

Оглавление:

- Установка
 - Ограничения и установка Docker
 - Загрузка образа WebTopicMiner
 - Запуск WebTopicMiner
 - Использование WebTopicMiner
- Обзор интерфейса
- Вкладка Workspace
 - Работа с файлами tmla
 - Работа с файлами bin
 - Анализ текстов с помощью встроенных классификаторов
- Тематическое моделирование
 - Запуск новой задачи тематического моделирования
 - Просмотр выполняющихся и завершенных задач моделирования
 - Просмотр результатов
 - Сентимент-анализ тем
 - Отображение распределения темы на карте России
 - Просмотр распределения тем во времени

Введение

Веб-сайт WebTopicMiner является современной версией ПО TopicMiner для Windows, разработанного в Лаборатории Интернет-Исследований НИУ ВШЭ (<https://linis.hse.ru>).

В текущей версии поддерживается режим сэмплирования Гиббса с использованием данных в бинарных файлах формата TopicMiner и мультимодальные модели (BigARTM).

Установка

WebTopicMiner — сложный программный продукт, состоящий из нескольких компонентов. Поэтому предпочтительным является использование версии, расположенной на серверах ЛИНИС. Если же по каким-либо причинам вы хотите настроить свою локальную копию или просто оценить возможности WebTopicMiner, воспользуйтесь этой инструкцией.

Самый простой способ установить продукт — с помощью виртуальной машины Docker. Именно этот метод будет описан далее. Если вам требуется полноценная установка, попросите вашего системного администратора обратиться в ЛИНИС.

Ограничения и установка Docker

Если вы используете Linux, просто установите и настройте Docker, следуя инструкциям для своего дистрибутива. Если вы используете Mac, скачайте и установите Docker for Mac по следующей ссылке: <https://docs.docker.com/docker-for-mac/install/>.

У пользователей Windows существует две возможности. Если у вас достаточно свежая Windows 10 с поддержкой Hyper-V, попробуйте сначала установить Docker for Windows: <https://docs.docker.com/docker-for-windows/install/>. Если на вашей системе не доступен Hyper-V (об этом сообщит установщик Docker) или если вы используете Windows 7 или Windows 8, установите Docker Toolbox, в состав которого входит виртуальная машина VirtualBox: <https://docs.docker.com/toolbox/overview/>. Для большего удобства и производительности мы рекомендуем первый вариант — Windows 10 с Hyper-V. При установке для удобства отметьте пункт «Добавить docker в PATH»

После установки Docker переходите к следующему разделу.

Загрузка образа WebTopicMiner

Вам был предоставлен файл `topicminer.tar.bz2`, содержащий все необходимые для работы приложения компоненты. Чтобы добавить этот образ в ваш Docker, потребуется использовать командную строку. Данная инструкция предполагает наличие базовых навыков использования командной строки.

Перейдите в папку, где вы сохранили файл образа и выполните следующую команду:

```
docker load -i topicminer.tar.bz2
```

Запуск WebTopicMiner

Пользователям Mac и Windows, первый раз знакомящимся с Docker, рекомендуется установка через графический интерфейс. Продвинутые пользователи, а также пользователи Linux, могут воспользоваться командной строкой.

С помощью командной строки

Создайте папку, в которой WebTopicMiner будет хранить системные и пользовательские файлы и запомните её путь. Выберите номер порта, по которому будет доступен локальный веб-интерфейс. В примерах будет использоваться порт 5000.

Для Windows:

```
C:\> docker run -d -v C:\Users\путь\к\папке:/data -p 5000:8080 --name topicminer topicminer
```

Для Mac и Linux:

```
$ docker run -d -v /home/путь/к/папке:/data -p 5000:8080 -u 1000:1000 --name topicminer topicminer
```

Замените 1000:1000 на ваши user id и group id.

При запуске WebTopicMiner создаст по указанному пути необходимые ему файлы, а его веб-интерфейс будет доступен по адресу <http://127.0.0.1:5000/>.

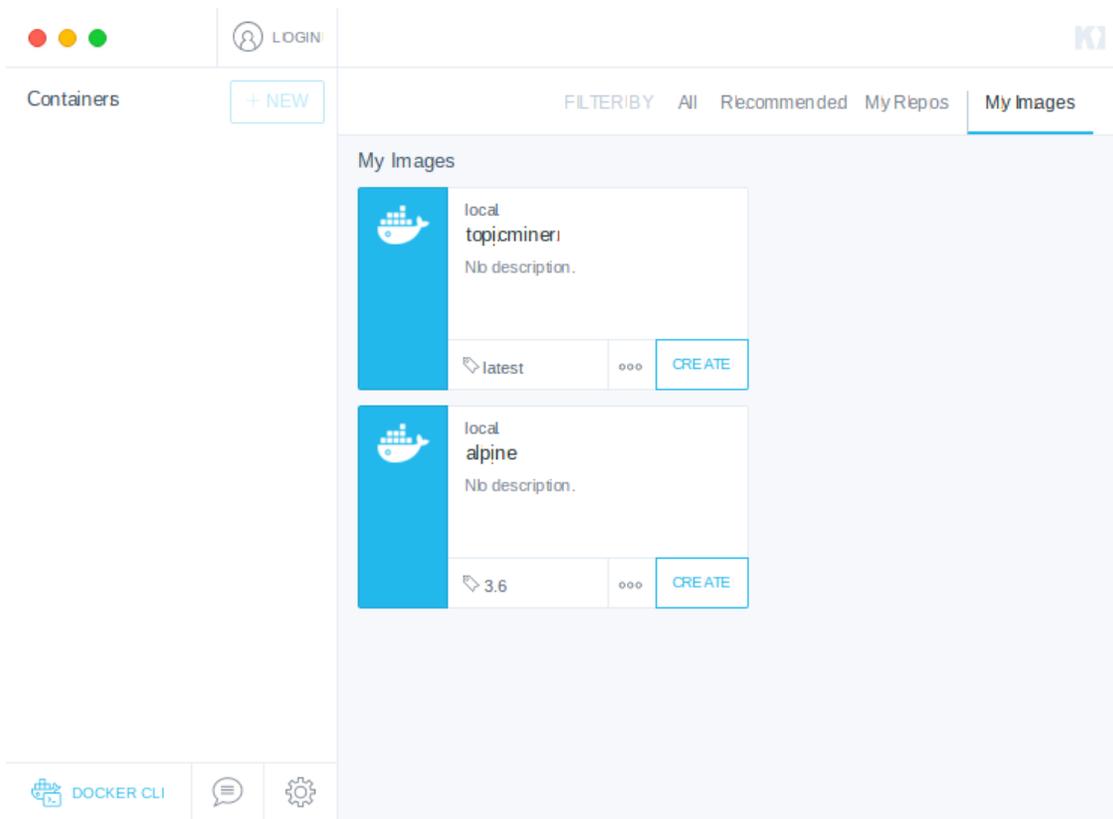
После перезапуска системы контейнер можно запустить с помощью

```
$ docker start topicminer
```

С помощью графического интерфейса Kitematic

Запустите Kitematic, либо с помощью контекстного меню Docker for Windows / Mac, либо из меню «Пуск» при использовании Docker Toolbox.

В разделе «My Images» должен отображаться образ WebTopicMiner, добавленный ранее:



Нажмите кнопку «Create».

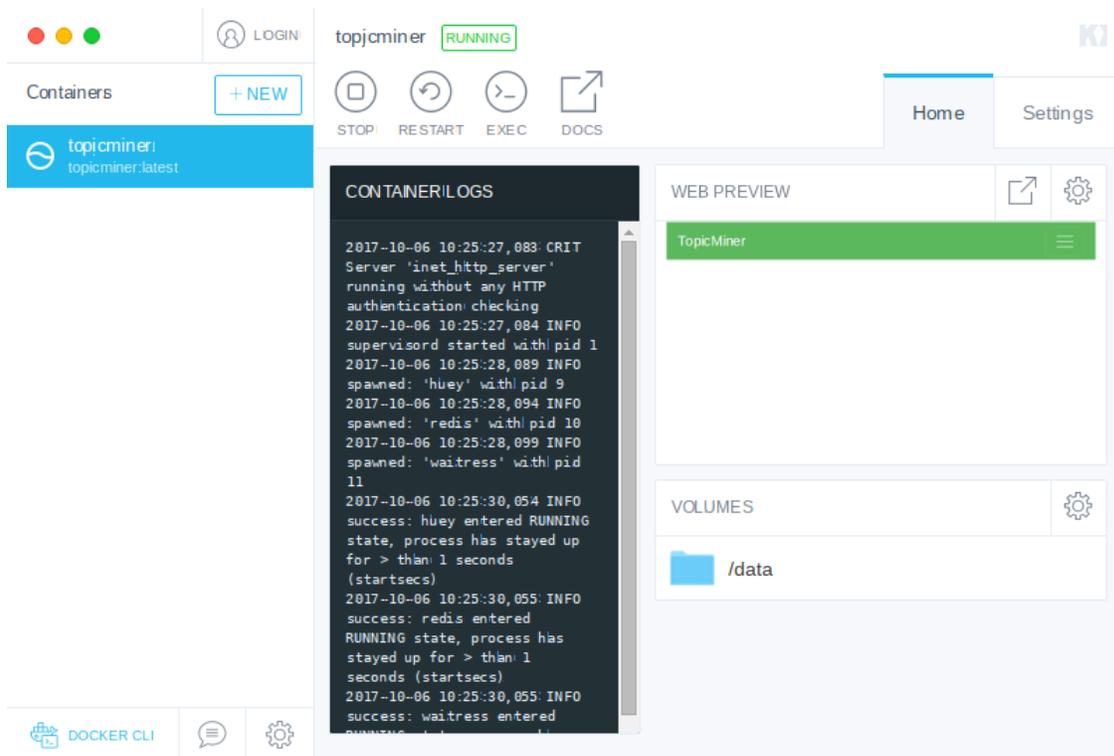
Примечание

Некоторые версии Kitematic содержат ошибку, из-за которой при нажатии на «Create» контейнер не создается. Если у вас такая версия, выполните в командной строке:

```
docker run -d --name topicminer topicminer
```

и вернитесь в графический интерфейс, запущенный контейнер должен там появиться.

