INTERNET STUDIES LAB

HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY
SAINT PETERSBURG

Руководство пользователя

Санкт - Петербург

2017

Оглавление:

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 каком-либо причинам вы хотите настроить свою локальную копию или просто оценить возможности WebTopicMiner, воспользуйтесь этой инструкцией.

Самый простой способ установить продукт — с помощью виртуальной машиной 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/.

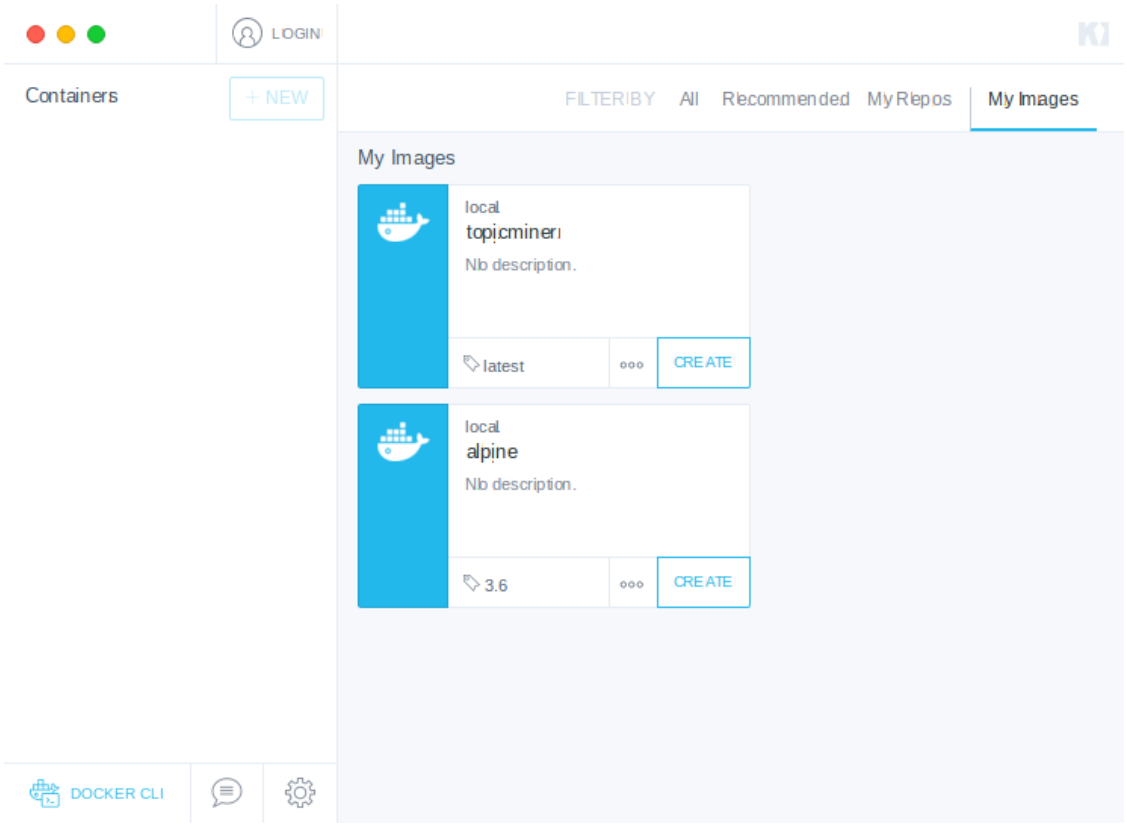
После перезапуска системы контейнер можно запустить с помощью

$ 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».

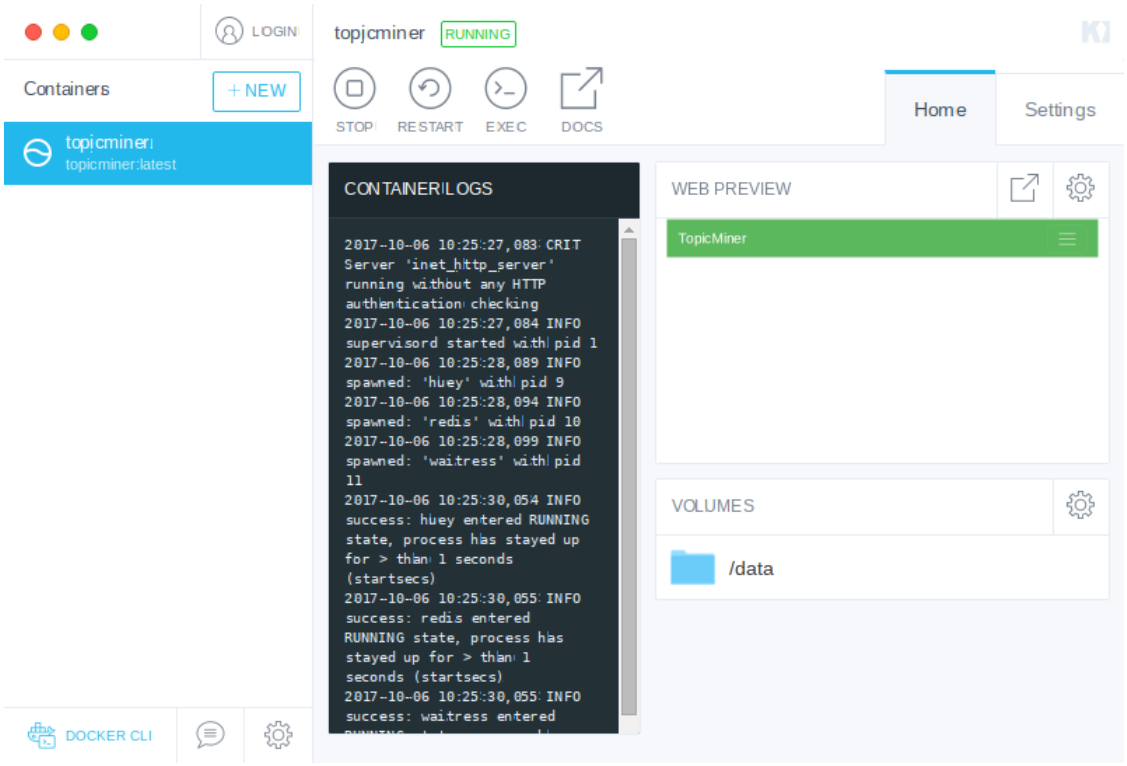
Примечание

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.

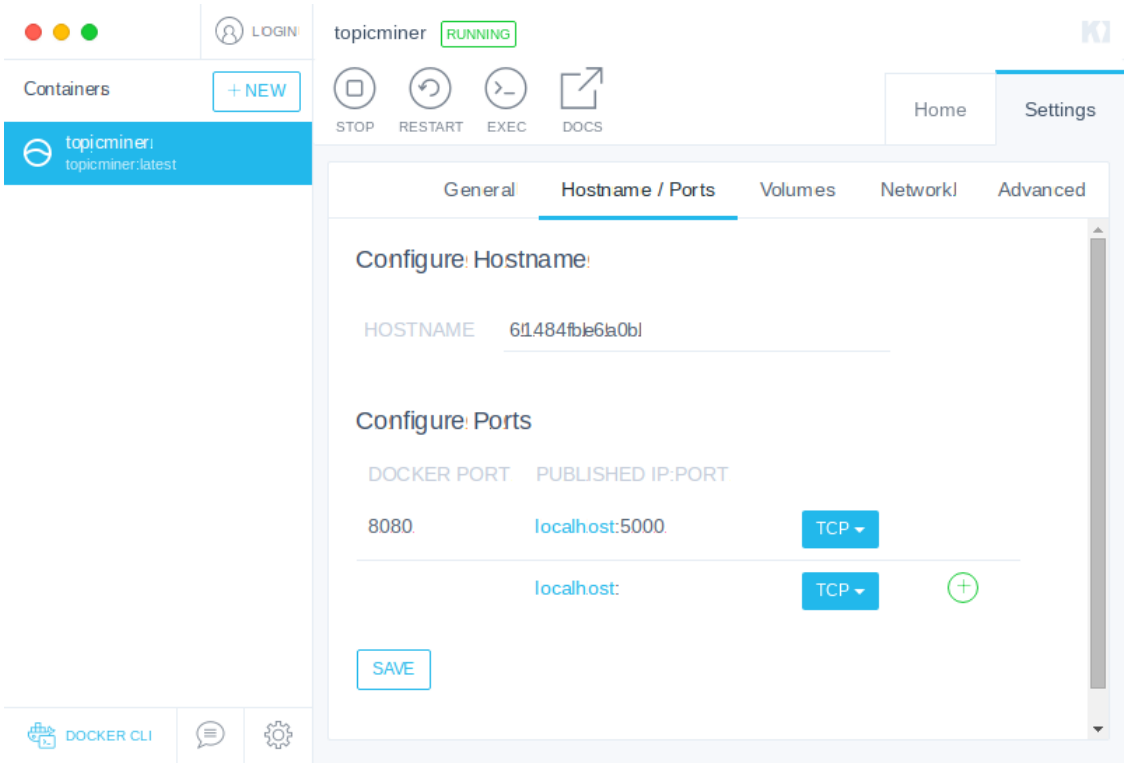
В появившемся окне открытого контейнера в первую очередь нажмите на папку «/data» — таким образом вы создадите системные файлы WebTopicMiner вне виртуальной машины.



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.

