Оригинал взят у в ЩеглишкоТак близко щеглы не подпускали меня с пленочных времен. Здесь будут	оригинал взять щеглишкотак близкий щегол подпускать мень пленочный время здесь показывать			
5	Митинги и акцииМировой судья Тверского района столицы Ольга Затомская оштрафовала на 1 тыс	митинг акциимировой судья тверской район столица ольга затомская оштрафовать тыс рубль			Column: Author •
6	Ну в общем я думаю так: менты на втором видео - переодетые в полицейскую форму уголовники,	ну общий думать мент второе видео переодевать полицейский форма уголовник надо взять охотничий			Words count: 0
7	Закон бутерброда Даже не намазанный маслом, кусок хлеба обязательно стремиться упасть туда,	закон бутерброд даже намазывать масло кусок хлеб обязательно стремиться упасть туда где грязно			Exclude short docs
<	·	·	· · · · ·	> -	Export to LDAXE2
Status: file is loa	aded	File name: DA3.tmlda		_	

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.

	Matrix data (T					
		AB) Matrix data (csv) 🧿				
Document ID	Original document осмотрите фотографию. Это ижегородец Николай Владимирович авинов. Это - Россия. Один из	Lematized document посмотреть фотография нижегородец, николай владимирович савин воссия космонавт видеть чж	Author	Field 1		Filters Data repair Mosco
Ор Иа М	ригинал взят у в Верной дорогой дете, товарищиВ пятницу источник в ВД сообщил, что, возможно,	оригинал взять верный дорога идти товарищив пятница источник мвд сообщать возможно подразделение				восток россия американский грант
Bo Bh Sa	опрос оказался суперлегким. место ответа процитирую книжку (а аодно и потестирую новый тэг	вопрос оказываться суперлегким вместо ответ процитировать книжка заодно потестирую новый тэг				федерация народ рФ жопа
В из лк	России, как это широкомасштабно звестно, предпринимателей не обят. В большинстве своем	россия широкомасштабный известно предприниматель любить большинство свое сталкиваться				
Οι δι πι	ригинал взяту в ЩеглишкоТак пизко щеглы не подпускали меня с пеночных времен. Здесь будут	оригинал взять щеглишкотак близкий щегол подпускать мень пленочный время здесь показывать				
Mi Te 3a	итинги и акцииМировой судья верского района столицы Ольга атомская оштрафовала на 1 тыс	митинг акциимировой судья тверской район столица ольга затомская оштрафовать тыс рубль				Column: Author
Ну вт пс	у в общем я думаю так: менты на гором видео - переодетые в олицейскую форму уголовники,	ну общий думать мент второе видео переодевать полицейский форма уголовник надо взять охотничий				Words count: 0
За на об	акон бутерброда. Даже не амазанный маслом, кусок хлеба бязательно стремиться упасть туда,	закон бутерброд даже намазывать масло кусок хлеб обязательно стремиться упасть туда где грязно				Exclude short docs
c				>	-	Export to LDAXE2

Fig. 2.3. Example of a loaded list of words

Uploading documents in tmlda format by word list. To reduce the collection of documents

Filter by words...

according to the list of downloaded words, you need to click the button

The program will create a file in tmlda 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.tmlda'. The file will contain only those documents in which there is at least one word from the downloaded list.

Uploading documents in the 'tmlda' 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 tmlda format without any empty documents, you need to click on

the button **Exclude empty docs...** The program will create a file in tmlda 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.tmlda'.

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 'tmlda' 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)

😍 TopicMiner S	entiment analysis ver. 88 (64 bit) LINIS	laboratory, HSE				↔ _ = ×
Lematization (Rus	ssian language) View of TMLDA file Git	bbs LDA sampling BigARTM Kullback-	Leibler Distar	nce, Jaccar	d index	
	sv Csv 🧭 Matrix data (1	(AB) Matrix data (csv) 🧿				
Document ID	Original document	Lematized document	Author	Field 1	A	Dicts for BigARTM
0	8 мая, средаСкорпионы ничего так сильно не хотят, как провести лето с любимым человеком.	май средаскорпионы сильно проводить лето любимая	56556171	Wall 5655617		Select field names (max 5);
1	Ренат, МиниТест «Какая тачка у тебя будет?» показал результат ниже. Много других тестов здесь -	ренат минитест какать тачка показывать результат тест minitest	169445299	Wall 1733110		Field 1
2	Странный праздник к нам идёт,Этот Старый Новый год!Снова полночь вместе встретимЗамечательно	странный праздник старый полночь встретимзамечательно отмечать веселый народв старый красивый	137183488	Wall 1249874		Field 4 Field 5 Field 6
3	С каждым такое было. Молча стоишь, улыбаешься, а сердце изнутри ломает рёбра.	такое молча стоить улыбаться сердце изнутри ломать ребро	159527866	Wall 1595278		Field 7 Field 8 Field 9
4	Умное лицо это еще не признак ума, а иногда это результат упорных тренировок перед зеркалом.	умный лицо признак ум иногда результат упорный тренировка зеркало	124987410	Wall 1249874		Field 10 Field 11
5	парни жалуются о том, что в наше время не осталось скромных девушек. парни, вы просто не	парень жаловаться время скромный девушка парень замечать скромный	159527866	Wall 1595278		
6	Срочно нужна 1 группа +, для новорожденного ребенка, админ опубликуй пожалуйста	срочно группа новорожденный ребенок админ опубликовывать	5304003	Wall 5304003		
7	Ну и что, что у меня нет парня, просто Бог посмотрев на меня сверху, сказал:" Пусть эта девочка	парень бог посмотреть сверху девочка доставаться достойный	159527866	Wall 1595278		Court di dianan
<				>	-	Save dictionary
Status: file is loa	aded	File name: region_test.tmlda				

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 Save dictionary...

				DA (I		Y						
Lemati	zation (Russia	an language)	View of TMI	DA hie Gib	bs LDA samp	ling BigART	M Kullback-L	eibler Distan	ce, Jaccar	d inde	×	
		csv	Ø	Matrix data (T	AB) Matrix (data (csv)	0					
hor	Field 1	Field 2	Field 3	Field 4	Field 5	Field 6	Field 7	Field 8	Field	-	Dicts for BigARTM	
171	Wall 56556171	post 7343	07.05.2013 21:05:12	Виктория	Ломанова	Улан-Удэ	Бурятия				Select field names	(max 5):
5299	Wall 173311047	post 2	12.22.2012 2:26:24 PM	llya	Grivtsov						Field 1 Field 2 Field 3	Î
3488	Wall 124987410	post 29	1.13.2014 8:51:30 PM	Lyuda	Gorelova	Киров	Кировская область				 Field 4 ✓ Field 5 Field 6 	
7866	Wall 159527866	post 43	11.15.2012 10:02:44 PM	Anastasia	Uskova	Набережн ые Челны	Татарстан				✓ Field 7 ☐ Field 8 ☐ Field 9	
7410	Wall 124987410	post 14	7.12.2013 6:11:20 PM	Vladislav	Enoktaev	Набережн ые Челны	Татарстан				Field 10	-
7866	Wall 159527866	post 23	5.3.2012 9:14:13 PM	Anastasia	Uskova	Набережн ые Челны	Татарстан					
103	Wall 5304003	post 1135	6.25.2015 8:59:35 PM	Igor	Moiseev	Арск	Татарстан					
7866	Wall 159527866	post 25	5.3.2012 9:15:00 PM	Anastasia	Uskova	Набережн ые Челны	Татарстан					
<	1			1	1		1		>	÷	Save dictiona	ry
											<u></u>	

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 tmlda. 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).

Contemporary Conte	aboratory, HSE	_ = ×
Lematization (Russian language) View of TMLDA file Gib	bs LDA sampling BigARTM Kullback-Leibler Distance, Jaccard in	dex
🖸 - 💽 🔲 🎹	o 🗠 🛃 🖶 📂 🔘 —	
Parameters of simulation Alpha: 0,5 Beta: 0,1 Save step: 10	40 100 Calculation parameters OMP Number of threads: 8 Method:	€ Fix topics
Number of documents: 369 Number of words	in documents: 28110 Sentiment diet: E	::\TopicMiner_RNF_2015\TopicMin
	teration	
Status: Documents as bag of words are loaded	File name: DA3.tmlda	0%

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 tmlda (see Figure 3.2).

Sentimer Sentime	ent analysis ver. 78 (64	bit) LINIS laboratory, HSE				_ = X
Lematization (Russian la	nguage) View of TMLD	A file Gibbs LDA sampling	BigARTM Kullback	-Leibler Distance, Jaccard	l index	
• • •	Open file with data				×	
Parameters of simulat	Nan <u>k</u> a:	TopicMiner_LINIS		🔹 🌀 🤌 🔛 🖽]-	
Alpha: 0,5	4	Имя 🔺		Дата изменени	я Тип	
Beta: 0,1	Быстрый достив	example_dataset		24.11.2016 16:3	0 Папка с ф	
Source above 10	обстрый доступ	example_project		03.12.2016 12:4	1 Папка с ф	
Save step. 10		1.tmlda		31.08.2016 13:4	2 Файл "ТМ 2 Файа "ТМ	
Number of docume	Рабочий стол	3.tmlda		31.08.2016 13:4	2 Файл ТМ	_2015\TopicMin 📄
Ratio	-					
	Библиотеки					
	Этот компьютер					
υ						
귀 0 >	1					
	Сеть					
		4			•	
		Имя файла: 3		-	Открыть	
		Тип файлов: Торіс	Miner LDA files (*.tmlda	i) 🔻	Отмена	
			· · ·		, iii	
Status: Documents as	bag of words are load	ed File name: DA3	.tmlda			0%

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

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

Contemporary Conte			_ = X
Lematization (Russian language) View of TMLDA fil	Gibbs LDA sampling	BigARTM Kullback-Leibler Distance, Jaccard index	
🛛 - 🖸 🗖 TO	v 🖸 🔍	📩 🛱 🎓 🔘	
Parameters of simulation		Calculation parameters	
Alpha: 0,5 Number of to	pies: 40	OMP Number of threads: 8	
Beta: 0,1 Number of iter	ation: 100	Method: Fix topics.	
Save step: 10			
Number of documents: TopicMiner		x	er_RNF_2015\TopicMin
Ratio			
	Loadir	ng data from file	
		45%	
		0 UT	
≅ 0			
20			
		0 Iteration	
Status: Documents as bag of words are loaded	File name: 3.tm	nlda	45%

Fig. 3.3. Example of uploading a data file

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

Lematization (Russian language)	View of TMLDA file	Gibbs LDA sampling	BigARTM Kullback-Leibler Distance, Jaccard index
	T _D T _W	0	🛃 🔁 🥏 🔘
Parameters of simulation			Calculation parameters
Alpha: 0,5	Number of topic	40 Katalana	OMP Number of threads: 8
Beta: 0,1	Number of iterati	003 100	Method:
Save step: 10			
Number of documents: 52	Number of wo	rds in documents:	is: 994 Sentiment diet:

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

3) The number of iterations. The number of iterations can be set in the option: Number of iteration: 100. 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: Save step: 10

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.

Method:	•
	LDA
	ISLDA
	Granulate LDA

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 \bigcirc . 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 To. 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.



- prob - 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.

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
76	Проголосуем за Зюганова, 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
ие европейски рения и его ра у 3D, для кото ской ассоциац именно такое,	ие врачинокулисты(чуть было не написал окультисты :)) уте звития в целом. Да кто ж сомневался, оно и у взрослых в рого кстати не нужны те самые специальные стерео-ског ии окулистов четко поясняет, что для нормального здоро , которое спокойно воспринимается глазками детей А 3D	зерждают, ч ызывает го ические оч вого разви видео уж о	по вот длительный про оловные боли даже.Во ки. Известный специа тия зрения у детей, им чень губительно може	осмотр 3D-видео детьми является очень вредн обще в первую очередь, это конкретно относит лист Карен Спарроу окулист из официальной 1 нужно исключительно чистое изображение, т т повлиять на зрение у детей с возрастом до б