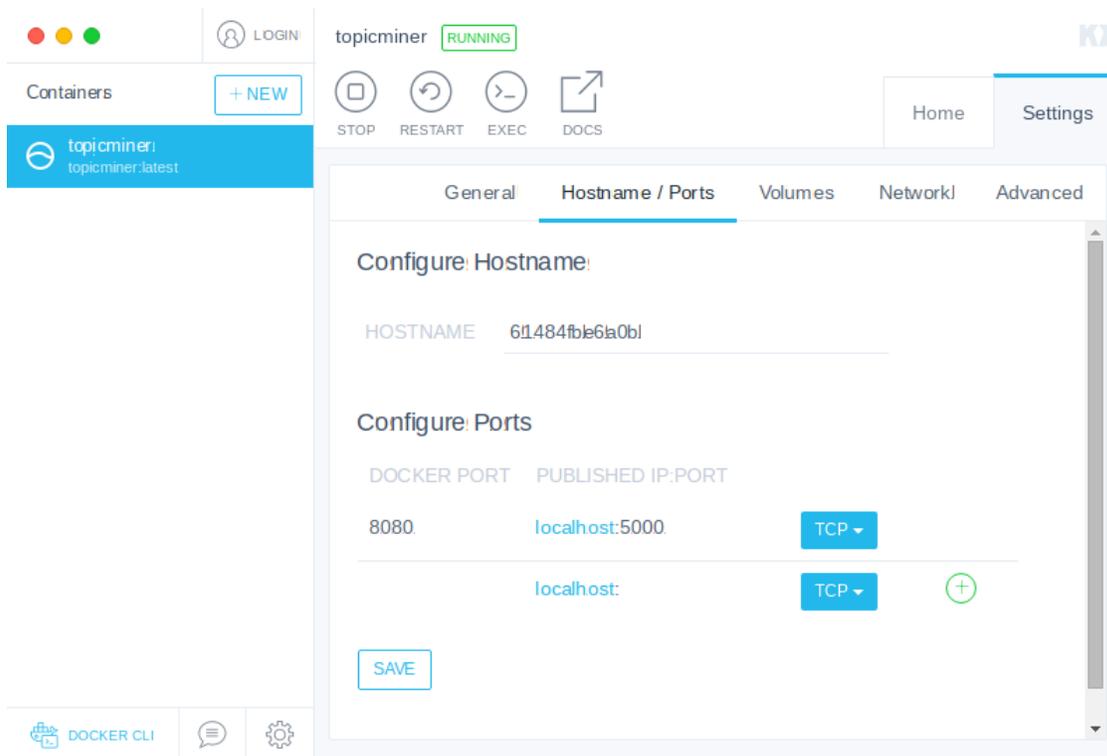
В появившемся окне открытого контейнера в первую очередь нажмите на папку «/data» — таким образом вы создадите системные файлы WebTopicMiner вне виртуальной машины.



Примечание

В Mac и Linux вам скорее всего придётся установить для созданной папки права на запись для всех (chmod o+w), так как Kitematic не имеет средств для запуска контейнера под заданным пользователем. После этого перезапустите контейнер.

Затем на вкладке «Hostname / Ports» раздела «Settings» установите желаемый номер порта в колонке «Published IP:PORT», например, 5000:



Для открытия интерфейса в браузере можно использовать кнопку рядом с надписью «Web preview» на странице контейнера либо пройти по ссылке <http://127.0.0.1:5000/>.

После перезапуска системы контейнер также можно запустить из интерфейса Kitematic.

Использование WebTopicMiner

Для входа в интерфейс используйте логин «admin» и пароль «zmeULF6kEv». Рекомендуем сменить пароль сразу же после установки WebTopicMiner.

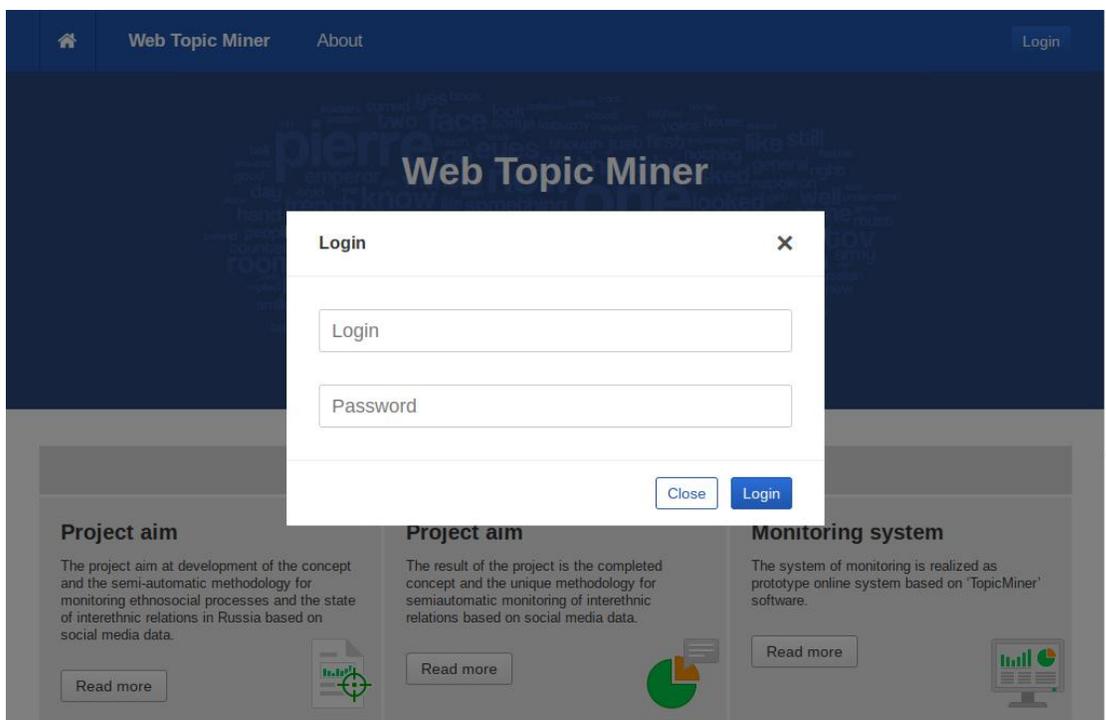
Файлы, отображаемые на вкладке «Workspace», будут храниться в подпапке «workspace/admin/» указанной вами или созданной через Kitematic папки. В подпапке «logs/» хранятся лог-файлы, содержащие отладочную информацию о работе сервера.

Обзор интерфейса

При первом открытии WebTopicMiner отображается приветственная страница и кнопка «Login».



При нажатии на «Login» появляется форма входа. На данный момент учётную запись можно зарегистрировать только обращением в Лабораторию Интернет-Исследований.



После входа на сайт на месте кнопки «Login» в правом верхнем углу будет указано имя текущего пользователя. Нажатие на имя выводит две ссылки: «Change

password» и «Logout». Первая ссылка позволяет изменить пароль, а вторая — выйти из системы.

Также всем пользователям доступен раздел «About» содержащий краткую информацию о продукте.

Вкладка Workspace

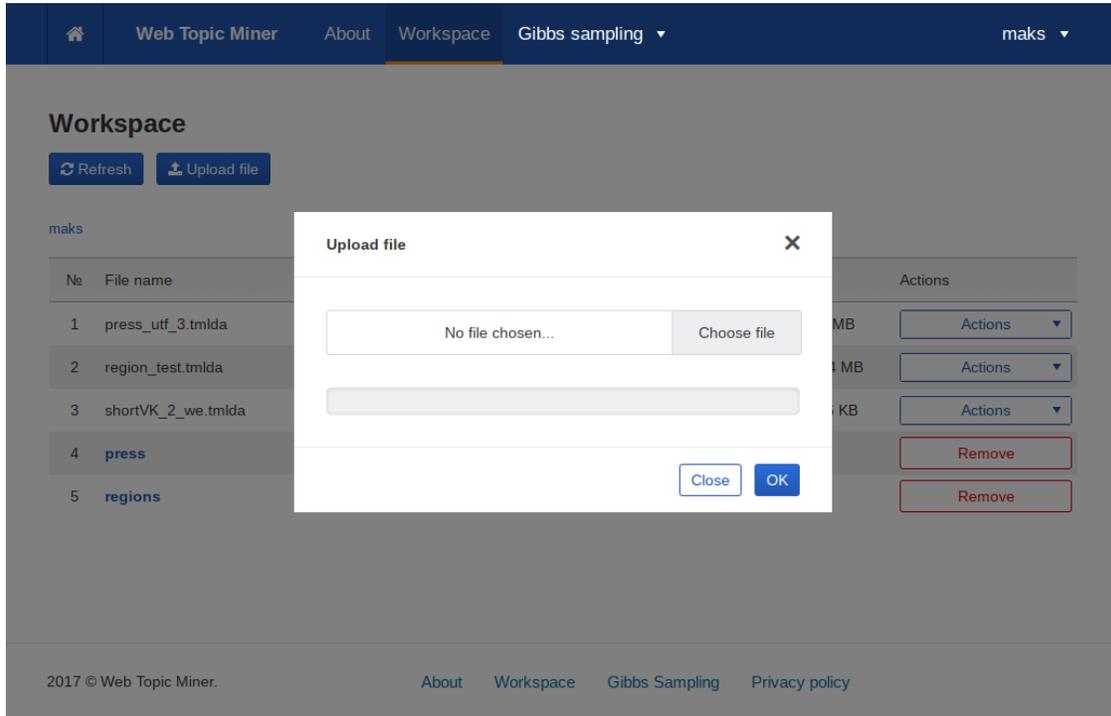
На вкладке Workspace находятся ваши личные файлы и каталоги. Для файлов выводится их тип и размер. На данный момент поддерживаются следующие типы файлов:

- tmla — бинарный формат документов, поддерживаемый ПО [TopicMiner](https://linis.hse.ru/soft-linis#TopicMiner) [https://linis.hse.ru/soft-linis#TopicMiner] для Windows.
- bin — бинарный формат таблиц, также поддерживаемый [TopicMiner](https://linis.hse.ru/soft-linis#TopicMiner) [https://linis.hse.ru/soft-linis#TopicMiner]. В этом формате записываются все результаты тематического моделирования.
- json — текстовый формат хранения произвольных объектов. В этом формате сохраняются данные распределения темы по регионам России.