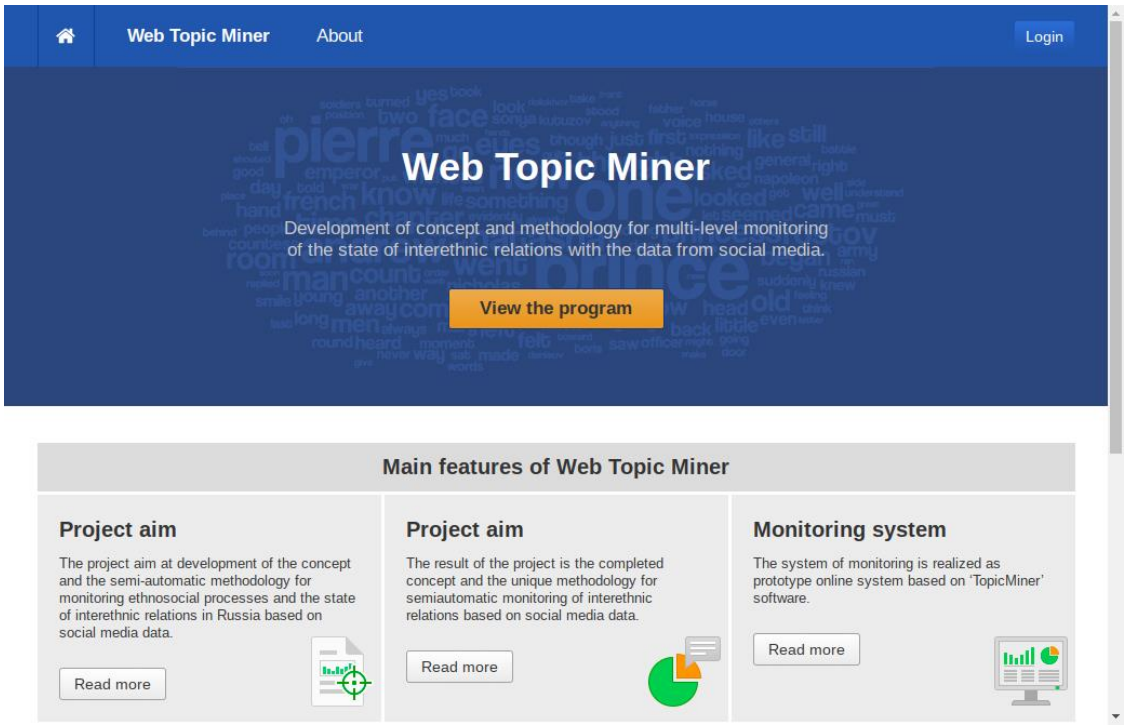
Использование WebTopicMiner

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

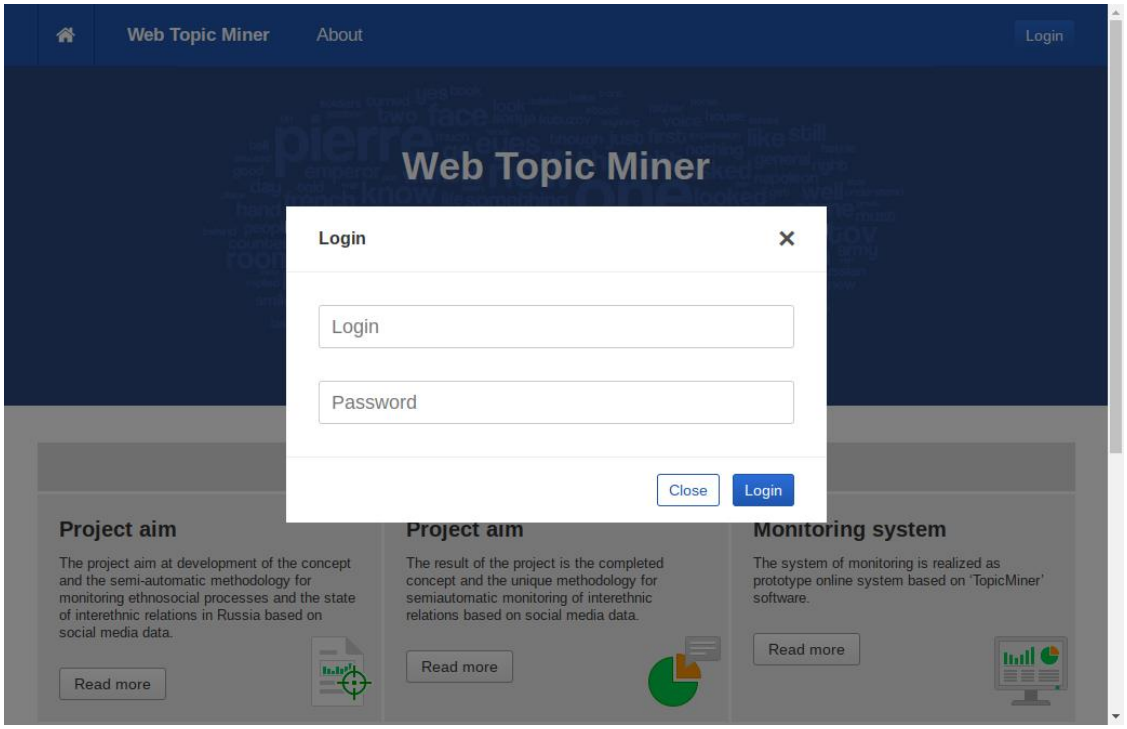
Файлы, отображаемые на вкладке «Workspace», будут храниться в подпапке «workspace/admin/» указанной вами или созданной через Kitematic папки. В подпапке «logs/» хранятся лог-файлы, содержащие отладочную информацию о работе сервера.

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.

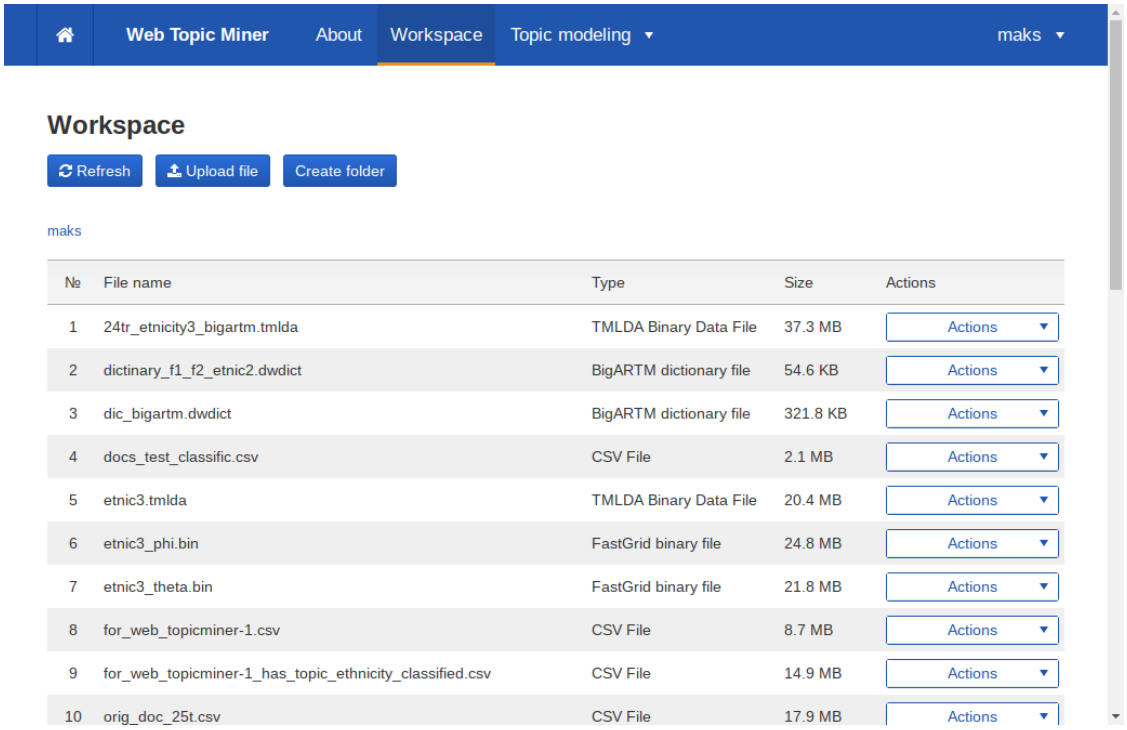
Также всем пользователям доступен раздел «About» содержащий краткую информацию о продукте.

Вкладка Workspace

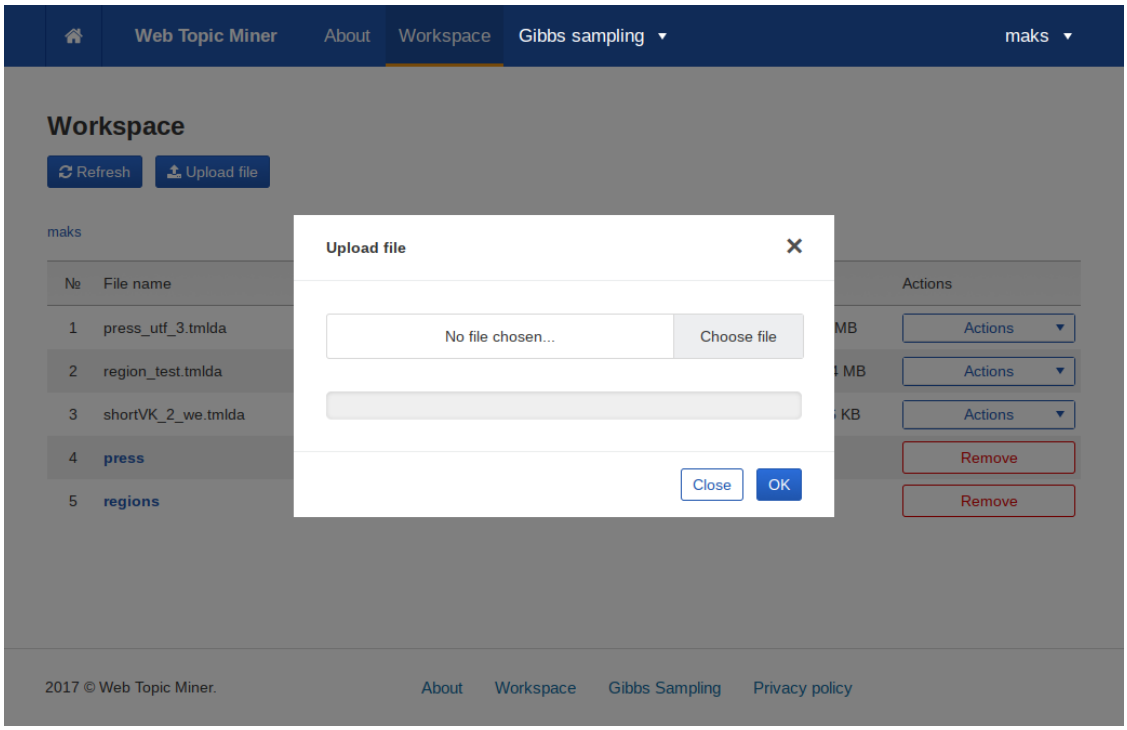
На вкладке Workspace находятся ваши личные файлы и каталоги. Для файлов выводится их тип и размер. На данный момент поддерживаются следующие типы файлов:

- tmlda — бинарный формат документов, поддерживаемый ПО TopicMiner [https://linis.hse.ru/soft-linis#TopicMiner] для Windows.
- bin — бинарый формат таблиц, также поддерживаемый 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.