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

-	2	3	4	5	6	7	8	9	10	11	12	13
0,00108	0,00181	0,00036	0,00036	0,00253	0,00036	0,00036	0,00181	0,01265	0,00036	0,00036	0,00108	0,00036
0,00103	0,00103	0,00103	0,00308	0,00103	0,00103	0,00308	0,00103	0,00103	0,00103	0,00103	0,00103	0,00103
0,00515	0,01546	0,00515	0,01546	0,01546	0,04639	0,02577	0,00515	0,00515	0,00515	0,00515	0,00515	0,05670
0,00460	0,24080	0,00460	0,00460	0,00767	0,00153	0,00153	0,00153	0,00153	0,01074	0,01380	0,00153	0,01074
0,01654	0,00184	0,00184	0,00551	0,00184	0,00184	0,00184	0,03493	0,00919	0,00184	0,00184	0,00184	0,00184
0,00676	0,00676	0,02027	0,00676	0,06081	0,04730	0,04730	0,00676	0,00676	0,04730	0,00676	0,00676	0,02027
0,00191	0,00573	0,00573	0,00191	0,00191	0,01336	0,00191	0,00573	0,00191	0,00191	0,00191	0,00954	0,00191
0,00221	0,00221	0,00221	0,00221	0,00221	0,33407	0,00664	0,00221	0,00221	0,00221	0,01991	0,01106	0,00664
0,00446	0,00446	0,01339	0,00446	0,03125	0,01339	0,01339	0,00446	0,01339	0,00446	0,00446	0,00446	0,00446
0,00260	0,00260	0,00260	0,01823	0,00260	0,01823	0,00260	0,00260	0,00260	0,00260	0,04427	0,00260	0,01302
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
0,00235	0,01643	0,00235	0,00704	0,00235	0,00235	0,01174	0,00235	0,00704	0,00235	0,00235	0,00704	0,00704
0,00526	0,03684	0,01579	0,02632	0,01579	0,00526	0,00526	0,00526	0,03684	0,00526	0,00526	0,01579	0,02632
0,02215	0,02215	0,00316	0,00316	0,00316	0,00949	0,00316	0,00949	0,00316	0,00316	0,00949	0,00316	0,00316
0,00236	0,00236	0,00236	0,00236	0,00236	0,00236	0,00708	0,00236	0,00236	0,00236	0,00236	0,00236	0,01651
0,00862	0,00862	0,00287	0,00287	0,00287	0,03736	0,00862	0,00287	0,01437	0,02011	0,00287	0,00287	0,02011
0,00070	0,00070	0,00070	0,00765	0,00765	0,00348	0,00070	0,00070	0,00070	0,00209	0,00070	0,00904	0,00209
0,00739	0,00246	0,00246	0,00246	0,00246	0,00246	0,02217	0,00246	0,03695	0,01232	0,07635	0,00246	0,00246
0,00187	0,00187	0,00187	0,00936	0,00187	0,01311	0,00187	0,00187	0,00936	0,00562	0,00562	0,00187	0,00187
	,00515 ,00460 0,01654 0,00191 0,00221 0,00446 0,00260 0,01724 0,00235 0,00526 0,00526 0,00526 0,002215 0,00236 0,00862 0,00862 0,00873 9,00187	J,00515 0,01546 0,00515 0,01546 0,00460 0,24080 0,01654 0,00184 0,00676 0,00676 0,00191 0,00573 0,00221 0,00221 0,00240 0,00246 0,00240 0,00246 0,00250 0,00260 0,01724 0,01724 0,00556 0,03684 0,002215 0,02215 0,00236 0,00236 0,00070 0,00070 0,00079 0,00246 0,00187 0,00187	J,00515 0,01546 0,00515 0,00545 0,00460 0,24080 0,00460 0,01654 0,00184 0,00184 0,00184 0,00676 0,00227 0,00277 0,00221 0,00191 0,00573 0,00221 0,00221 0,00221 0,00221 0,00221 0,00221 0,00260 0,00260 0,00260 0,00260 0,01724 0,01724 0,01724 0,01724 0,00255 0,01643 0,00235 0,00364 0,002215 0,00236 0,00236 0,00236 0,00236 0,00236 0,00236 0,00236 0,00070 0,00070 0,00070 0,00070 0,000739 0,00246 0,00246 0,00246	J,00515 0,01546 0,00515 0,01546 0,00460 0,24080 0,00460 0,00460 0,01544 0,00184 0,00515 0,00460 0,01554 0,00184 0,00184 0,00551 0,00676 0,02027 0,00676 0,00191 0,00573 0,00191 0,00221 0,00221 0,00221 0,00246 0,01244 0,01339 0,00446 0,00260 0,00260 0,01244 0,01724 0,00255 0,01643 0,00235 0,00744 0,00256 0,03684 0,01579 0,02632 0,00236 0,00236 0,00236 0,00236 0,00236 0,00236 0,00236 0,00236 0,00070 0,00070 0,00070 0,00236 0,00070 0,00070 0,00246 0,00246 0,00187 0,00187 0,00936	0,00515 0,01546 0,00515 0,01546 0,01546 0,005015 0,01546 0,00515 0,01546 0,01546 0,00460 0,24080 0,00460 0,00460 0,00767 0,01544 0,00184 0,00551 0,00184 0,00571 0,00184 0,00676 0,00227 0,00676 0,00211 0,00211 0,00211 0,00221 0,00221 0,00221 0,00221 0,00221 0,00221 0,00446 0,01339 0,00446 0,01323 0,00260 0,01224 0,00260 0,00260 0,01244 0,01724 0,01724 0,01724 0,00235 0,01643 0,00235 0,00704 0,00235 0,00260 0,00236 0,00236 0,00236 0,00236 0,00236 0,02215 0,00316 0,00316 0,00236 0,00236 0,00236 0,00236 0,00236 0,00236 0,00236 0,00236 0,00236 0,00236 0,00236 0,00070	0,00515 0,01546 0,00515 0,01546 0,01546 0,04639 0,00600 0,24080 0,00460 0,00767 0,00153 0,01546 0,00184 0,00184 0,00515 0,00184 0,00184 0,00676 0,00184 0,00184 0,00184 0,00184 0,00184 0,00676 0,00227 0,00676 0,000191 0,01330 0,00191 0,00221 0,00221 0,00221 0,03125 0,01339 0,00446 0,01339 0,00446 0,01323 0,00260 0,01224 0,01724 0,01724 0,01724 0,00235 0,01643 0,00235 0,00744 0,01240 0,01242 0,00236 0,02215 0,00316 0,00316 0,00235 0,00236 0,00236 0,0236 0,00236 0,00236 0,00236 0,00236 0,00236 0,00236 0,00236 0,00236 0,00236 0,00236 0,00236 0,00236 0,00236 0,00236 0,00236 0,00236	0,00515 0,01546 0,00515 0,01546 0,01546 0,04639 0,02577 0,00610 0,24080 0,00460 0,00767 0,00153 0,00153 0,01546 0,01844 0,00184 0,00184 0,00184 0,00184 0,00676 0,00187 0,00184 0,00184 0,00184 0,00184 0,00676 0,00277 0,00676 0,06081 0,04730 0,04730 0,00191 0,00573 0,00221 0,00221 0,033407 0,00664 0,00221 0,00221 0,00221 0,01253 0,01339 0,01339 0,00260 0,00260 0,01823 0,00260 0,01323 0,00260 0,01724 0,01724 0,01724 0,01724 0,01724 0,01724 0,00235 0,01643 0,00235 0,00235 0,00235 0,01174 0,00236 0,0236 0,00235 0,01245 0,01724 0,01724 0,00236 0,0236 0,00236 0,00236 0,00236 0,00236	0,00515 0,01546 0,00515 0,01546 0,04639 0,02577 0,00515 0,00610 0,24080 0,00460 0,00767 0,00153 0,00153 0,00153 0,01546 0,01844 0,00184 0,00184 0,00184 0,00184 0,00184 0,00184 0,00676 0,00277 0,00676 0,00184 0,0184 0,00767 0,00676 0,00277 0,00676 0,00184 0,00184 0,00767 0,00191 0,00573 0,00191 0,01336 0,00191 0,00573 0,00221 0,00221 0,00221 0,03125 0,01339 0,00446 0,00260 0,00260 0,01823 0,00260 0,01244 0,01724	0,00515 0,01546 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00153 0,00153 0,00153 0,00153 0,00153 0,00153 0,00153 0,00153 0,00153 0,00164 0,00184 0,00184 0,00184 0,00184 0,00184 0,00184 0,00184 0,00184 0,00184 0,00184 0,00184 0,00184 0,00184 0,00184 0,00184 0,00184 0,00184 0,00184 0,00184 0,00191 0,00573 0,00191 0,00573 0,00191 0,00173 0,00191 0,00573 0,00191 0,00221 0,00221 0,00221 0,00221 0,00221 0,00221 0,00221 0,00260 0,00260 0,00260 0,01234 0,01244 0,01724 0,01724 0,01724 0,01724 0,01724 0,01724 0,01724 0,01724 0,01724 0,01724 <t< td=""><td>0,00515 0,01546 0,01546 0,04639 0,02577 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00153 0,00154 0,00211 0,00212 <t< td=""><td>0,00515 0,01546 0,01546 0,04639 0,02577 0,00515 0,00184 0,00191 0,00573 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00221 0,00221 0,00221 0,00221 0,00221 0,00221 0,00221 0,00221 0,00221 0,00221 0,00221 0,00221 0,00221 <t< td=""><td>0,00515 0,01546 0,01546 0,01546 0,04639 0,02577 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00184 0,00191 0,00184 0,00191 <t< td=""></t<></td></t<></td></t<></td></t<>	0,00515 0,01546 0,01546 0,04639 0,02577 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00153 0,00154 0,00211 0,00212 <t< td=""><td>0,00515 0,01546 0,01546 0,04639 0,02577 0,00515 0,00184 0,00191 0,00573 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00221 0,00221 0,00221 0,00221 0,00221 0,00221 0,00221 0,00221 0,00221 0,00221 0,00221 0,00221 0,00221 <t< td=""><td>0,00515 0,01546 0,01546 0,01546 0,04639 0,02577 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00184 0,00191 0,00184 0,00191 <t< td=""></t<></td></t<></td></t<>	0,00515 0,01546 0,01546 0,04639 0,02577 0,00515 0,00184 0,00191 0,00573 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00191 0,00221 0,00221 0,00221 0,00221 0,00221 0,00221 0,00221 0,00221 0,00221 0,00221 0,00221 0,00221 0,00221 <t< td=""><td>0,00515 0,01546 0,01546 0,01546 0,04639 0,02577 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00184 0,00191 0,00184 0,00191 <t< td=""></t<></td></t<>	0,00515 0,01546 0,01546 0,01546 0,04639 0,02577 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00515 0,00184 0,00191 0,00184 0,00191 <t< td=""></t<>

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

	А	В	С	D	E	F	G	Н	1	J	К	L	М	N	0	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://gu	utta-honey	livejourn	al.com/298	8516.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, насто	3	7 yuzilla	http://yu	uzilla.livej	ournal.com	/926282.h	tml	0,00515	0,01546	0,00515	0,01546	0,01546	0,04639	0,02577	0,00515	0,00515
5		3 Этот пятилетний мальчик, британец 3	2	5 yuzilla	http://yu	uzilla.livejo	ournal.com	n/923209.h	tml	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	6	1 alexlotov	http://al	exlotov.liv	/ejournal.o	om/35788	5.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	2	6 yuzilla	http://yu	uzilla.livejo	ournal.com	/923636.h	tml	0,00676	0,00676	0,02027	0,00676	0,06081	0,0473	0,0473	0,00676	0,00676
8		6 Проголосуем за Зюганова, чтобы он г	8	5 alexlotov	http://al	exlotov.liv	/ejournal.o	om/36388	0.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	3	8 yuzilla	http://yu	uzilla.livejo	ournal.com	/926693.h	tml	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:	6	2 alexlotov	http://al	exlotov.liv	/ejournal.o	om/35813	4.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	2	7 yuzilla	http://yu	uzilla.livejo	ournal.com	/923797.h	tml	0,0026	0,0026	0,0026	0,01823	0,0026	0,01823	0,0026	0,0026	0,0026
12		10 видео от grigoruk Оппозиция не про	6	3 alexlotov	http://al	exlotov.liv	/ejournal.o	om/35829	4.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	4	9 yuzilla	http://yu	uzilla.livejo	ournal.com	/929280.h	itml	0,00235	0,01643	0,00235	0,00704	0,00235	0,00235	0,01174	0,00235	0,00704
14		12 Явка обещает быть высокой, потому	8	6 alexlotov	http://al	exlotov.liv	/ejournal.c	om/36428	2.html	0,00526	0,03684	0,01579	0,02632	0,01579	0,00526	0,00526	0,00526	0,03684
15		13 Поставлена жирная точка в деле Юри	2	8 yuzilla	http://yu	uzilla.livejo	ournal.com	/924058.h	itml	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	3	9 yuzilla	http://yu	uzilla.livej	ournal.com	/926959.h	itml	0,00236	0,00236	0,00236	0,00236	0,00236	0,00236	0,00708	0,00236	0,00236
17		15 Совершенно очевидно, что поднима	6	4 alexlotov	http://al	exlotov.liv	/ejournal.o	om/35843	6.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://gu	utta-honey	livejourn	al.com/298	8998.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	2	9 yuzilla	http://yu	uzilla.livej	ournal.com	/924253.h	tml	0,00739	0,00246	0,00246	0,00246	0,00246	0,00246	0,02217	0,00246	0,03695
20		18 Меланизм преимущественное расп	5	0 yuzilla	http://yu	uzilla.livejo	ournal.com	/929771.h	tml	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:	4	0 yuzilla	http://yu	uzilla.livejo	ournal.com	/927036.h	tml	0,00211	0,00211	0,03165	0,00633	0,02321	0,27637	0,00211	0,00211	0,01899
22		20 Новая парадигма мировоззрения Кто	7	3 alexlotov	http://al	exlotov.liv	/ejournal.c	om/36089	9.html	0,00214	0,00071	0,00071	0,00071	0,00214	0,00071	0,00071	0,00071	0,00357
23		21 el_murid: Арабская весна создала вес	8	7 alexlotov	http://al	exlotov.liv	/ejournal.c	om/36444	0.html	0,05968	0,00161	0,00161	0,00161	0,00161	0,00484	0,00161	0,17258	0,00484
24		22 В случае конфликта, неудовлетворит	1	3 gutta_honey	http://gu	utta-honey	livejourn.	al.com/303	3091.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 взры	3	0 yuzilla	http://yu		ournal.com	/924445.h	tml	0,00256	0,00256	0,00256	0,00256	0,01795	0,00769	0,01282	0,00256	0,00256
26		24 Проводимые разными компаниями и	4	1 yuzilla	http://yu	uzilla.livejo	ournal.com	/927292.h	tml	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.

	Α	В	С	D	E	F	G	Н	1	J	K	L	М	N	0	Р	Q	R	S	Т
1	ID docume	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://gu	tta-honey.	livejourna.	1.com/298	516.html	0,00103	0,00103	0,00103	0,00308	0,00103	0,00103	0,00308	0,00103	0,00103	0,00103	0,00103
3	2	74	3	7 yuzilla	http://yu	zilla.livejo	urnal.com	/926282.h	tml	0,00515	0,01546	0,00515	0,01546	0,01546	0,04639	0,02577	0,00515	0,00515	0,00515	0,00515
4	3	308	2	5 yuzilla	http://yu	zilla.livejo	urnal.com	/923209.h	tml	0,0046	0,2408	0,0046	0,0046	0,00767	0,00153	0,00153	0,00153	0,00153	0,01074	0,0138
5	4	253	6	1 alexlotov	http://al	exlotov.liv	ejournal.co	om/35788	5.html	0,01654	0,00184	0,00184	0,00551	0,00184	0,00184	0,00184	0,03493	0,00919	0,00184	0,00184
6	5	53	2	6 yuzilla	http://yu	zilla.livejo	urnal.com	/923636.h	tml	0,00676	0,00676	0,02027	0,00676	0,06081	0,0473	0,0473	0,00676	0,00676	0,0473	0,00676
7	6	243	8	5 alexlotov	http://al	exlotov.liv	ejournal.co	om/36388	0.html	0,00191	0,00573	0,00573	0,00191	0,00191	0,01336	0,00191	0,00573	0,00191	0,00191	0,00191
8	7	211	3	8 yuzilla	http://yu	zilla.livejo	urnal.com	/926693.h	tml	0,00221	0,00221	0,00221	0,00221	0,00221	0,33407	0,00664	0,00221	0,00221	0,00221	0,01991
9	8	102	6	2 alexlotov	http://al	exlotov.liv	ejournal.co	om/35813	4.html	0,00446	0,00446	0,01339	0,00446	0,03125	0,01339	0,01339	0,00446	0,01339	0,00446	0,00446
10	9	172	2	7 yuzilla	http://yu	zilla.livejo	urnal.com	/923797.h	tml	0,0026	0,0026	0,0026	0,01823	0,0026	0,01823	0,0026	0,0026	0,0026	0,0026	0,04427
11	10	9	6	3 alexlotov	http://al	exlotov.liv	ejournal.co	om/35829	4.html	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	4	9 yuzilla	http://yu	zilla.livejo	urnal.com	/929280.h	tml	0,00235	0,01643	0,00235	0,00704	0,00235	0,00235	0,01174	0,00235	0,00704	0,00235	0,00235
13	12	78	8	6 alexlotov	http://al	exlotov.liv	ejournal.co	om/36428	2.html	0,00526	0,03684	0,01579	0,02632	0,01579	0,00526	0,00526	0,00526	0,03684	0,00526	0,00526
14	13	138	2	8 yuzilla	http://yu	zilla.livejo	urnal.com	/924058.h	tml	0,02215	0,02215	0,00316	0,00316	0,00316	0,00949	0,00316	0,00949	0,00316	0,00316	0,00949
15	14	194	3	9 yuzilla	http://yu	zilla.livejo	urnal.com	/926959.h	tml	0,00236	0,00236	0,00236	0,00236	0,00236	0,00236	0,00708	0,00236	0,00236	0,00236	0,00236
16	15	154	6	4 alexlotov	http://ale	exlotov.liv	ejournal.co	om/35843	5.html	0,00862	0,00862	0,00287	0,00287	0,00287	0,03736	0,00862	0,00287	0,01437	0,02011	0,00287
17	16	714		2 gutta_honey	http://gu	tta-honey.	livejourna.	1.com/298	998.html	0,0007	0,0007	0,0007	0,00765	0,00765	0,00348	0,0007	0,0007	0,0007	0,00209	0,0007
18	17	190	2	9 yuzilla	http://yu	zilla.livejo	urnal.com	/924253.h	tml	0,00739	0,00246	0,00246	0,00246	0,00246	0,00246	0,02217	0,00246	0,03695	0,01232	0,07635
19	18	247	5	0 yuzilla	http://yu	zilla.livejo	urnal.com	/929771.h	tml	0,00187	0,00187	0,00187	0,00936	0,00187	0,01311	0,00187	0,00187	0,00936	0,00562	0,00562
20	19	216	4	0 yuzilla	http://yu	zilla.livejo	urnal.com	/927036.h	tml	0,00211	0,00211	0,03165	0,00633	0,02321	0,27637	0,00211	0,00211	0,01899	0,02321	0,01899
21	20	688	7	3 alexlotov	http://ale	exlotov.liv	ejournal.co	om/36089	9.html	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	8	7 alexlotov	http://al	exlotov.liv	ejournal.co	om/36444	0.html	0,05968	0,00161	0,00161	0,00161	0,00161	0,00484	0,00161	0,17258	0,00484	0,00806	0,00161
23	22	1315	1	3 gutta_honey	http://gu	tta-honey.	livejourna.	I.com/303	091.html	0,00038	0,00038	0,0019	0,00267	0,00038	0,00038	0,00343	0,00038	0,00038	0,00495	0,00038
24	23	175	3	0 yuzilla	http://yu	zilla.livejo	urnal.com	/924445.h	tml	0,00256	0,00256	0,00256	0,00256	0,01795	0,00769	0,01282	0,00256	0,00256	0,00256	0,00256
25	24	172	4	1 yuzilla	http://yu	zilla.livejo	urnal.com	/927292.h	tml	0,01289	0,00773	0,00258	0,00258	0,00258	0,00258	0,00773	0,00258	0,01289	0,00773	0,01804
26	25	370	6	5 alexlotov	http://ale	exlotov.liv	ejournal.co	om/35871	0.html	0,0013	0,0039	0,00649	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: 100

3.4.2. Visualization of word distributions by topics.

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

Ҿ Words	- Topics	Distribut	ions											-	= x
	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	0,00
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	0,00210	0,00
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	0,00
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	0,00
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	0,00
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	0,00
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	0,00
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	0,00
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	0,00
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	0,00
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	0,00
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	0,00
< 17	1_	0.00000	0.00000	0.00010	0.00100	0.00000	0.00000	0.00005	0.00000	0.00011	0.00011	0.00000	0.00000	0.01000	×
Hide	selected row	/	Reset hide	den rows	E	port to Excel	I I	Number of	words for e	xport: 100		Boundar	y for proba	bility: 0,00	000

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 Hide selected row 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 Reset hidden rows	allows you to restore all previously hidden words.
------------------------------	--

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

Export to Excel...