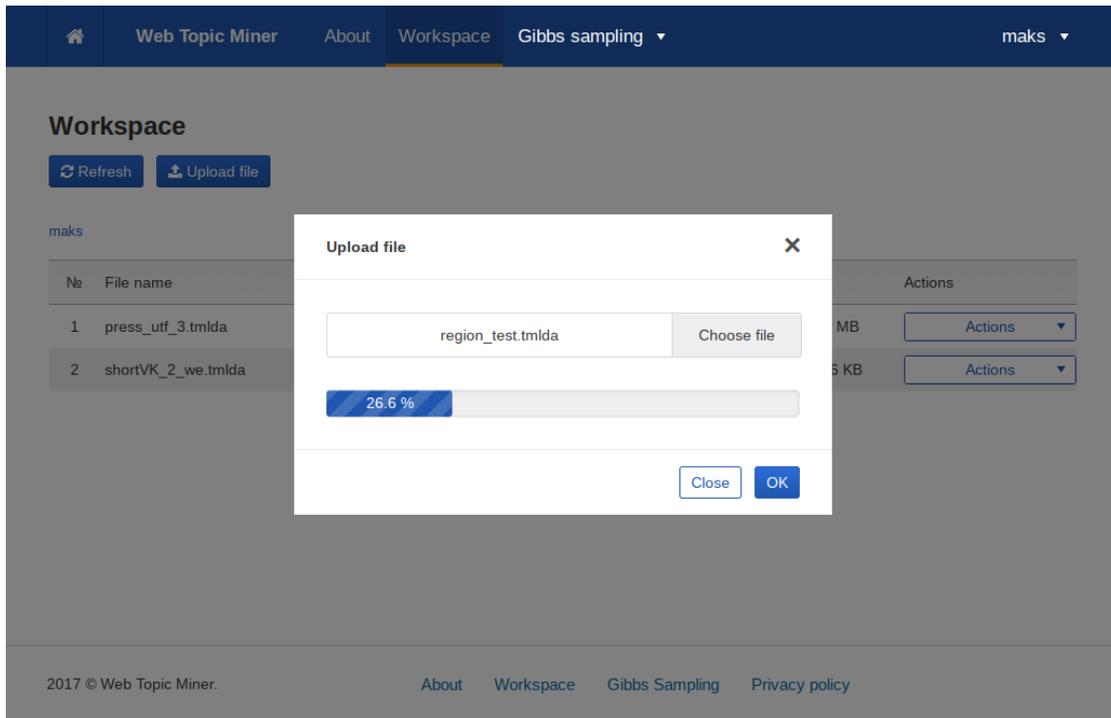
Войти в каталог можно, нажав на его название. Во всех каталогах кроме корневого присутствует кнопка «Back», позволяющая вернуться на уровень выше. Кнопка «Create folder» позволяет создать подкаталог в текущем.

№	File name	Type	Size	Actions
1	24tr_ethnicity3_bigartm.tmla	TMLDA Binary Data File	37.3 MB	Actions
2	dictionary_f1_f2_etic2.dwdict	BigARTM dictionary file	54.6 KB	Actions
3	dic_bigartm.dwdict	BigARTM dictionary file	321.8 KB	Actions
4	docs_test_classific.csv	CSV File	2.1 MB	Actions
5	ethnic3.tmla	TMLDA Binary Data File	20.4 MB	Actions
6	ethnic3_phi.bin	FastGrid binary file	24.8 MB	Actions
7	ethnic3_theta.bin	FastGrid binary file	21.8 MB	Actions
8	for_web_topicminer-1.csv	CSV File	8.7 MB	Actions
9	for_web_topicminer-1_has_topic_ethnicity_classified.csv	CSV File	14.9 MB	Actions
10	orig_doc_25t.csv	CSV File	17.9 MB	Actions

Вы можете загружать новые файлы на сервер, для этого предназначена кнопка «Upload File». При нажатии на неё отобразится форма выбора файла. На данный момент поддерживается только загрузка файлов tmla и bin.

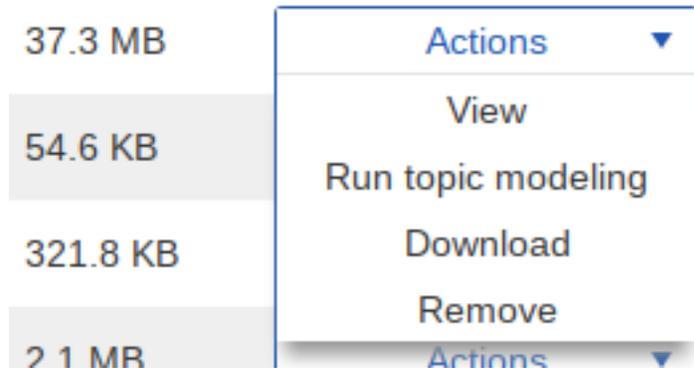


Выбрав файл, нажмите на «Ок» и дождитесь окончания загрузки.



Каждый файл вы можете просмотреть, нажав на кнопку «View». Подробнее о просмотре различных типов файлов смотрите далее.

Также вам доступно выпадающее меню действий:



Для каждого файла доступно два независимых от его типа действия: скачать файл на свой компьютер («Download») и удалить файл с сервера («Remove»). Действия, специфичные для каждого типа, будут описаны в соответствующих разделах.

Работа с файлами tmlda

Формат tmlda представляет собой контейнер для документов, представленных в нескольких видах:

- i. Оригинальный текст
- ii. Лемматизированный текст
- iii. Метаданные
- iv. Список слов, где каждое слово заменено на свой CRC32 хэш-код

Подробнее о том, как получить такой файл из набора текстовых документов, можно прочитать в документации к [TopicMiner](https://linis.hse.ru/soft-linis#TopicMiner) [https://linis.hse.ru/soft-linis#TopicMiner] для Windows.

Веб-сайт WebTopicMiner позволяет просмотреть каждый документ в виде оригинального и лемматизированного текстов, а также его метаданные. Представление в виде хэш-кодов используется только внутри программы и недоступно для просмотра.

Просмотреть документы вы можете, нажав на кнопку «View» на вкладке «Workspace». При этом открывается просмотрщик tmlda файлов, содержащий две вкладки: «Documents» и «Words».

Первая вкладка содержит таблицу с тремя колонками для каждого документа в файле: оригинальный текст, лемматизированный текст и метаданные.

View of file /press_utf_3.tmla

File contains 3997 binary LDA documents with 35095 words in the word map.

Documents	Words	
Original document	Lemmatized document	Document metadata
<p>When President Obama visits Saudi Arabia this week for a meeting with representatives from the Gulf Cooperation Council countries, he should avoid doing what he did at Camp David last May, the last time he met with them: promise more arms sales. Since Mr. Obama hosted that meeting, the United States has offered over \$33 billion in weaponry to its Persian Gulf allies, with the bulk of it going to Saudi Arabia. The results have been deadly. The Saudi-American arms deals are a continuation of a booming business that has developed between Washington and Riyadh during the Obama years. In the first six years of the Obama administration, the United States entered into agreements to transfer nearly \$50 billion in weaponry to Saudi Arabia, with tens of billions of dollars of additional offers in the pipeline. The Pentagon claims that these arms transfers to Saudi Arabia a??improve the security of an important partner which has been and continues to be an important force for political stability and economic progress in the Middle East.a?? Recent Saudi actions suggest</p>	<p>{when} {presid} {obama} {visit} {saudi} {arabia} {thi} {week} {for} {a} {meet} {with} {repres} {from} {the} {gulf} {cooper} {council} {countri} {he} {should} {avoid} {do} {what} {he} {did} {at} {camp} {david} {last} {mai} {the} {last} {time} {he} {met} {with} {them} {promis} {more} {arm} {sale} {sinc} {mr} {obama} {host} {that} {meet} {the} {unite} {state} {ha} {offer} {over} {33} {billion} {in} {weaponri} {to} {it} {persian} {gulf} {alli} {with} {the} {bulk} {of} {it} {go} {to} {saudi} {arabia} {the} {result} {have} {been} {deadli} {the} {saudi} {american} {arm} {deal} {ar} {a} {continu} {of} {a} {boom} {busi} {that} {ha} {develop} {between} {washington} {and} {riyadh} {dure} {the} {obama} {year} {in} {the} {first} {six} {year} {of} {the} {obama} {administr} {the} {unite} {state} {enter} {into} {agreement} {to} {transfer} {nearli} {50} {billion} {in} {weaponri} {to} {saudi} {arabia} {with} {ten} {of} {billion} {of} {dollar} {of} {addit} {offer} {in} {the} {pipelin} {the} {pentagon} {claim} {that} {these} {arm} {transfer} {to} {saudi} {arabia} {a} {improv} {the} {secur} {of} {an} {import} {partner} {which} {ha} {been} {and} {continu}</p>	<p>The Opinion Pages; WILLIAM D. HARTUNG; Obama Shouldn't Trade Cluster Bombs for Saudi Arabia's Friendship; http://www.nytimes.com/2016/04/20/opinion/obama-saudi-arabia-trade-cluster-bombs.html; 2016-04-20</p>

На второй вкладке находятся все слова, встречающиеся во всех документах в файле. TopicMiner для Windows создаёт список слов отсортированным по частоте слова.

