TopicMiner with Sentiment analysis

Руководство пользователя

Версия 96 (64 bits).

Санкт - Петербург

2017

Оглавление

Глава 1. Препроцессинг документов. 4
1.1 Процедура сборки и лемматизации документов. 4
1.2. Второй этап препроцессинга. 8
1.2.1. Создание списка стоп-слов. 9
1.3. Третий этап препроцессинга. 10
Глава 2. Просмотр файлов формата tmlda. 10
Загрузка файла tmlda. 11
Выгрузка оригинальных документов в формате csv. 11
Выгрузка лемматизированных документов в формате csv 11
Выгрузка лемматизированных документов в формате TAB. 11
Загрузка списка слов для фильтрации документов. 11
Выгрузка документов в формате tmlda по списку слов. 13
Выгрузка документов в формате ‘tmlda’ с удаленными пустыми документами. 13
Расчет term-document matrix (формат TAB). 13
Расчет term-document matrix (формат CSV). 13
Формирование файлов ‘tmlda’ для мультимодальных моделей (BigARTM). 13
Глава 3. Тематическое моделирование по модели сэмплирования Гиббса. 15
3.1. Интерфейс опции 'Gibbs LDA sampling'. 15
3.2. Загрузка документов для тематического моделирования. 16
3.3. Тематическое моделирование на основе сэмплирования Гиббса. 18
3.4. Визуализация результатов тематического моделирования. 20
3.4.1. Визуализация распределений документов по темам. 20
3.4.2. Визуализация распределений слов по темам. 22
3.4.3. Визуализация распределений отсортированных документов в темах. 23
3.4.2. Визуализация отсортированных распределений слов по темам. 24
3.5. Сохранение результатов тематического моделирования в виде проектного файла. 26
3.6. Загрузка результатов тематического моделирования из проектного файла. 27
Глава 4. Тематическое моделирование по моделям BigArtm (мультимодальное тематическое моделирование). 27
4.1. Задание параметров в моделях мультимодальной ТМ. 27
4.2. Визуализация результатов тематического моделирования. 29
4.3. Сохранения результатов тематического моделирования в виде проектного файла. 29
4.4. Расчет мультимодального варианта ТМ. 29
Глава 5. Анализ стабильности результатов моделирования. 31
5.1. Загрузка тематических решений. 31
5.2. Сравнение тематических решений. 33
5.2.1. Матрица ‘Kullback - Leibler distance’ 34

5.2.2. Сопоставление тем из разных решений. 35
Глава 6. Визуализация результатов тематического моделирования на карте Российской федерации. 36
6.1.Расчет распределений документов по регионам. 36
6.2.Визуализация распределения документов в Quantum GIS. 38
Глава 7. Анализ тональности текстов. 43
7.1. Введение 43
7.2. Подготовка словаря для сентимент анализа. 43
7.2. Подключение словаря к тематической модели. 44
7.3. Тональный расчет распределения слов по темам. 44
7.3.1. Выгрузка матрицы слова - темы с тональными оценками. 46
7.3.2. Подсказка тем. 46
7.4. Тональный расчет распределения документов по темам. 46
7.4.1. Выгрузка матрицы документы - темы с тональными оценками. 47
7.5. Тональный расчет распределения документов по темам для BigArtm. 48
Глава 8. Временные тренды в тематических моделях. 48
8.1. Унификация временных дат. 48
8.2. Построение временных трендов в моделях на основе мультимодального тематического моделирования 50
Заключение. 54

Введение.

Программа TopicMiner разработана в лаборатории Интернет исследований (http://linis.hse.ru/) с использованием внешних разработок, в том числе библиотеки алгоритмов BigARTM, которая входит в программу в виде DLL. Программа предназначена для тематического моделирования русскоязычных и англоязычных документов. Программа включает в себя: 1. Опция препроцессинга документов. 2. Опция тематического моделирования и визуализации результатов расчета. 3. Опция анализа стабильности результатов тематического моделирования. При публикации научных результатов, основанных на работе данной программы, необходимо ссылаться на лабораторию интернет-исследований, Высшая Школа Экономики.

Тематическое моделирование (topic modeling) – одно из современных приложений машинного обучения к анализу текстов, активно развивающееся с конца 1990-х годов. Тематическая модель (topic model) коллекции текстовых документов определяет, к каким темам относится каждый документ и какие слова (термины) образуют каждую тему. Каждый текст и слово принадлежат множеству тем – точнее, всем темам с разной вероятностью. Входными данными тематической модели является матрица (таблица) слов на документы, где элементы (ячейки) – частоты слов в документах. Выходными данными являются две матрицы меньшей размерности (меньшего размера): слова на темы и документы на темы, где элементы – вероятности принадлежности слов или документов к темам. Количество искомых тем устанавливается пользователем исходя из опыта.
В задачах машинного обучения для сокращения размерности матрицы обычно используется либо отбор признаков, приводящий к уменьшению числа параметров, либо регуляризация с помощью наложения дополнительных ограничений на параметры. В частности, байесовская регуляризация основана на введении априорного распределения вероятности в пространстве параметров. В данной программе используются два основных подхода к расчету распределений слов и документов по темам.

В данной версии реализованы следующие тематические модели:
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).

Кроме того, в данной версии ПО реализована процедура сентимент анализа на основе словарного подхода. В качестве русскоязычного словаря предлагается словарь полученный в результате выполнения проекта 'Разработка общедоступной базы данных и краудсорсингового веб-ресурса для создания инструментов сентимент-анализа', № 14-04-12031.

Глава 1. Препроцессинг документов.

Препроцессинг документов - существенная часть работы с документами. Препроцессирг состоит из трех этапов: 1. Процедура сборки комплекта документов в один файл и лемматизация. 2. Процедура расчета частот слов, выделение слов из скобок, и создание списка стоп-слов. 3. Удаление стоп-слов из лематизированных текстов.

1.1 Процедура сборки и лемматизации документов.

Входными данными для ПО TopicMiner является каталог с документами, в котором, каждый файл содержит один документ в формате txt. Кроме того, в данном каталоге

может лежать файл с метаданными, описывающий каждый файл. Пример такого файла приведен ниже. В каждой колонке находится отдельный атрибут метаданных.

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

В данном файле каждая строка содержит набор метаданных. Максимальное количество метаданных не может превышать 20 (20 колонок). В первой колонке находятся имена файлов, содержащие текст. Рекомендуется пронумеровать файлы и использовать их номера в качестве имён.

Первый этап пропроцессинга.

Общий вид окна препроцессинга приведен на рисунке 1.1.

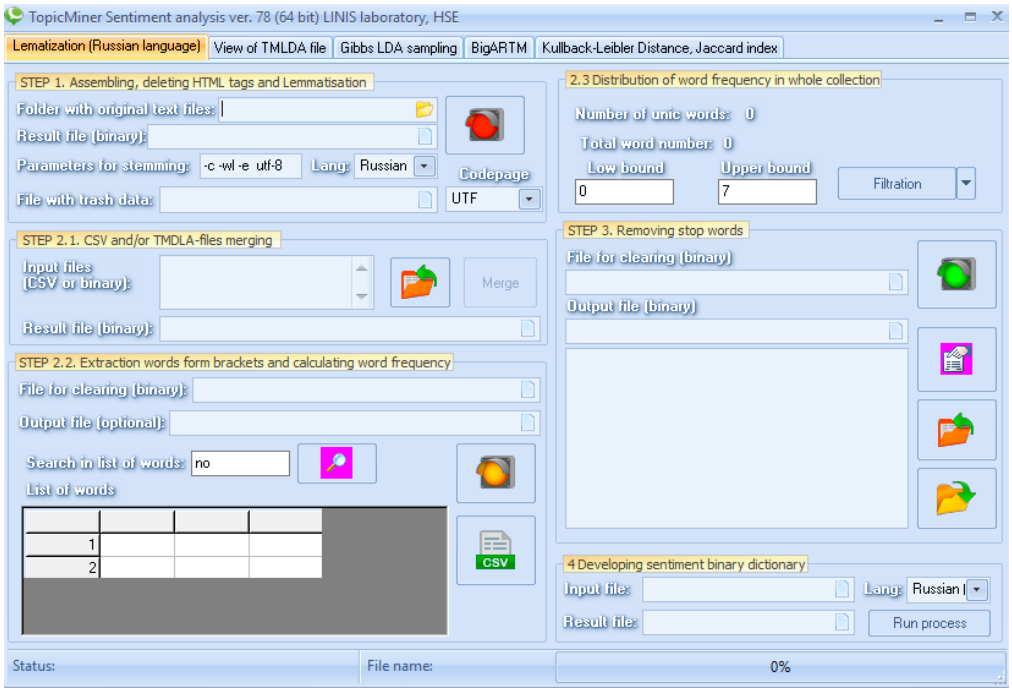
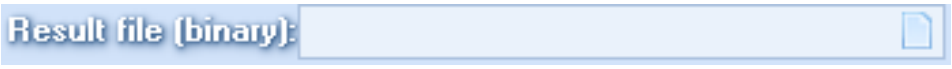


Рис. 1.1. Общий вид окна модуля русскоязычного препроцессинга.

Параметры первого этапа препроцессинга: 1. **Путь к каталогу с исходными данными**. Данный путь нужно указать в опции:

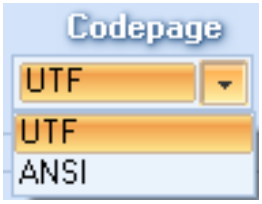
2. **Имя файла**, в котором будут находиться все оригинальные и лемматизированные тексты. Имя файла можно указать в следующей опции:



Достаточно указать лишь имя файла. Программа автоматически добавить расширение ‘tmlda’ (topic modeling LDA).

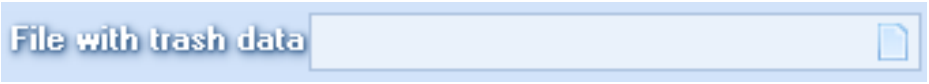
3. **Процедура лемматизации** основана на использовании лемматизатора ‘mystem.exe’ (разработка компании ‘Yandex’, https://tech.yandex.ru/mystem/), которая по условиям лицензии не может использоваться в коммерческих целях. Для работы программы ‘mystem.exe’ необходимо указать набор параметров. В программе TopicMiner эти параметры задаются автоматически, исходя из того, какой вариант кодировки выбран пользователем. Перечень параметров задан в строке ‘Parameters for stemming’.

Выбор типа кодировки для русскоязычных текстов. В данной программе реализованы два типы кодировки для исходных файлов.



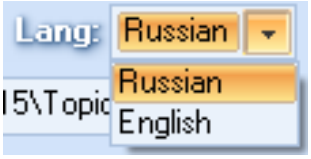
Пользователь может выбрать кодировку ‘UTF’ или ‘ANSI’.

4. **Файл со списком стоп-символов**. В оригинальных документах могут присутствовать символы и группы символом (например, html разметка, знаки препинания), которые мешают анализу и должны быть удалены из текстов. Для проведения первого этапа препроцессинга необходимо указать имя файла, в котором хранятся такие символы, и путь к нему. Это можно указать в следующей опции.



5. Выбор языка. В данной версии поддерживаются два языка, русский и английский.

Выбор осуществляется при помощи выпадающего списка:



Процедура лематизации осуществляется при помощи программ mystem и porter.

Заполненная таблица параметров для первого этапа препроцесинга может выглядеть следующим образом (пример):

После того как все параметры заполнены, для того что бы запустить процесс сборки и лемматизации нужно нажать на кнопку  . Процент исполнения первого этапа – см. рис. 1.2.

Внимание. Несмотря на то, что процесс лемматизации распараллелен, время исполнения первого этапа существенно зависит от числа исходных файлов и общего размера файлов (lkbys ntrcnjd). Например, для 9 миллионов коротких постов из социальной сети время лемматизации приблизительно 13 суток.

Результат препроцессинга после первого этапа.

Результатом работы опции препроцессинга после первого этапа является файл с расширением tmlda, в котором последовательно содержатся пары текстов в оригинальном и лематизированном виде. Пример содержимого такого файла приведен на рисунке 1.3. Программа ‘mystem.exe’ преобразует каждое слово в документах в начальную форму и помещает каждое слов в скобки.

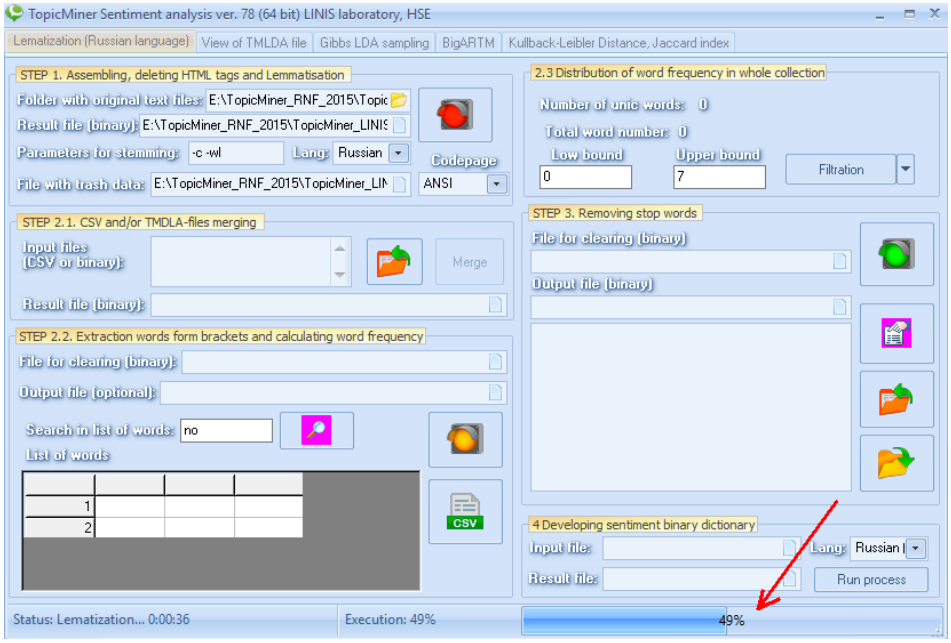


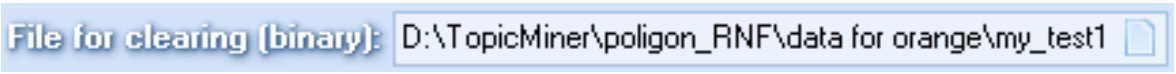
Рис. 1.2. Пример процесса лемматизации.

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=FB4EF53787D563FA8F1D6D2C3F205F0C.d01t01{новость} {наука} {о} {зависимость}: {Чантикс??} {средство} {против} {курение}, {показывать} {себе} {также} {положительно|положительный} {в} {отношение} {контроль} {над|нада} {прием} {алкоголь}. {тот}, {кто} {принимать} {прапарат??} {с} {цель} {бросать} {курить} {часто|частый} {сообщать}, {что} {у} {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??} {еще} {один} {механизм}, {который} {делать} {отказ} {от} {курение} {довольно|довольный} {трудный}. {в} {принцип}, {девать|дело} {вполне} {ожидать}. {отказ} {от} {курение} {приводить} {к} {падение} {уровень} {дофамин} {в} {система} {вознаграждение}, {что} {приводить} {к} {депрессия} {и} {к} {желание} {снова} {закуривать}. {подтверждение} {давать|данный} {механизм} {делать} {применение} {дофаминергических??} {препарат} {еще} {более|много}

Рис. 1.3. Пример результат препроцессинга после первого этапа.