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 \bigcirc . 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.

Cocume Docume	ents with hig	ıh probabil	ity																	-	= x
	1	2	3		4		5	6	7	8	9	10	11	12	13	14	15	16	17	18	19 ^
1	49: 0.61250	15: 0.89714	4 57: 0.3	70193 5	51: 0.44	273 12	2: 0.48748	89: 0.70118	33: 0.4824	3 25: 0.2399	2 14: 0.55332	2 7: 0.526868	3 29: 0.62096	74: 0.16298	44: 0.24714	22: 0.41896	56: 0.19898	3 21: 0.86398	3 24: 0.83972	16: 0.40666	34:
2	48: 0.54384	33: 0.05000	0 27: 0.2	26265 2	2: 0.270	658 9:	0.311845	13: 0.69335	38: 0.3216	2 55: 0.0725	8 49: 0.07250) 5: 0.470796	78: 0.07009	99: 0.10317	77: 0.06716	44: 0.39885	32: 0.14150	0 53: 0.13492	2 34: 0.11403	56: 0.19898	3 47:
3	31: 0.12013	56: 0.02551	1 29: 0.0	08871 8	34: 0.09	322 8:	0.302209	10: 0.56597	63: 0.0555	5 68: 0.0714	2 37: 0.06547	7 6: 0.460181	82: 0.05172	47: 0.08646	71: 0.06164	52: 0.12500	41: 0.1163	5 71: 0.10274	46: 0.06363	26: 0.11029	3 45:
4	79: 0.10215	61: 0.02455	5 92: 0.0	07746 7	71: 0.04	794 10	0: 0.24020	88: 0.53427	89: 0.0502	9 56: 0.0357	1 72: 0.04639	3 3: 0.216886	68: 0.04285	46: 0.08181	82: 0.05172	20: 0.12199	39: 0.08148	6 99: 0.08730	97: 0.05597	84: 0.09322	2 41:
5	71: 0.08904	71: 0.02054	4 34: 0.0	06140 7	77: 0.03	731 3:	0.204752	25: 0.52197	11: 0.0451	3 100: 0.034	8 90: 0.04487	7 39: 0.19943	98: 0.04166	11: 0.07291	68: 0.04285	5 41: 0.10377	26: 0.0808	8 46: 0.06363	8 63: 0.05555	47: 0.07894	42:
6	34: 0.07894	63: 0.01851	1 22: 0.0	D4310 5	56: 0.03	571 1:	3: 0.15589	72: 0.52061	68: 0.0428	5 67: 0.0325	4 78: 0.04205	5 9: 0.120021	99: 0.03968	64: 0.04421	39: 0.03089	30: 0.08631	80: 0.04779	9 77: 0.03731	93: 0.04411	28: 0.05395	5 11:
7	88: 0.05040	86: 0.01764	4 55: 0.0	02419 4	46: 0.02	727 1	1: 0.05902	8: 0.515060	31: 0.0422	0 64: 0.0306	1 69: 0.03333	3 83: 0.05814	71: 0.03424	75: 0.04310	86: 0.02941	53: 0.07142	81: 0.03848	6 60: 0.03679	87: 0.03992	86: 0.05294	4 93:
8	46: 0.04545	82: 0.01724	4 77: 0.0	02238 7	79: 0.02	688 2:	0.056492	38: 0.50000	52: 0.0350	0 96: 0.0251	6 46: 0.02727	7 8: 0.057229	86: 0.02941	40: 0.03787	46: 0.02727	26: 0.06617	92: 0.0352	1 31: 0.03571	90: 0.03205	71: 0.04794	4 37:
9	83: 0.04263	51: 0.01541	1 26: 0.0	02205 9	97: 0.02	611 33	2: 0.04717	9: 0.491090	83: 0.0271	3 45: 0.0200	0 79: 0.02688	3 43: 0.05605	6 46: 0.02727	86: 0.02941	30: 0.02678	3 90: 0.04487	94: 0.03393	3 69: 0.03333	84: 0.02542	45: 0.04666	3 86:
10	84: 0.04237	52: 0.01500	0 42: 0.0	02058 1	1: 0.02	430 93	3: 0.04411	3: 0.479778	75: 0.0258	6 63: 0.0185	1 34: 0.02631	25: 0.05311	34: 0.02631	62: 0.02577	26: 0.02205	5 97: 0.04104	46: 0.02723	7 45: 0.03333	8 52: 0.02500	62: 0.04639	3 79:
11	11: 0.03819	67: 0.01479	9 60: 0.0	01948 6	53: 0.01	851 68	3: 0.04285	2: 0.478078	80: 0.0257	3 17: 0.0179	7 35: 0.02559	3 11: 0.05208	62: 0.02577	1: 0.025301	8: 0.021084	81: 0.03846	34: 0.0263	64: 0.03061	99: 0.02381	68: 0.04285	j 67:
12	100: 0.0307	93: 0.01470	0 83: 0.0	01938 1	7: 0.01	797 3	7: 0.04166	6: 0.475780	69: 0.0238	1 86: 0.0176	4 100: 0.0225	5 13: 0.03979	35: 0.02559	94: 0.02488	75: 0.02011	77: 0.03731	62: 0.02577	7 32: 0.02830	53: 0.02381	69: 0.03333	3 74:
13	95: 0.02674	68: 0.01428	3 20: 0.0	01890 8	32: 0.01	724 83	3: 0.03488	74: 0.47237	45: 0.0200	0 11: 0.0173	6 85: 0.02153	3 4: 0.036751	53: 0.02381	55: 0.02419	45: 0.02000	56: 0.03571	77: 0.02238	3 42: 0.02647	78: 0.02336	67: 0.03254	4 55:
14	74: 0.02486	78: 0.01401	63: 0.0	01851 5	52: 0.01	500 88	6: 0.02941	61: 0.46651	90: 0.0192	3 82: 0.0172	4 63: 0.01851	47: 0.03383	100: 0.0225	53: 0.02381	63: 0.01851	28: 0.03237	71: 0.02054	\$ 56: 0.02551	45: 0.02000	11: 0.03125	i 69:
15	53: 0.02381	98: 0.01388	37: 0.0	01785 9	33: 0.01	470 40): 0.02777	4: 0.457804	22: 0.0189	6 43: 0.0163	6 86: 0.01764	45: 0.03333	45: 0.02000	77: 0.02238	37: 0.01785	5 34: 0.02631	47: 0.01879	9 51: 0.02422	83: 0.01938	80: 0.02573	3 27:
16	78: 0.02336	81: 0.01282	2 36: 0.0	01750 6	S8: 0.01	428 48	6: 0.02727	5: 0.432528	40: 0.0176	7 79: 0.0161	2 11: 0.01736	§ 90: 0.03205	83: 0.01938	67: 0.02071	40: 0.01767	78: 0.02338	63: 0.0185	87: 0.02091	47: 0.01879	52: 0.02500	J 85:
17	26: 0.02205	48: 0.01153	3 82: 0.0	01724 9	38: 0.01	388 21	3: 0.02518	60: 0.43073	57: 0.0172	8 51: 0.0154	1 82: 0.01724	1: 0.022892	47: 0.01879	65: 0.01923	56: 0.01530	16: 0.02000	50: 0.0185	74: 0.01933	82: 0.01724	49: 0.02250) 71: 🧹
<																					>
Только сейч мона по пле просила эта	нас и только чо или ток л и девушкаН	для вас, на ицо.Любой lo ее превр	астоящі каприз атили в	ие идол з за вац з часы	ты совр ии симг	іемені ты!Сда	ности, про елайте с э	сто Боги ф той фоткой	отошопа. Т если можн	акое нельз о что типо я	я было обойт і нюхаю красі	и стороной, ивый цветок	смотрим и г . или че то в	поднимаем н этом роде F	настроение. ′омашки спр	Завожу боя ятались, по	н)Вырежете никли люти	: мну от сюд. киИстория у	а плизз (я с) же не сохран	тевого края) іила, что	ı). 🔶
Text	Doc.	ID -						W_lis	E			Numbe	r of docum	ante Luce							
prob	· pro	b Sentar	nalysis					D_list	В				for exp	iont 100		Boundary	ior probabi	lity: 0,001			

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.

	А	В	С	D	E	F	G	Н
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 *w*. 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: 100

Boundary for probability: 0,001

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.

	С	D	E	F	G	н	L. L.	J. J.	К	L	М	N	0	Р	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://yu	uzilla.livejo	urnal.com/	/922914.ht	ml	395	0,788783	53	54	yuzilla	http://y	/uzilla.livejo	urnal.com/	930651.hti	nl
3	alexlotov	http://al	exlotov.liv	ejournal.co	m/363587	.html	46	0,084615	62	43	yuzilla	http://y	/uzilla.livejo	urnal.com/	927854.hti	nl
4	yuzilla	http://yu	uzilla.livejo	urnal.com/	/927501.ht	ml	174	0,071782	3	25	yuzilla	http://y	/uzilla.livejo	urnal.com/	923209.hti	nl
5	alexlotov	http://al	exlotov.liv	ejournal.co	0m/364440	.html	294	0,059677	32	42	yuzilla	http://y	/uzilla.livejo	urnal.com/	927501.hti	nl
6	alexlotov	http://al	exlotov.liv	ejournal.co	om/359355	.html	176	0,054124	12	86	alexlotov	http://a	alexlotov.liv	ejournal.co	m/364282	.html
7	alexlotov	http://al	exlotov.liv	ejournal.co	om/360194	.html	66	0,052326	78	99	alexlotov	http://a	alexlotov.liv	ejournal.co	m/367512	.html
8	alexlotov	http://al	exlotov.liv	ejournal.co	om/361912	.html	59	0,04321	40	33	yuzilla	http://y	/uzilla.livejo	urnal.com/	925211.hti	nl
9	alexlotov	http://al	exlotov.liv	ejournal.co	0m/365095	.html	75	0,038889	13	28	yuzilla	http://y	/uzilla.livejo	urnal.com/	924058.hti	nl
10	alexlotov	http://al	exlotov.liv	ejournal.co	0m/357587	.html	258	0,038182	99	11	gutta_hor	http://g	gutta-honey	livejourna	l.com/3023	69.html
11	alexlotov	http://al	exlotov.liv	ejournal.co	m/366522	.html	267	0,030357	81	100	alexlotov	http://a	alexlotov.liv	ejournal.co	m/367668	.html
12	gutta_hor	http://gu	utta-honey.	livejourna	l.com/3023	369.html	148	0,026786	33	53	yuzilla	http://y	/uzilla.livejo	urnal.com/	930544.hti	nl
13	yuzilla	http://yu	uzilla.livejo	urnal.com/	/928575.ht	ml	35	0,025862	10	63	alexlotov	http://a	alexlotov.liv	ejournal.co	m/358294.	.html
14	yuzilla	http://yu	uzilla.livejo	urnal.com/	/922308.ht	ml	311	0,025	63	83	alexlotov	http://a	alexlotov.liv	ejournal.co	m/363461	.html
15	yuzilla	http://yu	uzilla.livejo	urnal.com/	/925448.ht	ml	43	0,02459	36	90	alexlotov	http://a	alexlotov.liv	ejournal.co	m/365095	.html
16	alexlotov	http://al	exlotov.liv	ejournal.co	om/363461	.html	129	0,02349	11	49	yuzilla	http://y	/uzilla.livejo	urnal.com/	929280.hti	nl
17	yuzilla	http://yu	uzilla.livejo	urnal.com/	/931203.ht	ml	88	0,023148	46	3	gutta_hor	http://g	gutta-honey	livejourna	l.com/2993	20.html
18	yuzilla	http://yu	uzilla.livejo	urnal.com/	/930817.ht	ml	43	0,022727	61	82	alexlotov	http://a	alexlotov.liv	ejournal.co	m/363073.	.html
19	yuzilla	http://yu	uzilla.livejo	urnal.com/	/924058.ht	ml	138	0,022152	2	37	yuzilla	http://y	/uzilla.livejo	urnal.com/	926282.hti	nl
20	alexlotov	http://al	exlotov.liv	ejournal.co	om/364963	.html	187	0,021845	77	98	alexlotov	http://a	alexlotov.liv	ejournal.co	m/367193	.html
21	yuzilla	http://yu	uzilla.livejo	urnal.com/	/926122.ht	ml	218	0,018519	52	93	alexlotov	http://a	alexlotov.liv	ejournal.co	m/365861	.html
22	alexlotov	http://al	exlotov.liv	ejournal.co	0m/358294	.html	9	0,017241	39	68	alexlotov	http://a	alexlotov.liv	ejournal.co	m/359528	.html
23	alexlotov	http://al	exlotov.liv	ejournal.co	0m/357885	.html	253	0,016544	48	78	alexlotov	http://a	alexlotov.liv	ejournal.co	m/362198	.html
24	yuzilla	http://yu	uzilla.livejo	urnal.com/	/928978.ht	ml	143	0,016447	60	81	alexlotov	http://a	alexlotov.liv	ejournal.co	m/362988	.html
25	alexlotov	http://al	exlotov.liv	ejournal.co	0m/363073	.html	11	0,016129	70	58	yuzilla	http://y	/uzilla.livejo	urnal.com/	931721.hti	nl
26	alexlotov	http://al	exlotov.liv	ejournal.co	om/367193	.html	14	0,013889	30	32	yuzilla	http://y	/uzilla.livejo	urnal.com/	925144.hti	nl