View of file /press_utf_3.tmla

File contains 3997 binary LDA documents with 35095 words in the word map.

Documents	Words		
CRC32 Code	Word	Word frequency	TF-IDF
520675445	trump	8588	0.2091217425486855
3400726751	thi	8542	0.09420502465403893
2542907489	ha	8157	0.10489584084643983
2556329580	from	7664	0.08885514695665915
1851262458	mr	7111	0.22148078391213516
730787779	at	7099	0.08609492559262526
3605796537	an	7088	0.08451489622168527
563602355	who	6836	0.09554958624201683
2390616909	thei	6757	0.11027044888823605
1550720220	...	6727	0.09560560909270

Внизу каждой вкладки находятся кнопки, позволяющие переключать страницы с документами или словами.

Работа с файлами bin

Формат bin является бинарным форматом для таблиц, которые могут содержать текст и вещественные числа. В этом формате WebTopicMiner сохраняет матрицы распределений, получающиеся при тематическом моделировании. Открыть такой файл тоже можно нажав на кнопку «View».

View of file `press/press_utf_3_phi_sorted.bin`

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12
new: 0.02823	we: 0.068731	car: 0.015943	articl: 0.020516	washington: 0.0088913	law: 0.021985	polit: 0.015099	trump: 0.071168	more: 0.01952	black: 0.036238	worker: 0.014375	wom: 0.082
media: 0.017646	our: 0.056	park: 0.0099044	[: 0.018618	roosevelt: 0.0080453	state: 0.01253	an: 0.012662	clinton: 0.029167	thei: 0.015349	white: 0.029669	servic: 0.011859	abo: 0.022
report: 0.012544	us: 0.016618	road: 0.0093442	post: 0.014149	museum: 0.0079607	feder: 0.010505	or: 0.0098346	republican: 0.025414	at: 0.014384	american: 0.018248	compani: 0.011377	sexu: 0.018
journalist: 0.01211	my: 0.014852	citi: 0.0071031	game: 0.0117	kennedi: 0.0077915	or: 0.01026	religi: 0.0079638	campaign: 0.016569	their: 0.014028	hate: 0.014711	work: 0.010825	mei: 0.017
time: 0.010373	thi: 0.01413	mile: 0.0069163	april: 0.0094958	at: 0.0074532	case: 0.01016	right: 0.007855	candid: 0.014952	percent: 0.013182	racial: 0.01274	inform: 0.010687	se: 0.015
fox: 0.0098298	their: 0.012909	travel: 0.0060448	at: 0.0085162	book: 0.007284	would: 0.0099873	muslim: 0.00742	donald: 0.014606	than: 0.012885	african: 0.011729	or: 0.0096535	wom: 0.015
press: 0.0094499	peopl: 0.011882	town: 0.0060448	sport: 0.0083326	centuri: 0.0067764	right: 0.0095275	moral: 0.007333	hillari: 0.013096	american: 0.012751	at: 0.010011	their: 0.0096535	whi: 0.013
about: 0.01211	will: 0.014852	into: 0.0071031	team: 0.0117	art: 0.0077915	thi: 0.01026	who: 0.0079638	who: 0.014606	from: 0.012885	group: 0.011729	job: 0.0096535	right: 0.015

Download matrix as CSV Download matrix as CSV with splitted fields

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

2017 © Web Topic Miner. About Workspace Topic modeling User manual

На открывшейся странице будет отображён заголовок таблицы и её строки с возможностью переключения страниц.

Под таблицей так же расположена кнопка «Download matrix as CSV». С помощью этой кнопки можно запустить первые несколько строк bin-файла в формате CSV.

Многие матрицы, создаваемые WebTopicMiner, имеют в каждой ячейке несколько полей, разделённых двоеточием. Кнопка «Download matrix as CSV with splitted fields» позволяет скачать первые несколько строк матрицы в виде CSV файла, в котором каждый столбец исходной матрицы будет разбит на столько столбцов, сколько в нем содержится полей.

Анализ текстов с помощью встроенных классификаторов

WebTopicMiner имеет возможность классифицировать тексты с помощью нескольких классификаторов, разработанных в лаборатории ЛИНИС.

На данный момент доступны следующие классификаторы:

- Содержится ли в тексте межэтнический конфликт? («Whether text has ethnic conflict»)

- Содержится ли в тексте тема этничности? («Whether text is about ethnic topic»)
- Имеет ли текст отрицательную эмоциональную окраску? («If text has negative sentiment»)

Чтобы воспользоваться классификатором, загрузите файл csv с исходными текстами в кодировке UTF-8. WebTopicMiner поддерживает два формата входных файлов:

- На каждой строке файла находится отдельный текст целиком
- Тексты находятся в первой колонке файла CSV, остальные колонки содержат различные данные

Для того, чтобы запустить классификацию, выберите пункт «Run classifier» в выпадающем меню файла. В открывшемся диалоге выберите желаемый классификатор и формат файла из описанных выше. Классификация начнётся автоматически, а на экран будет окно с сообщением, которое закроется по окончании классификации. Если закрыть окно, то необходимо самостоятельно ждать появления файла с результатами.

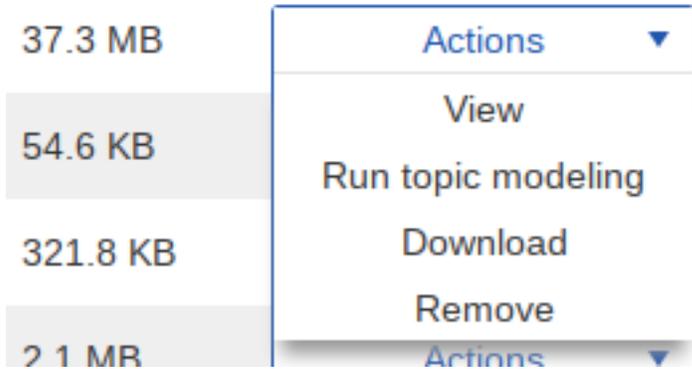
Результат классификации будет сохранен в файл с таким же именем, как исходный плюс краткое имя классификатора и суффикс «_classified» перед расширением. В этом файле после всех столбцов исходного файла будет добавлены ещё два: лемматизированный текст и оценка классификатора — 0 или 1. Для лемматизации текстов используется библиотека [PyMorphy2](https://pymorphy2.readthedocs.io/en/latest/) [https://pymorphy2.readthedocs.io/en/latest/].

Тематическое моделирование

Запуск новой задачи тематического моделирования

WebTopicMiner поддерживает два алгоритма тематического моделирования: сэмплирование Гиббса и [BigARTM](http://bigartm.org/) [http://bigartm.org/].

Для того, чтобы запустить моделирование, предназначена кнопка «Run topic modeling», находящаяся в меню действий для tmla файлов на вкладке «Workspace».



Нажав на эту кнопку вы попадаете в интерфейс настроек моделирования.

2017 © Web Topic Miner. [About](#) [Workspace](#) [Topic modeling](#) [User manual](#)

На этой странице доступен выбор алгоритма: переключатель «Gibbs LDA Sampling» и «BigARTM».

Моделирование с помощью сэмпирования Гиббса

В этом алгоритме для настройки доступны следующие параметры:

- «Alpha», «Beta» — коэффициенты, определяющие поведения алгоритма сэмпирования. Рекомендуется использовать значения по умолчанию
- «Number of topics» — число тем, для которых будет проводиться моделирование
- «Number of iterations» — число итераций алгоритма сэмпирования
- «Save step» — число итераций, по истечении которых текущий результат будет сохраняться и отображаться на этой странице. Не рекомендуется использовать маленькие значения, так как частое сохранение результатов замедляет работу моделирования

- «OpenMP threads» — число параллельных потоков, используемых для моделирования. Рекомендуются значения 4 или 8
- «Output directory name» — название каталога, в который будут сохранены результаты моделирования. Этот каталог появится среди ваших файлов на вкладке «Workspace»

Параметр «Mode» задаёт тип тематической модели. WebTopicMiner поддерживает три типа:

- «LDA» — стандартная модель
- «Supervised LDA» — модель, позволяющая фиксировать начальные темы для нескольких слов
- «Granulated LDA» — гранулированное сэмплирование, учитывающие несколько соседних слов

В режиме «ISLDA» доступна форма ввода фиксированной темы:

The screenshot shows a web interface for configuring the ISLDA mode. At the top, there is a 'Mode:' label followed by a dropdown menu currently showing 'Supervised LDA'. Below this is a section titled 'Fixed topics:' which contains a large text input field with the word 'Word' entered. Underneath this field are five smaller input fields, each labeled 'Topic 1', 'Topic 2', 'Topic 3', 'Topic 4', and 'Topic 5'. Below these fields is a grey button with a plus sign and the text '+ Add'. At the bottom of the configuration area is a prominent blue button labeled 'Run sampling'.

Каждому слову можно назначить до 5 тем в полях «Topic 1», ..., «Topic 5», нумерация тем начинается с 1.