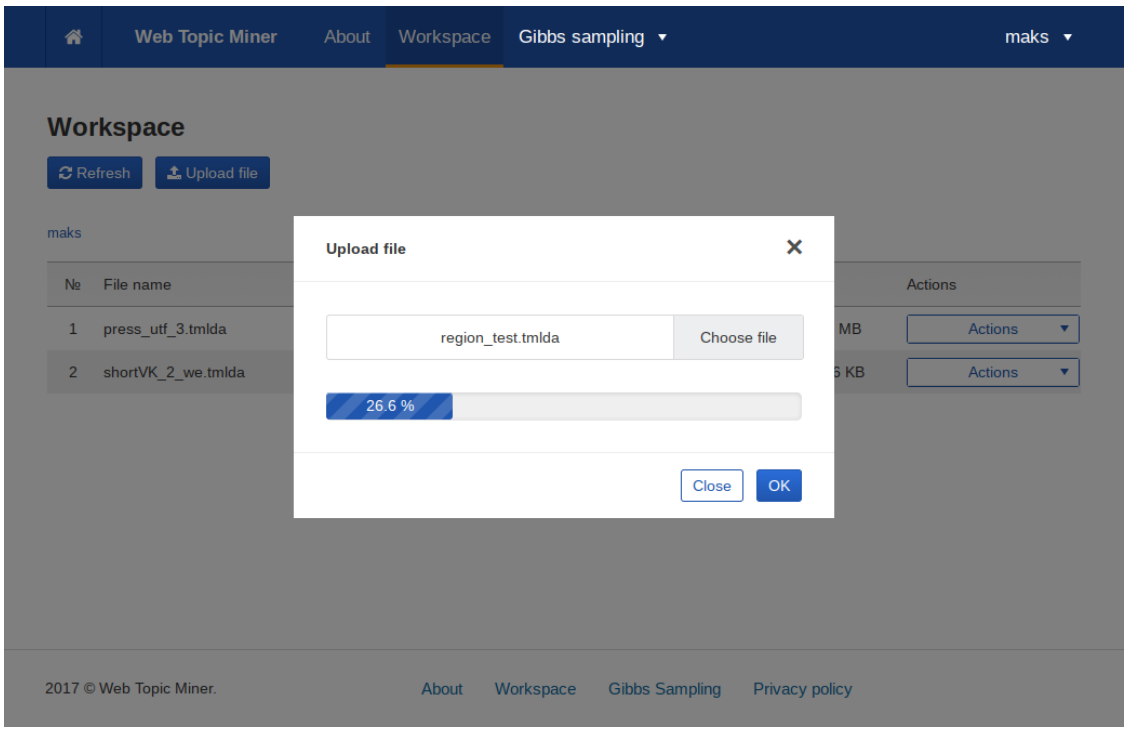
Войти в каталог можно, нажав на его название. Во всех каталогах кроме корневого присутствует кнопка «Back», позволяющая вернуться на уровень выше. Кнопка «Create folder» позволяет создать подкаталог в текущем.



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

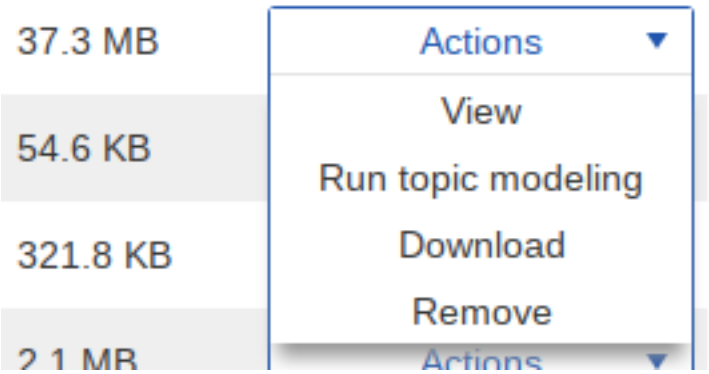


Выбрав файл, нажмите на «Ok» и дождитесь окончания загрузки.



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 независящих 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

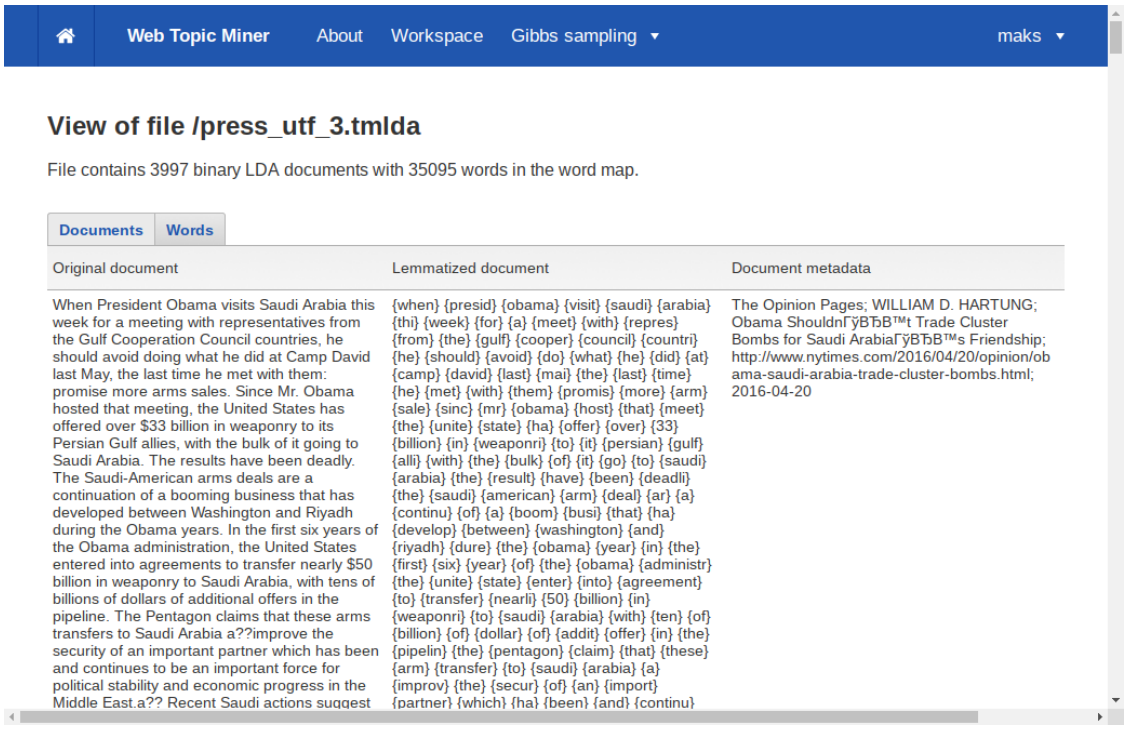
iv. Список слов, где каждое слово заменено на свой CRC32 хэш-код

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

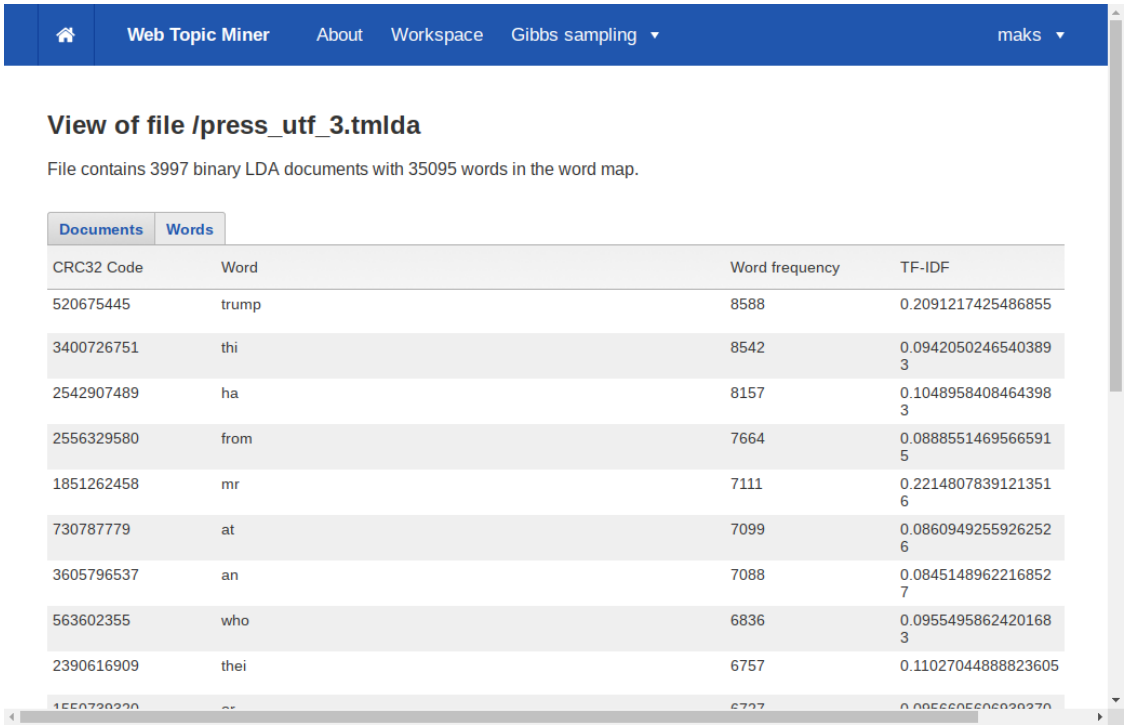
Веб-сайт WebTopicMiner позволяет просмотреть каждый документ в виде оригинального и лемматизированного текстов, а также его метаданные. Представление в виде хэш-кодов используется только внутри программы и недоступно для просмотра.

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.



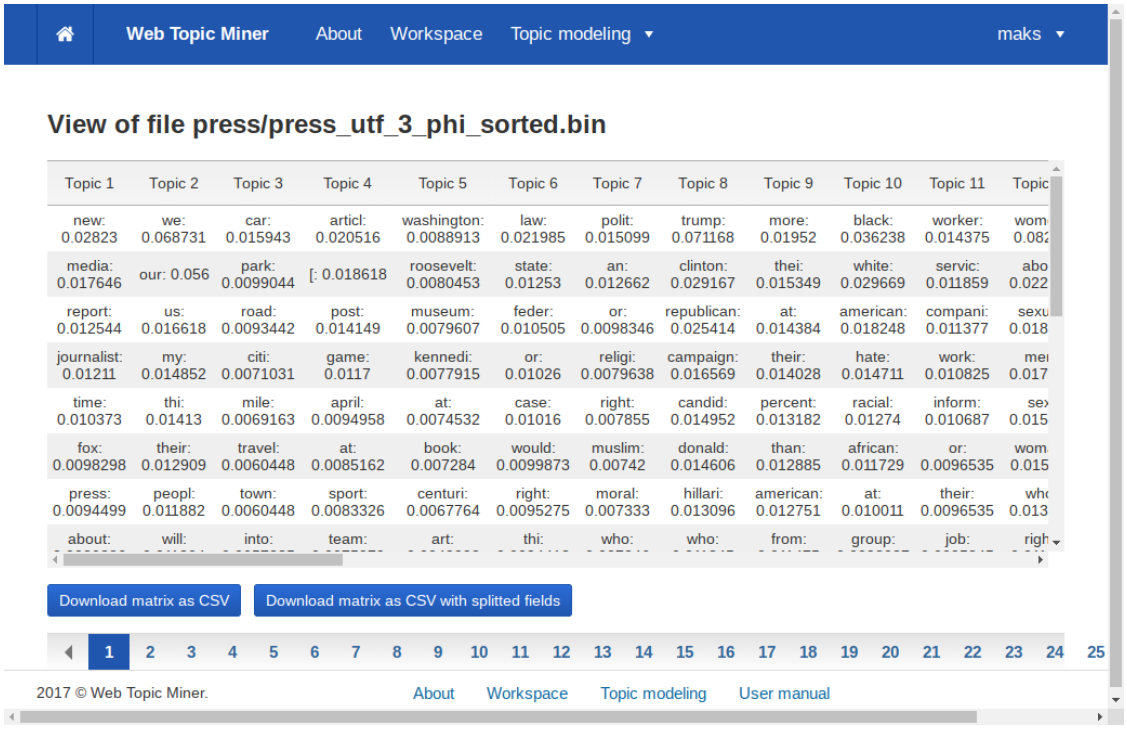
На второй вкладке находятся все слова, встречающиеся во всех документах в файле. TopicMiner для Windows создаёт список слов отсортированным по частоте слова.