1.2. Второй этап препроцессинга.

На втором этапе препроцессинга производится выделение слов из скобок (смотри рис. 1.3) и подсчет частот слов по всем документам. Входными данными для второго этапа является файл, полученный после первого этапа. Необходимо задать имя и путь к данному файлу в опции ‘File for clearing (binary)’ (например):



Кроме этого, следует задать имя файла, в котором будут храниться результаты второго этапа препроцессинга. Это нужно сделать в следующей опции ‘Output file’ (например):

Результатом второго этапа препроцессинга является создание частотного словаря уникальных слов и преобразование лемматизированных документов в цифровой формат. В данном цифровом формате слова в документах заменены на цифровые коды (IDs) слов из списка уникальных слов. Для того, чтобы запустить второй этап препроцессинга нужно нажать на кнопку  . В результате работы, новые данные (частотный словарь уникальных слов и цифровые документы) будут добавлены в файл с расширением tmlda. Пример работы приведен на рисунке 1.4.

Внимание. В данной версии реализован расчет TF-IDF, однако данная опция еще полностью не оттестированна.

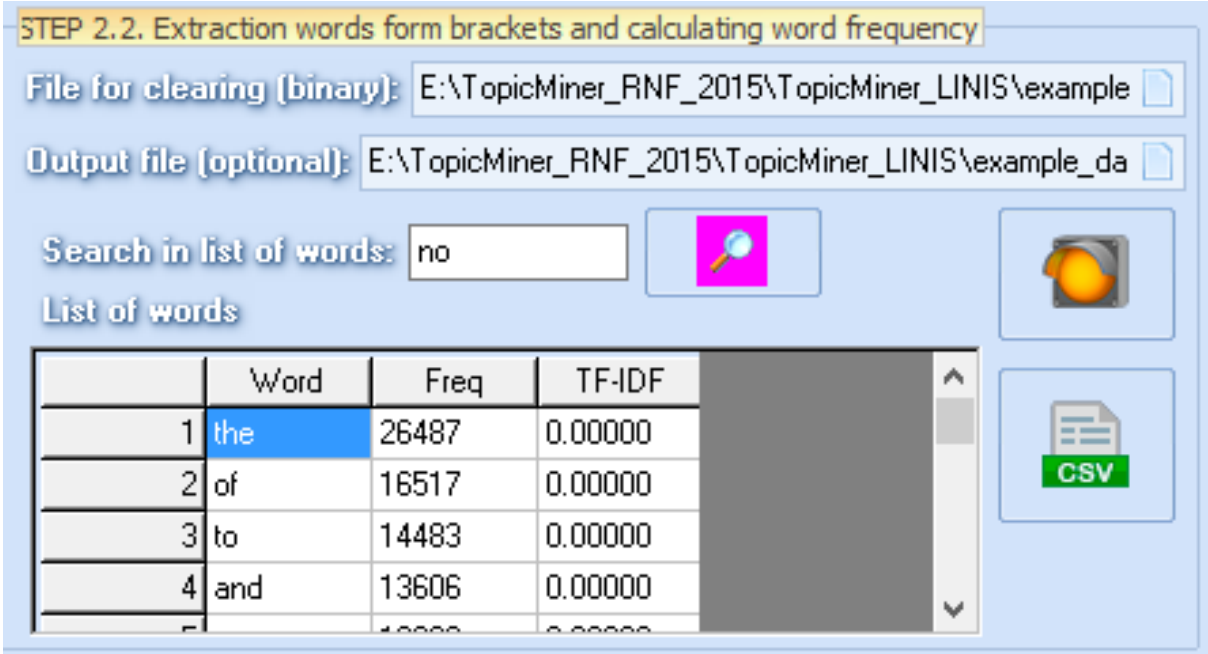


Рис. 1.4. Пример результат препроцессинга после второго этапа.

Частотный словарь можно выгрузить в формате csv во внешний файл. Для этого нужно нажать на кнопку  и указать имя файла. Если нужно найти слово в списке уникальных слов, нужно указать его в окне ‘Search in list of words’ и нажать на кнопку  . Пример результата представлен на Рис 1.5.

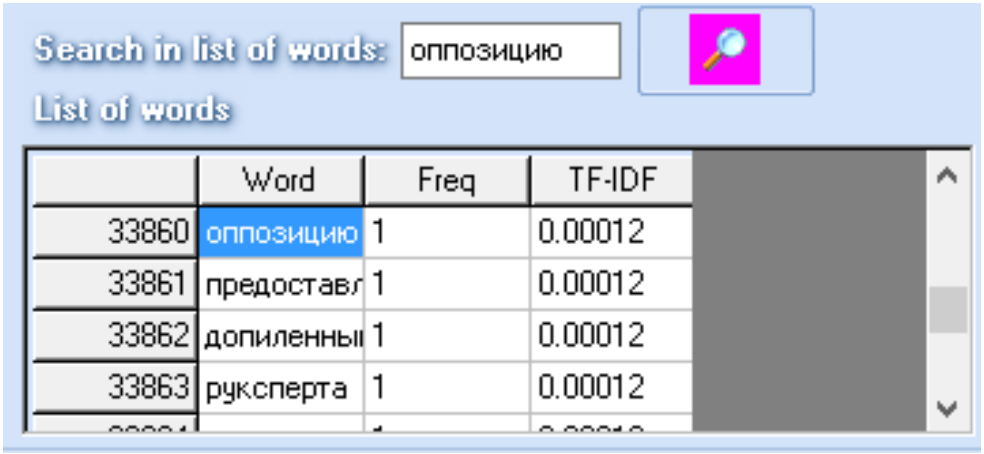


Рис. 1.5. Пример результата препроцессинга после второго этапа.

1.2.1. Создание списка стоп-слов.

На втором этапе препроцессинга можно сформировать список стоп-слов на основе списка частот уникальных слов. Для этого нужно указать верхнюю и нижнюю границы по частотам из списка уникальных слов в опции ‘Distribution of word frequency in whole collection’:

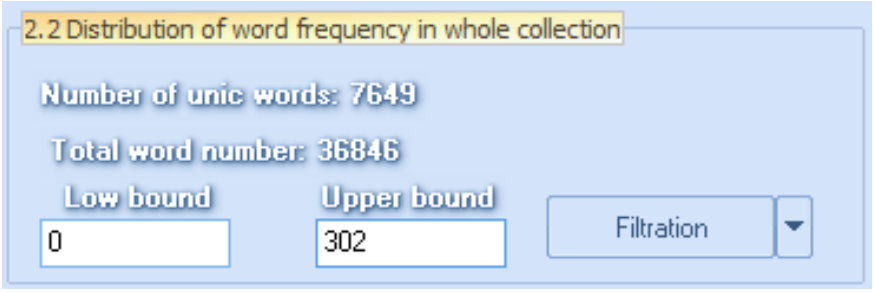


Рис. 1.6. Опция для создания списка стоп-слов.

После нажатия кнопки ‘Filtration’ откроется окно, в котором нужно указать имя файла, где будет храниться список стоп-слов. Туда будут сохранены слова, чьи частоты находятся за указанными пределами. В этом примере пределами являются числа ‘0’ и ‘302’.

Результатом препроцессинга после второго этапа является файл, который содержит оригинальные, лемматизированные и оцифрованные тексты.

1.3. Третий этап препроцессинга.

Здесь происходит удаление стоп-слов из оцифрованных документов. Входными данными является файл, который получился на выходе из второго этапа; его нужно указать. Затем нужно указать имя выходного файла, в котором будут находится оригинальные, лематизированные и оцифрованные тексты с удаленными стоп-словами. Кроме того, в данной опции нужно загрузить список стоп-слов из текстового файла. Это может быть файл, созданный на втором этапе, или внешний файл с любым другим списком слов, или файл, содержащий и то, и другое.

В данной опции присутствуют следующие кнопки:

1. Кнопка  : Очистка поля для списка стоп-слов.
2. Кнопка  : Загрузка стоп-слов из текстового файла.
3. Кнопка  : Сохранение списка стоп-слов в текстовый файл.

Третья кнопка нужна, если пользователь вводит стоп-слова в поле ТопикМайнера вручную. Процент выполненной работы по удалению стоп-слов показывается в том же месте, что и процент исполнения в на первом этапе препроцессинга.

Внимание. Необходимо пройти все три этапа процедуры препроцессинга.

Глава 2. Просмотр файлов формата tmlda.

В программе TopicMiner реализована возможность просмотра файлов формата tmlda, а также опция выгрузки текстов (оригинальных и лемматизированных) в файл формата csv. Опция просмотра полезна, так как позволяет посмотреть, какие стоп-слова еще не удалены из документов. Здесь же искать документы по списку ключевых слов и удалять пустые документы. Это позволяет существенно уменьшить размер коллекции и, соответственно, увеличить скорость тематического моделирования. Общий вид опции ‘View of tmlda files’ приведен на рисунке 2.1.

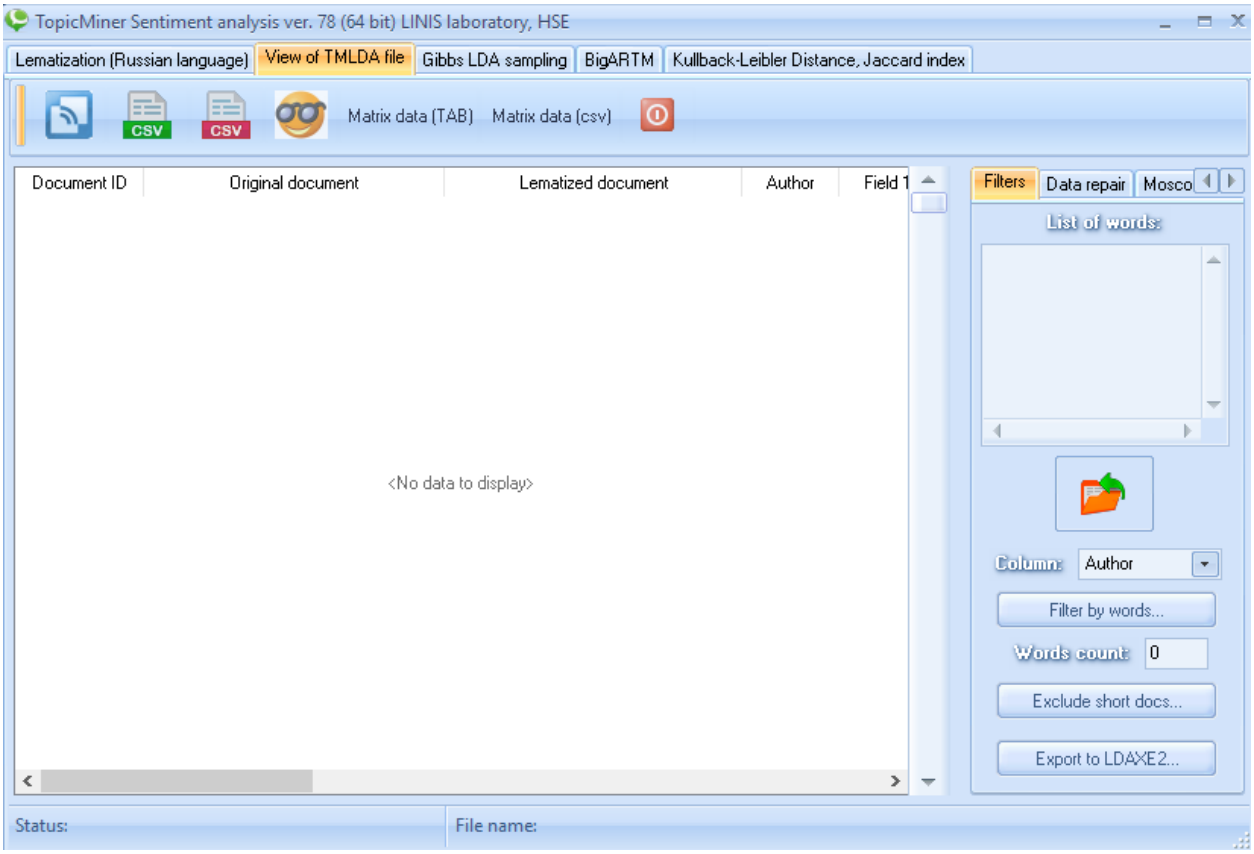


Рис. 2.1. Опция для просмотра tmlda файлов.

Загрузка файла tmlda. Для того, что бы загрузить файл в формате tmlda нужно нажать на кнопку  . В появившемся окне нужно указать имя файла. В результате указанный файл будет отображен в таблице (пример приведен на рисунке 2.2). В ней есть следующие столбцы: 1. Столбец с оригинальными документами. 2. Столбец с лемматизированными документами. 3. Набор столбцов с метаданными. Формат метаданных описан в главе 1.

Выгрузка оригинальных документов в формате csv. Формат csv поддерживается множеством внешних программ, в частности, Excel (если данные не очень велики). Для выгрузки в формате csv нужно нажать на кнопку  и указать имя файла.

Выгрузка лемматизированных документов в формате csv. Следует нажать на кнопку  и указать имя файла.

Выгрузка лемматизированных документов в формате TAB. Формат TAB поддерживается рядом внешних программных продуктов, в частности, статистическим пакетом Orange. Для выгрузки в формате TAB нужно нажать на кнопку  и указать имя файла.

Загрузка списка слов для фильтрации документов. Для загрузки списка слов нужно нажать на кнопку  и указать имя файла.

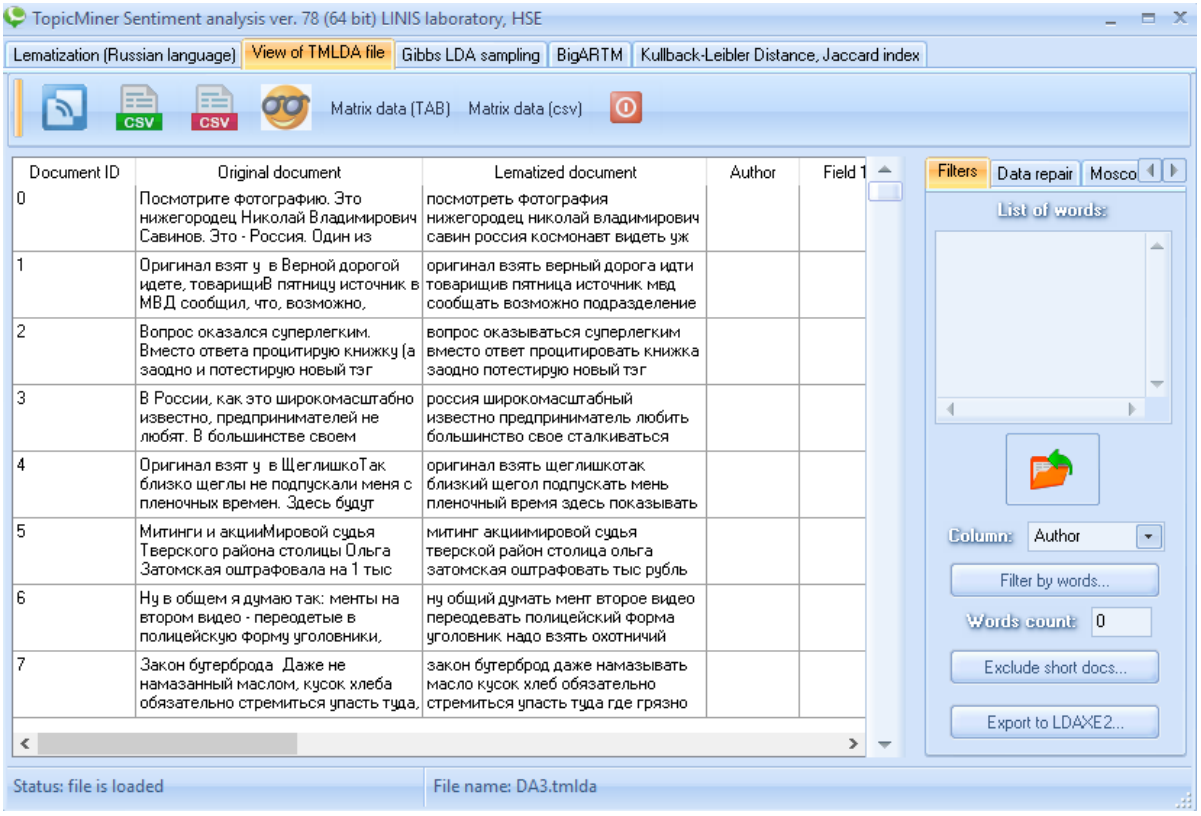


Рис. 2.2. Пример загруженного файла.