Поле ввода слова будет автоматически предлагать слова, встречающиеся в TMLDA файле:

Mode:

Fixed topics:

Word

trump thi ha from mr at an who thei or would we more state their t
will about presid new if peopl our year all when than what been
you which american us like were can had so no other

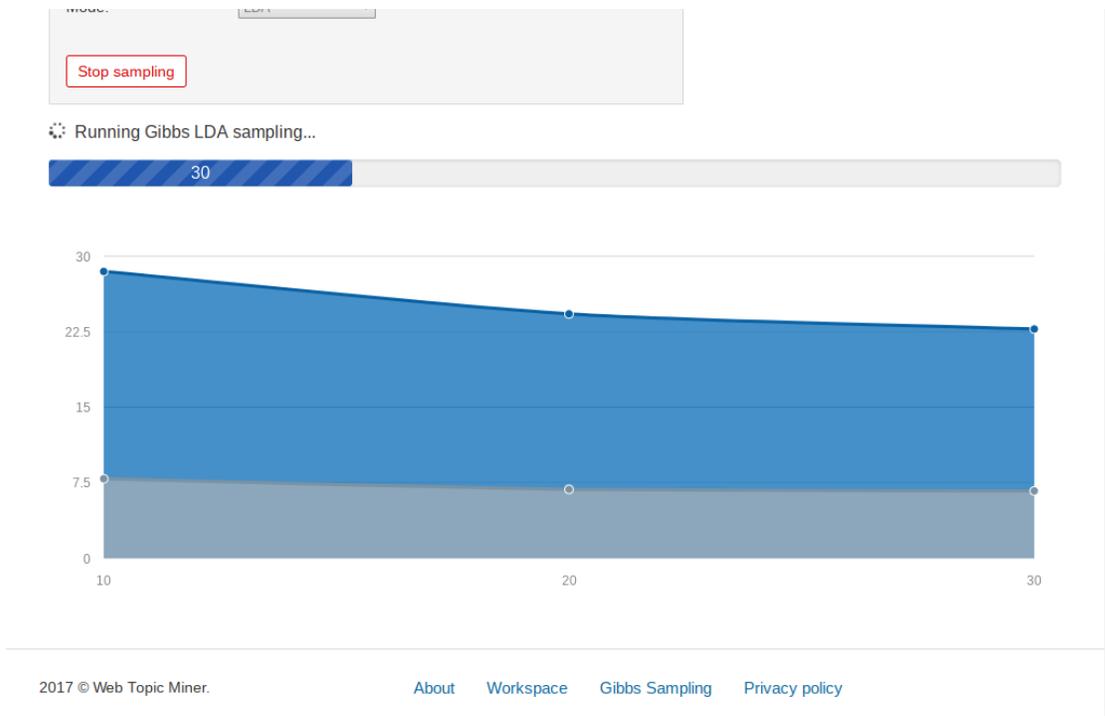
Для того, чтобы запомнить слово, нажмите кнопку «Add».

В режиме «GLDA» доступна поле выбора размера окна:

Mode:

Granulate window:

После ввода всех параметров нажмите на кнопку «Run sampling». После этого все поля ввода станут неактивными, а внизу страницы будет отображаться текущий статус и результаты сэмплирования.



Моделирование с помощью BigARTM

Web Topic Miner [About](#) [Workspace](#) [Topic modeling](#) maks

Topic modeling for /press_utf_3.tmla

Gibbs LDA Sampling BigARTM

Number of topics: Number of iterations:

Save step: Threads:

Output directory name:

BigARTM text options:

BigARTM dictionary file:

BigARTM additional dictionary:

Description of sampling options:

- Number of topics — number of topics to look for in the documents
- Number of iterations — number of iterations of the sampling algorithm to run. The more iterations the better results
- Save step — save the sampling progress each specified number of iterations
- Threads — number of parallel threads to use for calculations
- Output directory name — directory where to store sampling results
- BigARTM text options — arbitrary text configuration for ARTM regularizers, etc. See in the documentation
- BigARTM dictionary file — a file with additional dictionaries for BigARTM multimodal mode

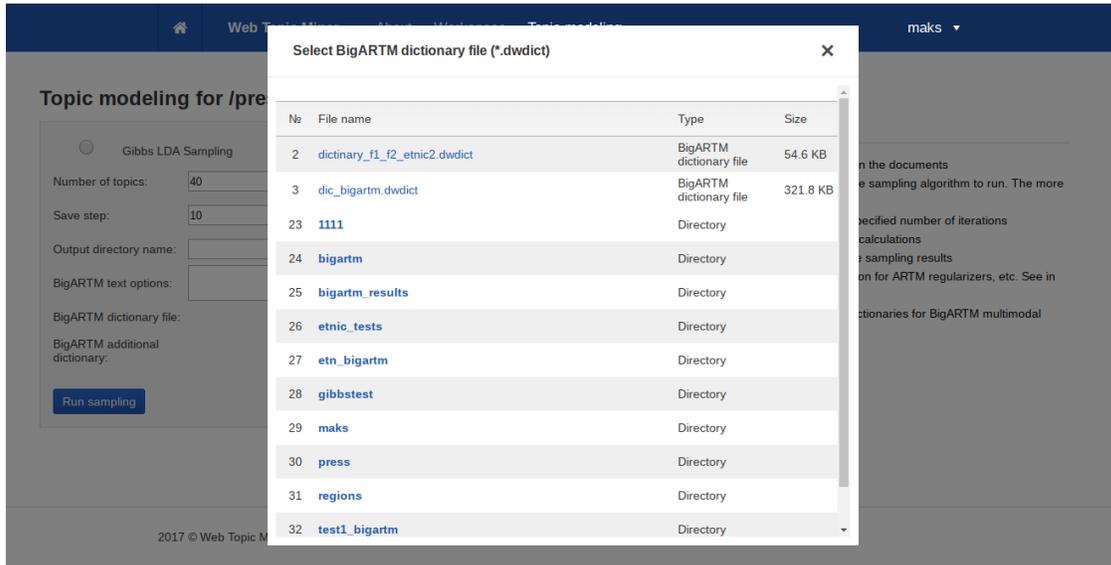
2017 © Web Topic Miner. [About](#) [Workspace](#) [Topic modeling](#) [User manual](#)

В режиме BigARTM опции «Number of topics», «Number of iterations», «Savestep», «Threads» и «Output directory name» имеют тот же смысл, что и для сэмплирования Гиббса. Кроме этого доступны следующие опции:

- «BigARTM text options» — дополнительные параметры регуляторов BigARTM, передаваемые в библиотеку. Может быть пустым.
- «BigARTM dictionary file» — обязательный параметр. Файл со словарем для дополнительных модальностей BigARTM, полученный с помощью TopicMiner.

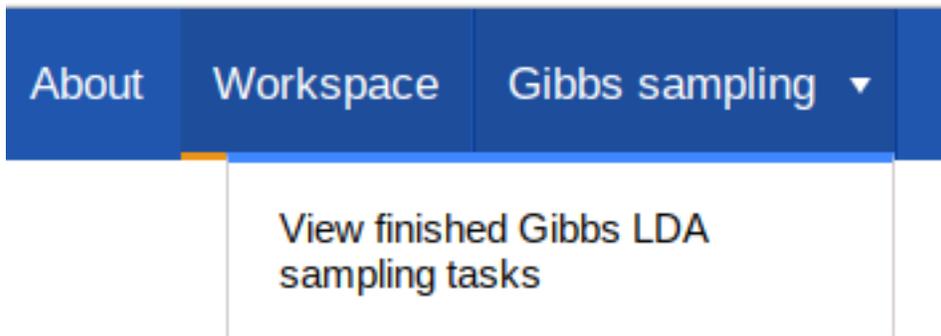
- «BigARTM additional dictionary» — текстовый файл в кодировке UTF-8, содержащий дополнительный словарь для BigARTM, например словарь этнонимов. Каждое слово должно находиться на отдельной строке.

Выбрать файловые параметры можно с помощью кнопки «Browse», которая отображает все подходящие файлы, которые вы загрузили на сервер.



Просмотр выполняющихся и завершенных задач моделирования

Все задачи, которые выполняются на сервере в текущий момент, представлены в выпадающем меню вкладки «Topic modeling»:



Каждая задача представлена названием своего tmlda файла. Ссылка «View finished topic modeling tasks» открывает таблицу, содержащие все завершенные задачи.

Finished Gibbs sampling tasks

No	File name	Iterations	Topics	Creation date	
1	press_utf_3.tmla	100	40	25.10.2017, 18:10:08	

Здесь кроме имени файла указано число итераций, число тем и время создания задачи, для упрощения поиска нужной задачи среди большого числа похожих.

Нажатие на имя файла ведёт на страницу результатов тематического моделирования.

Просмотр результатов

Страница результатов содержит обзор параметров, с которыми запускалось моделирование. Внизу страницы находится график, показывающий сходимость процесса. Синий и серый график показывают соответственно процент документов и слов во всех темах, вероятности которых выше средней. В конце процесса моделирование этот процент не должен сильно меняться от итерации к итерации.



Над графиком находятся две кнопки, позволяющие просмотреть распределение слов и документов по темам.

[Web Topic Miner](#) [About](#) [Workspace](#) [Gibbs sampling](#) maks ▾

Word-topic distributions

[Show sorted distributions](#)

Word	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10	Topic11
trump	0.00027682	4.9769e-05	6.2253e-06	6.1224e-06	8.4599e-06	1.4368e-06	4.5681e-05	0.071168	1.4843e-06	5.0534e-06	3.0173e-05
thi	0.0019052	0.01413	0.0028699	0.00080203	0.00076985	0.0094413	0.0067456	0.00015316	0.0033411	0.0055133	0.0002286
ha	0.0041849	0.00046719	0.00025524	6.1224e-06	9.3059e-05	0.009312	0.005658	0.0069268	0.009664	5.0534e-06	0.0002286
from	0.00098244	0.0089922	0.00031749	6.1224e-06	8.4599e-06	0.0084355	0.0060278	0.0026124	0.011475	0.0022286	0.0002286
mr	5.4279e-06	1.6055e-06	6.2253e-06	6.1224e-06	8.4599e-06	3.0173e-05	2.1753e-06	0.00030416	1.6327e-05	5.0534e-06	3.0173e-05
at	0.0042934	0.004818	0.00019298	0.0085162	0.0074532	0.004355	0.00080704	0.0034752	0.014384	0.010011	0.0002286
an	0.00065677	0.00051535	0.005609	0.0040469	9.3059e-05	0.0092258	0.012662	0.00032574	0.0052409	0.0014706	0.0002286
who	5.4279e-06	0.0088156	6.8478e-05	6.1224e-06	8.4599e-06	0.0058349	0.007246	0.011845	0.0057307	0.0022791	3.0173e-05

2017 © Web Topic Miner. [About](#) [Workspace](#) [Gibbs Sampling](#) [Privacy policy](#)

В матрице слова-темы для каждого слова указана вероятность принадлежности этого слова каждой теме.