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».



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 открывшимся диалоге выберите желаемый классификатор и формат файла из описанных выше. Класссификация начнётся автоматически, а на экран будет окно с сообщением, которое закроется по окончанию классификации. Если закрыть окно, то необходимо самостоятельно ждать появления файла с результатами.

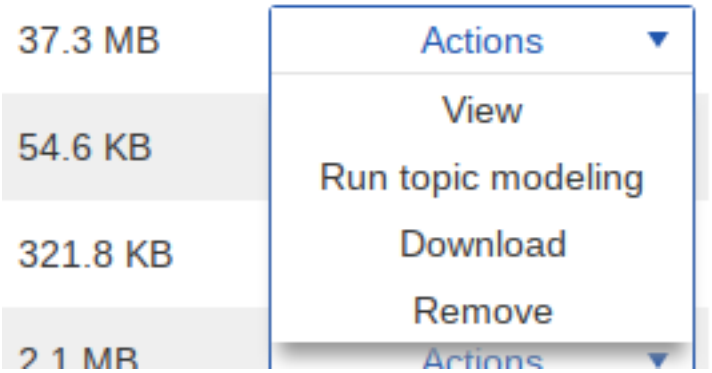
Результат классификации будет сохранен в файл с таким же именем, как исходный плюс краткое имя классификатора и суффикс «_classified» перед расширением. В этом файле после всех столбцов исходного файла будет добавлены ещё два: лемматизированный текст и оценка классификатора — 0 или 1. Для лемматизации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

Тематическое моделирование

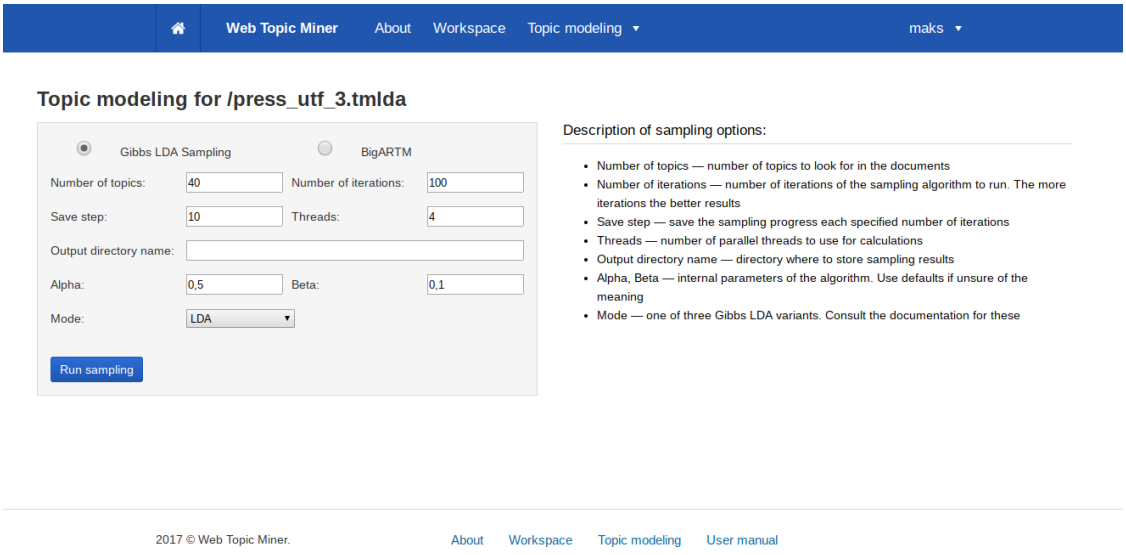
Запуск новой задачи тематического моделирования

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

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



Нажав на эту кнопку вы попадаете в интерфейс настроек моделирования.



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

Моделирование с помощью сэмплирования Гиббса

В этом алгоритме для настройки доступны следующие параметры:

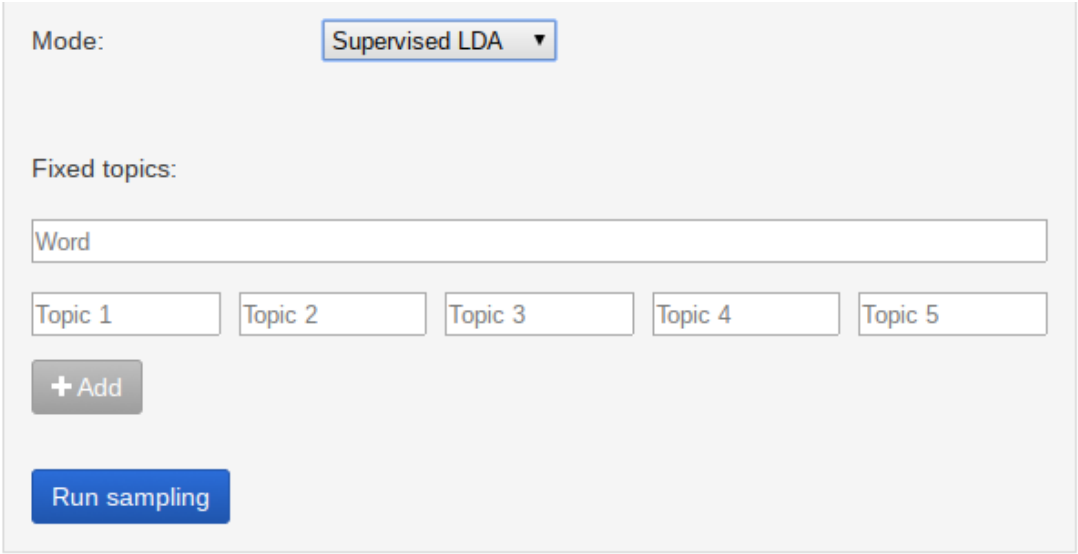
- «Alpha», «Beta» — коэффициенты, определяющие поведения алгоритма сэмплирования. Рекомендуется использовать значения по умолчанию
- «Number of topics» — число тем, для которых будет проводиться моделирование
- «Number of iterations» — число итераций алгоритма сэмплирования
- «Save step» — число итераций, 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:

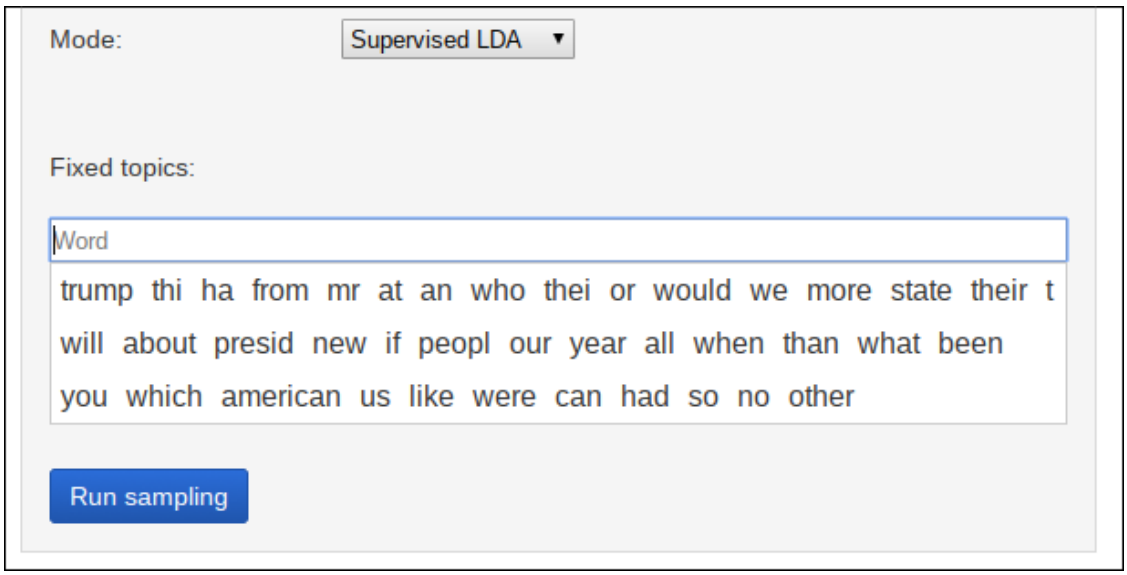
- «LDA» — стандартная модель
- «Supervised LDA» — модель, позволяющая фиксировать начальные темы для нескольких слов
- «Granulated LDA» — гранулированное сэмплирование, учитывающие несколько соседних слов

В режиме «ISLDA» доступна форма ввода фиксированной темы:



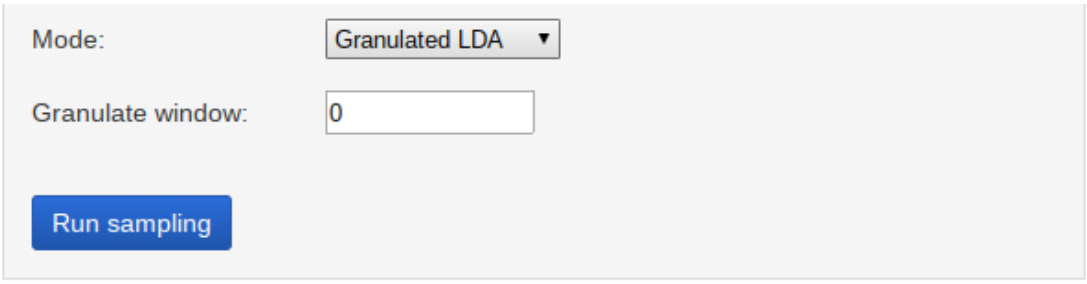
Каждому слову можно назначить до 5 тем в полях «Topic 1», …, «Topic 5», нумерация тем начинается с 1.