Внимание: слова в текстовом файле должны быть представлены по одному в строке. Пример списка загруженных слов приведен на рисунке 2.3 в правой части.

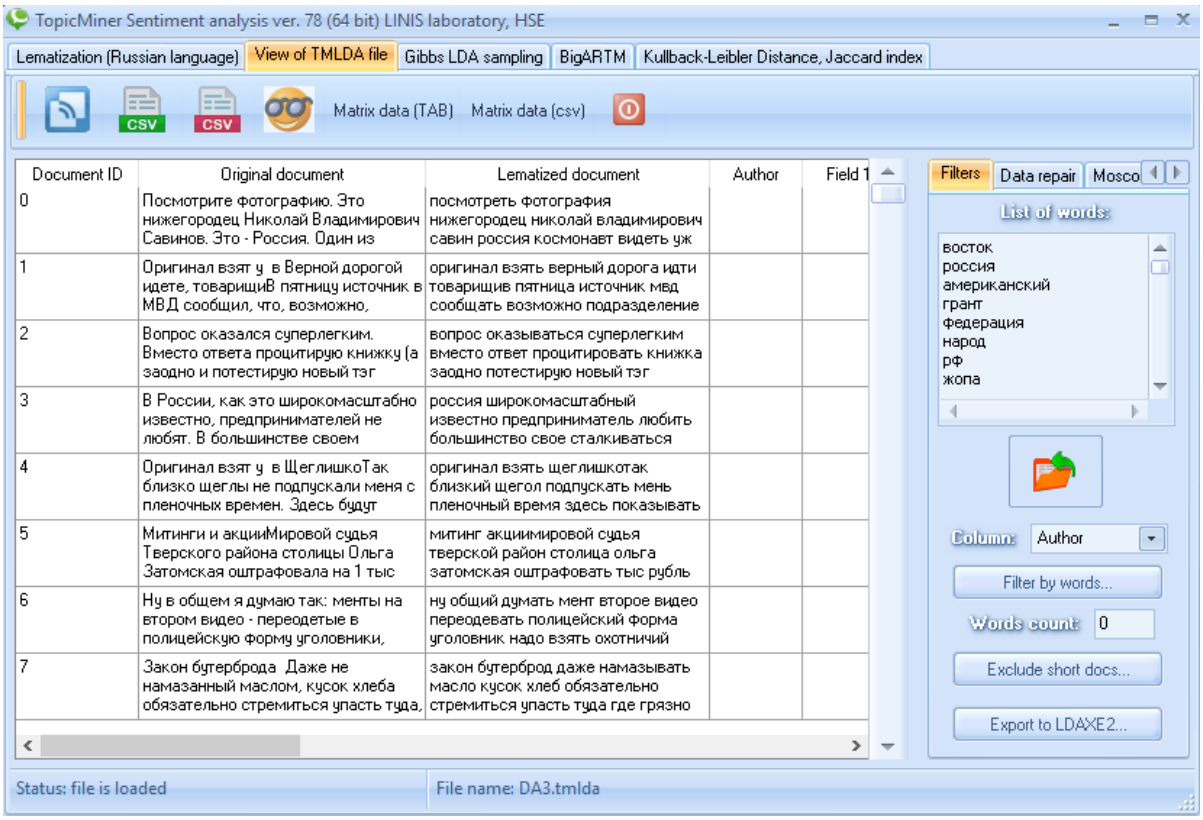


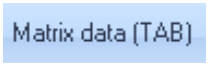
Рис. 2.3. Пример загруженного списка слов.

Выгрузка документов в формате tmlda по списку слов. Чтобы уменьшить коллекцию документов в соответствии со списком загруженных слов, нужно нажать на кнопку 

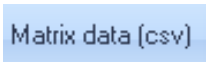
. Программа создаст файл в формате tmlda с именем изначально загруженного файла, однако к имени файла будет добавлена комбинация букв '_ww'. Например, 'my_test2_ww.tmlda'. В файле будут только те документы, в которых встречается хотя бы одно слово из загруженного списка.

Выгрузка документов в формате 'tmlda' с удаленными пустыми документами. Документы могут оказаться пустыми в результате удаления стоп-слов либо изначально – например, это записи в соцсетях, содержащие только фотографию. Для уменьшения времени моделирования такие документы рекомендуется удалять. Чтобы создать файл в формате tmlda без пустых докуентов, нужно нажать на кнопку 

. Программа создаст файл в формате tmlda, с именем изначально загруженного файла, однако к имени файла будет добавлена комбинация букв '_we'. Например, 'my_test2_we.tmlda'.

Расчет term-document matrix (формат TAB). При нажатии на кнопку 

производится расчет частот списка слов, загруженных в данную опцию, и выгрузка матрицы частот для использования этой матрицы в статистическом пакете 'Orange' (разделитель TAB). Данная матрица может быть использована для обучения классификаторов типа 'Naïve Bayes', 'KNN', 'SVM'.

Расчет term-document matrix (формат CSV). При нажатии на кнопку 

производится расчет частот списка слов, загруженных в данную опцию, и выгрузка матрицы частот в формате CSV. Данная матрица может быть использована для обучения классификаторов типа 'Naïve Bayes', 'KNN', 'SVM'.

Формирование файлов 'tmlda' для мультимодальных моделей (BigARTM). Мультимодальная схема тематического моделирования включает в себя использование полей метаданных (не более 5 штук полей). В результате расчета мультимодальных схем получаются дополнительные матрицы распределения выбранный полей метаданных по темам. Для того что бы сформировать данные для модели BigARTM нужно выбрать опцию 'Dict for BigARTM' (смотри рис. 2.4)

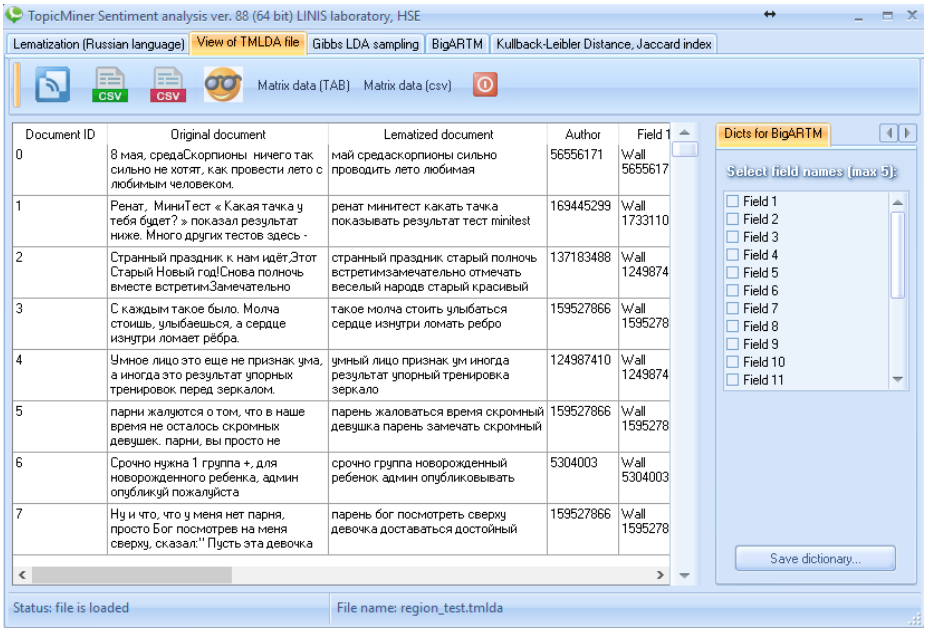


Рис. 2.4. Пример опции формирования данных для моделей BigArtm.

В списке полей необходимо указать нужные поля, например, поле 5 (фамилия автора поста) и поле 7 (геотег). После этого необходимо на жать на кнопку 

.

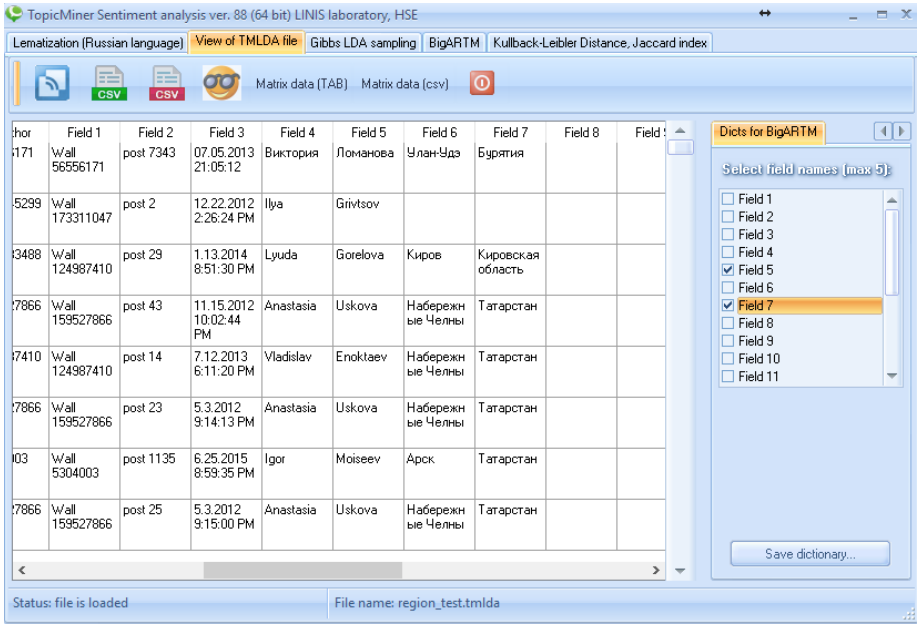


Рис. 2.5. Пример опции формирования данных для моделей BigArtm.

В результате запустится процес создания двух файлов: 1. Файл Tmlda, в котором сформированы выбранные поля метаданных. 2. Файл со словарями метаданных. **Внимание**, данный процесс занимает много времени, так как происходит процедура лематизации выбранных полей, конвертация лематизированных данных в формат crc32 и создание списка уникальных слов по выбранным полям.

Глава 3. Тематическое моделирование по модели сэмплирования Гиббса.

3.1. Интерфейс опции 'Gibbs LDA sampling'.

Результатом препроцессинга является файл с расширением tmlda. Он содержит лемматизированные, оригинальные документы и документы в цифровой форме. Каждый из документов имеет свой ID (ID лемматизированных и оригинальных документов одинаковы). Лемматизированые документы используются непосредственно для тематического моделирования, а оригинальные документы удобны для чтения.

Интерфейс опции 'Gibbs LDA sampling' выглядит следующим образом (см. Рис. 3.1).

 - кнопка загрузки данных для тематического моделирования.

 - кнопка запуска тематического моделирования.

 - кнопка остановки тематического моделирования.

 - кнопка просмотра матрицы распределения документов по темам (не отсортированный вариант матрицы).

 - кнопка просмотра матрицы распределений слов по темам (не отсортированный вариант матрицы).

 - кнопка просмотра матрицы распределения документов по темам (документы отсортированы по вероятности в каждой теме в порядке убывания).

 - кнопка просмотра матрицы распределения слов по темам (слова отсортированы по вероятности в каждой теме в порядке убывания).

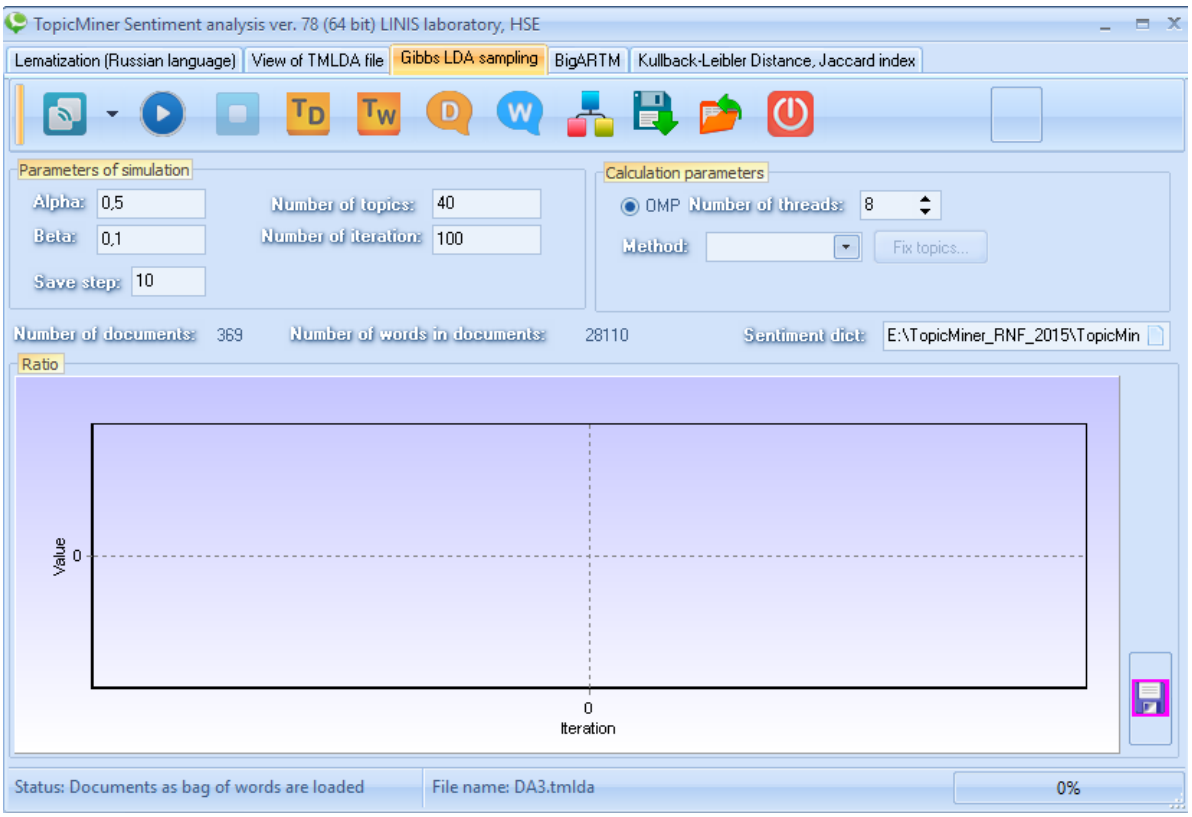


Рис. 3.1. Интерфейс опции 'Gibbs LDA sampling'

3.2. Загрузка документов для тематического моделирования.