Web Topic Miner About Workspace Gibbs sampling maks

Document-topic distributions

Show sorted distributions Show one topic on the map

ID	Original text	Nick	Field 1	Field 2	Field 3	Field 4	Field 5
1	When President Obama visits Saudi Arabia this week for a meeting	The Opinion Pages	WILLIAM D. HARTUNG	Obama Shouldn't Trade Cluster Bombs for Saudi Arabia	http://www.nytimes.com/2016/04/20/opinion/obama-saudi-arabia-tra	2016-04-20	
2	Saturday marks President Trump's 100th day in office. We asked	The Opinion Pages	THE EDITORS	How Trump's First 100 Days Have Changed You	https://www.nytimes.com/2017/04/28/opinion/how-trumps-first-100-	2017-04-28	
	The European Union to its	The	THE	Europe Takes a Braver	http://www.nytimes.com/2016/12/28/opinion/europe-	2016-	

2017 © Web Topic Miner. About Workspace Gibbs Sampling Privacy policy

В матрице документы-темы для каждого документа также указана вероятность принадлежности темам. При нажатии на строку открывается всплывающее окно, содержащие полный текст выбранного документа. Дополнительно в этой матрице представлены метаданные каждого документа в файле.

Обе матрицы можно отсортировать для удобства чтения. При этом сортированная матрица слова-темы для каждой темы будет содержать ячейку вида «слово: вероятность». Слова в каждом столбце упорядочены по убыванию вероятности.

Word-topic distributions

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12
new: 0.02823	we: 0.068731	car: 0.015943	articl: 0.020516	washington: 0.0088913	law: 0.021985	polit: 0.015099	trump: 0.071168	more: 0.01952	black: 0.036238	worker: 0.014375	women: 0.08219
media: 0.017646	our: 0.056	park: 0.0099044	[: 0.018618	roosevelt: 0.0080453	state: 0.01253	an: 0.012662	clinton: 0.029167	thei: 0.015349	white: 0.029669	servic: 0.011859	abort: 0.02269
report: 0.012544	us: 0.016618	road: 0.0093442	post: 0.014149	museum: 0.0079607	feder: 0.010505	or: 0.0098346	republican: 0.025414	at: 0.014384	american: 0.018248	compani: 0.011377	sexual: 0.01894
journalist: 0.01211	my: 0.014852	citi: 0.0071031	game: 0.0117	kennedi: 0.0077915	or: 0.01026	religi: 0.0079638	campaign: 0.016569	their: 0.014028	hate: 0.014711	work: 0.010825	men: 0.01757
time: 0.010373	thi: 0.01413	mile: 0.0069163	april: 0.0094958	at: 0.0074532	case: 0.01016	right: 0.007855	candid: 0.014952	percent: 0.013182	racial: 0.01274	inform: 0.010687	sex: 0.01537
fox: 0.0098298	their: 0.012909	travel: 0.0060448	at: 0.0085162	book: 0.007284	would: 0.0099873	muslim: 0.00742	donald: 0.014606	than: 0.012885	african: 0.011729	or: 0.0096535	woman: 0.01531
press: 0.0094499	peopl: 0.011882	town: 0.0060448	sport: 0.0083326	centuri: 0.0067764	right: 0.0095275	moral: 0.007333	hillari: 0.013096	american: 0.012751	at: 0.010011	their: 0.0096535	who: 0.01334
about: -----	will: -----	into: -----	team: -----	art: -----	thi: -----	who: -----	who: -----	from: -----	group: -----	job: -----	right: -----

В сортированной матрице документы-темы аналогично содержатся упорядоченные ячейки вида «id документа: вероятность». При нажатии на каждую ячейку открывается полный текст указанного в ней документа.

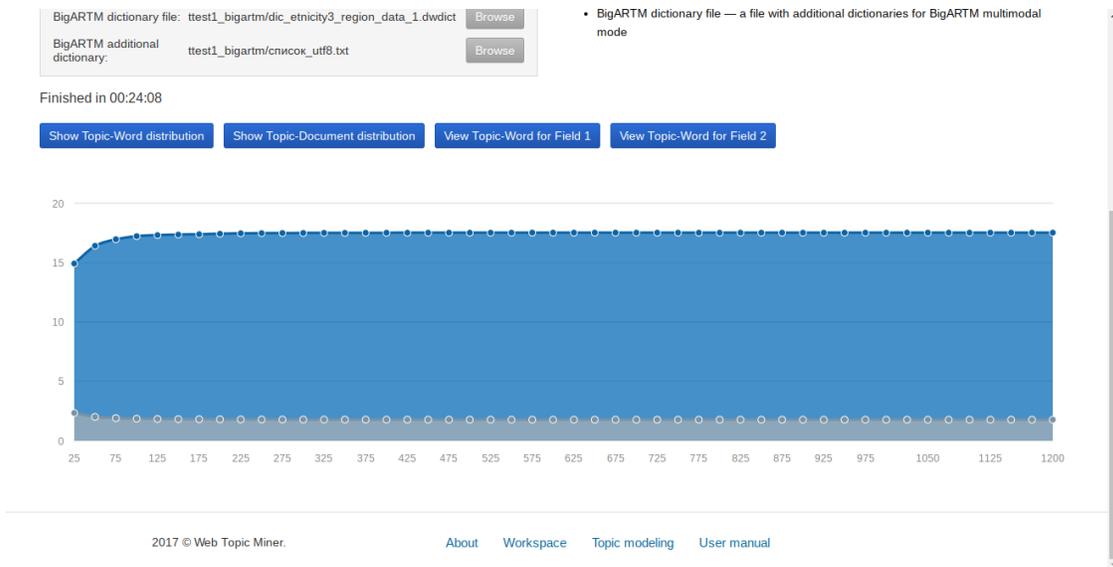
Document-topic distributions

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15
1649: 0.62088	1855: 0.44149	966: 0.38021	1927: 0.5625	2334: 0.59474	3411: 0.74668	2997: 0.56239	1712: 0.53049	9: 0.5318	3042: 0.52067	3703: 0.49332	158: 0.44667	1444: 0.51974	3782: 0.72529	3066: 0.57083
1601: 0.5989	2363: 0.43788	412: 0.33529	2124: 0.54327	2525: 0.58421	2249: 0.69894	2756: 0.48656	2006: 0.46567	2779: 0.52645	3691: 0.42746	704: 0.43981	2580: 0.42	1685: 0.50658	1275: 0.71524	2105: 0.55515
1588: 0.56593	668: 0.43357	1602: 0.32486	1953: 0.53365	2526: 0.55263	3847: 0.66559	1436: 0.47292	2194: 0.45335	2228: 0.45944	3097: 0.34058	2179: 0.43959	32: 0.40667	1538: 0.49342	1441: 0.69121	3020: 0.53649
2383: 0.34302	554: 0.42091	1066: 0.32178	1778: 0.5315	604: 0.50503	1728: 0.56464	3396: 0.47175	1981: 0.4375	3079: 0.45322	872: 0.33938	2772: 0.4287	684: 0.33146	3996: 0.30561	497: 0.65206	3178: 0.52405
1910: 0.32937	2176: 0.41204	2637: 0.31341	1464: 0.52114	2858: 0.46802	1316: 0.56436	3425: 0.44113	1690: 0.43293	499: 0.44987	2251: 0.3155	736: 0.39872	2901: 0.27384	1706: 0.3	1077: 0.64148	2856: 0.51382
1062: 0.3097	3262: 0.40833	1009: 0.29609	1626: 0.5	3181: 0.46482	2814: 0.56095	2267: 0.43785	3715: 0.40153	3906: 0.40424	1832: 0.3154	3043: 0.39242	1647: 0.25789	672: 0.26247	1389: 0.6335	606: 0.48969
755: 0.3022	798: 0.39881	60: 0.2931	2508: 0.49522	2964: 0.45031	2724: 0.5559	284: 0.43701	1904: 0.35393	416: 0.40094	2243: 0.31308	1415: 0.38567	1965: 0.24805	1201: 0.23534	1820: 0.61712	2573: 0.48301
3474: 0.3022	2595: 0.39881	2325: 0.2931	1816: 0.49522	906: 0.45031	3774: 0.5559	3743: 0.43701	2192: 0.35393	290: 0.40094	374: 0.31308	241: 0.38567	1639: 0.24805	1544: 0.23534	2874: 0.61712	513: 0.48301

Original text of the document

Regarding the May 8 Metro article "A veteran neighbor moves to Philly": I chose to interpret Anne Seymour's sign in the accompanying photograph, reading "Anchors Away," as a clever play on words and not ignorance. To weigh anchor is to raise the anchor from the seafloor and pull it onto the ship. The anchor is then aweigh, and the ship is able to move on. But I guess it did go away, so perhaps . . . Pen Suritz , Arlington

Для задач в режиме BigARTM также доступны дополнительные кнопки, показывающие распределения тем для всех дополнительных модальностей.



С этими матрицами можно производить те же действия, что и с обычной матрицей Word-Topic.