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

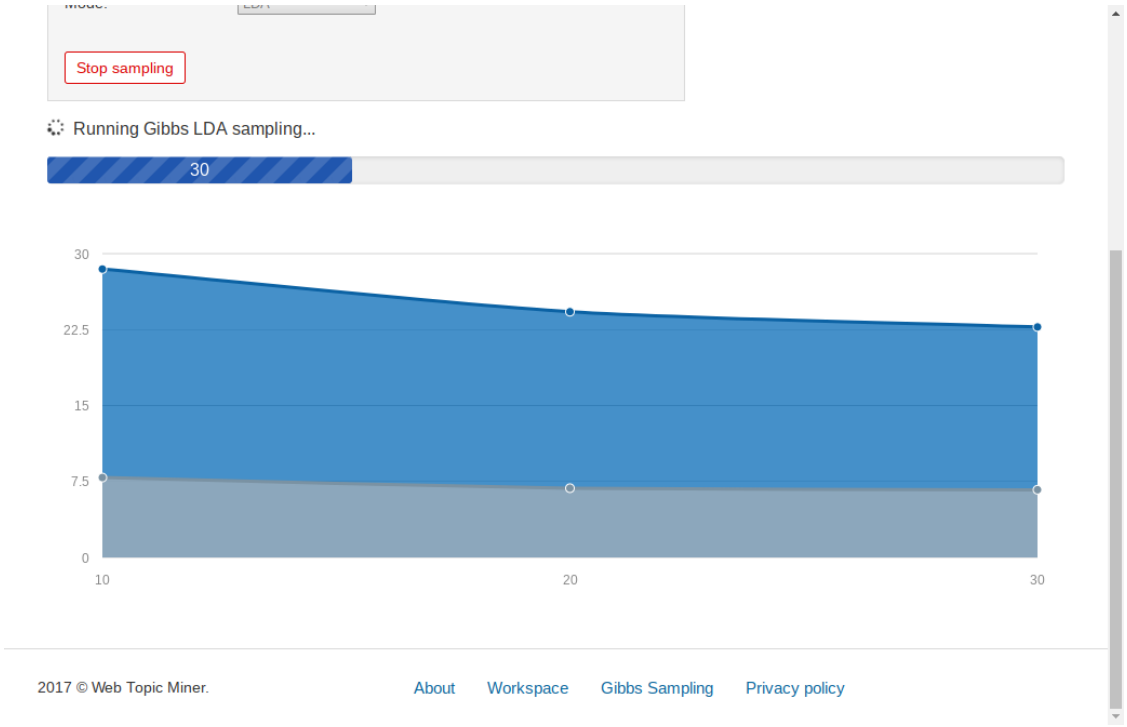


Для того, чтобы запомнить слово, нажмите кнопку «Add».

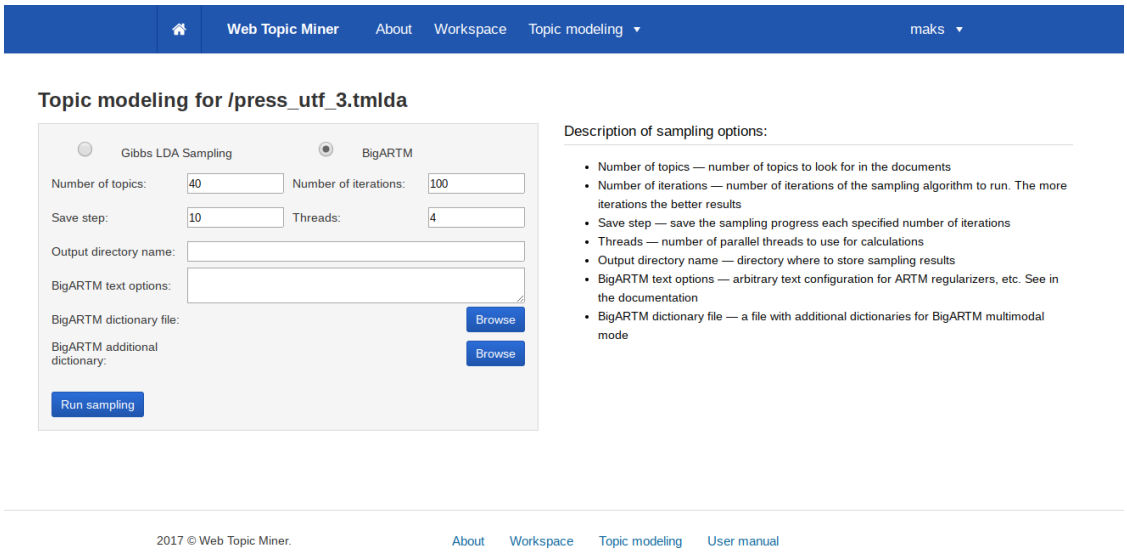
В режиме «GLDA» доступна поле выбора размера окна:



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

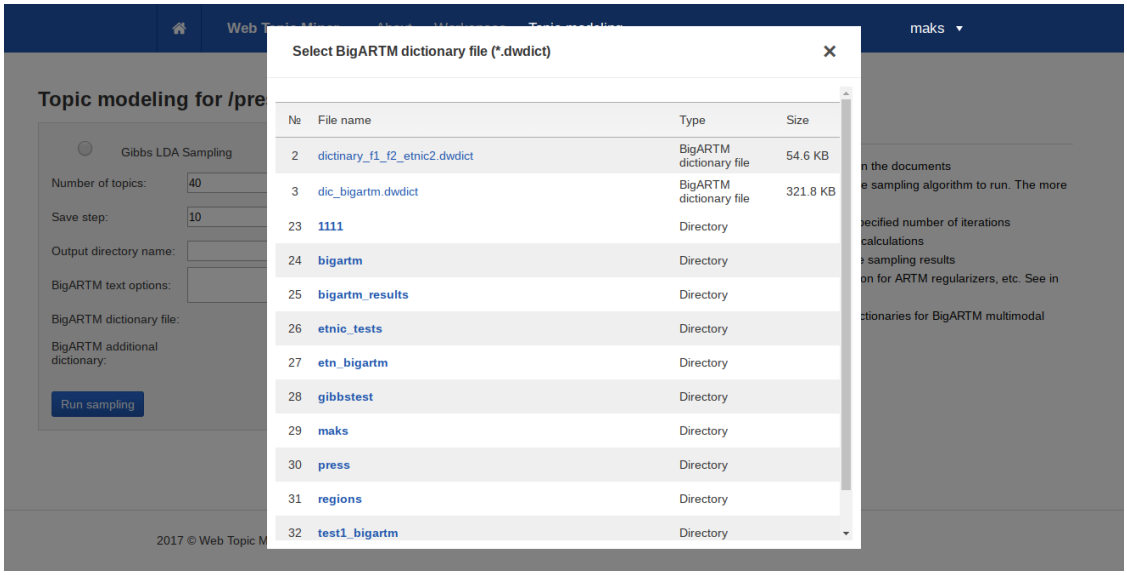


В режиме BigARTM опции «Number of topics», «Number of iterations», «Savestep», «Threads» и «Output directory name» имеют тот же смысл, что и для сэмплирования Гиббса. Кроме этого доступны следующие опции:

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

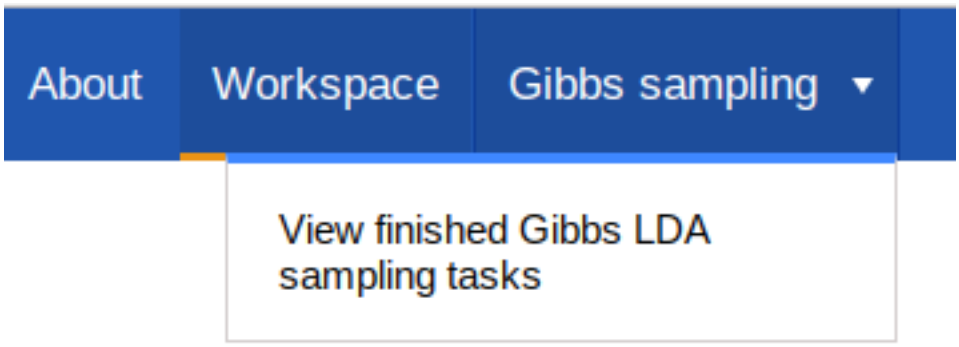
- «BigARTM additional dictionary» — текстовый файл в кодировке UTF-8, содержащий дополнительный словарь для BigARTM, например словарь этнонимов. Каждое слово должно находиться на отдельной строке.

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

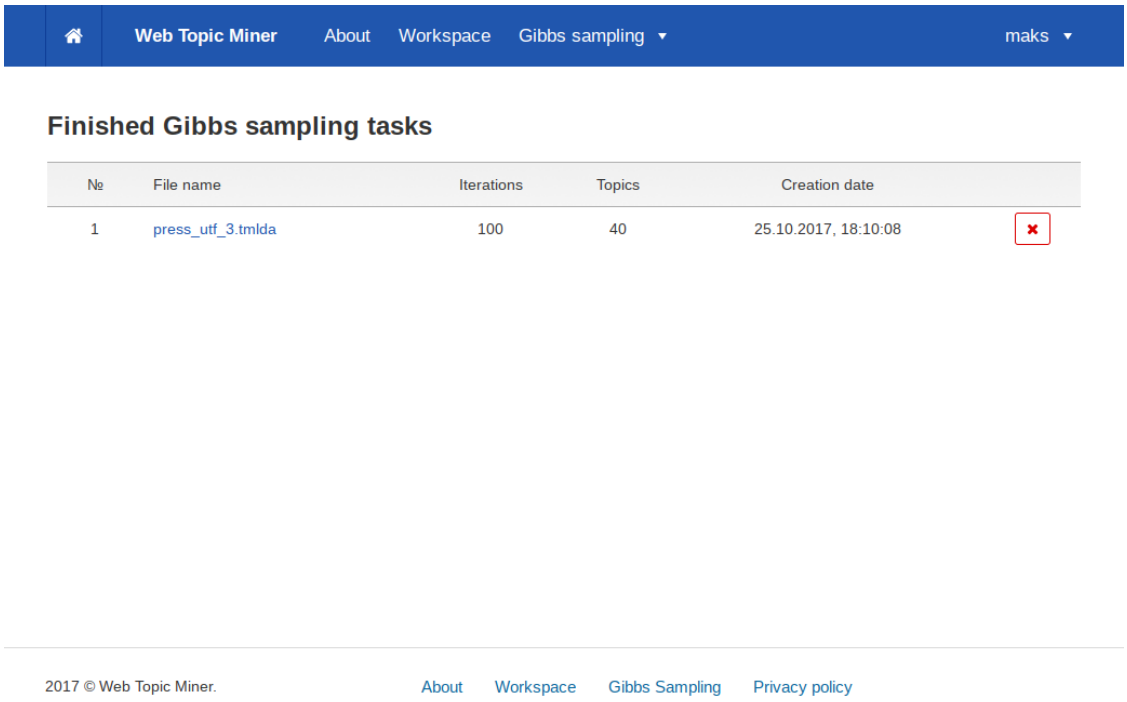


Просмотр выполняющихся и завершенных задач моделирования

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 завершённые задачи.