Чтобы загрузить документы в программу для моделей на основе сэмплирования Гиббса нужно нажать на кнопку  , и в появившемся окне указать файл с расширением tmlda (см. Рис. 3.2).

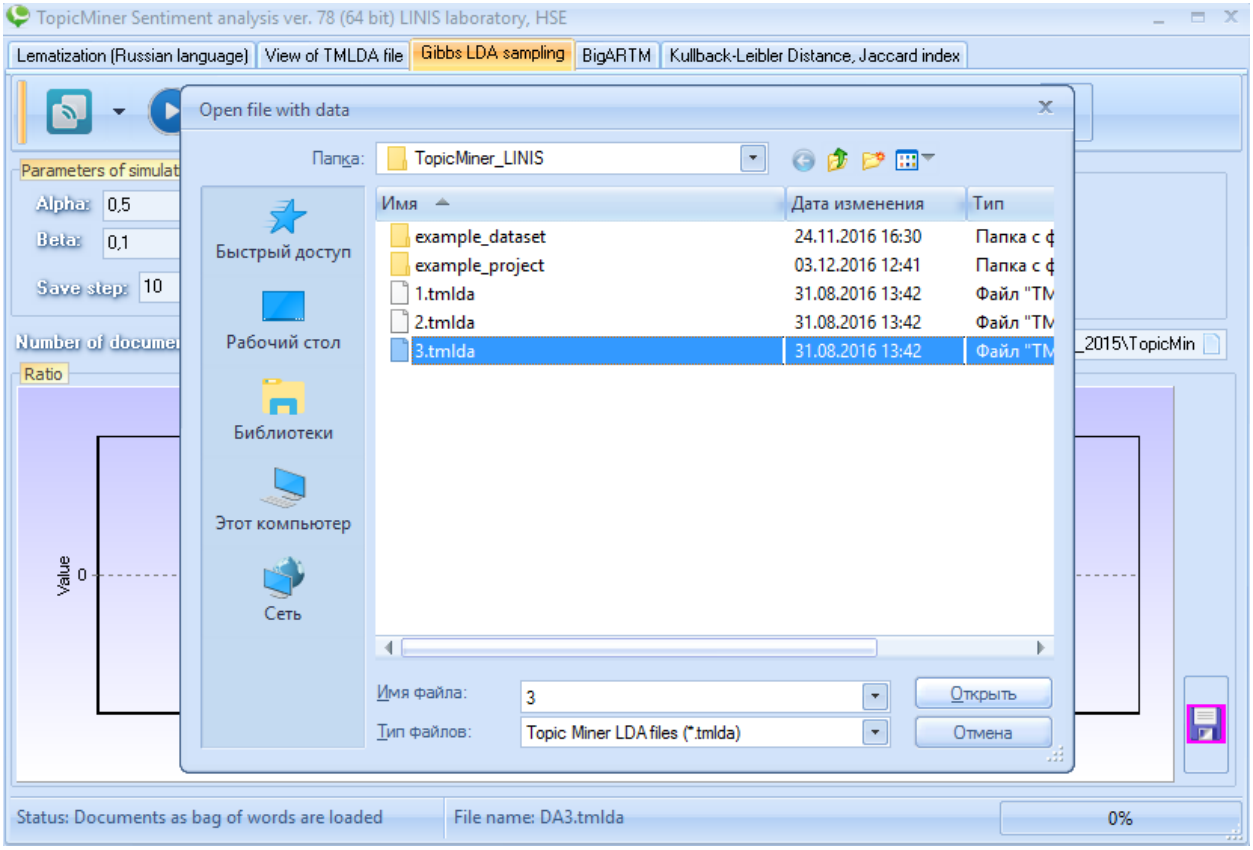


Рис. 3.2. Пример загрузки файла с данными.

Пример процесса загрузки данных показан на рисунке 3.3.

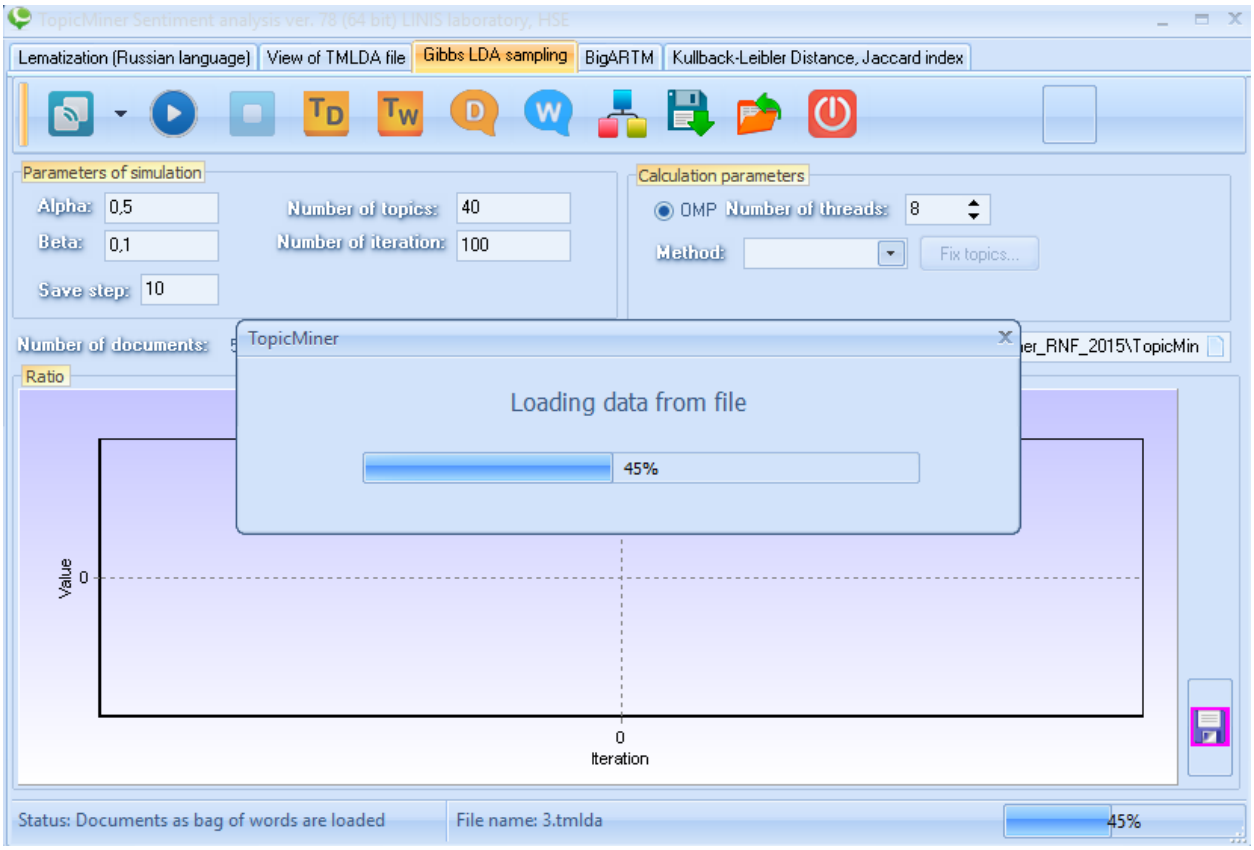


Рис. 3.3. Пример загрузки файла с данными.

После загрузки программа покажет статистику по документам и словам (см. рис. 3.4).

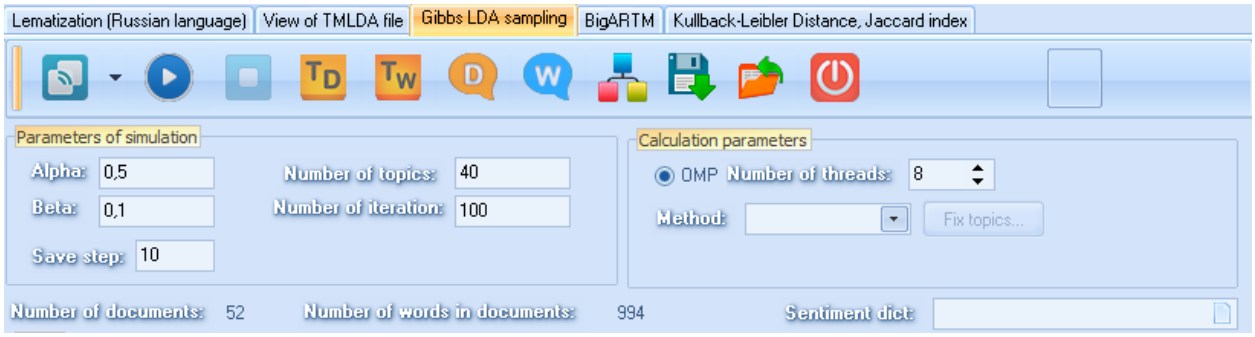


Рис. 3.4. Пример загрузки файла с данными.

Number of documents - число документов в коллекции (число документов в файле tmlda).

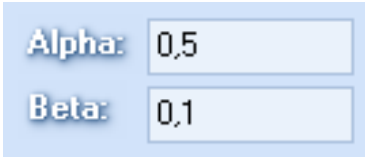
Number of words in documents: - число уникальных слов в коллекции.

Загруженная коллекция документов может использоваться в тематическом моделировании.

3.3. Тематическое моделирование на основе сэмплирования Гиббса.

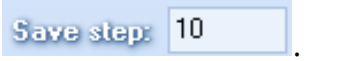
Перед запуском моделирования необходимо указать следующие параметры моделирования:

1. Коэффициенты α, β. Значение по умлочанию: α=0,5, β=1. Начинающие пользователи могут пользоваться значениями по умолчанию.

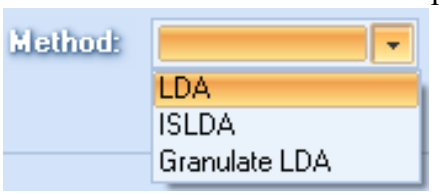


2. Number of topics. Число тем можно установить в опции: 

. Значение по умолчанию: 40 тем, однако всем пользователям рекомендуется экспериментировать с числом тем – как правило, в большую сторону от установленного по умолчанию.
3. Число итераций. Число итераций можно установить в опции: 

. По умолчанию стоит величина 100. Начинающие пользователи могут пользоваться этим значением.
4. Save step. Данный параметр показывает шаг по итерациям, который устанавливает, через какой шаг нужно визуализировать результаты расчета. По умолчанию стоит величина 10. Изменить величину можно в следующей опции: 

.
5. Тип модели. В данной версии реализованы три вида моделей (стандартная модель LDA, модель ISLDA и гранулированный метод сэмплирования GLDA). Рекомендовано опытным пользователям. Выбор модели осуществляется из выпадающего списка.



6. Число потоков. В данной программе реализовано распараллеливание тематической модели на основе сэмплирования Гиббса по технологии OpenMP. Число потоков можно указать в следующей опции: 

После установки параметров нужно нажать на кнопку  . Процесс вычисления (номер итерации) показывается в нижнем левом углу окна (см. рис. 3.5).

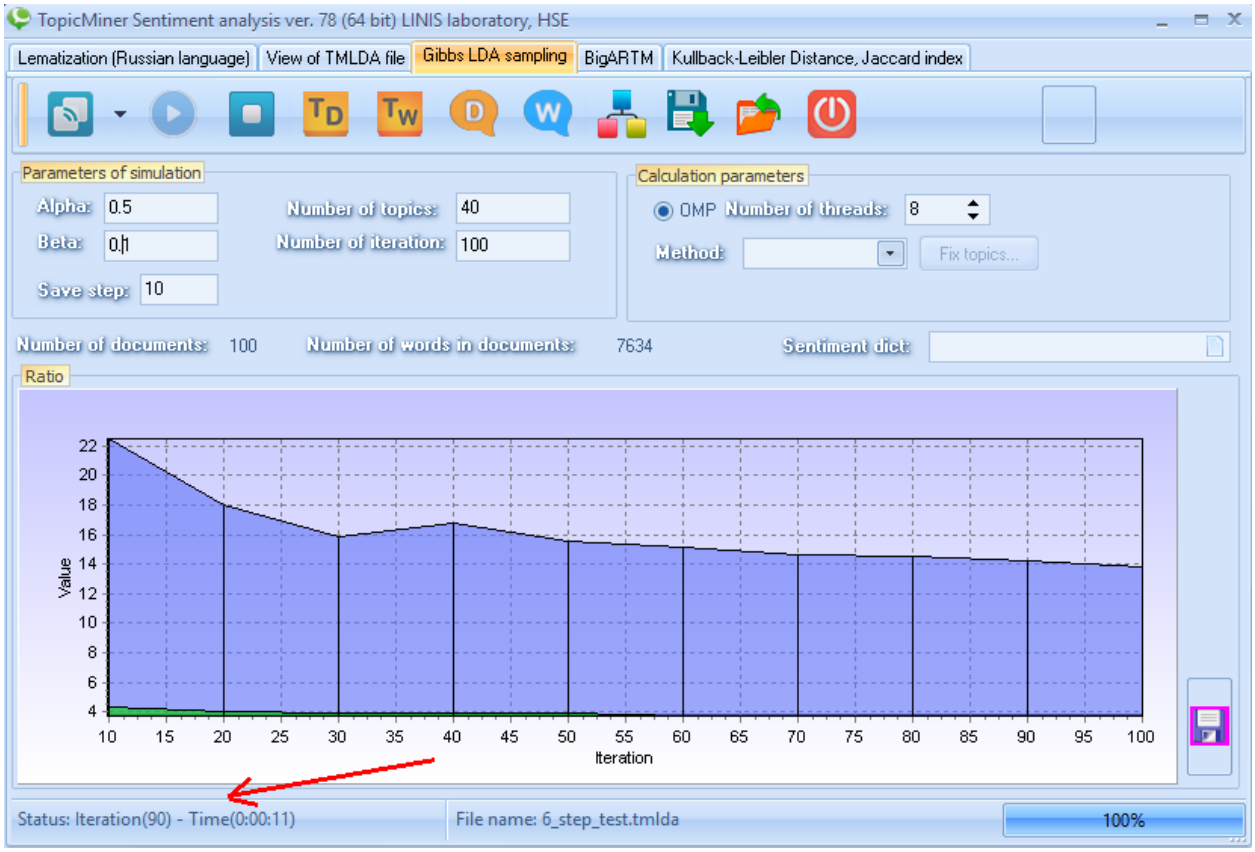


Рис. 3.5. Процесс исполнения тематической модели.

В ходе тематического моделирования программа производит вычисление доли слов и документов, у которых вероятность выше среднего. Графики вероятностей в ходе итераций показаны на графике (см. рис. 3.6).