Сентимент-анализ тем

Для документов на русском языке доступно проведение сентимент-анализа тем. Для этого предназначена кнопка «Add sentiment data for words in each topic», размещённая на странице отсортированной матрицы слова-темы.

[Web Topic Miner](#) [About](#) [Workspace](#) [Gibbs sampling](#) [maks](#)

Word-topic distributions

[Add sentiment data for words in each topic](#)

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
цвет: 0.015131	хотеться: 0.031553	красота: 0.046292	женщина: 0.092523	мама: 0.067325	деньга: 0.010463	мир: 0.026601	вод: 0.017987	спешивать: 0.079925	алексей: 0.016309
размер: 0.010831	забывать: 0.023542	услуга: 0.040482	мужчина: 0.080062	ребенок: 0.051991	рубль: 0.0097596	бог: 0.01506	пить: 0.013205	помощь: 0.058097	такать: 0.009227
ряд: 0.0095796	ждать: 0.02317	цена: 0.03648	жена: 0.028849	дитя: 0.046076	автомобиль: 0.0091582	земля: 0.01424	продукт: 0.011734	подарок: 0.057823	александ: 0.008739
лицо: 0.0079767	время: 0.021916	салон: 0.030619	муж: 0.028805	девочка: 0.020296	работа: 0.0071724	свет: 0.0098599	чай: 0.011722	прийти: 0.056214	пойти: 0.008382
см: 0.0077291	момент: 0.018665	акция: 0.030199	женский: 0.010749	пап: 0.019855	машин: 0.0071044	жизнь: 0.0080105	час: 0.010827	подарить: 0.055956	хуй: 0.007326
малыш: 0.0072991	иногда: 0.01692	казань: 0.025455	секс: 0.0099814	сын: 0.018174	компания: 0.0069455	ангел: 0.0068842	сок: 0.0093557	собирать: 0.05322	блин: 0.006985
бумага: 0.0070384	верить: 0.015464	рассказывать: 0.023709	мужской: 0.0073388	родители: 0.018138	информация: 0.006208	смерть: 0.0064671	мед: 0.0090002	нажимать: 0.051288	пиво: 0.00674
ребенок: 0.0070384	помнить: 0.015464	whatsapp: 0.023709	находить: 0.0073388	малыш: 0.018138	телефон: 0.006208	умирать: 0.0064671	день: 0.0090002	loc: 0.046265	сух: 0.00674

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

2017 © Web Topic Miner. [About](#) [Workspace](#) [Gibbs Sampling](#) [Privacy policy](#)

При нажатии на эту кнопку можно выбрать, какое количество самых вероятных слов в каждой теме учитывать. После того, как процесс будет закончен, в каждую

ячейку кроме слова и его вероятности добавится новое число — сентиментная оценка слова по шкале от -2 до 2. Для удобства анализа положительные слова окрашены в зелёный цвет, а отрицательные — в красный. Нейтральные слова цвета не имеют.

Word-topic distributions

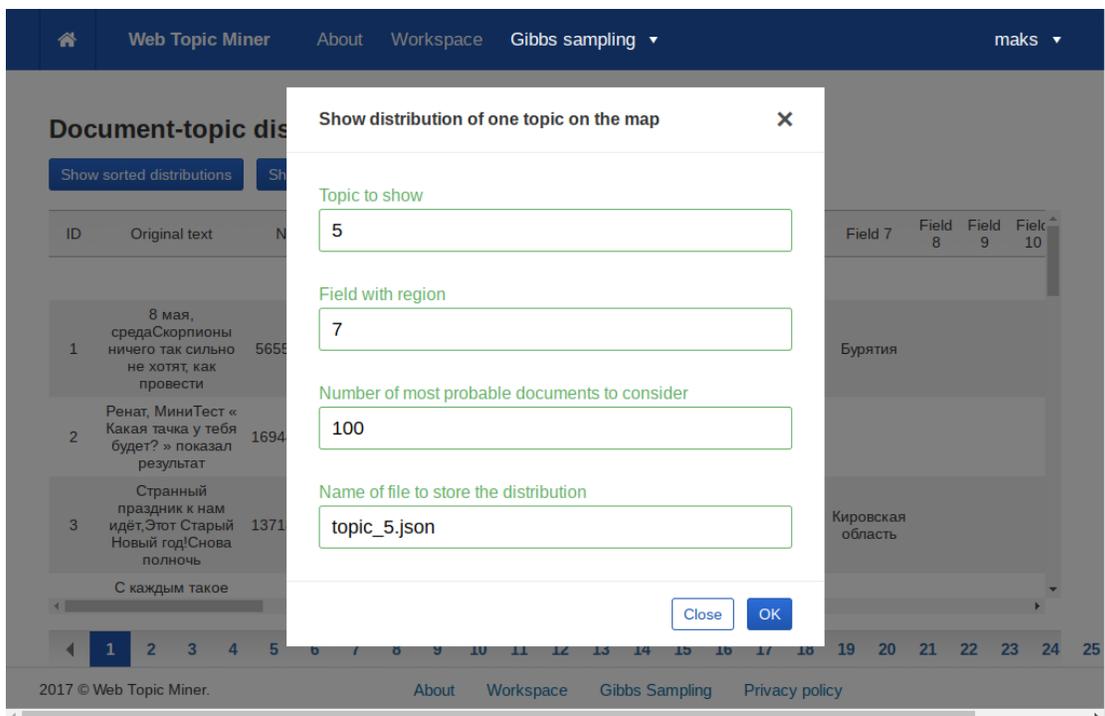
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
цвет: 0.015131: 0	хотеться: 0.031553: 0	красота: 0.046292: 0	женщина: 0.092523: 0	мама: 0.067325: 0	деньга: 0.010463: 0	мир: 0.026601: 0	вод: 0.017987: 0	спешивать: 0.079925: 0	алексей: 0.016309: 0
размер: 0.010831: 0	забывать: 0.023542: 0	услуга: 0.040482: 0	мужчина: 0.080062: 0	ребенок: 0.051991: 0	рубль: 0.0097596: 0	бог: 0.01506: 0	пить: 0.013205: 0	помощь: 0.058097: 1	такать: 0.0092271: 0
ряд: 0.0095796: 0	ждать: 0.02317: 0	цена: 0.03648: 0	жена: 0.028849: 0	дитя: 0.046076: 1	автомобиль: 0.0091582: 0	земля: 0.01424: 0	продукт: 0.011734: 0	подарок: 0.057823: 0	александр: 0.0087398: 0
лицо: 0.0079767: 0	время: 0.021916: 0	салон: 0.030619: 0	муж: 0.028805: 0	девочка: 0.020296: 0	работа: 0.0071724: 0	свет: 0.0098599: 0	чай: 0.011722: 0	прийти: 0.056214: 0	пойти: 0.0083825: 0
см: 0.0077291: 0	момент: 0.018665: 0	акция: 0.030199: 0	женский: 0.010749: 0	пап: 0.019855: 0	машин: 0.0071044: 0	жизнь: 0.0080105: 0	час: 0.010827: 0	подарить: 0.055956: 0	хуй: 0.0073268: -1
мальш: 0.000014: 0	иногда: 0.000014: 0	казань: 0.000014: 0	секс: 0.000014: 0	сын: 0.018174: 0	компания: 0.000014: 0	ангел: 0.000014: 0	сок: 0.000014: 0	собирать: 0.000014: 0	блин: 0.000014: 0

2017 © Web Topic Miner. About Workspace Gibbs Sampling Privacy policy

На данный момент WebTopicMiner использует встроенный сентимент-словарь для русского языка, разработанный в Лаборатории Интернет-Исследований.

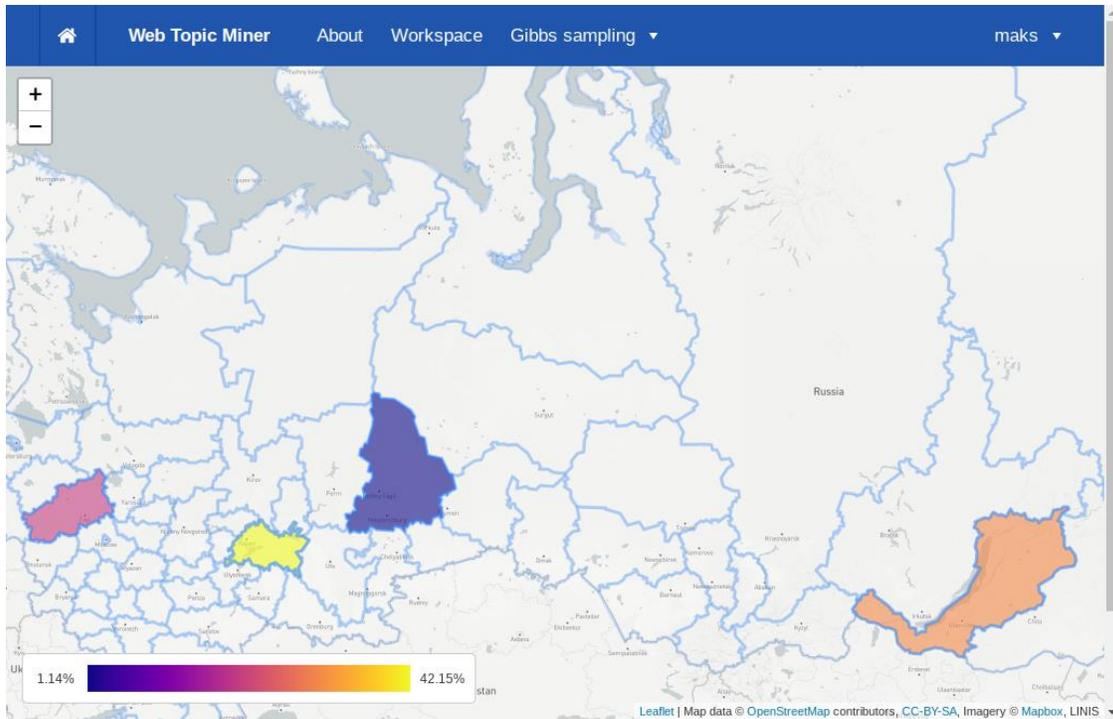
Отображение распределения темы на карте России

Если какой-то из столбцов метаданных документа содержит названия субъекта Российской Федерации, к которому относится документ, возможно отобразить распределение произвольной темы на карте России. Для этого предназначена кнопка «Show one topic on the map», при нажатии на которую открывается форма выбора темы.