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

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Год: 0.04	<mark>18</mark> модель: О.	С фотографи	информаци	человек: О.	весь: 0.024	хостинг: 0.0	(зак: 0.0063	я: 0.023270	ребенок: О.	индонезия:	сотовый: О	считаться:	из: 0.01236	офис: 0.00	5 инвалид: О
2 дерево: І).0 reuters: 0.0	12 сделать: О.	лошадь: 0.0	довольно: (то: 0.02301	сайт: 0.004	девочка: 0.1	легкий: 0.0	граница: О.	западный:	бизнес: 0.0	загадываты	желание: О	книга: 0.0)! полицейск
3 бумага: І).0 показ: 0.0	2 парк: 0.008	астро: 0.00	схема: 0.01	быть: 0.021	наряд: 0.00) мальчик: О.	нужно: 0.00	родители: (ява: 0.0043	бессмысле	исполнение	павлин: 0.0	google: 0.0	С февраль: С
4 принима	гь: коллекция	т националы	информаци	себе: 0.014	a: 0.019352	надежный:	слово: 0.00	плевра: 0.0	должный: С) храм: 0.004	избиратель	удача: 0.00	просить: О.	окулист: О	.(ла-пас: 0.0
5 ислам: О	01 неделя: О.	0 эта: 0.0071	трясина: О.	лицо: 0.012	как: 0.0190	день: 0.003	поддержив	пневмотор	если: 0.009	восточный:	мертвый: С	верона: 0.0	находиться	шрифт: 0.0	О боливия: О
6 ap: 0.006	55 упасть: 0.0) фотограф:	книга: 0.00	личность: О	который: О.	наоми: 0.00	назад: 0.00	грудной: О.	ты: 0.00983	океан: 0.00	вой: 0.0022	самый: 0.0	(житель: 0.0	кредит: 0.0	0 mercado: 0
7 photo: 0.1	006 мода: 0.01	Зафрика: 0.0	николь: 0.0	конфликт: (я: 0.017392	премия: 0.0	специалист	быстро: 0.0	родитель: (пляж: 0.003	покер: 0.00	колонна: О.	место: 0.00	телеведуц	(david: 0.00
3 имя: 0.00	58 во: 0.0087	1 отдыхать: (избыток: О	общество:	но: 0.01602	встречать:	содеять: 0.	ребро: 0.00	делать: 0.0	путих: 0.002	средний: О.	алтарь: 0.0	башня: 0.00	рознить: О	і фотографи
Экраска: І).0 время: 0.0	IC заповедни	перегрузка	а разный: 0.0	они: 0.0132	телек: 0.00	одри: 0.002	больная: О.	дитя: 0.007	nanya: 0.00	вводить: О.	менгир: 0.0	мост: 0.006	набирать:	Creuters: 0.0
) взять: О.	ОО подиум: О.	0 медведь: О	постоянная	чувство: 0.0	или: 0.0116	доминанта	: диагноз: 0.	клетка: 0.0	себе: 0.007	кавах: 0.00	писать: 0.0	мечеть: 0.0	где: 0.0067	труд: 0.002	2 автомобил
1 мир: 0.00	ІЗЄ творение:	Синдий: 0.00	успевать: (т: 0.005867	человек: О.	прием: 0.00	перерезыв	она: 0.0041	т: 0.006570	джокьякар	дух: 0.0022	каменный:	один: 0.006	время: 0.0	С путь: 0.007
2 религия:	0.(падение: 0), рассказыв	стараться:	идентичнос	для: 0.0110	мечтать: О.	бросаться:	игла: 0.004	e: 0.005916	боробудур:	заключаты	честь: 0.00	дом: 0.0062	глазок: 0.0) набрасыва
3 акрилові	ый: после: 0.0	0 род: 0.0043	вовремя: О	причина: 0.1	так: 0.0109	марш: 0.00	апроцент: 0.1	воздух: 0.0	(нужно: 0.00) тропически	разум: 0.00) знамениты	улица: 0.00	партия: 0.0	О костыль: С
4 дейв: 0.0	02 оступатьс	я ттица: 0.00	ужасный: С	какой-то: О	ero: 0.0109	звезда: 0.0) идентичнос	маленький	что-то: 0.00	курорт: 0.0	устремлен	кора: 0.003	век: 0.0045	окружать:	(плаз: 0.003
б организа	щи платье: 0.1	DI южный: 0.0	засасыват	дефект: 0.0	чтобы: 0.01	например:	вести: 0.00	одышка: О.	(разный: О.С	чубайс: 0.0	заток: 0.00	джульетта:	построить:	глянцевый	і: столица: О
блэйки: І).0 лондонски	ю под: 0.0035	усталость:	начинать: О	из: 0.00989	джоди: 0.00	три: 0.0021	отмечать: (играть: 0.0	(http: 0.0022	оседать: О.	сучжоу: 0.0	колодец: О.	паника: О.	О перегораж
7 фольга:).0 дизайнер:	Cinational: 0.0	временить	ты: 0.00461	свой: 0.008	давно: 0.00) история: 0.1	жить: 0.003	воспитание	загонять: С	перехлесть	любовь: 0.0	камень: 0.0	обнародов	а патрульны
3 ко: 0.002	95 пытаться:	(тупик: 0.00	хозяйка: О.	иметь: 0.00	же: 0.00810	эрин: 0.002	2 обычный: О	пункция: О.	стол: 0.005	православн	екатеринбу	написать: (дерево: 0.0	тренировк	а направлят
Э стоун: О.	00 нью-йорко	к слон: 0.003	дочь: 0.002	эмоция: 0.0	один: 0.007	мил: 0.0028	ж: 0.002189	год: 0.0034	ситуация: С	император:	этнический	пояс: 0.003	достоприм	янг: 0.001	1 тысяча: 0.0
) tip: 0.002	23 представл	е лес: 0.0028	снижение:	осознавать	другой: 0.0	цапля: 0.00	способный:	рахель: 0.0	довольно: І	иерархия: (alone: 0.00*	благополуч	район: 0.00	соседка: С), площадь; (
1 q: 0.0022	33 запутыват	ъ сова: 0.002	планирова	жизнь: 0.00	мы: 0.0075	бекап: 0.00) великобрит	пар: 0.0034	такой: 0.00	прикладной	пиксельно	город: 0.00	местный: О	умственны	ы видео: 0.00
2 джексон	0. suzanne: 0	.(остер: 0.00	грязь: 0.00	e: 0.004202	если: 0.007	awards: 0.0	(поставлять	давление: (нельзя: 0.0	звоевали: (навевать: (ритуал: 0.0	легенда: О.	гром: 0.00	1 главный: С
3 али: 0.00	22 plunkett: 0.	(неподалеку	грэм: 0.001	возникать:	вы: 0.00742	раздевать:	развлекать	синагога: С	друг: 0.004	nino6pect: I	креатива: (центр: 0.00	сегодняшн	юсуф: 0.00	1 разбивать
4 нация: О.	00 помогать:	Ссеренгети:	однажды: (понимание:	сам: 0.0071	портал: 0.0	I мумбаи: 0.0	край: 0.002	надо: 0.004	волейбольн	гэпа: 0.001	прикасатьс	вод: 0.0034	портман: (), который: С
5 включаті	: Свставать:	0 рак: 0.0028	вытягивать	тело: 0.003	такой: 0.00	гвинет: 0.0	(Tar: 0.00114	больной: О.	ну: 0.00395	отряд: 0.00	калека: 0.0	ица: 0.0023	некоторые:	случайный	і: армас: 0.0
1															
	_								_						

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

3.4.2.1. Export sort the results in csv file format.

	А	В	С	D	E	F	G	Н	I.	J	К	L	М
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	стрит	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	жених	0,008893	нестабилі	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	качествен	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	младене	0,00536	динамиче	0,001293	быстрый	0,002357	reuters	0,008166	надежны	0,003681	житель
12	афганский	0,004853	религиоз	0,00536	агонизиро	0,001293	pasquini6	0,001235	автомоби	0,006549	магазин	0,003681	исполнен
13	демонстр	0,004853	дитя	0,004771	замешате	0,001293	nino6pect	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	построить
15	сша	0,004853	ортодокса	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	jesseглазг	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	южноаме	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	находиты
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

Topic Distribution the button . 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 the probable words graph also shows 6 most in each topic. Topic distribution according to words



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 tmlda (e.g., 2_step_test.tmlda). 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.tmproj), which will contain the paths to the source data (tmlda) 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.tmlda</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: 40 This parameter is used in calculations with a fixed number of topics.

This parameter is used in calculations with a fixed number of topic

2. Number of itarations. Number of iteration: 100

. 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:

Contemporation and the second	+	-	Ξ X
Lematization (Russian language) View of TMLDA file Gibbs LDA sampling BigARTM Kullback-Leibler Distance, Jaccard index			
🖸 - 🖸 🗖 🔟 - 🔍 🔍 📑 🔘			
Number of iteration: 100	Save step: 10)	
Word diets.:			
Number of topies: 40			
Multiple topics simulation:			
Number of documents: 0 Number of words in documents: 0 End of topics: 200	Topic step:	10	

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 tmlda (e.g., 2_step_test.tmlda). 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.tmproj), which will contain the paths to the source data (tmlda) 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 'tmlda' 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.

Words -	· Topics Distributions										~
	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	Grivtsov	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			
	1.										
Hide	selected row	Reset hi	dden rows		Export to Exc	el	Number of	i words for (export: 100	Boundary for probability:	0,0000

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.

1			1					
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 \square . 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.

😍 TopicMiner RNF ver. 1 (64 bit) LINIS labo	oratory, HSE		_ = ×
Lematization (Russian language) View of TMLDA file Gi	ibbs LDA sampling BigARTM	Kullback-Leibler Distance, Jaccard index	
First duster solution		Second duster solution	
Joint Dictionary	Distance		Depth for Jaccard Index: 100 \$ Export to CSV
Status:	File name:		0%

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

😍 TopicMiner RNF ve	er. 1 (64 bit)) LINIS la	borator	y, HSE						. =	х
Lematization (Russian langua	age) View of 1	TMLDA file	Gibbs LD.	A sampling	BigARTM	Kullback-Leibl	er Distance, Jai	ccard index			
First cluster solution					-	Second d	uster solution				5
CRC32	Word T	fopic 1	Topic 2	Торі 🔨							
1 <mark>1308918612</mark> в	; 0,0)2077 0,	00006	0,00			1				
2 -136200747 и	0,0	00008 0,	00006	0,00			2				
3 1001429846 н	ie 0,0	00008 0,	00006	0,00							
4 1272606681 н	ia 0,0	00008 0,	00006	0,00							
<u>5</u> -776787691 ч	то 0,0	00008 0,	00006	0,00	2,						
6 -601576253 a	то 0,0	00008 0,	00006	0,00	_						
7 -893730166 c	: 0,0	00008 0,	00006	0,00							
8 1794372915 в	есь 0,0	00008 0,	00065	0,00	CSV						
9 1631079488 TI	o 0,0	00008 0,	00006	0,00							
10 1171490660 6	іьть 0,0	00008 0,	00006	0,00							
11 507920551 o	н 0,0	00008 0,	00006	0,00							
12 -116046694 к	ак 0,0	00008 0,	00006	0,00 🗸 🔰	Ο						
<				>	-						
Joint Dictionary	Kı	ullback - Leib	oler Distanc	e					pic to Topic		
										~	
1											
2											
								<		- ×	
									D		
									Jaccard index: 100	\$	
										1	
	<u> </u>								Export to CSV		
Status:			File	name:					0%		

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

TopicMi	iner RNF v	ver. 1 (64	bit) LINIS	laborat	tory, HSE								
ematization (Russian lang	uage) View	of TMLDA fi	le Gibbs	LDA sampling	BigARTM	Kul	lback-Leibler D	istance, Jac	card index			
First duster s	solution					-		Second clust	er solution				
	CRC32	Word	Topic 1	Topic 2	2 Topi 🔨				CRC32	Word	Topic 1	Topic 2	Topi 🔨
1	1308918612	в	0,02077	0,00006	0,00			1	1308918612	в	0,00093	0,00006	0,00
2	-136200747	и	0,00008	0,00006	0,00			2	-136200747	и	0,00008	0,00006	0,00
3	1001429846	не	0,00008	0,00006	0,00			3	1001429846	не	0,00008	0,00006	0,00
4	1272606681	на	0,00008	0,00006	0,00			4	1272606681	на	0,00008	0,00006	0,00
5	-776787691	что	0,00008	0,00006	0,00	$\sum_{i=1}^{n}$		5	-776787691	что	0,00008	0,00006	0,00
6	-601576253	это	0,00008	0,00006	0,00			6	-601576253	это	0,00008	0,00006	0,00
7	-893730166	с	0,00008	0,00006	0,00			7	-893730166	с	0,00008	0,00006	0,00
8	1794372915	весь	0,00008	0,00065	0,00	CSV		8	1794372915	весь	0,00008	0,00069	0,00
9	1631079488	то	0,00008	0,00006	0,00			9	1631079488	то	0,00008	0,00069	0,00
10	1171490660	быть	0,00008	0,00006	0,00			10	1171490660	быть	0,00008	0,00006	0,00
11	507920551	он	0,00008	0,00006	0,00	_		11	507920551	он	0,00008	0,00006	0,00
12	-116046694	как	0,00008	0,00006	0,00 🗸			12	-116046694	как	0,00008	0,00006	0,00 🗸
<					>			<					>
Joint Diction	hary		Kullback - L	eibler Dist.	ance						ppic to Topic		
	1	1											~
<u> </u>	1												
· · ·	4												. ×
											Depth for Jaccard in	100	\$
											Justice of the	tient.	
<		>									Expo	rt to CSV	
tatus:				F	ile name:							0%	

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 2. 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.

🜻 TopicMi										+		. = X
Lematization (F	Russian language) 🛛 Viev	v of TMLDA fi	ile Gibbs L[DA sampling	BigARTM Ku	llback-Leibler	Distance, Jaco	card index				
First cluster s	olution					Second clus	ter solution					
	CRC32 Word	Topic 1	Topic 2	Торі 🔨			CRC32	Word	To	pic 1	Topic 2	Topi 🔨
1	1308918612 <mark>в</mark>	0,02077	0,00006	0,00		1	1308918612	в	0,00	093	0,00006	0,00
2	-136200747 и	0,00008	0,00006	0,00		2	-136200747	и	0,00	008	0,00006	0,00
3	100142984€ не	0,00008	0,00006	0,00		3	1001429846	не	0,00	008	0,00006	0,00
4	1272606681 на	0,00008	0,00006	0,00		4	1272606681	на	0,00	008	0,00006	0,00
5	-776787691 что	0,00008	0,00006	0,00	2	5	776787691	что	0,00	008	0,00006	0,00
6	-601576253 это	0,00008	0,00006	0,00		6	-601576253	это	0,00	008	0,00006	0,00
7	-893730166 c Topic	: Miner							x	908	0,00006	0,00
8	1794372915 ве									208	0,00069	0,00
9	1631079488 то			Calculat	ion of KI B (listance				208	0,00069	0,00
10	1171490660 бы									208	0,00006	0,00
11	507920551 он				108/					208	0,00006	0,00
12	-116046694 ка				1976					208	0,00006	0,00 🗸
<												>
Joint Diction	ary	Kullback - I	eibler Distar	nce				T	opic to	Topic		
			Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	^				^
1		Topic 1	0,114321	1,34517	7 0,662613	0,726218	1,031062					
2	2	Topic 2	1,357674	0,19486	0,829697	0,902407	1,350087	<				~
		Topic 3	0,984088	1,09490	0,392517	0,460537	0,948441		Dept	h for	100	
		Topic 4	1,279706	1,33173	0,722061	0,751408	1,233179	~	Jaco	ard in	itea: 100	•
<	>	<				1	2			Expo	rt to CSV	
Status:			File	name:							0%	

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

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

Joint Diction	ary		Kullback	- Leibler Di	stance							Topic to Topic	
	CRC32	Wore ^		Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	^	38 89,4487772173344 28 39 68 988183622686 19	^
	1 984961486	a	Topic 1	95,942	52,245	76,477	74,219	63,397	63,414	55,234		40 95,1161384148887 40	
	2 -145893245	abby abstr	Topic 2	51,802	93,082	70,545	67,964	52,071	58,780	49,725		4 3	~
	4 109801287	4 adhe	Topic 3	65,064	61,130	86,065	83,651	66,330	73,477	61,583		Douth for	
	5 442695966	aero:	Topic 4	54,570	52,723	74,366	73,325	56,221	62,524	51,369		Jaccard index: 100	
<	21710040000	>	<							>	Ť	Export to CSV	

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

	А	В	С	D	E	F	G	Н	1	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:

Topic to Topic		
38 89,4487772173344 28		^
39 68,988183622686 19		
40 95,1161384148887 40		
		\mathbf{v}
<	≥	

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: Depth for Jaccard index: 100

. 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

Export to CSV...

. In the appeared window it is

necessary to specify a file name. An example of such an unloading is shown in Figure 5.6.