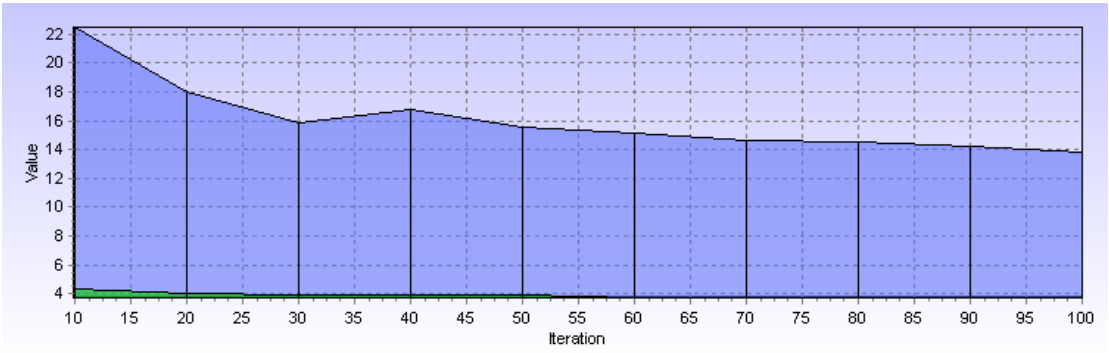


Рис. 3.6. Процесс исполнения тематической модели.

Синий график показывает долю документов, зеленый показывает долю слов. Например, для документов из социальной сети Живой журнал, типичное количество документов с вероятностью выше средней величины порядка 11%.

3.4. Визуализация результатов тематического моделирования.

Визуализация тематического моделирования состоит из следующих пунктов:

1. Визуализация распределения документов по темам.
2. Визуализация слов по темам.
3. Визуализация сортированных распределений документов по темам
4. Визуализация сортированных распределений слов по темам.

Запуск модулей визуализации осуществляется при помощи кнопок 

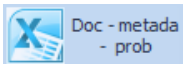
.

3.4.1. Визуализация распределений документов по темам.

Для визуализации распределения документов по темам нужно нажать на кнопку 

. Появится окно (см. рис. 3.7 и 3.8). В таблице каждая строка представляет текст документа (столбец ‘Orig text’), его метаданные (начиная со столбца ‘Nick’ и заканчивая столбцом ‘Field 20’) и вероятности принадлежности к темам. Таким образом, TopicMiner позволяет использовать 21 столбец для метаданных (см. рис. 3.7). Распределение документов по темам приводится в столбцах, начиная со столбца ‘1’ и заканчивая номером темы, которая задана в параметре ‘Number of topic’.

В этом окне также есть ряд кнопок, которые позволяют сделать выгрузку результатов тематического моделирования в файлы формата csv.



- Выгрузка результатов тематического моделирования в формате csv в виде: оригинальный текст – метаданные – вероятности по всем темам. Пример такой выгрузки приведен на рисунке 3.9.

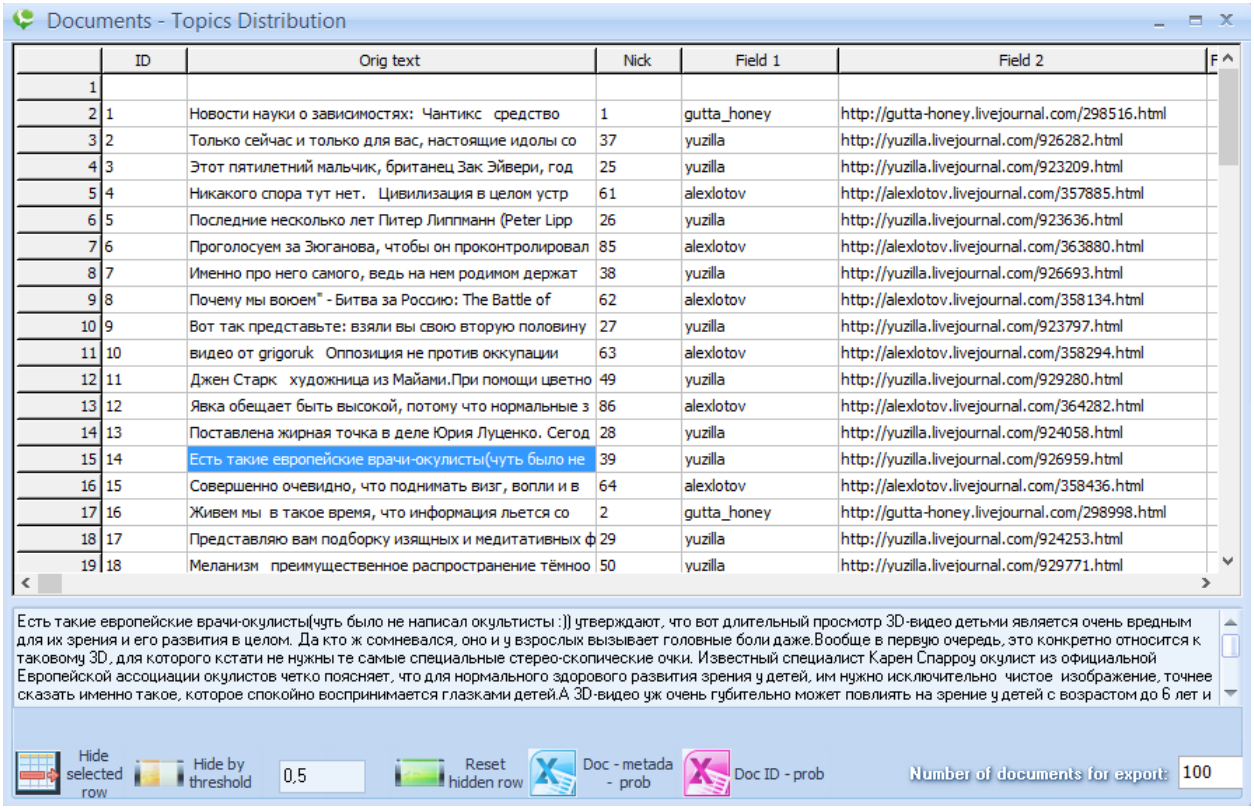
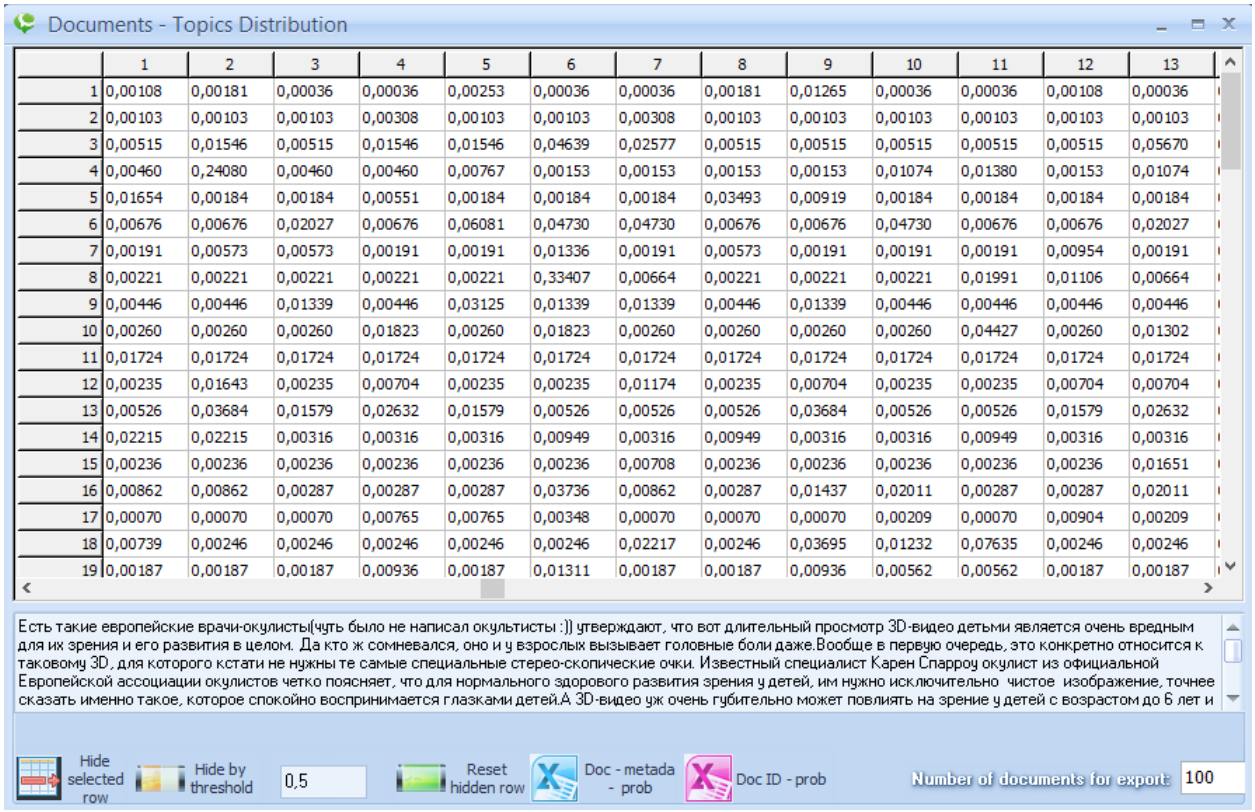


Рис. 3.7. Визуализация распределений документов по темам (первая часть).

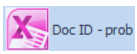


	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0,00108	0,00181	0,00036	0,00036	0,00253	0,00036	0,00036	0,00181	0,01265	0,00036	0,00036	0,00108	0,00036
2	0,00103	0,00103	0,00103	0,00308	0,00103	0,00103	0,00308	0,00103	0,00103	0,00103	0,00103	0,00103	0,00103
3	0,00515	0,01546	0,00515	0,01546	0,01546	0,04639	0,02577	0,00515	0,00515	0,00515	0,00515	0,00515	0,05670
4	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
5	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
6	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
7	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
8	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
9	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
10	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
11	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724
12	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
13	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
14	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
15	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
16	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
17	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
18	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
19	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

Рис. 3.8. Визуализация распределений документов по темам (вторая часть).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	ID	Orig text	Nick	Field 1	Field 2	Field 3	Field 4	Field 5	Field 6	1	2	3	4	5	6	7	8	9
2										0,00108	0,00181	0,00036	0,00036	0,00253	0,00036	0,00036	0,00181	0,01265
3	1	Новости науки о зависимостях: Чанти	1	gutta_honey	http://gutta-honey.livejournal.com/298516.html					0,00103	0,00103	0,00103	0,00308	0,00103	0,00103	0,00308	0,00103	0,00103
4	2	Только сейчас и только для вас, насто	37	yuzilla	http://yuzilla.livejournal.com/926282.html					0,00515	0,01546	0,00515	0,01546	0,01546	0,04639	0,02577	0,00515	0,00515
5	3	Этот пятилетний мальчик, британец	25	yuzilla	http://yuzilla.livejournal.com/923209.html					0,0046	0,2408	0,0046	0,0046	0,00767	0,00153	0,00153	0,00153	0,00153
6	4	Никакого спора тут нет. Цивилизаци	61	alexlotov	http://alexlotov.livejournal.com/357885.html					0,01654	0,00184	0,00184	0,00551	0,00184	0,00184	0,00184	0,03493	0,00919
7	5	Последние несколько лет Питер Лип	26	yuzilla	http://yuzilla.livejournal.com/923636.html					0,00676	0,00676	0,02027	0,00676	0,06081	0,0473	0,0473	0,00676	0,00676
8	6	Проголосуем за Зюганова, чтобы он г	85	alexlotov	http://alexlotov.livejournal.com/363880.html					0,00191	0,00573	0,00573	0,00191	0,00191	0,01336	0,00191	0,00573	0,00191
9	7	Именно про него самого, ведь на нем	38	yuzilla	http://yuzilla.livejournal.com/926693.html					0,00221	0,00221	0,00221	0,00221	0,00221	0,33407	0,00664	0,00221	0,00221
10	8	Почему мы воюем - Битва за Россию:	62	alexlotov	http://alexlotov.livejournal.com/358134.html					0,00446	0,00446	0,01339	0,00446	0,03125	0,01339	0,01339	0,00446	0,01339
11	9	Вот так представьте: взяли вы свою в	27	yuzilla	http://yuzilla.livejournal.com/923797.html					0,0026	0,0026	0,0026	0,01823	0,0026	0,01823	0,0026	0,0026	0,0026
12	10	видео от grigoruk Оппозиция не про	63	alexlotov	http://alexlotov.livejournal.com/358294.html					0,01724	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724
13	11	Джен Старк художница из Майами.Г	49	yuzilla	http://yuzilla.livejournal.com/929280.html					0,00235	0,01643	0,00235	0,00704	0,00235	0,00235	0,01174	0,00235	0,00704
14	12	Явка обещает быть высокой, потому	86	alexlotov	http://alexlotov.livejournal.com/364282.html					0,00526	0,03684	0,01579	0,02632	0,01579	0,00526	0,00526	0,00526	0,03684
15	13	Поставлена жирная точка в деле Юри	28	yuzilla	http://yuzilla.livejournal.com/924058.html					0,02215	0,02215	0,00316	0,00316	0,00316	0,00949	0,00316	0,00949	0,00316
16	14	Есть такие европейские врачи-окули	39	yuzilla	http://yuzilla.livejournal.com/926959.html					0,00236	0,00236	0,00236	0,00236	0,00236	0,00236	0,00708	0,00236	0,00236
17	15	Совершенно очевидно, что поднима	64	alexlotov	http://alexlotov.livejournal.com/358436.html					0,00862	0,00862	0,00287	0,00287	0,00287	0,03736	0,00862	0,00287	0,01437
18	16	Живем мы в такое время, что инфор	2	gutta_honey	http://gutta-honey.livejournal.com/298998.html					0,0007	0,0007	0,0007	0,00765	0,00765	0,00348	0,0007	0,0007	0,0007
19	17	Представляю вам подборку изящных	29	yuzilla	http://yuzilla.livejournal.com/924253.html					0,00739	0,00246	0,00246	0,00246	0,00246	0,00246	0,02217	0,00246	0,03695
20	18	Меланизм преимущественное расп	50	yuzilla	http://yuzilla.livejournal.com/929771.html					0,00187	0,00187	0,00187	0,00936	0,00187	0,01311	0,00187	0,00187	0,00936
21	19	Самое дорогое путешествие на двои	40	yuzilla	http://yuzilla.livejournal.com/927036.html					0,00211	0,00211	0,03165	0,00633	0,02321	0,27637	0,00211	0,00211	0,01899
22	20	Новая парадигма мировоззрения Кто	73	alexlotov	http://alexlotov.livejournal.com/360899.html					0,00214	0,00071	0,00071	0,00071	0,00214	0,00071	0,00071	0,00071	0,00357
23	21	el_murid: Арабская весна создала вес	87	alexlotov	http://alexlotov.livejournal.com/364440.html					0,05968	0,00161	0,00161	0,00161	0,00161	0,00484	0,00161	0,17258	0,00484
24	22	В случае конфликта , неудовлетворит	13	gutta_honey	http://gutta-honey.livejournal.com/303091.html					0,00038	0,00038	0,0019	0,00267	0,00038	0,00038	0,00343	0,00038	0,00038
25	23	Сегодня в Астрахани произошел взры	30	yuzilla	http://yuzilla.livejournal.com/924445.html					0,00256	0,00256	0,00256	0,00256	0,01795	0,00769	0,01282	0,00256	0,00256
26	24	Проводимые разными компаниями и	41	yuzilla	http://yuzilla.livejournal.com/927292.html					0,01289	0,00773	0,00258	0,00258	0,00258	0,00258	0,00773	0,00258	0,01289

Рис. 3.9. Выгрузка результатов тематического моделирования в формате ‘Оригинальный текст – метаданные – вероятности’.