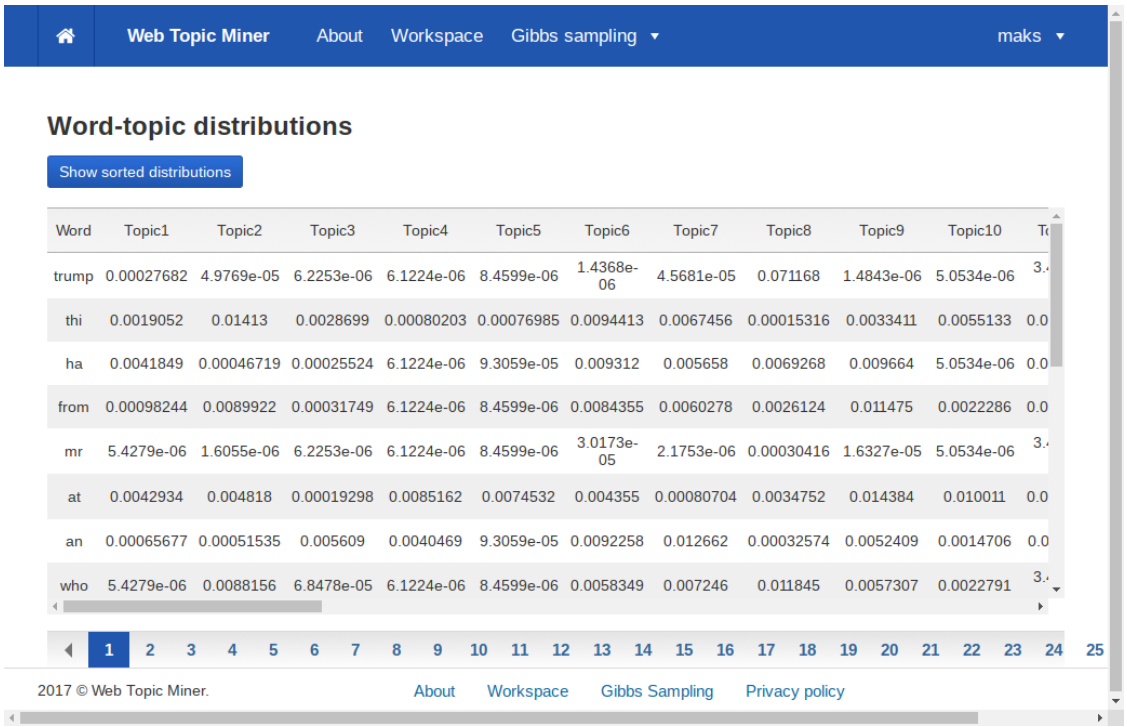
Здесь кроме имени файла указано число итераций, число тем и время создания задачи, для упрощения поиска нужной задачи среди большого числа похожих.

Нажатие на имя файла ведёт на страницу результатов тематического моделирования.

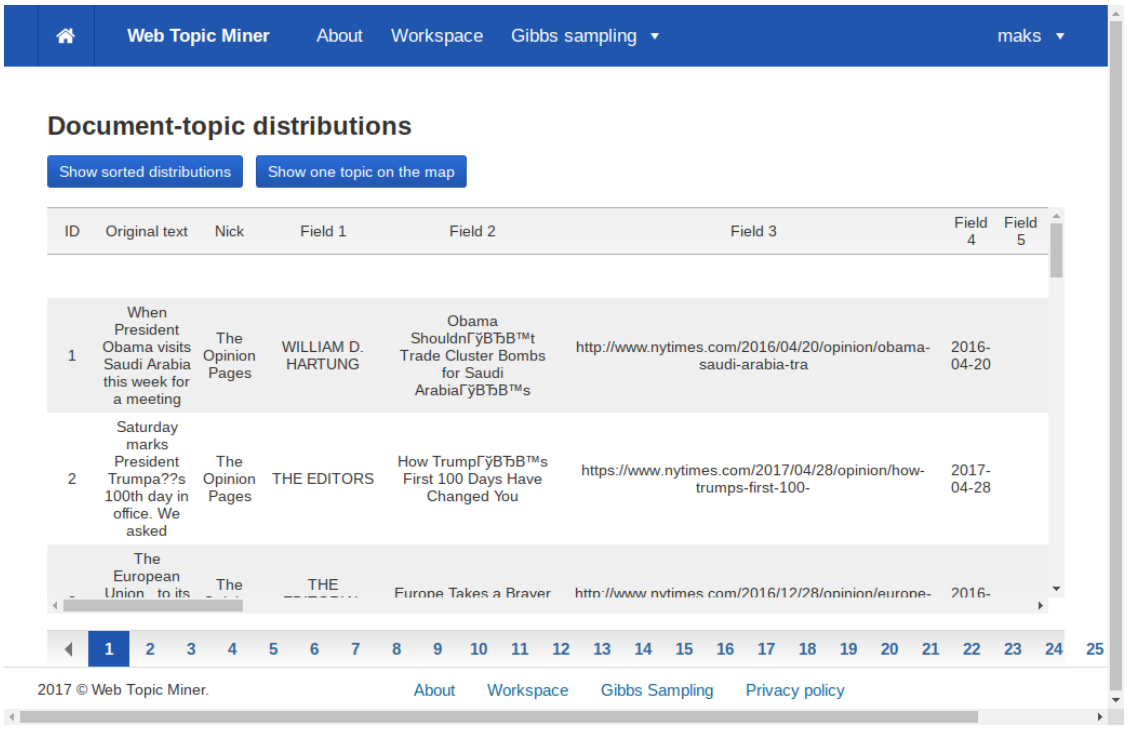
Просмотр результатов

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 менятся от итерации к итерации.

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

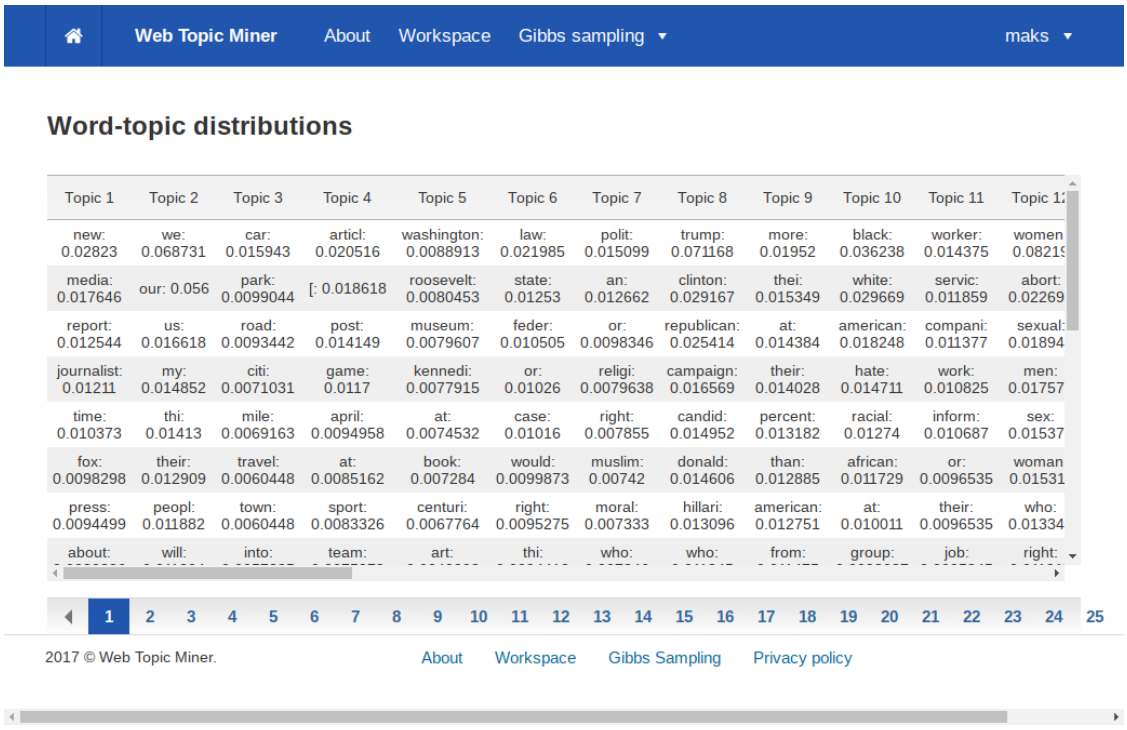


В матрице слова-темы для каждого слова указана вероятность принадлежности этого слова каждой теме.