	А	В	С	D	E	F	G	Н
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	почувство	немало	почувствс
4	кабул	afp	невеста	невеста	р	населени	также	населени
5	afp	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	отчетлив	быстрый	отчетлив
9	демонстр	тысяча	девушка	девушка	интересо	религия	смеяться	религия
10	полицейс	полицейск	родители	религиоз	род	такать	стрит	такать
11	тысяча	демонстра	религиоз	младенег	сильный	источникт	дурак	источникт
12	сша	афганский	младенег	ортодокса	замешате	действо	создавать	действо
13	демонстр	getty	дитя	живой	сантьяго	погон	источники	погон
14	афганский	нато	реборн	реборн	креативн	празднов	интерпре	празднов
15	военный	ар	ортодокса	сначала	канители	сеть	адекватні	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	закрывать	объезжат	закрывать
20	оскорбля	marai	также	будущее	диктатор	логотип	подписые	логотип
21	ар	атаковать	мама	малыш	альфред	купиться	safina	купиться
22	километр	сша	мальчик	родители	тема	интуиция	ужасный	интуиция
23	провинци	провинция	свадебны	договор	лет	ложиться	гои	ложиться
24	баграм	военный	этап	знакомсте	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 $N \ge 1$ of the first decision and the topic $N \ge 1$ 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 systemQuantumGIS(downloadthehttp://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.

🜻 Docu	Documents - Topics Distribution																
	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	1	люблю	244159479	Wall 244159	post 2	22.02.2014	Олеся	Малодзинск	Улан-Удэ	Бурятия							
3	2	С днем рож	15538973	Wall 124987	post 33	4.1.2014 9:	Alsu	Asapova	Набережны	Татарстан							
4	3	Новый год?	150682985	Wall 124987	post 27	12.31.2013	Lyudmila	Trofimova	Набережны	Татарстан							
5	4	?? Вам откр	137183488	Wall 124987	post 36	4.20.2014	Lyuda	Gorelova	Киров	Кировская (
6	5	?? Вам откр	137183488	Wall 124987	post 35	4.1.20146:	Lyuda	Gorelova	Киров	Кировская (
7	6	лишь 2% лн	56556171	Wall 565561	post 7348	08.05.2013	Виктория	Ломанова	Улан-Удэ	Бурятия							
8	7	что или кто	0	Wall 124987	comments o	7.22.20138	8 Vladislav	Enoktaev	Набережны	Татарстан							
9	8	Отправлено	237671145	Wall 124987	post 34	4.1.2014 2:	Ruslan	Enoktaev	Москва								
10	9	?Отправлен	97170510	Wall 124987	post 8	5.8.2013 1	l Ruslan	Enoktaev									
11	10	С НАСТУПА	22306292	Wall 124987	post 28	12.31.2013	Vera	Yasnikovska	Набережны	Татарстан							
12	11	С НАСТУПА	63412409	Wall 124987	post 26	12.30.2013	Irinochka	Aldemirova	Уржум	Кировская (
13	12	не прикалы	124987410	Wall 124987	post 22	7.25.20132	2 Vladislav	Enoktaev	Набережны	Татарстан							
14	13	?Отправлен	137183488	Wall 124987	post 31	2.14.20148	Lyuda	Gorelova	Киров	Кировская (
15	14	Братишка п	200339270	Wall 124987	post 2	3.18.2013 9	Alexander	Makarov	Тюмень	Тюменская							
16	15	?Отправлен	97170510	Wall 124987	post 7	5.5.2013 1:	Ruslan	Enoktaev									
17	16	мы ждем др	56556171	Wall 565561	post 7345	07.05.2013	Виктория	Ломанова	Улан-Удэ	Бурятия							
18	17	?Отправлен	97170510	Wall 124987	post 6	5.4.2013 10	Ruslan	Enoktaev									
19	18	Лови позит	171916464	Wall 173311	post 4	5.18.2013	l Azalia	Giniatullina	Бавлы	Татарстан							×
																	_
																	_
																	1
Hid	e ted	Hide by	0.5		Reset	Doc	- metada 👔	Doc ID	- 6	0015	Mar	nher of dee	aumonte for	wmmth 1	00		
rov	V V	threshold	0,5		hidden row		prob	prob	P.	2213	5101		achiterito 10	Tanta I	••		

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 **ID**. 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).

	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
1																
2 1	L	люблю	244159479	Wall 244159	post 2	22.02.2014	Олеся	Малодзинск	Улан-Удэ	Бурятия						
3 2	(С днем рож	15538973	Wall 124987	post 33	4.1.2014 9:	Alsu	Asapova	Набережны	Татарстан						
4 3	ł	Новый год?	150682985	Wall 124987	post 27	12.31.2013	Lyudmila	Trofimova	Набережнь	Татарстан						
5 4	1	?? Вам откр	137183488	Wall 124987	nost 36	4.20.2014	11 vuda	Gorelova	Киров	Кировская						
6 5	1	?? Вам откр	13718348	Export	to QQIS								×			
7 6	L	лишь 2% лн	56556171													
8 7		что или кто	0	Topic:		1										
9 8	(Отправлен	23767114			100										
10 9		?Отправлен	97170510	Docume	nts:	100										
11 10	(С НАСТУПА	22306292			E 112										
12 11	(С НАСТУПА	63412409	Field Wi	in regions:	Field /		-								
13 12		не прикалы	12498741													
14 13	1	?Отправлен	13718348	DBF File	e Name:											
15 14	8	Братишка п	20033927													
16 15	-	?Отправлен	97170510	Regions	Matching	File:										
17 16		мы ждем др	56556171													
18 17		?Отправлен	97170510													
19 18	1	Лови позиті	17191646		-	Export				Cancel						

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'.

Export to QQIS		X
Topic: Documents:	1 100	
Field with regions:	Field 7	
DBF File Name:	E:\TopicMiner_RNF_2015\Vk_data_example\regions2010_sib_5.dbf	
Regions Matching F	File: E:\TopicMiner_RNF_2015\Vk_data_example\vktm.dbf	
	Export Cancel	

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.

<i>4</i>	QGIS2.12.1-Lyon - full_project	- 🗆 ×
Проекты Правка Вид Слой Установки Модули В	🔏 Свойства слоя — regions2010_sib_5 Стиль	? ×
Image: Second	Obuve Image: ObuveLoil 3H8k Image: ObuveLoil 3H8k Image: ObuveLoil 3H8k	ь библиотеку
Yes Yes <	Спрененные Сохранить Сохранить	лянгельно • = 0 • • • Справка

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.

×.	Свойства слоя
🔀 Общие	🔁 Обычный знак 🔍 🔻
V	Обычный знак
💽 Стиль	👱 Уникальные значения
(аbc Подписи	Градуированный знак Правила Точки со смещением
Поля	Инвертированные полигоны Создание теплокарт

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).

sa an	Ce	зойства слоя — regions20	10_sib_5 Стиль
🔀 Общие	🍷 Уникальные значения	•	
😻 Стиль Г			3
(abc) Подписи	HAK PERIMETER REGION		чайные цвета
Поля	GOST779B 3HaK UNIVERSAL Topic		
🎸 Отрисовка			
🗭 Вывод			
💭 Действия			
• ┥ Связи			
💯 Диаграммы			
🥡 Метаданные			
8 Переменные			
	Классифицировать Добави	ить Удалить Удалить	все

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).

Á	Свойства слоя — regions2010_sib_5 Стиль ?	×
🔀 Общие	2 Уникальные значения	
😻 Стиль	Tone Topic	
(abc) Подписи	Знак Изменить Градиент Случайные цвета 🔻 Изменить 🖸 Об	ратить
Поля	Знак $ abla$ Значение Легенда	
🎸 Отрисовка	x 0 0 x 2.692000009 2.692000009	
🧭 Вывод	143.9300417 143.9300417 153.6289691 153.6289691	
🔅 Действия		
• Связи		
阿 Диаграммы		
🥡 Метаданные		
8 Переменные		
	Классифицировать Добавить Удалить Удалить все Дополните	ельно 🔻
	• Отрисовка	
	Прозрачность 0	
	Режим смешивания слоя Обычный 🔻 Режим смешивания объектов Обычный	_
	Draw effects	
	Стиль • ОК Отмена Применить Сп	равка

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

×.			Свойства слоя -	- regions20)10_sib_	5 Стиль		
Х Общие	붙 Уник	альные значения	•			_		
😻 Стиль	Поле Тор	pic			3	J		
авс Подписи	Знак	Из	менить	Градиент	Случайны	ые цвета	-	Изменить
						Greys		
Поля	Знак	🗸 Значение	Легенда			OrRd		
	×					Oranges		
	×	0	0			PRGn		
🔷 Отрисовка	×	2.692000009	2.692000009			PiYG		
<u> </u>	×	143.9300417	143.9300417			PuBu		
- Вывод	: *	153.6289691	153.6289691			PuBuGn		
i	×					PuOr		
🔆 Действия						PuRd		
1						Purples		
• 🤤 Связи								
	8							
💯 Диаграммы								

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 N_2 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 lematized dictionary. Preparing the dictionary is as follows. The 'Lematization' tab provides the following option (see Figure 7.1)

4 Developing	sentiment binary dictionary	
Input file:		Lang: Russian 🗸
Result file:		Run process

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).

4 Developing	sentiment binary dictionary	
Input file:		ang: Russian 🔽
Result file:		Bi Russian (UTF)
Succession states		
r		English (UTF)
	0%	English (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:

Sentiment dict:	E:\TopicMiner RNF 2015\TopicMin	٦
-----------------	---------------------------------	---

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 Sentiment. As a result, the tonality calculation starts. However, in the case of an unconnected dictionary, then a warning window will appear (see below):

Warning		X
	Dictionary is not specified.	
	ОК	

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

Topic Miner	
Casting at an electric in any second	
Senument analysis in progress	
25%	

오 Words v	vith high probability							_ = ×
	1	2	3	4	5	6	7	8 ^
1	вот: 0.013934	млн: 0.014051	русская: 0.011952	объявлять: 0.006512	скатерть: 0.001542	http: 0.019630	россия: 0.013715	корабль: 0.0159:
2	фонд: 0.009691	российский: 0.011751	ракета: 0.008167	шутка: 0.006022	заезжать: 0.001240	ru: 0.013461	заявлять: 0.005458	проект: 0.01125:
3	только: 0.009521	федерация: 0.009573	самолет: 0.006420	апрель: 0.005408	ой: 0.001240	www: 0.010297	президент: 0.005344	удк: 0.007285
4	бы: 0.008333	народ: 0.009573	км: 0.006420	статья: 0.005408	мужик: 0.001240	news: 0.006343	рф: 0.005155	линкор: 0.00662-
5	деньга: 0.007485	рф: 0.007273	истребитель: 0.005983	страна: 0.004673	штатный: 0.001240	com: 0.005078	страна: 0.004852	тип: 0.006293
6	восток: 0.006976	россия: 0.007152	километр: 0.005255	время: 0.004427	жопа: 0.000937	христиания: 0.003812	также: 0.004587	мм: 0.005853
7	россия: 0.006976	русский: 0.006910	группа: 0.004673	черви: 0.004427	водолей: 0.000937	парковка: 0.003496	партия: 0.004549	фрунз: 0.005742
8	нужный: 0.006806	миграционный: 0.006426	миг: 0.004091	эта: 0.003937	пуаро: 0.000937	html: 0.003338	российский: 0.004511	флот: 0.005522
9	если: 0.006806	трудовой: 0.006305	рэб: 0.004091	журнал: 0.003814	разность: 0.000937	международный: 0.0028	путин: 0.004398	морской: 0.0055
10	давать: 0.006466	прирост: 0.006063	идти: 0.003800	спагетти: 0.003691	обижать: 0.000937	беларусь: 0.002863	глава: 0.004170	скорость: 0.005(
11	пономарев: 0.00629	работодатель: 0.005942	пво: 0.003654	the: 0.003324	лесной: 0.000937	двор: 0.002863	власть: 0.004132	вариант: 0.0048(
12	правда: 0.006297	минус: 0.005700	зур: 0.003363	тот: 0.003324	ока: 0.000937	дорожка: 0.002705	акция: 0.003943	при: 0.004640
13	сказать: 0.006127	составлять: 0.005579	дальность: 0.003217	эфир: 0.003201	задний: 0.000937	город: 0.002705	будет: 0.003905	м: 0.004420
14	сибирь: 0.005109	часть: 0.005337	удар: 0.003217	bell: 0.003078	торгово-промышленный: 0.0	улица: 0.002705	развитие: 0.003867	новое: 0.004199
15	мень: 0.004769	русская: 0.005216	хорнетов: 0.002926	новость: 0.003078	гладкость: 0.000635	житель: 0.002705	москва: 0.003829	тонна: 0.003979
16	американский: 0.00	статья: 0.004732	корабль: 0.002926	число: 0.003078	побояться: 0.000635	v: 0.002547	принимать: 0.003792	восстановление
17	грант: 0.004769	район: 0.004369	соединение: 0.002926	после: 0.002833	щегол: 0.000635	копенгаген: 0.002547	правительство: 0.003754	ход: 0.003758
18	ресурс: 0.004769	работник: 0.004248	бой: 0.002781	новое: 0.002710	финал: 0.000635	s: 0.002388	решение: 0.003678	вмф: 0.003648
19	организация: 0.0045	компенсация: 0.004248	супер: 0.002781	большой: 0.002710	мыть: 0.000635	id: 0.002230	должный: 0.003564	сила: 0.003538
20	дальний: 0.004599	проезд: 0.004006	патруль: 0.002781	телеканал: 0.002710	выбиваться: 0.000635	сша: 0.002230	предлагать: 0.003413	завод: 0.003428
21	вы: 0.004430	численность: 0.004006	cy: 0.002781	многие: 0.002710	бля: 0.000635	наркотик: 0.002072	совет: 0.003413	крейсер: 0.0033
22	гражданский: 0.004	расход: 0.004006	целить: 0.002489	несколько: 0.002588	талон: 0.000635	провод: 0.002072	проц: 0.003375	башня: 0.003318
23	японский: 0.004430	север: 0.003885	воздух: 0.002489	новый: 0.002588	рубашка: 0.000635	магазин: 0.002072	система: 0.003337	подводной: 0.00
24	общество: 0.003751	конституционный: 0.003885	гарри: 0.002344	появляться: 0.002588	рогопилы: 0.000635	жилье: 0.001914	группа: 0.003110	главное: 0.0030
25	тема: 0.003751	население: 0.003885	рлс: 0.002344	письмо: 0.002465	тормозной: 0.000635	org: 0.001914	гражданин: 0.003072	уза: 0.002987
<	1				i			>
	S	entiment Sentiment to Ex	cel Number of word	is for export: 100	Boundary for prot	pability: 0,001	Colors	

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)..