- Выгрузка данных в виде: id документа – количество слов в документе – метаданные – вероятности. Данные выгружаются в формате csv. Пример подобной выгрузки приведен на рисунке 3.10.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	ID docum	Number o	Nick	Field1	Field2	Field3	Field4	Field5	Field6	Topic(1)	Topic(2)	Topic(3)	Topic(4)	Topic(5)	Topic(6)	Topic(7)	Topic(8)	Topic(9)	Topic(10)	Topic(11)
2	1	447	1	gutta_honey	http://gutta-honey.livejournal.com/298516.html					0,00103	0,00103	0,00103	0,00308	0,00103	0,00103	0,00308	0,00103	0,00103	0,00103	0,00103
3	2	74	37	yuzilla	http://yuzilla.livejournal.com/926282.html					0,00515	0,01546	0,00515	0,01546	0,01546	0,04639	0,02577	0,00515	0,00515	0,00515	0,00515
4	3	308	25	yuzilla	http://yuzilla.livejournal.com/923209.html					0,0046	0,2408	0,0046	0,0046	0,00767	0,00153	0,00153	0,00153	0,00153	0,01074	0,0138
5	4	253	61	alexlotov	http://alexlotov.livejournal.com/357885.html					0,01654	0,00184	0,00184	0,00551	0,00184	0,00184	0,00184	0,03493	0,00919	0,00184	0,00184
6	5	53	26	yuzilla	http://yuzilla.livejournal.com/923636.html					0,00676	0,00676	0,02027	0,00676	0,06081	0,0473	0,0473	0,00676	0,00676	0,0473	0,00676
7	6	243	85	alexlotov	http://alexlotov.livejournal.com/363880.html					0,00191	0,00573	0,00573	0,00191	0,00191	0,01336	0,00191	0,00573	0,00191	0,00191	0,00191
8	7	211	38	yuzilla	http://yuzilla.livejournal.com/926693.html					0,00221	0,00221	0,00221	0,00221	0,00221	0,33407	0,00664	0,00221	0,00221	0,00221	0,01991
9	8	102	62	alexlotov	http://alexlotov.livejournal.com/358134.html					0,00446	0,00446	0,01339	0,00446	0,03125	0,01339	0,01339	0,00446	0,01339	0,00446	0,00446
10	9	172	27	yuzilla	http://yuzilla.livejournal.com/923797.html					0,0026	0,0026	0,0026	0,01823	0,0026	0,01823	0,0026	0,0026	0,0026	0,0026	0,04427
11	10	9	63	alexlotov	http://alexlotov.livejournal.com/358294.html					0,01724	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724	0,01724
12	11	247	49	yuzilla	http://yuzilla.livejournal.com/929280.html					0,00235	0,01643	0,00235	0,00704	0,00235	0,00235	0,01174	0,00235	0,00704	0,00235	0,00235
13	12	78	86	alexlotov	http://alexlotov.livejournal.com/364282.html					0,00526	0,03684	0,01579	0,02632	0,01579	0,00526	0,00526	0,00526	0,03684	0,00526	0,00526
14	13	138	28	yuzilla	http://yuzilla.livejournal.com/924058.html					0,02215	0,02215	0,00316	0,00316	0,00316	0,00949	0,00316	0,00949	0,00316	0,00316	0,00949
15	14	194	39	yuzilla	http://yuzilla.livejournal.com/926959.html					0,00236	0,00236	0,00236	0,00236	0,00236	0,00236	0,00708	0,00236	0,00236	0,00236	0,00236
16	15	154	64	alexlotov	http://alexlotov.livejournal.com/358436.html					0,00862	0,00862	0,00287	0,00287	0,00287	0,03736	0,00862	0,00287	0,01437	0,02011	0,00287
17	16	714	2	gutta_honey	http://gutta-honey.livejournal.com/298998.html					0,0007	0,0007	0,0007	0,00765	0,00765	0,00348	0,0007	0,0007	0,0007	0,00209	0,0007
18	17	190	29	yuzilla	http://yuzilla.livejournal.com/924253.html					0,00739	0,00246	0,00246	0,00246	0,00246	0,00246	0,02217	0,00246	0,03695	0,01232	0,07635
19	18	247	50	yuzilla	http://yuzilla.livejournal.com/929771.html					0,00187	0,00187	0,00187	0,00936	0,00187	0,01311	0,00187	0,00187	0,00936	0,00562	0,00562
20	19	216	40	yuzilla	http://yuzilla.livejournal.com/927036.html					0,00211	0,00211	0,03165	0,00633	0,02321	0,27637	0,00211	0,00211	0,01899	0,02321	0,01899
21	20	688	73	alexlotov	http://alexlotov.livejournal.com/360899.html					0,00214	0,00071	0,00071	0,00071	0,00214	0,00071	0,00071	0,00071	0,00357	0,00071	0,00071
22	21	294	87	alexlotov	http://alexlotov.livejournal.com/364440.html					0,05968	0,00161	0,00161	0,00161	0,00161	0,00484	0,00161	0,17258	0,00484	0,00806	0,00161
23	22	1315	13	gutta_honey	http://gutta-honey.livejournal.com/303091.html					0,00038	0,00038	0,0019	0,00267	0,00038	0,00038	0,00343	0,00038	0,00038	0,00495	0,00038
24	23	175	30	yuzilla	http://yuzilla.livejournal.com/924445.html					0,00256	0,00256	0,00256	0,00256	0,01795	0,00769	0,01282	0,00256	0,00256	0,00256	0,00256
25	24	172	41	yuzilla	http://yuzilla.livejournal.com/927292.html					0,01289	0,00773	0,00258	0,00258	0,00258	0,00258	0,00773	0,00258	0,01289	0,00773	0,01804
26	25	370	65	alexlotov	http://alexlotov.livejournal.com/358710.html					0,0013	0,0039	0,00649	0,0013	0,0013	0,0013	0,0013	0,0013	0,0013	0,0039	0,0013

Рис. 3.10. Выгрузка результатов тематического моделирования в формате ‘Id документа– метаданные – вероятности’.

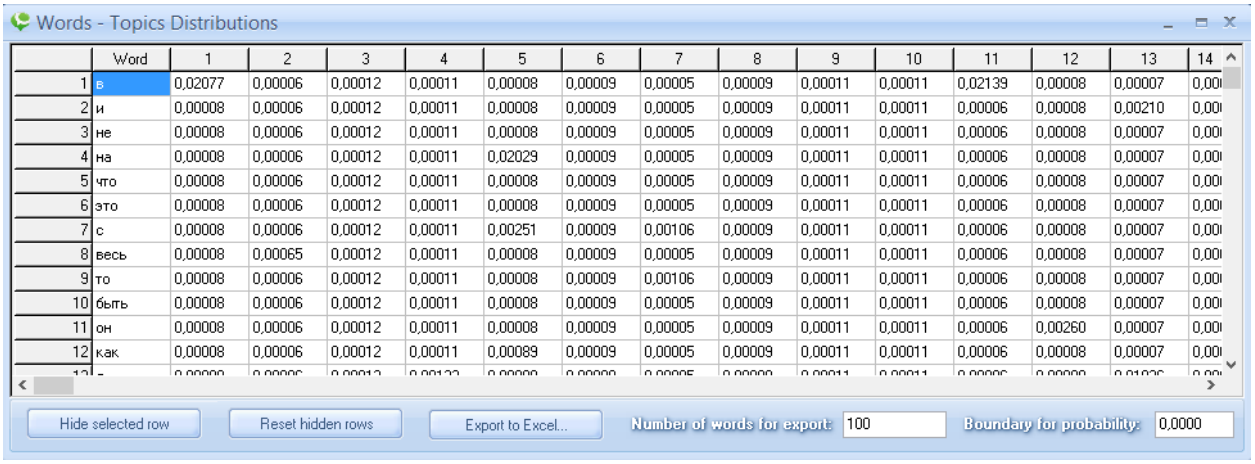
В данных выгрузках число выгружаемых документов можно задавать в опции: 

.

3.4.2. Визуализация распределений слов по темам.

Чтобы визуализировать распределение слов по темам, нужно нажать на кнопку 

. При нажатии на эту кнопку появится окно (см. рисунок 3.11).



	Word	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	в	0,02077	0,00006	0,00012	0,00011	0,00008	0,00009	0,00005	0,00009	0,00011	0,00011	0,02139	0,00008	0,00007	0,00
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

Рис. 3.11. Пример визуализации распределений слов по темам.

Размер выгрузки (количество документов) регулируется двумя параметрами: 1. ‘Number of words for export’. 2.‘Boundary for probability’ (см. рис. 3.11). Первый параметр регулирует количество слов для экспорта в формате csv, второй указывает вероятность слова в теме, минимально необходимую для попадания слова в выгрузку. Слова с более низкими вероятностями не выгружаются.

Кнопка 

позволяет скрыть выбранную строку в таблице. Скрытое слово не участвует в выгрузке в формате csv.

Кнопка 

позволяет восстановить все скрытые ранее слова.

Чтобы выгрузить распределения слов по темам в файл формата csv, нужно нажать на кнопку 

и в появившемся окне указать имя файла.

Внимание: эта выгрузка полезна при исследовании стабильности тематического моделирования или при сравнении работы нескольких моделей между собой. Подобное сравнение обсуждается в главе 5.

3.4.3. Визуализация распределений отсортированных документов в темах.

Чтобы открыть окно, в котором представлены распределения документов по темам в порядке убывания вероятностей, нужно нажать на кнопку  . В результате появится окно; сортировка в нем производится по вероятности, таким образом, что бы наверху (в каждой теме) оказался документ с наибольшей вероятностью принадлежности этой теме. Пример подобной сортировки приведен на рисунке 3.12 В каждой ячейке данной таблицы лежит номер документа и его вероятность. Если кликнуть на выбранную ячейку, то в нижней части экрана отобразится оригинальный текст.

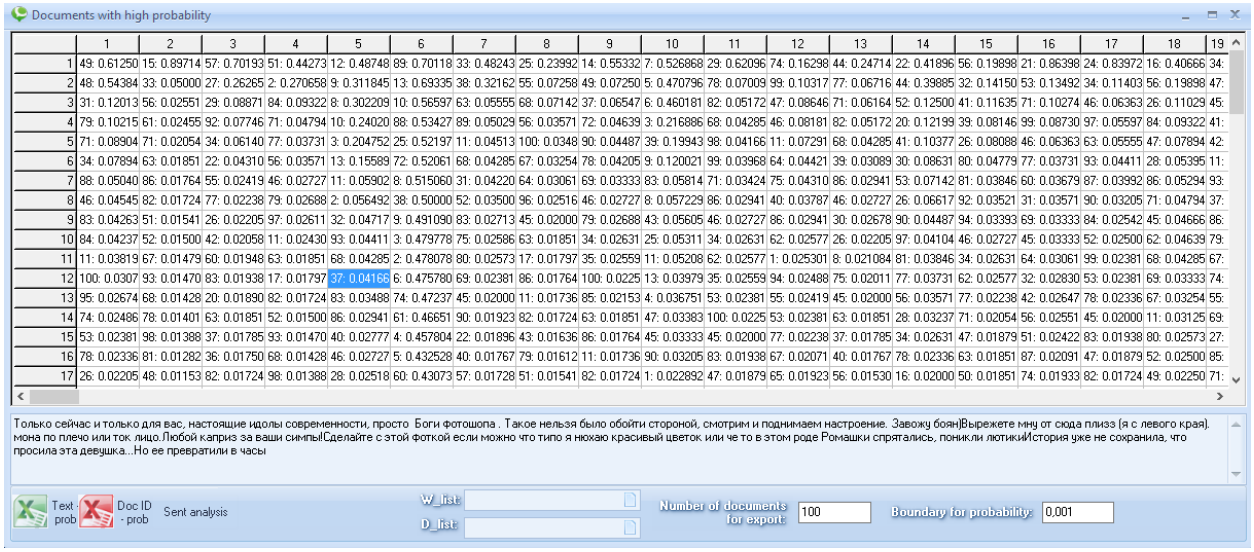
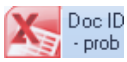


Рис. 3.12. Пример визуализации распределений документов по темам.

Выгрузка отсортированных результатов.

В этом окне реализованы несколько вариантов выгрузки отсортированных данных в файле формата csv.

- выгружаются тексты документов и их вероятности.



- выгружаются id документов и их вероятности. Пример такой выгрузки показан на рисунке 3.13.

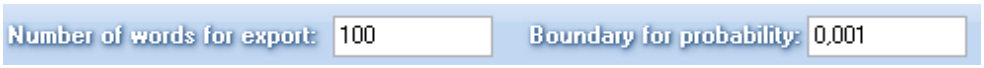
	A	B	C	D	E	F	G	H
1	ID doc(Topic 1)	Prob. of Doc.(Topic 1)	ID doc(To	Prob. of D	ID doc(To	Prob. of D	ID doc(To	Prob. of D
2	94	0,788783	53	0,39734	59	0,106481	86	0,134868
3	65	0,084615	62	0,307404	19	0,031646	31	0,048673
4	32	0,071782	3	0,240798	31	0,022124	32	0,042079
5	21	0,059677	32	0,14604	42	0,021186	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

Рис. 3.13. Пример выгрузки распределений документов по темам (id-probability).

Внимание, описание сентимент анализа приведено в главе 7.

3.4.2. Визуализация отсортированных распределений слов по темам.

Чтобы открыть окно, в котором представлены распределения слов по темам в порядке убывания вероятностей, нужно нажать на кнопку  . В результате появится окно, в котором, произведена сортировка по вероятности таким образом, что наверху (в каждой теме) находится слово с наибольшей вероятностью принадлежности каждой теме. Пример подобной сортировки приведен на рисунке 3.15. В данном окне также сделана возможность выгрузить результаты сортировки в файл формата csv. Размер выгрузки регулируется двумя параметрами: 1. Количество слов для выгрузки. 2. Граница по вероятности.



Первый параметр определяет максимальное число слов, которое нужно выгрузить. Второй параметр определяет границу. Слова с вероятностями ниже заданной границы выгружаться не будут.