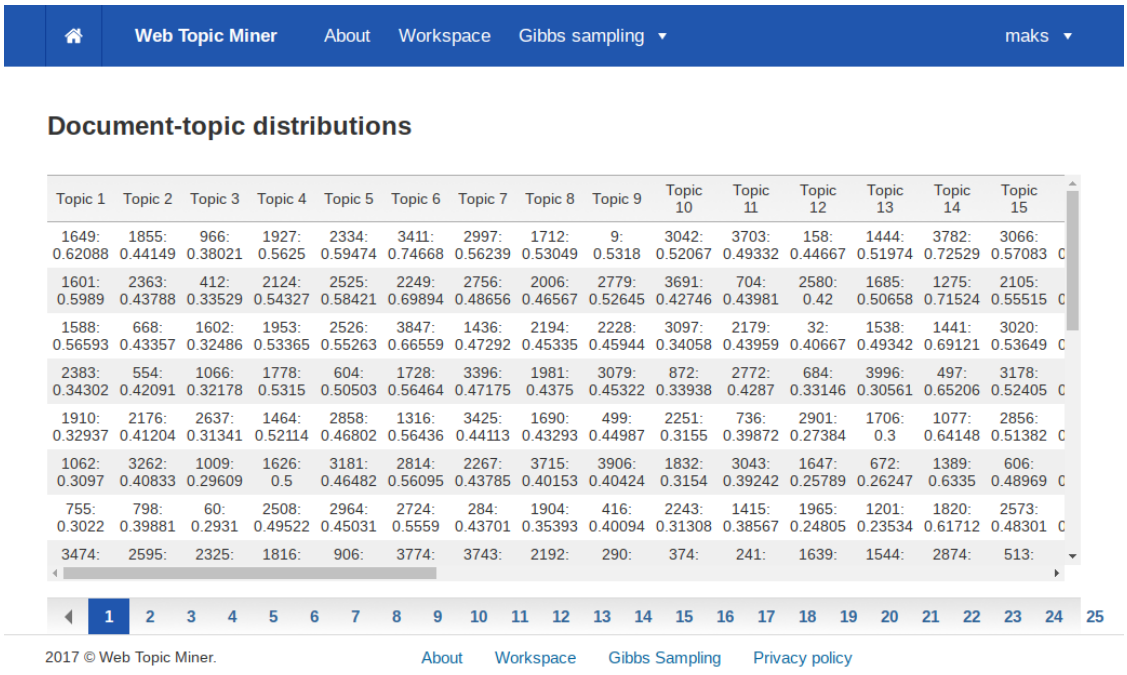


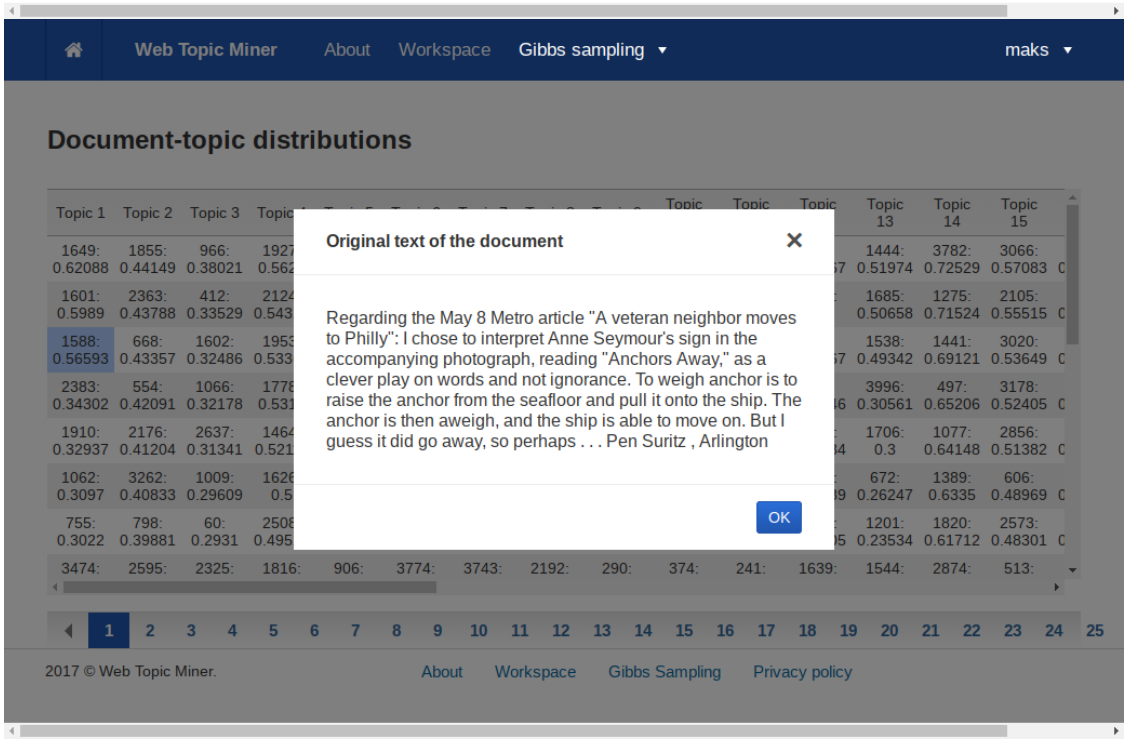
В матрице документы-темы для каждого документа также указана вероятность принадлежности темам. При нажатии на строку открывается всплывающее окно, содержащие полный текст выбранного документа. Дополнительно в этой матрице представлены метаданные каждого документа в файле.

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.



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.



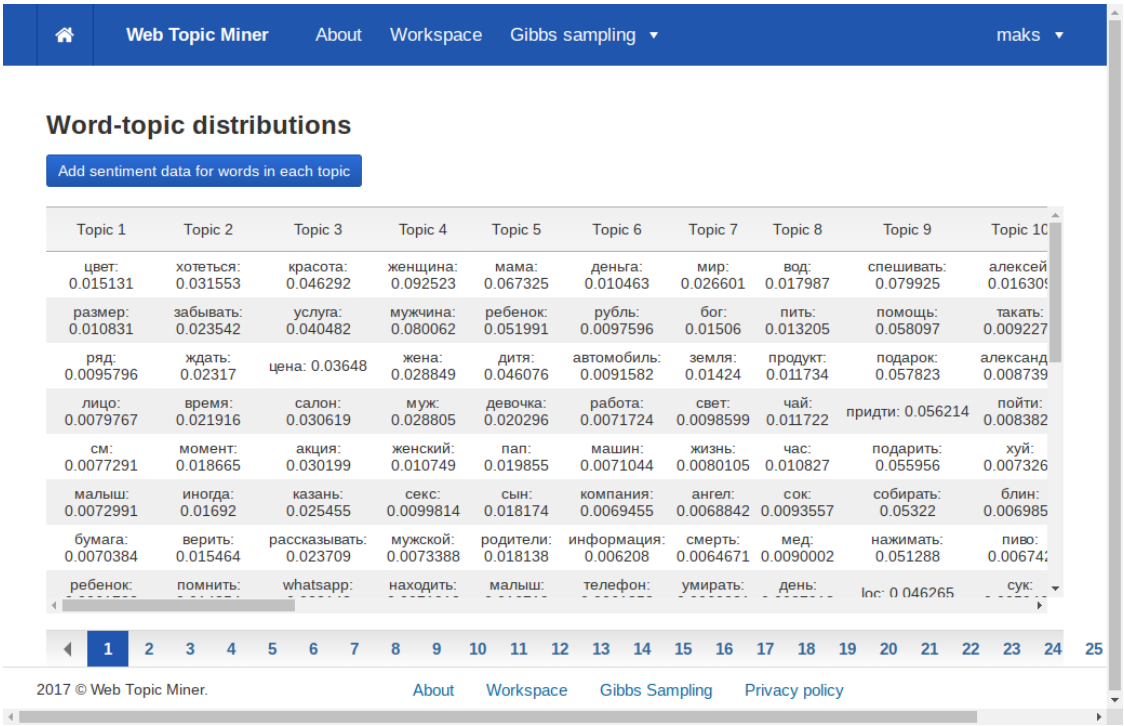


Для задач в режиме BigARTM также доступны дополнительные кнопки, показывающие распределения тем для всех дополнительных модальностей.

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

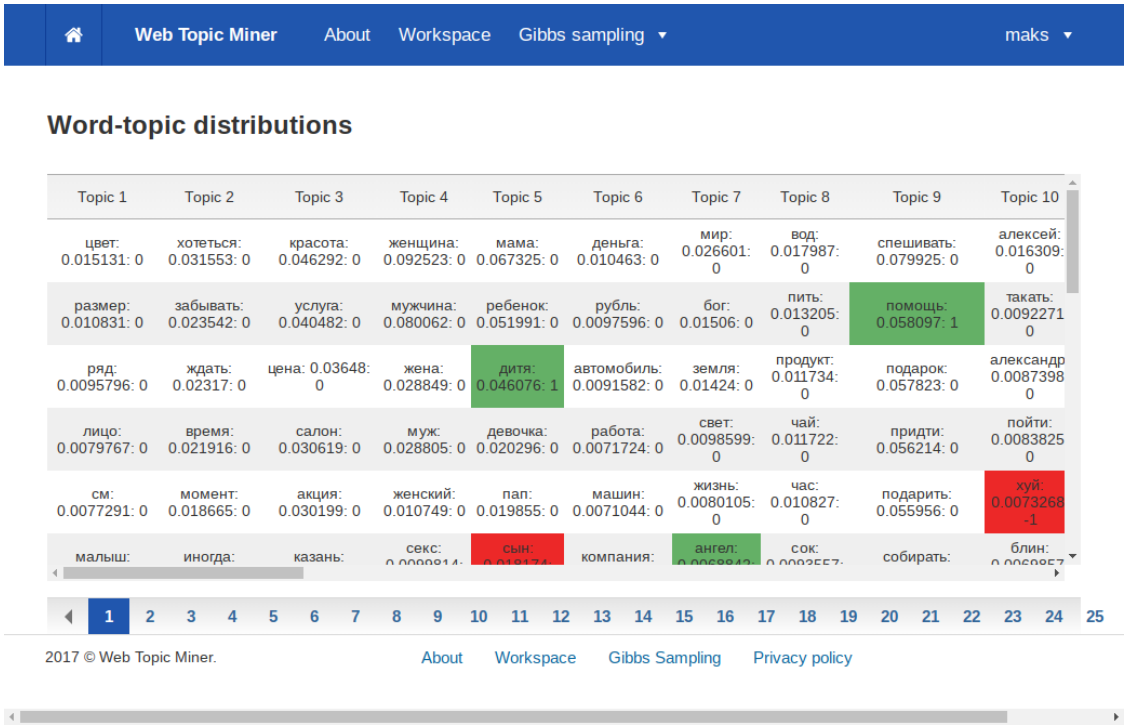
Сентимент-анализ тем

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.



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

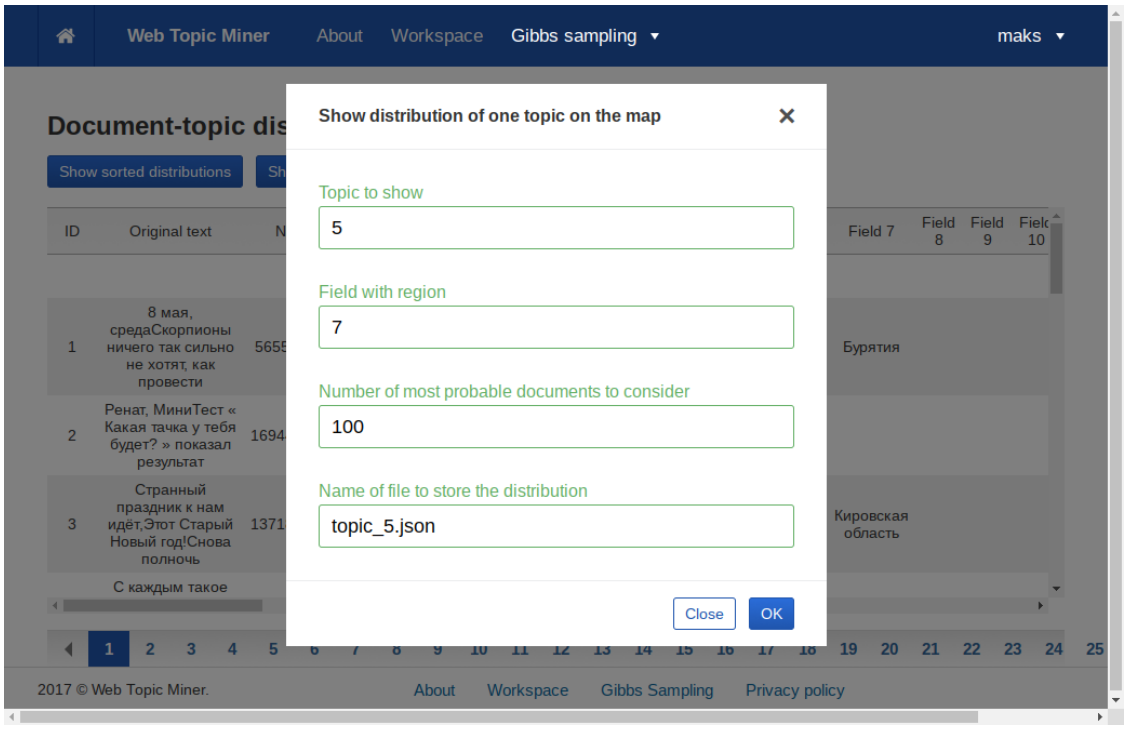
ячейку кроме слова и его вероятности добавится новое число — сентиментная оценка слова по шкале от -2 до 2. Для удобства анализа положительные слова окрашены в зелёный цвет, а отрицательные — в красный. Нейтральные слова цвета не имеют.