	0	1	2	3	4	5	6
1	вот: 0.013934: 0	млн: 0.014051: 0	русская: 0.011952: 0	объявлять: 0.006512: 0	скатерть: 0.001542: 0	http: 0.019630: 0	россия: 0.013715
2	фонд: 0.009691: 0	российский: 0.011751: 0	ракета: 0.008167: 0	шутка: 0.006022: 0	заезжать: 0.001240: 0	ru: 0.013461: 0	заявлять: 0.0054
3	только: 0.009521: 0	федерация: 0.009573: 0	самолет: 0.006420: 0	апрель: 0.005408: 0	ой: 0.001240: 0	www: 0.010297: 0	президент: 0.005
4	бы: 0.008333: 0	народ: 0.009573: 0	км: 0.006420: 0	статья: 0.005408: 0	мужик: 0.001240: 0	news: 0.006343: 0	рф: 0.005155: 0
5	деньга: 0.007485: 0	рф: 0.007273: 0	истребитель: 0.005983: 0	страна: 0.004673: 0	штатный: 0.001240: 0	com: 0.005078: 0	страна: 0.004852
6	восток: 0.006976: 0	россия: 0.007152: 0	километр: 0.005255: 0	время: 0.004427: 0	жопа: 0.000937: -2	христиания: 0.003812: 0	также: 0.004587:
7	россия: 0.006976: 0	русский: 0.006910: 0	группа: 0.004673: 0	черви: 0.004427: 0	водолей: 0.000937: 0	парковка: 0.003496: 0	партия: 0.004549
8	нужный: 0.006806: 1	миграционный: 0.006426: 0	миг: 0.004091: 0	эта: 0.003937: 0	пуаро: 0.000937: 0	html: 0.003338: 0	российский: 0.00
9	если: 0.006806: 0	трудовой: 0.006305: 0	рэб: 0.004091: 0	журнал: 0.003814: 0	разность: 0.000937: 0	международный: 0.002863: 0	путин: 0.004398:
10	давать: 0.006466: 0	прирост: 0.006063: 0	идти: 0.003800: 0	спагетти: 0.003691: 0	обижать: 0.000937: -1	беларусь: 0.002863: 0	глава: 0.004170:
11	пономарев: 0.00629	работодатель: 0.005942: 0	пво: 0.003654: 0	the: 0.003324: 0	лесной: 0.000937: 0	двор: 0.002863: 0	власть: 0.004132
12	правда: 0.006297: 0	минус: 0.005700: 0	зур: 0.003363: 0	тот: 0.003324: 0	ока: 0.000937: 0	дорожка: 0.002705: 0	акция: 0.003943:
13	сказать: 0.006127: (составлять: 0.005579: 0	дальность: 0.003217: 0	эфир: 0.003201: 0	задний: 0.000937: 0	горад: 0.002705: 1	будет: 0.003905:
14	сибирь: 0.005109: 0	часть: 0.005337: 0	удар: 0.003217: 0	bell: 0.003078: 0	торгово-промышленный: 0.000635: 0	улица: 0.002705: 0	развитие: 0.0038
15	мень: 0.004769: 0	русская: 0.005216: 0	хорнетов: 0.002926: 0	новость: 0.003078: 0	гладкость: 0.000635: 0	житель: 0.002705: 0	москва: 0.00382
16	американский: 0.00	статья: 0.004732: 0	корабль: 0.002926: 0	число: 0.003078: 0	побояться: 0.000635: 0	v: 0.002547: 0	принимать: 0.003
17	грант: 0.004769: 0	район: 0.004369: 0	соединение: 0.002926: -1	после: 0.002833: 0	щегол: 0.000635: 0	копенгаген: 0.002547: 0	правительство: (
18	pecypc: 0.004769: 0	работник: 0.004248: 0	бой: 0.002781: 0	новое: 0.002710: 0	финал: 0.000635: 0	s: 0.002388: 0	решение: 0.0036
19	организация: 0.0045	компенсация: 0.004248: 0	cynep: 0.002781: 1	большой: 0.002710: 0	мыть: 0.000635: 0	id: 0.002230: 0	должный: 0.0035
20	дальний: 0.004599: (проезд: 0.004006: 0	патруль: 0.002781: 0	телеканал: 0.002710: 0	выбиваться: 0.000635: 0	сша: 0.002230: 0	предлагать: 0.00
21	вы: 0.004430: 0	численность: 0.004006: 0	cy: 0.002781: 0	многие: 0.002710: 0	бля: 0.000635: 0	наркотик: 0.002072: 0	совет: 0.003413:
22	гражданский: 0.004	расход: 0.004006: 0	целить: 0.002489: 0	несколько: 0.002588: 0	талон: 0.000635: 0	провод: 0.002072: 0	проц: 0.003375: 0
23	японский: 0.004430:	север: 0.003885: 0	воздух: 0.002489: 0	новый: 0.002588: 0	рубашка: 0.000635: 0	магазин: 0.002072: 0	система: 0.00333
24	общество: 0.003751	конституционный: 0.003885: 0	гарри: 0.002344: 0	появляться: 0.002588: 0	рогопилы: 0.000635: 0	жилье: 0.001914: 0	rpynna: 0.003110
25	тема: 0.003751: 0	население: 0.003885: 0	рлс: 0.002344: 0	письмо: 0.002465: 0	тормозной: 0.000635: 0	org: 0.001914: 0	гражданин: 0.003
				1			

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: 100

. 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 Sentiment to Excel... (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: 100 and Boundary for probability: 0,001 . 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.

Search Words with high pr	obability											×
21	22	23	24	25	26	27	28	29	30	31	32	33 ^
3 фомин: 0.0	0 мвд: 0.012553; 0	космонавт: 0.007160: 0	договор: О.	уж: 0.00806	американс	комаңдир:	массимо: С	соотечеств	вертолет: () бой: 0.0022	животное: 0.005038: 0	произе
4 наблюдате	» сотрудник: 0.012064: 0	новое: 0.007160: 0	трудовой: С	этот: 0.007	армия: 0.01	сержант: 0.	. книга: 0.00	русский: О.	r: 0.005392	ленинский:	череп: 0.004202: 0	летать
5 комиссия:	отдел: 0.011901: 0	программа: 0.005255: 0	н: 0.005569	или: 0.0075	сша: 0.0096	мень: 0.004	the: 0.00429	деревня: О.	система: О	кг: 0.00200	температура: 0.003992: 0	dc: 0.0
6 мой: 0.003	девать: 0.010923; 0	космос: 0.004937: 0	срок: 0.005	сам: 0.0071	войска: 0.0	армия: 0.00	кияшко: 0.0	российски	машин: 0.0	(четверг: 0.)	размер: 0.003783: 0	пасса»
7 ермолаев:	дальний: 0.009185: 0	космонавтика: 0.004773	генеральны	бы: 0.00624	японский: (пилот: 0.00	штаб: 0.003	страна: 0.0	ми: 0.00460	армейский	titanoboa: 0.003574: 0	страна
8 участок: О.	0 март: 0.007936: 0	луна: 0.004461: 0	директор: (если: 0.006	япония: 0.0	экипаж: 0.0	андрей: 0.0	этнический	комплекс:	первенство	очень: 0.003365: 0	полет:
9 стол: 0.002	2 уголовный: 0.007827: 0	экспедиция: 0.004144: 0	хусаиновна	наш: 0.0061	советский:	летчик: 0.0	(март: 0.003	гражданин	мм: 0.0044	соревнова	рептилия: 0.003156: 0	авиала
10 володя: 0.0) татарстан: 0.007012; 0	планет: 0.004144: 0	ангелин: О.	когда: 0.00	россия: 0.0	полк: 0.003	8 кубанский:	граждансті	т: 0.004127	: днк: 0.0017	ученый: 0.003156: 0	легкий
11 com: 0.002	Следственный: 0.007012: 0	полет: 0.004144: 0	ф: 0.003719	вы: 0.00561	ссер: 0.005	война: 0.00) поэт: 0.003	нация: 0.00	http: 0.0041	рукопашнь	йоко: 0.003156: 0	лодка:
12 html: 0.002	5 казань: 0.006687: 0	миссия: 0.003985: 0	p: 0.003534	давать: 0.0	r: 0.005642:	п: 0.003508	3 автор: 0.00	документ: І	пушка: 0.00) изображат	анн: 0.002947: 0	боинг:
13 марат: 0.0	0 задержанный: 0.006198: 0	расход: 0.003667: 0	чухраёв: 0.0	становитьс	народ: 0.00	т: 0.003349	атаман: 0.0	институт: О	название: І	генерал: О.	министр: 0.002738: 0	рынок:
14 livejournal:	🕻 бывший: 0.005872: 0	русская: 0.003667: 0	больная: О.	кто: 0.0052	страна: 0.0	училище: О.	. зал: 0.0028	фмс: 0.002	ракета: 0.0	уборка: 0.0	жила: 0.002738: 0	вто: 0.1
15 sociopat: 0.	(пытка: 0.005274: 0	россия: 0.003667: 0	система: 0.	другой: 0.0	китай: 0.00-	йорк: 0.003	3 писатель: (народ: 0.00	борт: 0.003	феномен: (глобальный: 0.002738: 0	км: 0.С
16 dairy: 0.002	суд: 0.004840: 0	r: 0.003508: 0	учреждение	даже: 0.005	граница: 0.1	полет: 0.00	i like: 0.0026	рф: 0.00240	экипаж: 0.0	соль: 0.001	блоха: 0.002529: 0	сша: О.
17 президент.	февраль: 0.004514: 0	киборг: 0.003350: 0	работа: 0.0	время: 0.00	государств	фронт: 0.00) начальник:	рвп: 0.0024	лодка: 0.00	юноша: 0.0	тропический: 0.002529: 0	линия:
18 мама: 0.00	0 январь: 0.004460: 0	пилотировать: 0.003191	нарушение:	чтобы: 0.00	американе	аэродром:	войска: 0.0	язык: 0.002	испытание	турнир: 0.0	узнавать: 0.002320: 0	скорос
19 бюллетень	спреступление: 0.004460: -1	ракета: 0.003191: 0	т: 0.002794	она: 0.0049	европа: 0.0	6: 0.002714	казачий: О.	очередь: О.	фото: 0.003	фаза: 0.001	гигантский: 0.002320: 0	дально
20 медведь: 0), область: 0.004460: 0	центр: 0.003191: 0	теперь: 0.0	хотеть: 0.00	территория	час: 0.0025	in: 0.002335	население:	снаряд: 0.0	евсюков: О	окаменелость: 0.002111: 0	рф: 0.C
21 сидеть: 0.0) летний: 0.004242; 0	проект: 0.002715: 0	минздрав:	эта: 0.0048	сторона: 0.1	лейтенант:	песнь: 0.00	культура: С	российски	спорт: 0.00	потепление: 0.002111: 0	нынеш
22 детская: О	возбуждать: 0.003971: 0	корабль: 0.002715: 0	лечение: О.	мочь: 0.004	германий: (c6: 0.00255	і литературн	квота: 0.00	польский: І	силовик: О.	полагать: 0.002111: 0	новый:
23 двери: 0.00) управление: 0.003916: 0	варварство: 0.002397: 0	министр: 0.	во: 0.00481	наш: 0.0034	посадка: О.	русская: О.	получение:	украинский	логичный: І	варвар: 0.002111: 0	авиаци
24 дверь: 0.00	О отношение: 0.003916: 0	план: 0.002397: 0	назначение	вот: 0.0047	солдат: 0.0	го: 0.00223	ee: 0.00197	старик: 0.0	подлодка: І	занимать:	млн: 0.002111: 0	война:
25 собака: 0.0) мужчина: 0.003808; 0	носитель: 0.002397: 0	м: 0.002424	два: 0.0046	америка: О.	пулемет: О.	that: 0.0019	снг: 0.0018	установка:	пожелание	растение: 0.002111: 0	виктор
26 коридор: О	после: 0.003808: 0	земля: 0.002397: 0	койка: 0.00	при: 0.0044	русский: О.	красная: О.	н: 0.001978	ресурс: 0.0	шасси: 0.00	видное: 0.0	прогноз: 0.001902: 0	завод:
27 396: 0.0016	начальник: 0.003699: 0	отряд: 0.002238: 0	дип: 0.0020	ee: 0.00438	после: 0.00	истребител	войсковой:	принадлеж	производст	гнаводить: (нынешний: 0.001693: 0	ч: 0.00: 🗸
<												>
	Sentiment	ent to Excel Numbe	er of words	for export:	100	Bo	undary for	probability:	0,001		Colors	

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 ⁴. As a result, the following window will appear (see Figure 7.6)

Cocume	nts with hi	gh probabili	ty															-	= x
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19 ^
1	139: 0.7582	293: 0.8156	283: 0.6372	90: 0.97812	326: 0.2865	5 36: 0.62887	223: 0.7543	3 75: 0.75493	156: 0.7385	i 92: 0.80277	151: 0.5288	64: 0.97244	271: 0.4688	290: 0.5537	291: 0.6368	322: 0.6525	190: 0.4667	289: 0.488	3 122
2	119: 0.7506	226: 0.7301	285: 0.5348	254: 0.9780	256: 0.2364	30: 0.59677	306: 0.7281	110: 0.5839	166: 0.7376	227: 0.5797	296: 0.4811	17: 0.97136	228: 0.4449	53: 0.45332	132: 0.6352	43: 0.54172	197: 0.4366	155: 0.459	0 324
3	113: 0.7344	194: 0.1354	95: 0.33914	253: 0.9761	164: 0.1688	8 286: 0.5060	288: 0.7272	2 46: 0.38526	168: 0.7296	222: 0.5701	243: 0.3916	52: 0.96997	333: 0.4029	325: 0.2956	299: 0.5998	318: 0.5287	28: 0.40860	32: 0.3952	8 85:
4	116: 0.7122	287: 0.1106	332: 0.1433	255: 0.9748	67: 0.08189	102: 0.4727	154: 0.7105	5 201: 0.1605	96: 0.06756	91: 0.52994	205: 0.3164	31: 0.96558	23: 0.38228	160: 0.1998	202: 0.4988	361: 0.3997	278: 0.2402	72: 0.3282	2 86:
5	120: 0.6871	180: 0.0935	234: 0.1139	257: 0.9628	368: 0.0803	3 147: 0.4347	138: 0.7076	6 126: 0.1554	157: 0.0641	61: 0.21851	258: 0.0882	19: 0.96492	127: 0.2660	248: 0.1424	65: 0.47164	179: 0.3200	344: 0.2367	225: 0.303	2 206
6	109: 0.6711	292: 0.0738	201: 0.1117	250: 0.1181	3: 0.071759	334: 0.3571	358: 0.7051	369: 0.1542	39: 0.05523	8 276: 0.0641	309: 0.0575	i 1: 0.964185	246: 0.1524	206: 0.1000	157: 0.3062	362: 0.2715	196: 0.2254	104: 0.269	8 263
7	114: 0.6531	91: 0.07172	80: 0.07723	86: 0.08593	80: 0.06097	87: 0.18865	25: 0.69789	9 351: 0.0973	206: 0.0529	224: 0.0580	37: 0.04296	12: 0.96153	221: 0.0889	186: 0.0998	233: 0.2876	210: 0.2565	181: 0.2012	317: 0.210	2 191
8	130: 0.5317	367: 0.0691	149: 0.0619	301: 0.0789	82: 0.05681	186: 0.1589	203: 0.6822	2 152: 0.0816	232: 0.0492	2 72: 0.02760	178: 0.0388	13: 0.95985	98: 0.08333	239: 0.0955	224: 0.1205	284: 0.2060	209: 0.1615	256: 0.073	6 80:
9	79: 0.34269	345: 0.0572	100: 0.0582	232: 0.0595	5: 0.046610	176: 0.1363	131: 0.6591	93: 0.07985	194: 0.0467	146: 0.0268	192: 0.0340	7: 0.954235	5: 0.063559	341: 0.0919	40: 0.09000	292: 0.1382	239: 0.1292	263: 0.072	3 18:
10	106: 0.3352	128: 0.0500	331: 0.0570	202: 0.0530	14: 0.04545	5 82: 0.10984	220: 0.6539	9 341: 0.0744	176: 0.0454	68: 0.02530	218: 0.0321	37: 0.14453	364: 0.0619	349: 0.0737	134: 0.0789	339: 0.0681	8: 0.125917	312: 0.069	2 337
11	208: 0.3346	284: 0.0481	93: 0.05370	150: 0.0507	79: 0.03932	2 342: 0.1060	348: 0.6287	7 11: 0.06842	221: 0.0398	3 29: 0.02482	40: 0.03000	244: 0.1175	146: 0.0483	330: 0.0609	326: 0.0770	206: 0.0647	137: 0.1189	23: 0.0661	4 276
12	315: 0.3291	278: 0.0454	55: 0.04747	79: 0.05056	2: 0.034091	332: 0.0923	182: 0.6086	6 96: 0.06756	342: 0.0353	3 215: 0.0245	158: 0.0279	344: 0.1010	14: 0.04545	212: 0.0600	262: 0.0764	265: 0.0409	213: 0.1127	297: 0.062	5 145
13	319: 0.3038	328: 0.0443	297: 0.0446	111: 0.0445	87: 0.03298	8 107: 0.0903	93: 0.60016	6 132: 0.0605	67: 0.03017	187: 0.0237	368: 0.0267	169: 0.0903	87: 0.03825	141: 0.0541	327: 0.0514	96: 0.04054	343: 0.1098	192: 0.034	0 157
14	309: 0.2675	277: 0.0411	98: 0.03787	56: 0.03869	58: 0.03191	276: 0.0897	6: 0.593281	234: 0.0598	40: 0.03000	167: 0.0223	136: 0.0251	192: 0.0568	158: 0.0378	15: 0.05188	39: 0.04941	239: 0.0393	225: 0.1082	262: 0.029	4 82:
15	128: 0.2642	130: 0.0396	212: 0.0333	327: 0.0367	84: 0.03132	86: 0.08906	68: 0.59075	5 331: 0.0578	180: 0.0294	292: 0.0218	63: 0.02371	2: 0.056818	192: 0.0340	209: 0.0504	263: 0.0486	212: 0.0333	142: 0.1026	164: 0.029	0 364
16	345: 0.2552	146: 0.0376	9: 0.032710	181: 0.0317	251: 0.0300	228: 0.0877	108: 0.5752	2 95: 0.05527	339: 0.0277	⁷ 56: 0.02083	159: 0.0235	365: 0.0567	133: 0.0338	170: 0.0468	297: 0.0446	86: 0.03281	9: 0.098131	157: 0.027	8 40:
17	121: 0.1896	163: 0.0342	141: 0.0291	318: 0.0314	259: 0.0297	47: 0.07974	38: 0.57149	9 262: 0.0529	188: 0.0255	3: 0.020833	98: 0.02272	262: 0.0529	8: 0.033007	176: 0.0454	8: 0.040342	56: 0.03273	246: 0.0914	87: 0.0277	0 363 🗸
<																			>
(No original te	ext available)																		* *
Text prob	Doc • pr	ID Sentar	nalysis			W_lis D_lis	ile B			Number	r of docum for exp	ents 100 port:		Boundary I	or probabil	ity: 0,001			

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 As a result (if a dictionary is connected), a window will appear (see below), which reflects the calculation process.