	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Field1	Field2	Field3	Field4	Field5	Field6	Number words in doc(1)	Probability(1)	Documen	Nick	Field1	Field2	Field3	Field4	Field5	Field6
2	yuzilla	http://yuzilla.livejournal.com/922914.html					395	0,788783	53	54	yuzilla	http://yuzilla.livejournal.com/930651.html				
3	alexlotov	http://alexlotov.livejournal.com/363587.html					46	0,084615	62	43	yuzilla	http://yuzilla.livejournal.com/927854.html				
4	yuzilla	http://yuzilla.livejournal.com/927501.html					174	0,071782	3	25	yuzilla	http://yuzilla.livejournal.com/923209.html				
5	alexlotov	http://alexlotov.livejournal.com/364440.html					294	0,059677	32	42	yuzilla	http://yuzilla.livejournal.com/927501.html				
6	alexlotov	http://alexlotov.livejournal.com/359355.html					176	0,054124	12	86	alexlotov	http://alexlotov.livejournal.com/364282.html				
7	alexlotov	http://alexlotov.livejournal.com/360194.html					66	0,052326	78	99	alexlotov	http://alexlotov.livejournal.com/367512.html				
8	alexlotov	http://alexlotov.livejournal.com/361912.html					59	0,04321	40	33	yuzilla	http://yuzilla.livejournal.com/925211.html				
9	alexlotov	http://alexlotov.livejournal.com/365095.html					75	0,038889	13	28	yuzilla	http://yuzilla.livejournal.com/924058.html				
10	alexlotov	http://alexlotov.livejournal.com/357587.html					258	0,038182	99	11	gutta_hor	http://gutta-honey.livejournal.com/302369.html				
11	alexlotov	http://alexlotov.livejournal.com/366522.html					267	0,030357	81	100	alexlotov	http://alexlotov.livejournal.com/367668.html				
12	gutta_hor	http://gutta-honey.livejournal.com/302369.html					148	0,026786	33	53	yuzilla	http://yuzilla.livejournal.com/930544.html				
13	yuzilla	http://yuzilla.livejournal.com/928575.html					35	0,025862	10	63	alexlotov	http://alexlotov.livejournal.com/358294.html				
14	yuzilla	http://yuzilla.livejournal.com/922308.html					311	0,025	63	83	alexlotov	http://alexlotov.livejournal.com/363461.html				
15	yuzilla	http://yuzilla.livejournal.com/925448.html					43	0,02459	36	90	alexlotov	http://alexlotov.livejournal.com/365095.html				
16	alexlotov	http://alexlotov.livejournal.com/363461.html					129	0,02349	11	49	yuzilla	http://yuzilla.livejournal.com/929280.html				
17	yuzilla	http://yuzilla.livejournal.com/931203.html					88	0,023148	46	3	gutta_hor	http://gutta-honey.livejournal.com/299320.html				
18	yuzilla	http://yuzilla.livejournal.com/930817.html					43	0,022727	61	82	alexlotov	http://alexlotov.livejournal.com/363073.html				
19	yuzilla	http://yuzilla.livejournal.com/924058.html					138	0,022152	2	37	yuzilla	http://yuzilla.livejournal.com/926282.html				
20	alexlotov	http://alexlotov.livejournal.com/364963.html					187	0,021845	77	98	alexlotov	http://alexlotov.livejournal.com/367193.html				
21	yuzilla	http://yuzilla.livejournal.com/926122.html					218	0,018519	52	93	alexlotov	http://alexlotov.livejournal.com/365861.html				
22	alexlotov	http://alexlotov.livejournal.com/358294.html					9	0,017241	39	68	alexlotov	http://alexlotov.livejournal.com/359528.html				
23	alexlotov	http://alexlotov.livejournal.com/357885.html					253	0,016544	48	78	alexlotov	http://alexlotov.livejournal.com/362198.html				
24	yuzilla	http://yuzilla.livejournal.com/928978.html					143	0,016447	60	81	alexlotov	http://alexlotov.livejournal.com/362988.html				
25	alexlotov	http://alexlotov.livejournal.com/363073.html					11	0,016129	70	58	yuzilla	http://yuzilla.livejournal.com/931721.html				
26	alexlotov	http://alexlotov.livejournal.com/367193.html					14	0,013889	30	32	yuzilla	http://yuzilla.livejournal.com/925144.html				

Рис. 3.14. Пример выгрузки распределений документов по темам (метаданные – число слов в документе – вероятность документа).

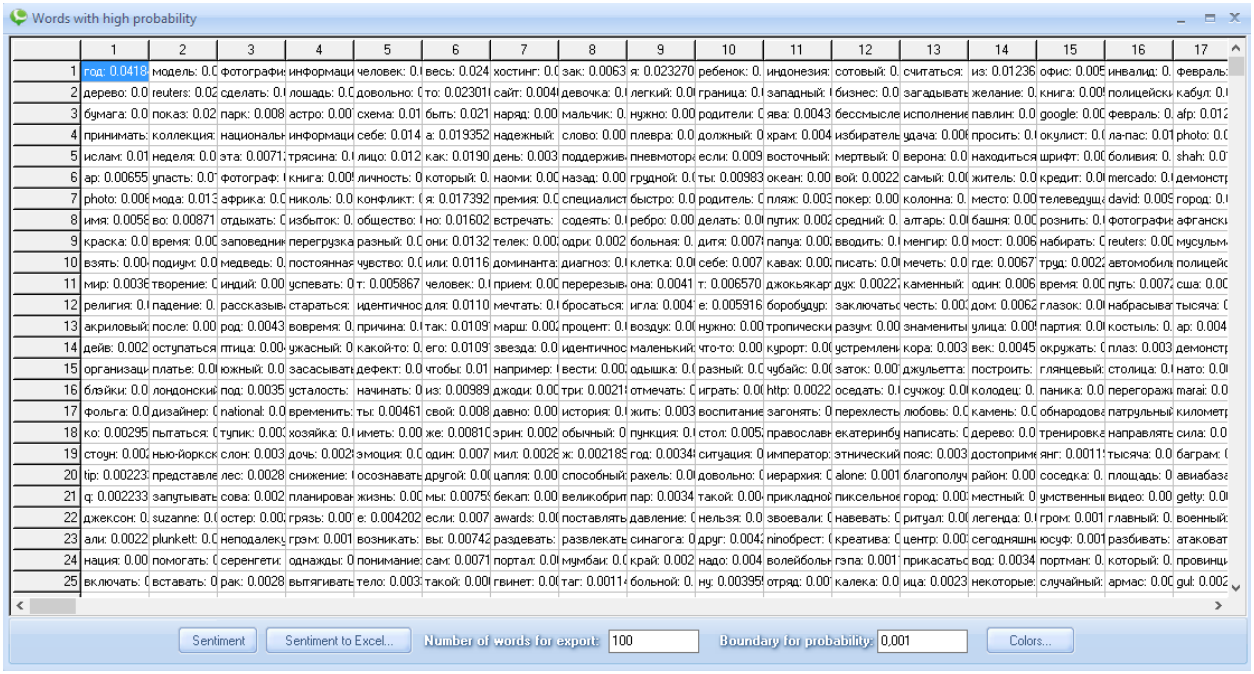
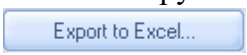


Рис. 3.15. Пример визуализации распределений слов по темам.

3.4.2.1. Экспорт результатов сортировки в файл формата csv.

Чтобы выгрузить результаты сортировки в файл формата csv, нужно нажать кнопку 

. В появившемся окне нужно указать имя файла. Пример подобной выгрузки приведен на рисунке 3.16.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	1	prob	2	prob	3	prob	4	prob	5	prob	6	prob	7
2	в	0,020765	кукла	0,020673	р	0,002468	северный	0,004602	на	0,020293	интернет	0,01176	просить
3	февраль	0,017583	семья	0,014194	ji	0,002468	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	нестабил	0,001293	также	0,002357	ла-пас	0,011399	отдых	0,005476	место
7	shah	0,009627	свадьба	0,008893	альфред	0,001293	смеяться	0,002357	полицейс	0,011399	сайт	0,005476	из
8	демонстр	0,008036	молодая	0,007715	прислуга	0,001293	два	0,002357	боливия	0,009783	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	младенец	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	pasquiniб	0,001235	автомоби	0,006549	магазин	0,003681	исполнен
13	демонстр	0,004853	дитя	0,004771	замешате	0,001293	ninобрест	0,001235	фотограф	0,006549	мобильны	0,003681	считаться
14	запад	0,004853	реборн	0,004771	хлопья	0,001293	leon	0,001235	путь	0,00574	машин	0,003681	построит
15	сша	0,004853	ортодокс	0,004771	отличите	0,001293	жанейро	0,001235	набрасыв	0,004932	мегафон	0,002783	удача
16	военный	0,004853	живой	0,004771	мстить	0,001293	jesseглаз	0,001235	костыль	0,004123	страна	0,002783	пора
17	оскорбля	0,004058	женщина	0,004182	уоллес	0,001293	насквозь	0,001235	район	0,004123	турист	0,002783	верона
18	ар	0,004058	зак	0,004182	актерский	0,001293	росиии	0,001235	из	0,003315	любой	0,002783	город
19	километр	0,004058	мальчик	0,003593	определе	0,001293	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	находить
21	нато	0,004058	мама	0,003593	джа	0,001293	источник	0,001235	численно	0,002506	бесплатн	0,002783	колодец

Рис. 3.16. Пример выгрузки распределений слов по темам в формате csv.

3.4.2.2. Визуализация распределений по весу темы.

Внимание, данная опция временно отключена, так как предполагается модернизация опции. Предполагается добавления визуализации распределений по сентимент весу.

Как правило, важна возможность быстрой оценки суммы весов всех вероятностей в заданной теме (в рамках заданного количество слов) и сортировка всех тем по весу. Это можно сделать, нажав на кнопку 

. В результате появится окно, в котором визуализировано отсортированное распределение тем по весам. Пример такого распределения приведен на рисунке 3.17. На графике также выводятся 6 наиболее вероятностных слов в каждой теме.

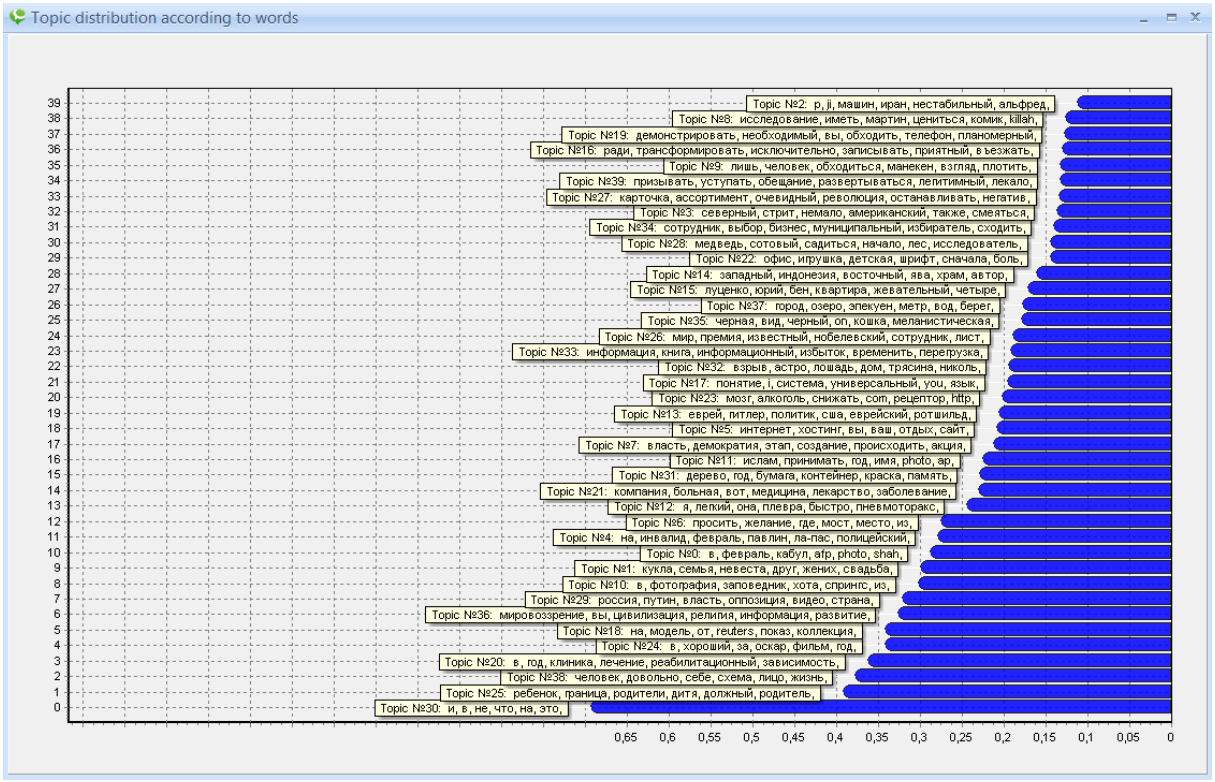


Рис. 3.17. Пример визуализации распределения тем по весу темы.

3.5. Сохранение результатов тематического моделирования в виде проектного файла.

Тематическое моделирование проводится на основе данных, загруженных из файла с расширением tmlda (например, 2_step_test.tmlda). В результате тематического моделирования создаются две матрицы: 1. Матрица распределения документов по темам (матрица phi). 2. Матрица распределения слов по темам (матрица theta). То есть в каталоге появляются два дополнительных файла: 2_step_test_phi.bin и 2_step_test_theta.bin. Таким образом, необходимо всегда хранить комбинацию: исходные данные плюс результаты моделирования. Это можно сделать нажав на кнопку  . В появившемся окне необходимо указать имя файла. Программа создаст проектный файл (например, my_test.tmproj), в котором будут прописаны пути к исходным данным (tmlda) и результатам тематического моделирования, то есть матрицам _phi.bin и _theta.bin.

Пример такого файла приведен ниже:

<?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>

Это позволит загружать для последующего анализа только один проектный файл, а программа автоматически подгрузит все остальные файлы. Проектный файл представляет собой текстовый файл, который легко изменить в случае переноса проекта на другой компьютер или в другой каталог.

3.6. Загрузка результатов тематического моделирования из проектного файла.

Чтобы загрузить полученные ранее результаты тематического моделирования, нужно нажать на кнопку  . В появившемся окне нужно указать имя проектного файла. Программа автоматически подгрузит все необходимые файлы на основе путей, указанных в проектном файле.

Глава 4. Тематическое моделирование по моделям BigArtm (мультимодальное тематическое моделирование).

4.1. Задание параметров в моделях мультимодальной ТМ.

Мультимодальный вариант тематического моделирования основан на процедуре регуляризации BigARTM, в котором задействованы метаданные, например, дата поста или геотег поста. Кроме того в версии 88 реализована возможность указания диапазона тем, что позволяет расчитать оптимальное число тем. Тематическое моделирование на основе аддитивной регуляризации и мультимодальные схемы реализованы на вкладке ‘BigArtm’. Пример интерфейса ‘BigArtm’ приведен на рисунке 4.1.