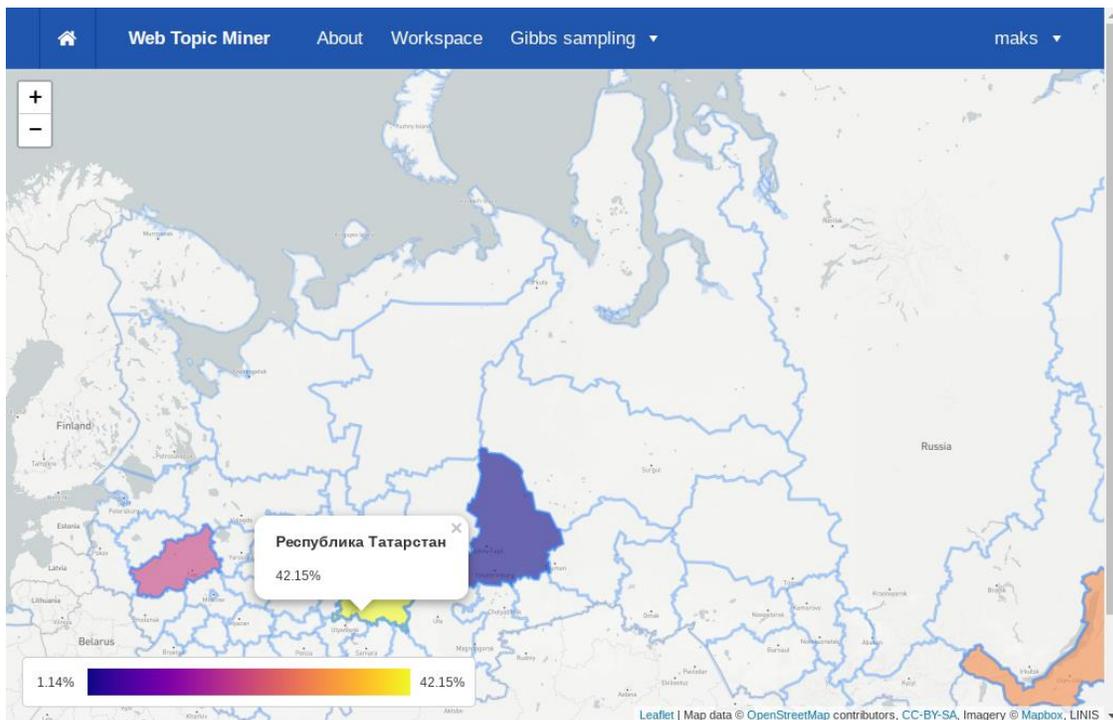
Выберите номер темы, которую требуется отобразить, номер поля метаданных, в котором содержится регион, и число самых вероятных документов темы, которые требуется учитывать. В последнем поле укажите имя json-файла, в которой будет сохранён результат. После нажатия на «Ok» вы будете перенаправлены в каталог, указанный как «Output directory» для задачи сэмплирования. Через небольшое время в этом каталоге появится json-файл с указанным вами названием.

Для того, чтобы посмотреть карту, нажмите на ссылку «View on map» в меню действий этого файла. Откроется карта России, на которую нанесены контуры регионов.

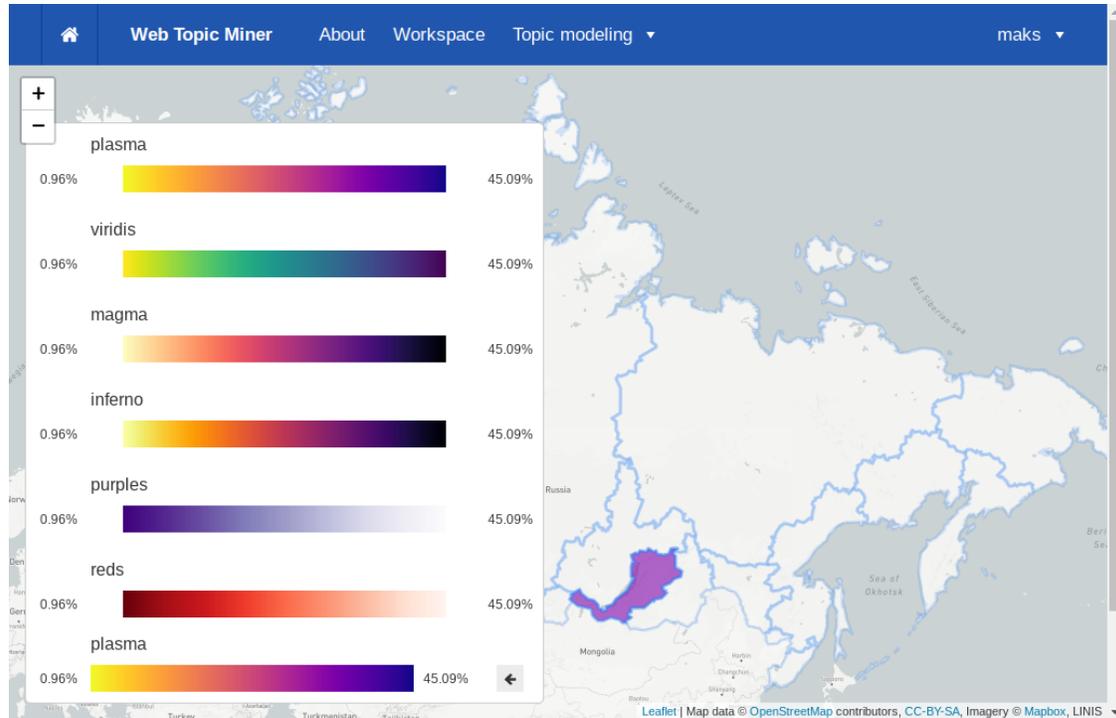


Регионы, в которых выбранной темы нет, отображены только контуром. Остальные регионы окрашены по цветовой шкале от синего до оранжевого в зависимости от доли темы в этом регионе.

При нажатии на регион показывается его название и процент темы.



В левом нижнем углу карты находится индикатор цветовой шкалы, который показывает, в каких пределах находятся вероятности тем и каким им соответствуют цвета. Нажав на индикатор, можно изменить цветовую шкалу на любую из встроенных.



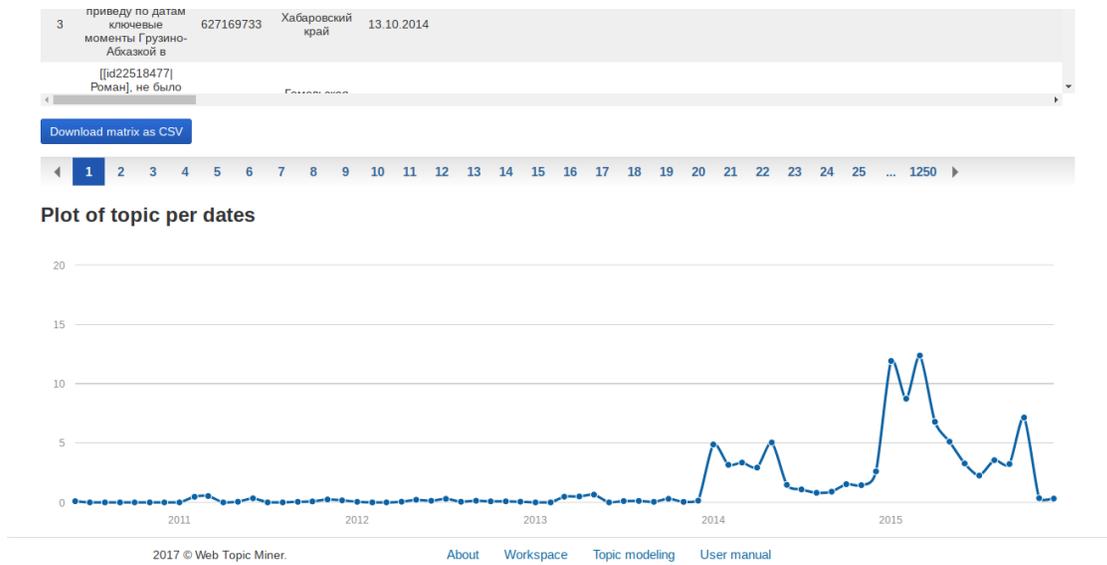
Кнопка со стрелкой позволяет инвертировать направление цветовой шкалы (от темного к светлому и наоборот).

Просмотр распределения тем во времени

Если какое-то поле метаданных файла TMLDA содержит дату написания документа в формате «ДД.ММ.ГГГГ», то можно построить график распределения вероятности выбранной темы во времени.

Для этого надо нажать кнопку «Plot topic probabilities per time interval» на странице матрицы Topic-Document или дополнительной матрицы BigARTM для поля, содержащего дату. В открывающемся диалоге можно указать номер поля с датой, интересующий топик, интервал дат (месяц или неделя) и количество самых вероятных документов в теме, которые требуется рассмотреть.

После ввода всех параметров график будет отображён в нижней части той же страницы.



По горизонтальной оси отложены первые даты выбранного периода (понедельник каждой недели или первое число каждого месяца), по вертикальной — суммарная вероятность (в процентах) документов, написанных в этот период. Вероятности нормализованы так, чтобы сумма по всем периодам равнялась 100%.