Topic Miner	X
Sentiment analysis in progress	
10%	

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).

Cocumo	ents with high	gh probabil	ity															-	= x
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18 ^
1	139: 0.7582	293: 0.8156	283: 0.6372	90: 0.97812	2 326: 0.2865	5 36: 0.62887	223: 0.7543	3 75: 0.75493	8 156: 0.7385	92: 0.80277	7 151: 0.5288	3 64: 0.97244	271: 0.4688	290: 0.5537	291: 0.6368	322: 0.6525	190: 0.4667	289: 0.488	13 122
2	119: 0.7506	226: 0.7301	285: 0.5348	254: 0.9780	256: 0.2364	4 30: 0.59677	306: 0.7281	110: 0.5839	9 166: 0.7376	227: 0.5797	7 296: 0.4811	17: 0.97136	228: 0.4449	53: 0.45332	132: 0.6352	2 43: 0.54172	197: 0.4366	155: 0.459	0 324
3	113: 0.7344	194: 0.1354	95: 0.33914	253: 0.9761	164: 0.1686	6 286: 0.5060	288: 0.7272	2 46: 0.38526	6 168: 0.7296	222: 0.5701	243: 0.3916	52: 0.96997	333: 0.4029	325: 0.2956	299: 0.5998	318: 0.5287	28: 0.40860	32: 0.3952	.8 85:
4	116: 0.7122	287: 0.1106	332: 0.1433	255: 0.9748	8 67: 0.08189	9 102: 0.4727	154: 0.7105	5 201: 0.1605	5 96: 0.06756	91: 0.52994	1 205: 0.3164	31: 0.96558	23: 0.38228	160: 0.1998	202: 0.4988	361: 0.3997	278: 0.2402	72: 0.3282	2 86:
5	120: 0.6871	180: 0.0935	5 234: 0.1139	257: 0.9628	368: 0.0803	3 147: 0.4347	138: 0.7076	126: 0.1554	157: 0.0641	61: 0.21851	258: 0.0882	2 19: 0.96492	127: 0.2660	248: 0.1424	65: 0.47164	179: 0.3200	344: 0.2367	225: 0.303	2 206
6	109: 0.6711	292: 0.0738	8 201: 0.1117	250: 0.1181	3: 0.071759	334: 0.3571	358: 0.7051	369: 0.1542	2 39: 0.05523	276: 0.0641	309: 0.0575	5 1: 0.964185	246: 0.1524	206: 0.1000	157: 0.3062	362: 0.2715	196: 0.2254	104: 0.269	18 263
7	114: 0.6531	91: 0.07172	80: 0.07723	86: 0.08593	80: 0.06097	7 87: 0.18865	25: 0.69789	351: 0.0973	3 206: 0.0529	224: 0.0580	37: 0.04296	6 12: 0.96153	221: 0.0889	186: 0.0998	233: 0.2876	210: 0.2565	181: 0.2012	317: 0.210	12 191
8	130: 0.5317	367: 0.0691	149: 0.0619	301: 0.0789	9 82: 0.05681	186: 0.1589	203: 0.6822	2 152: 0.0816	6 232: 0.0492	72: 0.02760	0 178: 0.0388	3 13: 0.95985	98: 0.08333	239: 0.0955	224: 0.1205	5 284: 0.2060	209: 0.1615	256: 0.073	16 80:
9	79: 0.34269	345: 0.0572	2 100: 0.0582	232: 0.0595	5 5: 0.046610	0 176: 0.1363	131: 0.6591	93: 0.07985	5 194: 0.0467	146: 0.0268	3 192: 0.0340	7: 0.954235	5: 0.063559	341: 0.0919	40: 0.09000	292: 0.1382	239: 0.1292	263: 0.072	3 18:
10	106: 0.3352	128: 0.0500	331: 0.0570	202: 0.0530	0 14: 0.04545	5 82: 0.10984	220: 0.6539	341: 0.0744	176: 0.0454	68: 0.02530	0 218: 0.0321	37: 0.14453	364: 0.0619	349: 0.0737	134: 0.0789	339: 0.0681	8: 0.125917	312: 0.069	12 337
11	208: 0.3346	284: 0.0481	93: 0.05370	150: 0.0507	79: 0.03932	2 342: 0.1060	348: 0.6287	11: 0.06842	2 221: 0.0398	29: 0.02482	2 40: 0.03000	0 244: 0.1175	146: 0.0483	330: 0.0609	326: 0.0770	206: 0.0647	137: 0.1189	23: 0.0661	4 276
12	315: 0.3291	278: 0.0454	55: 0.04747	79: 0.05056	6 2: 0.034091	332: 0.0923	182: 0.6086	96: 0.06756	6 342: 0.0353	215: 0.0245	5 158: 0.0279	9 344: 0.1010	14: 0.04545	212: 0.0600	262: 0.0764	265: 0.0409	213: 0.1127	297: 0.062	5 145
13	319: 0.3038	328: 0.0443	297: 0.0446	111: 0.0445	5 87: 0.03298	8 107: 0.0903	93: 0.60016	132: 0.0605	5 67: 0.03017	187: 0.0237	7 368: 0.0267	7 169: 0.0903	87: 0.03825	141: 0.0541	327: 0.0514	96: 0.04054	343: 0.1098	192: 0.034	0 157
14	309: 0.2675	277: 0.0411	98: 0.03787	56: 0.03869	9 58: 0.03191	276: 0.0897	6: 0.593281	234: 0.0598	8 40: 0.03000	167: 0.0223	3 136: 0.0251	192: 0.0568	158: 0.0378	15: 0.05188	39: 0.04941	239: 0.0393	225: 0.1082	262: 0.029	14 82:
15	128: 0.2642	130: 0.0396	212: 0.0333	327: 0.0367	7 84: 0.03132	2 86: 0.08906	68: 0.59075	331: 0.0578	8 180: 0.0294	292: 0.0218	3 63: 0.02371	2: 0.056818	192: 0.0340	209: 0.0504	263: 0.0486	212: 0.0333	142: 0.1026	164: 0.029	10 364
16	345: 0.2552	146: 0.0376	9: 0.032710	181: 0.0317	7 251: 0.0300	228: 0.0877	108: 0.5752	95: 0.05527	7 339: 0.0277	56: 0.02083	3 159: 0.0235	5 365: 0.0567	133: 0.0338	170: 0.0468	297: 0.0446	6 86: 0.03281	9: 0.098131	157: 0.027	'8 40:
17	121: 0.1896	163: 0.0342	2 141: 0.0291	318: 0.0314	259: 0.0297	7 47: 0.07974	38: 0.57149	262: 0.0529	9 188: 0.0255	3: 0.020833	3 98: 0.02272	2 262: 0.0529	8: 0.033007	176: 0.0454	8: 0.040342	2 56: 0.03273	246: 0.0914	87: 0.0277	'0 363 🗸
<																			>
(No original t	ext available																		*
Text prot	Doc pr	ID Sentar	nalysis			W_lis D_lis	ale le			Numbe	r of docum for exp	ents 100 port:		Boundary f	or probabi	ity: 0,001			

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



Doc ID

of evaluation. In order to get such an unloading you need to click on the button ^{prob}. 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 is a prob. 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):

Number of documents	100
for export:	100
and any second	

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

```
Boundary for probability: 0,001
```

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 tmlda' implements the option of unifying dates. To start the unification process, open the 'View of tmlda' tab, load the data containing the timestamps (see Figure 8.1).

Next, you need to open the 'Data / Time repear' option (see figure 8.2), specify the field number that contains the timestamps. In this example, this field is \mathbb{N} 2. Then you need to click on the 'Save as TMLDA' button and specify the name of the new file..

TopicMiner Sentiment analysis ver	r. 96 (64 bit) LINIS laboratory, HSE					↔ _ = ×
Lematization (Russian language) View	of TMLDA file Gibbs LDA sampling Big	ARTM Kull	back-Leibler D	istance, Jacca	ard index	
	Matrix data (TAB) Matrix data (csv	a) 🚺				
Original document	Lematized document	Author	Field 1	Field 2	-	Dicts for BigARTM
Гость писал(а):Не верю,что цены когда нибудь опустятся,они себе сделали рай а намПравильно	гость писать верить цена когда нибудь опускаться сделать правильно делать верить	718206380	Московска я область	2015-01-22 06:10:00		Select field names (max 5);
Две пираньи, добивающие Грецию. Европейцы против такой политики. Кому нужен греческий труп??Читать	пиранья добивать греция европеец против политик нужный греческий труп читать дальшесвернутьмеркель	963263342	Москва	2015-07-14 00:00:00		Field 1 Field 2 Field 3 Field
Американцы, после метеорита, очень дивились, что почти в каждой русской машине стоит	американец после метеорит удивляться почесть каждая русская машин стоить видеорегистратор	757465969	Курганская область	2015-02-12 04:01:46		 Field 4 Field 5 Field 6
(іd95525685/Дмитрий), надеюсь ты толучишь свое и те кто тебя высрал тоже за рождения такого козла))	дмитрий надеяться получать свое высрал рождение козел когда быть ичкерия точно также нацики	928747104	Санкт-Пете рбург	2015-06-08 18:40:46		 Field 7 Field 8 Field 9
У меня есть дома десяток яиц и я рчеень люблю их совать в рот собаке потому что у м еня ее нет азазаз у	быть десятка яйцо очеень любить совать собака потому азазаз очинь харошие друг лаффкиии скамейка	716129928	Пермский край	2015-01-21 07:40:04		Field 10 Field 11
Строительство дома. Обычаи и приметы кабардинских народов ‡строительство	строительство обычай примета кабардинский народ строительство tkpss blog stroitels	865091594	Санкт-Пете рбург	2015-04-17 00:00:00		Column with Date/Time: None •
(id231058098)Ирина], насмотрелся канала россия 24 и возненавидел эсе российское. Сначала критиковал	ирин насмотреться канал россия возненавидеть российский сначала критиковать излишний	770509083	Новосибир ская область	2015-02-19 21:51:03		
"Как говорят сербы: Мы с русскими один народ, правительство - другой народ" Видео на 3 минутки: почему	говорить серб русская народ правительство другой народ видео минутка почему украина можно	791800299	Красноярс кий край	2015-03-05 00:49:20		Caus distingan
٢				>	-	Save dictionaly
Status: file is loaded	File name: 24tr_etn	iicity3.tmlda				

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 tmlda file. To make sure that all timestamps are unified enough to load a new file.

Next, you need to create a special tmlda 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.

TopicMiner Sentiment analysis v	ver. 96 (64 bit) LINIS laboratory, HSE					↔ _ = ×
Lematization (Russian language)	w of TMLDA file Gibbs LDA sampling	BigARTM K	ullback-Leible	r Distance, Jaco	ard inde	x
	🤠 Matrix data (TAB) Matrix data (c	:sv) 🧿				
Original document	Lematized document	Author	Field 1	Field 2		Dicts for BigARTM
1178370173 Міа], но у вас же бред писан. Какими русскими? Вообще там грузины воевали с Абхазами.	бред написать русская вообще там грузин воевать абхаз исторический факт статья может быть правдивый					Select field names (max 5);
11725211Юлег), существую и не ликоросс я, посмотри мой город, у с свой НЕ русский язык , который	олег существовать великоросс посмотреть город русский язык никогда смочь розуметь	716699232	Коми	21.01.2015		 ✓ Field 1 ✓ Field 2 ✓ Field 3
ія примера приведу по датам ючевые моменты узино-Абхазкой войны в	пример приводить ключевой момент грузино абхазка война хронологический порядок	627169733	Хабаровск ий край	13.10.2014		 Field 4 Field 5 Field 6
122518477 Роман], не было бы ссии не было бы Украины. Не ло бы России грузины бы казнили	роман было россия было украина было россия грузин казнить абхаз было россия чечня дагестан уровень	668964756	Гомельска я область	05.12.2014		 Field 7 Field 8 Field 9
ительница:- Вовочка, у тебя личное сочинение, но все-таки я ставила тебе 4 за тавтологию в	учительница вовочка отличный сочинение все-таки поставлять тавтология предложение	849433	Свердловс кая область	27.02.2014		Field 10 Field 11
)РУССИЯВ историческом тонем мане мы,А пора уже знать авянину,Что Поруссия стала	поруссияв исторический тонуть обман пора уж знать славянин поруссия становиться германия	848471	Ленинград ская область	29.06.2015		Column with Date/Time: Field 2 🔹
к сообщает Росстат, по личеству безработных Иркутская ласть лидирует среди регионов	сообщать росстат количество безработная иркутский область лидировать среди регион сибирский	742379825	Тыва	04.02.2015		
14404307 Александр], еще бы был против)) в нашем городе)) охуел раки совсем)) Ето как в Молдове	александр еще быть против город охуевать русак совсем молдова бельц русский город бельц русский	837669729	Волгоградс кая область	30.03.2015		
<				3	• -	Save dictionary
Status: file is loaded	File name: 24tr_e	tnicity3_tim	e.tmlda			



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).