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

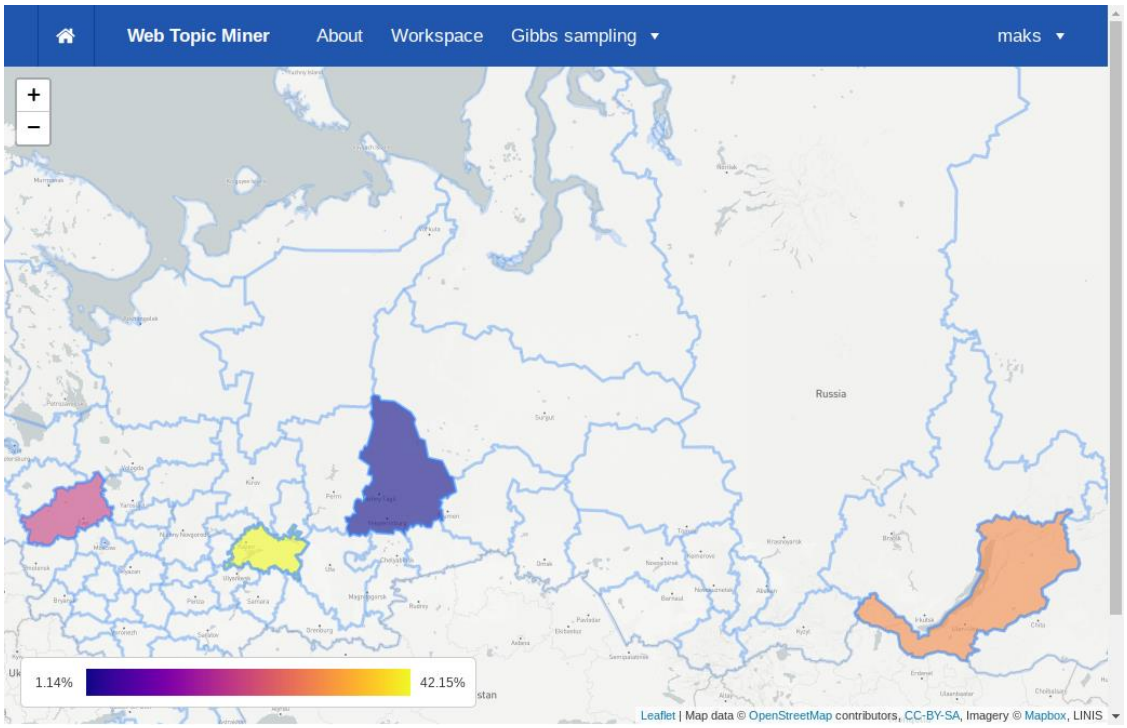
Отображение распределения темы на карте России

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.



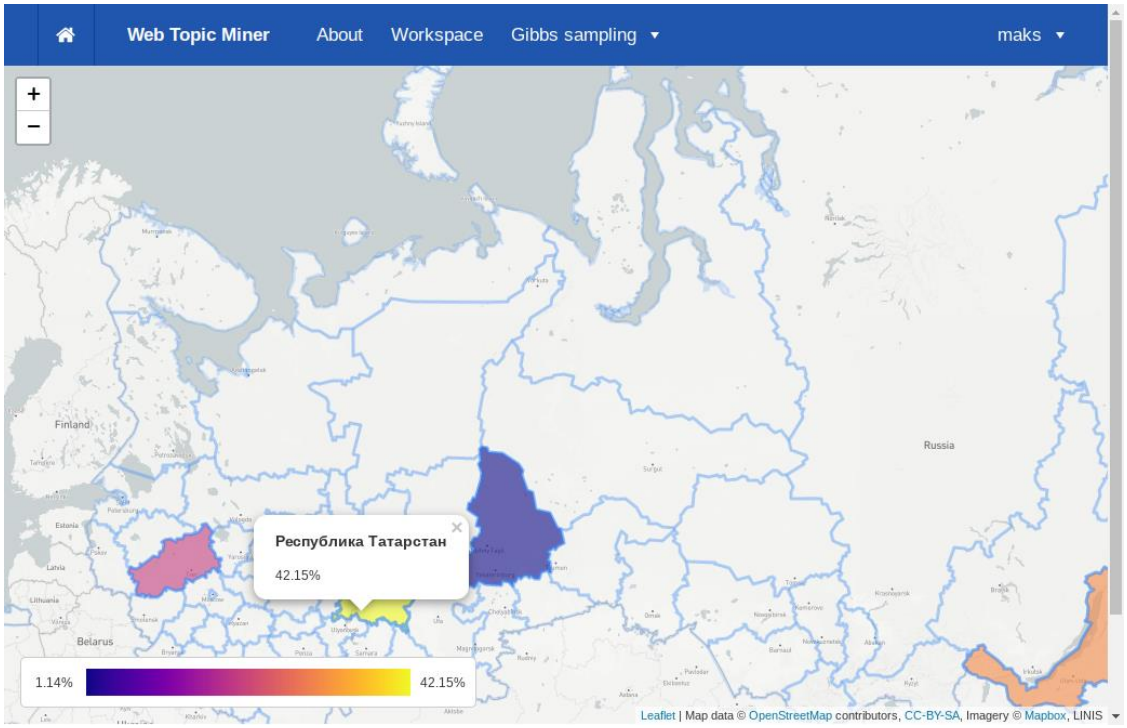
Выберите номер темы, которую требуется отобразить, номер поля метаданных, в котором содержится регион, и число самых вероятных документов темы, которые требуется учитывать. В последнем поле укажите имя json-файла, в которой будет сохранён результат. После нажатия на «Ok» вы будете перенаправлены в каталог, 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.

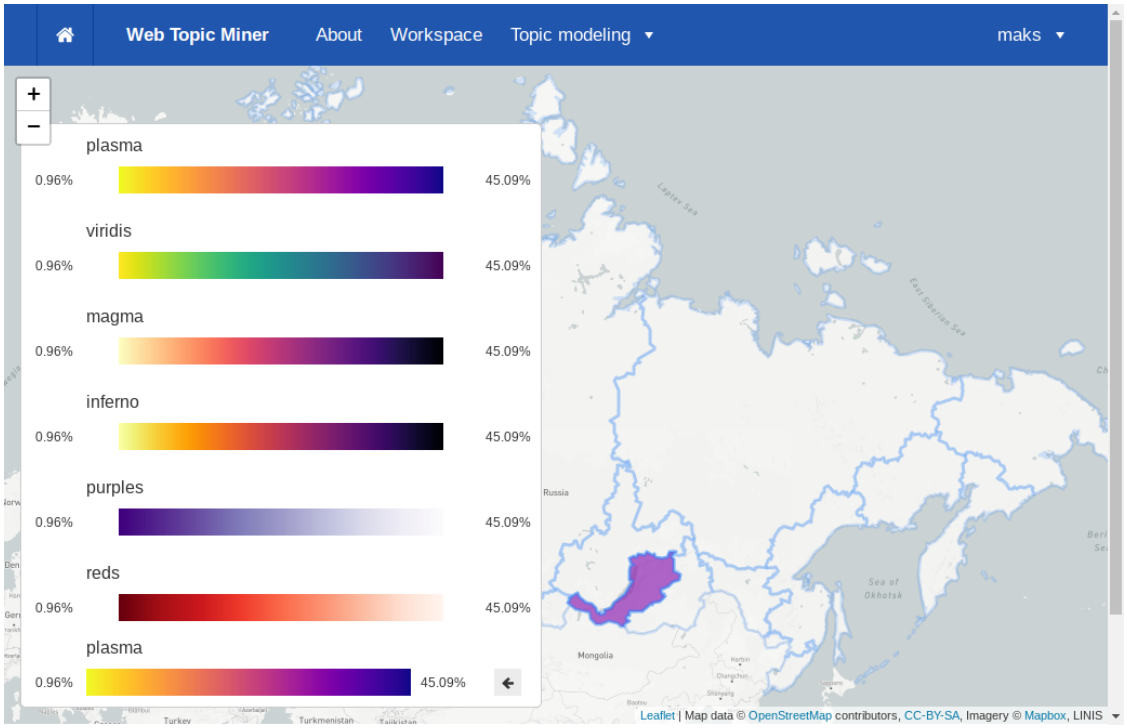


Регионы, в которых выбранной темы нет, отображены только контуром. Остальные регионы окрашены по цветовой шкале от синего до оранжевого в зависимости от доли темы в этом регионе.

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.



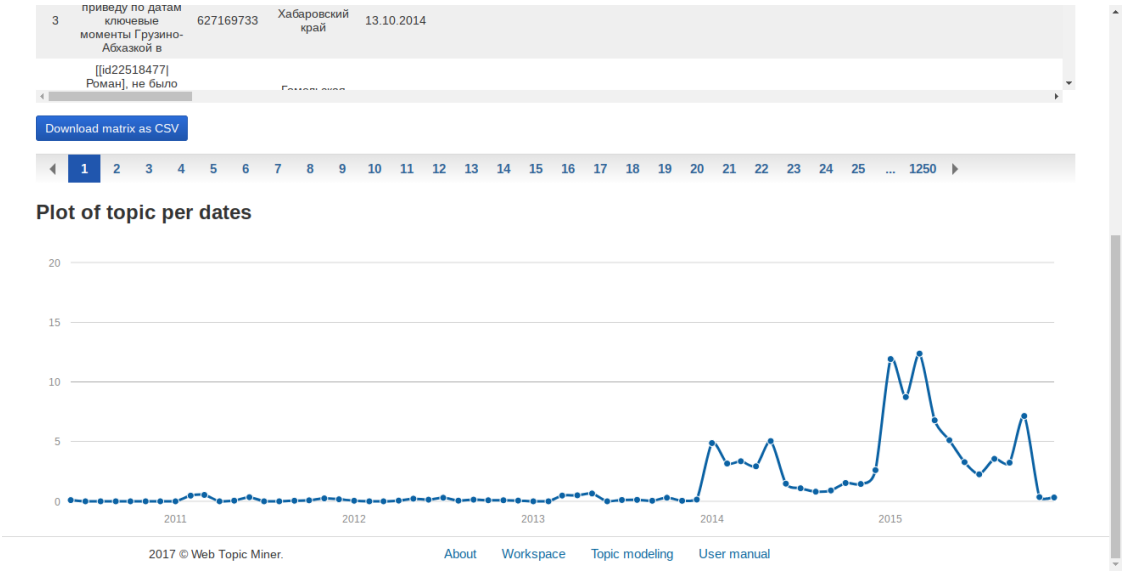
Кнопка со стрелкой позволяет инвертировать направление цветовой шкалы (от темного к светлому и наоборот).

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%.