	1	2		3	4		5	6	7	- 1	8	9		10	11	1	12	13	14	1	15	16	17		18	
1	2955: 0.999	4002:	0.983	6890: 0.987	6085:	0.959	24899: 0.5	9 7236: 0.99	9 5990: 0).999	20183: 0.96	16386: 0.3	89 2091	0: 0.99	23930: 0.	99 9453	3: 0.906	23741: 0.9	8 24844	0.90	24111: 0.9	8 5025: 0.97	7 18717:	0.96	12515:	ō
2	10911: 0.91	5903:	0.939	1910: 0.844	4735:	0.958	331: 0.995	9 13109: 0.9	9 5836: 0	0.999	681: 0.9698	774: 0.88	41 2064	18: 0.99	6742: 0.9	29 164:	0.8489	24412: 0.8	9 18785	0.89	19309: 0.9	8 16329: 0.9	4 18877:	0.95	1233: 0	1.5
3	12435: 0.75	15007	: 0.87	17314: 0.79	9 7103:	0.958	8754: 0.99	5 10896: 0.9	6 6089: 0	0.994	4089: 0.920	7336: 0.8	50 1763	32: 0.99	920: 0.92	96 1674	12: 0.83	2032: 0.87	0 21679	0.89	277: 0.982	7 17120: 0.9	3 15150:	0.94	7728: 0	Ū,
4	6390: 0.745	6357:	0.823	7742: 0.777	7 13658	: 0.95	10999: 0.9	9 11001: 0.9	6 6188: 0	0.990	4227: 0.916	10105: 0.3	85 1963	33: 0.99	20919: 0.	31 1244	11: 0.83	2027: 0.84	1 7049:	0.888	11899: 0.9	6 12600: 0.9	3 22889:	0.91	14188:	C
5	2006: 0.726	13310	: 0.72	19400: 0.77	7 11306	0.95	1837: 0.99	4 10851: 0.9	6 6048: 0	0.973	22361: 0.91	5054: 0.8	23 212	8: 0.99	12864: 0.	30 1942	2: 0.829	17535: 0.8	3 8321:	0.877	24397: 0.9	6 12079: 0.9	3 24178:	0.91	504: 0.9	9
6	3283: 0.707	12572	: 0.70	24243: 0.75	5 23849	t 0.95	18395: 0.9	9 16121: 0.9	6 6247: 0	0.969	2448: 0.907	15784: 0.1	77 1994	45: 0.99	12065: 0.	80 4477	7: 0.796	8880: 0.82	8 3318:	0.862	9656: 0.96	8 13924: 0.9	3 17994:	0.89	12781:	0
7	4606: 0.706	12527	: 0.68	24425: 0.74	1 23664	: 0.95	18978: 0.5	9 11020: 0.9	6 10608:	0.96	19126: 0.89	21466: 0.	74 1818	33: 0.99	4828: 0.8	04 3524	1: 0.790	7367: 0.82	4 20458	0.86	18889: 0.9	5 23770: 0.9	2 20291:	0.86	14306:	0
8	19082: 0.65	7227:	0.672	2 12958: 0.74	4 24214	: 0.95	3367: 0.98	8 10965: 0.9	6 3849: 0	0.962	3770: 0.884	2550: 0.7	06 1972	23: 0.99	5475: 0.8	03 1311	1: 0.76	21380: 0.8	2 23019	0.83	23789: 0.9	4 2260: 0.92	6 8333: 0	.856	13874:	C
9	24483: 0.69	16199	: 0.66	552: 0.7430	0 13034	: 0.94	16209: 0.9	8 10871: 0.9	6 6104: 0	0.948	4352: 0.872	13789: 0.	70 178	55: 0.99	3446: 0.8	03 1172	27: 0.76	106: 0.818	9 9383:	0.827	4535: 0.93	1 3251: 0.91	6 8424: 0	.852	22211:	0
0	16489: 0.68	9108:	0.655	5 14620: 0.73	8 6870:	0.946	8432: 0.98	8 10156: 0.9	6 5845: 0	0.948	686: 0.8724	6696: 0.7	05 2071	9: 0.99	24923: 0.	30 8927	7: 0.752	22898: 0.8	1 8941:	0.819	10060: 0.9	2 7101: 0.91	2 15783:	0.83	17708:	0
1	23431: 0.68	14794	: 0.65	22414: 0.73	3 7874:	0.944	2178: 0.98	7 478: 0.959	2 6053: 0	0.947	5281: 0.869	6168: 0.6	94 2055	59: 0.99	15011: 0.1	76 1776	58: 0.74	1622: 0.79	0 21087	0.81	11412: 0.9	1 17870: 0.9	0 1552: 0	.838	6724: 0	. 9
2	16303: 0.65	9023:	0.644	16444: 0.72	2 24175	0.94	5244: 0.98	1 19201: 0.9	5 23324:	0.94	11947: 0.86	7549: 0.6	62 1792	29: 0.99	4435: 0.7	67 2813	3: 0.724	11671: 0.7	5 24153	0.81	5067: 0.91	7 1328: 0.90	0 16056:	0.82	21976:	0
3	18131: 0.63	16648	: 0.61	11173: 0.72	2 12321	: 0.91	23959: 0.9	7 17694: 0.9	1 13068:	0.94	1892: 0.858	20509: 0.	64 1737	70: 0.99	21343: 0.1	76 1998	32: 0.72	11603: 0.7	4 12488	0.81	23153: 0.9	1 9587: 0.88	5 1722: 0).814	10103:	0
4	18509: 0.62	21103	: 0.60	4399: 0.72	11542	0.89	5703: 0.97	6 12080: 0.8	8 2795: 0	0.941	3596: 0.845	17910: 0.	63 2025	59: 0.99	17968: 0.1	74 1142	22: 0.72	12402: 0.7	4 14644	0.80	17541: 0.9	1 10790: 0.8	8 8359: 0	.812	17311:	0
5	3692: 0.623	3924:	0.603	8 4115: 0.704	4921:	0.856	17125: 0.9	7 12083: 0.8	8 10679:	0.93	716: 0.8415	20691: 0.	62 1895	52: 0.99	10807: 0.	73 2045	53: 0.71	9244: 0.74	1 13070	0.80	15590: 0.9	D 19484: 0.8	8 8388: 0	1.810	52: 0.99	3
6	10597: 0.62	23641	: 0.59	8 1772: 0.70	19515	0.85	18844: 0.5	7 4122: 0.87	5 1593: 0	0.933	713: 0.8415	6329: 0.6	14 2044	17: 0.97	11286: 0.	71 4464	1: 0.707	24221: 0.7	3 118: 0	8024	12416: 0.9	3239: 0.88	3 1939: 0	.806	7621: 0	
7	1808: 0.604	14726	: 0.59	8 7560: 0.700	4261:	0.851	22327: 0.5	6 20821: 0.8	7 8215: 0	0.929	11895: 0.83	7347: 0.5	98 1833	30: 0.97	20308: 0.	59 1318	39: 0.69	5482: 0.73	7 24233	0.79	10446: 0.8	9 20950: 0.8	8 23841:	0.79	15415:	0
l tr	ext available)																									

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.

Cocum 🗢	ents with hi	igh probab	oility																		- = ×
	1	2	3	4	L	5	6	7	8		9	10	11	12	13	14	15	16	17	18	19 ^
1	21503: 0.9	<mark>e</mark> 23787: 0.	82 5771: 0.8	24 6971:	0.753	18931: 0.91	2146: 0.83	7 22946: 0.5	52 16121:	0.96 437	3: 0.622	15917: 0.90	5903: 0.96	6 9428: 0.71	2 24052: 0	.78 11001: 0	.93 20910: 0.9	97 21976: 0.9	99 5243: 0	941 18480:	0.80 102
2	2 20211: 0.9	9 24220: 0.	82 5596: 0.8	21 17525	0.69	17703: 0.87	9987: 0.57	9 16701: 0.5	52 21720:	0.93 838	1: 0.557	296: 0.5354	20645: 0.7	1 5879: 0.70	8 12435: 0	.73 10896: 0	.93 20648: 0.9	97 20050: 0.9	38 3618: 0	.941 18555:	0.79 781
3	3 22281: 0.9	9 19813: 0.	82 5694: 0.8	20 17634	: 0.64	6567: 0.860	12911: 0.5	7 8488: 0.52	23 24054:	0.93 145	62: 0.47	5465: 0.513	9120: 0.65	2 1140: 0.69	8 8424: 0.7	732 11020: 0	.91 19633: 0.9	97 17867: 0.9	95 304: 0.9	3412 19080:	0.79 212
4	11757: 0.9	9 14552: 0.	82 3353: 0.8	11 20812	2: 0.62	6518: 0.845	5 19400: 0.5	5 17803: 0.5	51 16149:	0.93 189	65: 0.46	20122: 0.48	22962: 0.6	0 15128: 0.6	6 8333: 0.7	731 10965: 0	.87 19945: 0.9	95 13874: 0.9	94 4414: 0	.939 18825:	0.79 106
5	5 3274: 0.99	4 23572: 0.	81 3534: 0.8	04 20530): 0.57	6476: 0.844	21142: 0.4	7 4078: 0.47	74 13600:	0.93 106	60: 0.45	10678: 0.42	20466: 0.5	9 5998: 0.66	1 20892: 0	.72 10871: 0	.87 21218: 0.9	95 22211: 0.9	33 24001:	0.92 18736	0.78 170
E	17667: 0.9	9 21985: 0.	81 5863: 0.8	04	Docur	ments - ara	up by date											+		. = x	1.77 341
7	22134: 0.9	9 24227: 0.	77 11625: 0.	80																	.77 218
	888: 0.994	6 9823: 0.6	35 5937: 0.8	02	Г															_	.77 218
	15827: 0.9	9 22959: 0.	62 3594: 0.7	95																	.77 218
- 10	23056: 0.9	9 23084: 0.	62 3657: 0.7	70																	.76 218
12	17010-0.0	4 8034: 0.6	10 23331: 0.	73																	76 218
12	20670-0.9	3 13340: 0. 9 2619: 0 6	53 15503: 0. 79 2672: 0.7	72																	76 210
14	ance: n ee	4 4644: 0.5	FE 21102-0	71																	72 115
15	14021-0.9	9 15642: 0	56 7128 0.5	95																	73 115
16	12438 0.9	9 10157: 0	55 5984 0.5	87																	61 151
17	23643 0.9	9 12621: 0	55 13400 0	58																	59 137
<	1			-																	>
					۳Ľ																
(No original	text available	e)																			^
																					-
I ext	h K Do	cID Sent	analysis																		
					L								-								
													12.1899								
				_		_				_	_									_	
				ţ	Lopic	1		Document	s count:	100		Time	period: 🚺	v/eek	•	Update	graph				
													V	/eek							
													N	tonth							
													Y	ear							

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 N_2 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_etnicity3_time_bigartm.tmlda' 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 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).

An

👂 Words v	vith high probability - Fiel	ld 2						÷ _	= x
	1	2	3	4	5	6	7	8	^
1	18.04.2015: 0.133450	28.02.2015: 0.068136	02.02.2015: 0.058923	10.02.2015: 0.207571	21.04.2015: 0.195884	05.01.2015: 0.051894	20.02.2015: 0.243819	04.05.2015: 0.108129	_
2	18.03.2015: 0.120320	26.01.2015: 0.065610	13.04.2015: 0.047438	23.03.2015: 0.183897	06.03.2014: 0.057535	18.11.2014: 0.047293	08.06.2015: 0.062857	20.05.2015: 0.056847	
3	30.04.2015: 0.038458	17.02.2015: 0.055930	05.04.2015: 0.040010	09.04.2015: 0.107579	15.03.2014: 0.044147	02.01.2015: 0.045747	19.06.2014: 0.061073	26.08.2015: 0.052674	
4	03.07.2014: 0.034420	04.03.2015: 0.048440	15.05.2015: 0.039527	15.01.2015: 0.039836	24.02.2015: 0.035007	02.04.2015: 0.025140	15.05.2014: 0.046717	03.12.2014: 0.048832	
5	26.04.2015: 0.034325	25.02.2015: 0.033704	22.04.2015: 0.037852	25.01.2015: 0.024845	23.10.2014: 0.032986	06.01.2015: 0.024899	30.01.2014: 0.026326	07.08.2015: 0.032263	
6	21.07.2015: 0.034099	26.02.2015: 0.031375	25.01.2015: 0.036698	12.05.2015: 0.022742	06.09.2014: 0.031522	23.01.2014: 0.024309	13.10.2015: 0.016656	04.04.2014: 0.031769	
7	21.06.2014: 0.030105	07.02.2015: 0.028153	21.05.2015: 0.036659	04.03.2015: 0.019013	08.08.2015: 0.025627	03.03.2015: 0.022999	17.04.2015: 0.015990	14.11.2014: 0.029541	
8	17.03.2015: 0.029587	01.06.2015: 0.024922	13.03.2015: 0.035113	22.05.2015: 0.018928	19.02.2015: 0.022286	14.12.2014: 0.022146	26.02.2015: 0.014816	24.05.2014: 0.026298	
9	05.05.2014: 0.026704	27.04.2015: 0.019187	19.01.2015: 0.030486	19.04.2015: 0.018261	07.06.2015: 0.021512	16.07.2014: 0.020815	05.04.2015: 0.013035	30.12.2014: 0.023940	
10	20.03.2015: 0.021668	16.03.2015: 0.018307	19.03.2015: 0.029981	14.02.2015: 0.017309	25.03.2015: 0.021135	16.04.2014: 0.016350	14.10.2015: 0.011144	06.04.2015: 0.023064	
11	29.01.2015: 0.021088	14.04.2015: 0.018010	28.04.2015: 0.026314	16.09.2014: 0.013838	13.08.2014: 0.020413	27.05.2014: 0.016110	13.10.2014: 0.011048	22.06.2015: 0.018792	
12	31.03.2014: 0.019021	21.01.2014: 0.017488	12.03.2015: 0.025972	05.03.2015: 0.012233	28.03.2015: 0.017009	21.02.2015: 0.015972	21.02.2015: 0.010853	04.10.2014: 0.017003	
13	13.05.2014: 0.016174	25.03.2015: 0.017262	08.02.2015: 0.023613	28.03.2015: 0.011855	21.02.2015: 0.016258	27.03.2015: 0.015811	16.07.2014: 0.010749	06.06.2015: 0.016246	
14	23.04.2015: 0.014222	07.04.2015: 0.016515	30.01.2015: 0.023485	22.07.2015: 0.011733	07.12.2014: 0.015662	03.04.2015: 0.015541	21.07.2014: 0.010450	10.02.2014: 0.015268	
15	03.06.2014: 0.013571	03.04.2015: 0.016511	10.04.2015: 0.020661	23.04.2015: 0.011669	14.04.2015: 0.014318	23.08.2014: 0.014815	11.01.2014: 0.010236	24.11.2014: 0.014775	
16	03.05.2014: 0.012667	20.03.2015: 0.015400	14.03.2015: 0.020147	05.02.2015: 0.011423	11.07.2015: 0.014095	26.04.2015: 0.013891	22.07.2014: 0.009839	10.08.2014: 0.013860	
17	14.06.2015: 0.012181	24.06.2015: 0.015325	13.02.2015: 0.019241	21.02.2015: 0.011169	11.02.2015: 0.013436	Свердловская область	21.01.2014: 0.009505	06.08.2014: 0.012580	
18	04.07.2015: 0.011682	29.01.2015: 0.015164	17.03.2015: 0.019084	14.05.2015: 0.010275	08.05.2014: 0.010500	05.08.2015: 0.013009	07.03.2014: 0.009268	26.04.2015: 0.012202	
19	10.10.2014: 0.011567	30.10.2015: 0.014184	19.04.2015: 0.018171	03.03.2015: 0.010140	31.03.2015: 0.010449	10.07.2015: 0.012595	14.02.2014: 0.009188	01.06.2015: 0.011619	
20	06.02.2015: 0.011335	13.03.2015: 0.014038	09.02.2015: 0.018046	16.05.2014: 0.009320	16.02.2015: 0.010435	02.03.2015: 0.012580	17.08.2015: 0.008877	20.12.2014: 0.010586	
21	06.01.2015: 0.010984	31.05.2015: 0.014007	30.04.2015: 0.017610	13.02.2015: 0.009172	18.02.2015: 0.009529	29.01.2015: 0.012551	20.04.2015: 0.008860	22.07.2015: 0.010210	
22	12.01.2014: 0.010974	27.10.2015: 0.013440	17.04.2015: 0.017179	04.02.2014: 0.008603	20.03.2015: 0.009400	07.05.2015: 0.012466	06.08.2015: 0.008309	09.01.2014: 0.010095	
23	03.10.2015: 0.009583	18.07.2015: 0.013333	04.04.2015: 0.017177	07.04.2015: 0.008343	23.04.2015: 0.009055	26.10.2015: 0.012438	19.02.2015: 0.008032	07.08.2014: 0.009890	
24	14.04.2015: 0.009515	27.02.2015: 0.013161	24.02.2015: 0.017079	03.04.2015: 0.008058	31.10.2015: 0.008819	03.06.2015: 0.010426	20.10.2015: 0.007499	13.12.2014: 0.009838	
25	14.03.2014: 0.009133	07.03.2015: 0.012800	07.02.2015: 0.017011	19.10.2015: 0.007715	13.07.2015: 0.008334	05.03.2015: 0.010394	21.09.2015: 0.007019	28.12.2014: 0.009774	—
<	1	1		1	1	1	1	1	>
-									
Grap	bhics Sentim	ient Sentiment to E	xcel Number of w	ords for export: 100	Boundary for	probability: 0,001	Colors		

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)