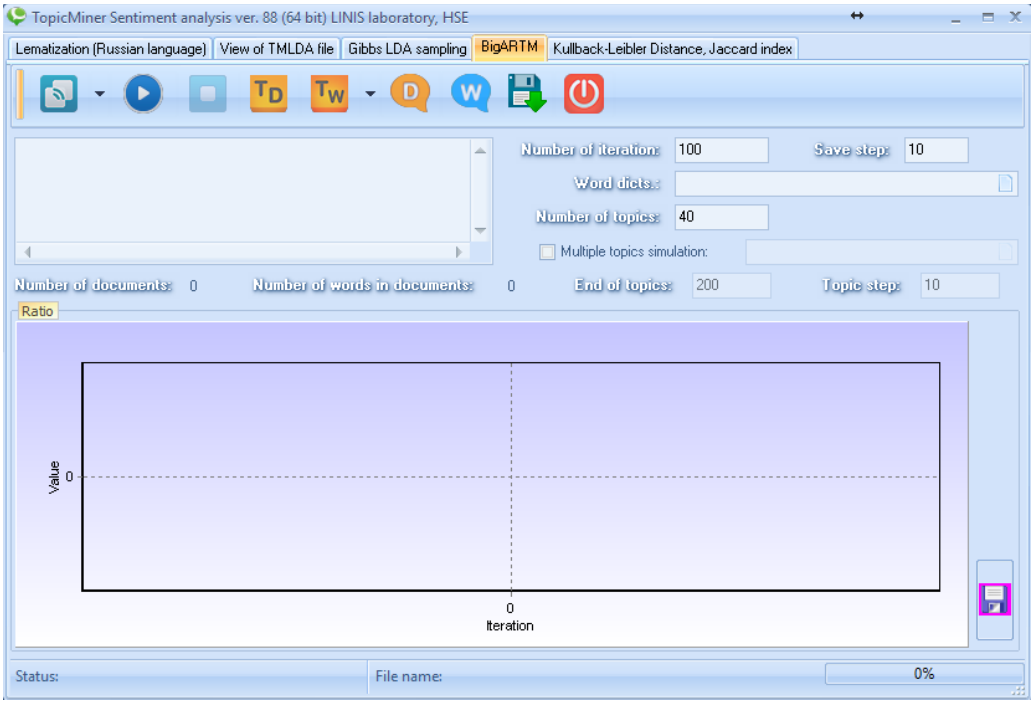


Рис. 4.1. Пример интерфейса ‘BigArtm’.

Модели ‘BigArtm’ характеризуются следующими параметрами (аналогичные моделям, основанные на сэмплировании Гиббса):

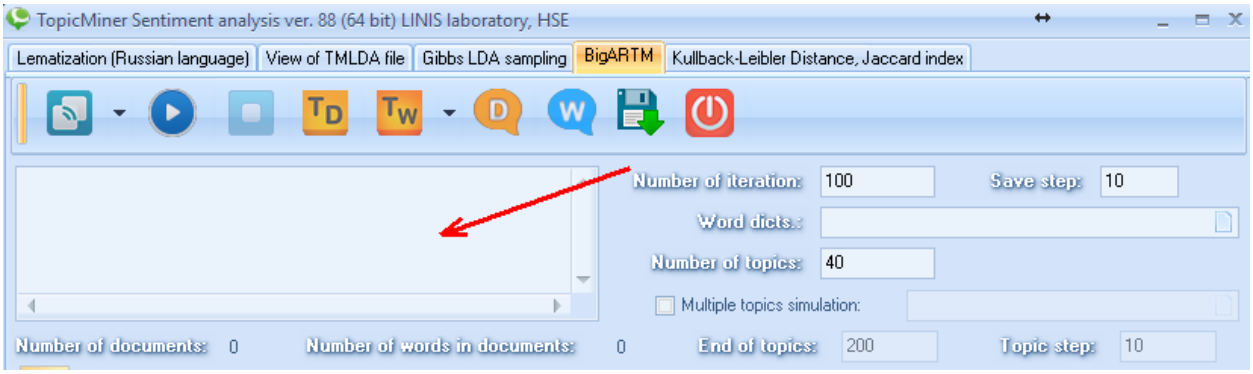
1. Количество тем. 

Данный параметр используется при расчетах с фиксированным количеством тем.
2. Количество итераций. 

3. 

. Шаг – количество итераций, после которых визуализируются результаты расчета.
4. Word dict – в данной опции, указывается словарь (в формате bin), которые содержит списки уникальных слов по выбранным полям метаданным (смотри главу 2 руководства пользователя).
5. Multiple topic simulation – в данной опции включается возможность задания диапазона числа тем.
6. End of topics – конечное число тем. Внимание, начальное число тем определяется параметром ‘Number of topics’.
7. Topic step - параметр, определяющий шаг по темам.

Существенным отличием от других моделей является способ задания регуляризаторов. Регуляризаторы задаются в виде текста в следующем окне:



В данной версии программного обеспечения заложены следующие возможности задания регуляризаторов:

1. Модель pLSA (не вводить никаких параметров).
2. Модель с очень разреженной матрицей Theta (Td) и плотной матрицей Phi (Tw). Пример задания регуляризатора: *--regularizer "0.2 SparseTheta* Величине регуляризатора (0.2) можно варьировать.
3. Модель с очень разреженной матрицей Phi (Tw) и плотной матрицей Theta (Td)
4. Пример задания регуляризатора: *--regularizer "0.5 SparsePhi"* Величину регуляризатора (0.5) можно варьировать.
5. Модель, в которой регуляризаторы применяются к фиксированным столбцами. Пример задания регуляризатора: *--topics obj:35,back:5 --regularizer "0.2 SmoothTheta #back" --regularizer "0.5 SparseTheta #obj"* => первые 35 столбцов матрицы Td разреженные, остальные пять – плотные.
6. Модель декорреляции тем. Пример задания регуляризатора: *--regularizer "1000 decorrelation"*. Величину регуляризатора (1000) можно менять.

Подробное описание моделей и регуляризаторов можно найти по адресу: http://bigartm.org/

Внимание, визуализация величин word_ratio и doc_ratio для BIGARTM не реализованы в данной версии.

4.2. Визуализация результатов тематического моделирования.

Визуализация тематического моделирования состоит из следующих пунктов:

1. Визуализация распределения документов по темам.
2. Визуализация распределения слов по темам.
3. Визуализация сортированных распределений документов по темам
4. Визуализация сортированных распределений слов по темам.

Запуск модулей визуализации осуществляется при помощи кнопок 

.

Действие кнопок аналогично действию кнопок в моделях на основе сэмплирования Гиббса.

4.3. Сохранения результатов тематического моделирования в виде проектного файла.

Тематическое моделирование проводится на основе данных, загруженных из файла с расширением tmlda (например, 2_step_test.tmlda). В результате тематического моделирования создаются две матрицы: 1. Матрица распределения документов по темам (матрица phi). 2. Матрица распределения слов по темам (матрица theta). Таким образом, в каталоге появляются два дополнительных файла: 2_step_test_phi.bin и 2_step_test_theta.bin. Необходимо всегда хранить комбинацию: исходные данные плюс результаты моделирования. Это можно сделать, нажав на кнопку  . В появившемся окне необходимо указать имя файла. Программа создаст проектный файл (например, my_test.tmproj), в котором будут прописаны пути к исходным данным (tmlda) и результатам тематического моделирования, то есть матрицам _phi.bin и _theta.bin.

Внимание: загрузить результаты расчета по модели BigArtm можно на вкладке ‘Gibbs LDa sampling’, в силу того, что модели на основе сэмплирования Гиббса и модели на основе аддитивной регуляризации аналогичны по структуре. Иными словами, в обоих случаях результатами являются матрицы _phi.bin и _theta.bin.

4.4. Расчет мультимодального варианта ТМ.

Для того, что бы запустить расчет мультимодьной тематической модели нужно, во-первых, загрузить файл ‘tmlda’ для BigARTM и словарь, содержащий уникальные слова по выбранным метаданным. После загрузки данных, нужно задать параметры и запустить расчет. В качестве примера можно использовать данные из каталога ‘test_bigartm’. Пример расчета приведен на рисунке 4.2.

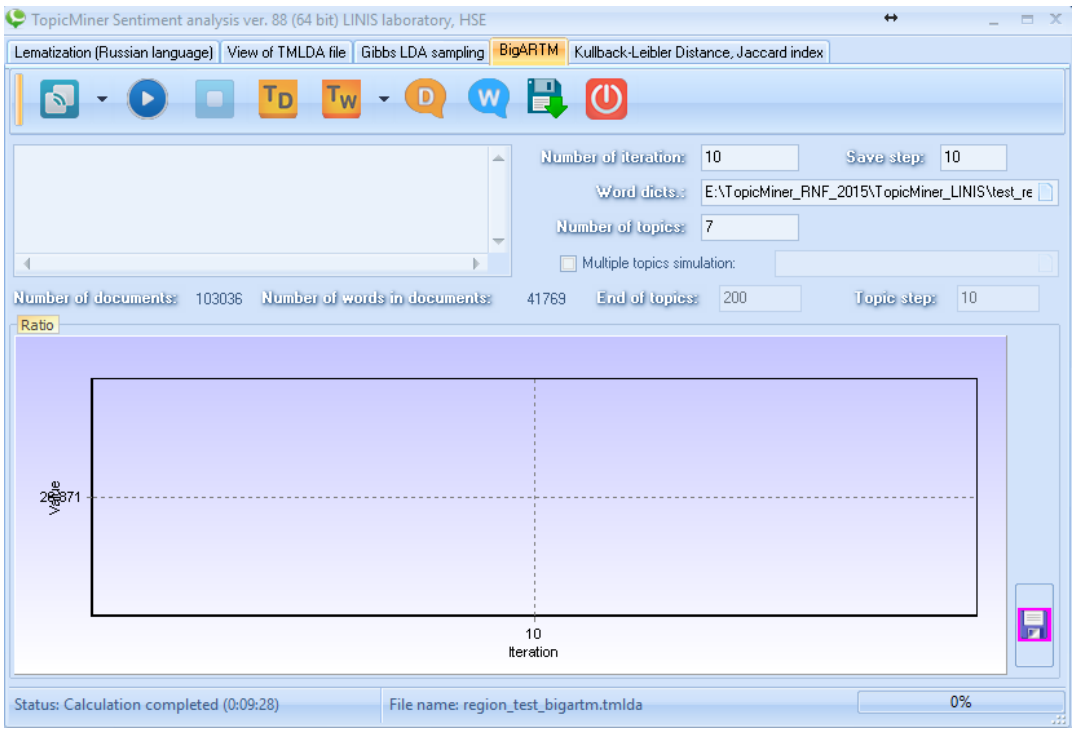
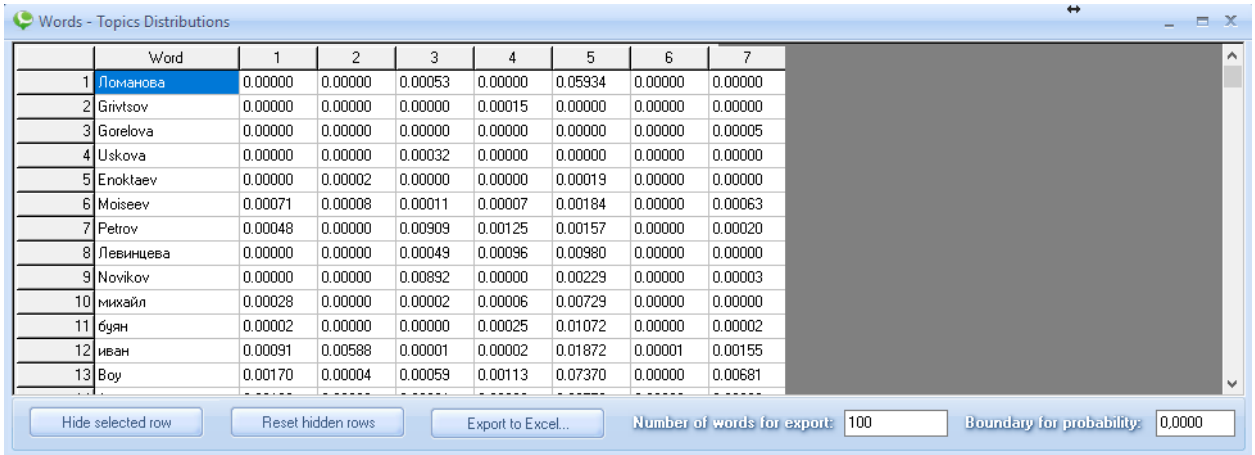


Рис. 4.2. Пример интерфейса ‘BigArtm’.

В результате мультимодального тематического моделирования появятся дополнительные матрицы распределения слов по темам. Данные матрицы представляют собой распределение слов из выбранных полей по темам. В данном примере были задействованы два поля: 1. Поле ‘фамилия автора поста’, 2. Поле ‘геотег автора поста’. На рисунке 4.3 приведен пример матрицы распределения фамилий автора по темам.



	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	Boy	0.00170	0.00004	0.00059	0.00113	0.07370	0.00000	0.00681

Рис. 4.3. Пример визуализации распределения фамилий по темам.

Пример распределения геотегов по темам приведен на рисунке 4.4.

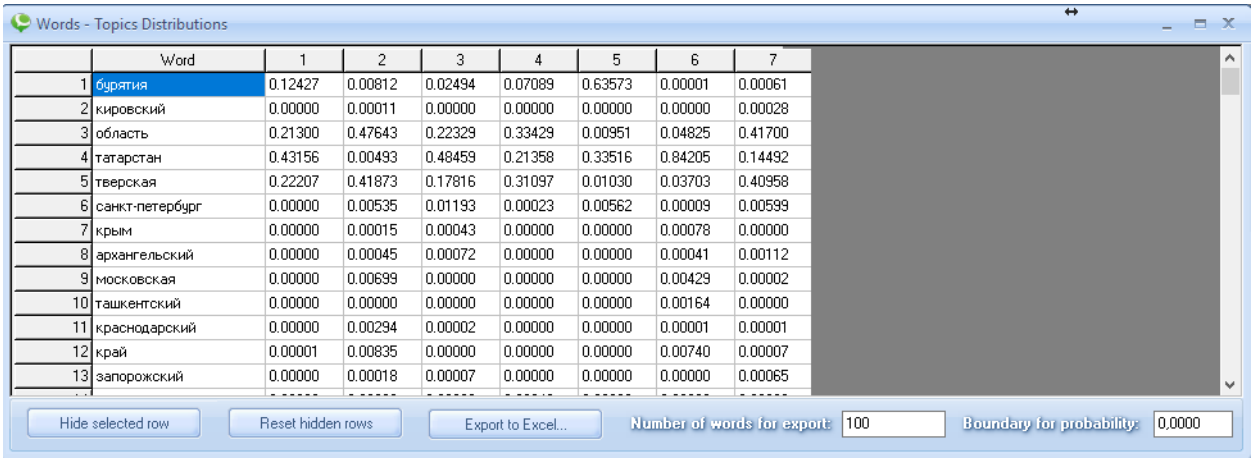


Рис. 4.4. Пример визуализации распределения фамилий по темам.

Полученные данные по разным авторам можно экспортировать в ‘csv’ формате. Для этого нужно использовать кнопку ‘’Export to Excel.

Глава 5. Анализ стабильности результатов моделирования.

При исследовании тематической структуры разными моделями, а также при анализе стабильности тематических моделей, необходимо сравнивать тематические решения между собой. В ПО реализована опция сравнения двух решений на основе двух мер: 1. Мера Кульбака – Лейблера. 2. Мера Жаккара. Общий вид данной опции приведен на рисунке 5.1.

5.1. Загрузка тематических решений.

Для сравнения двух решений их нужно предварительно загрузить. В качестве решения используется выгрузка распределения слов по темам (смотри параграф **‘3.4.2. Визуализация распределений слов по темам’**). Чтобы загрузить первое тематическое решение, нужно нажать на кнопку  . Появившемся окне необходимо указать имя файла. Пример загруженного первого решения приведен на рисунке 5.2.

Результатом загрузки является матрица, в которой в первом столбце находятся коды слов в формате crc32, а во втором - слова. Последующие столбцы содержат вероятности принадлежности слов к темам. Загрузка второго решения осуществляется при помощи кнопки  . Пример загрузки двух решений приведен на рисунке 5.3.

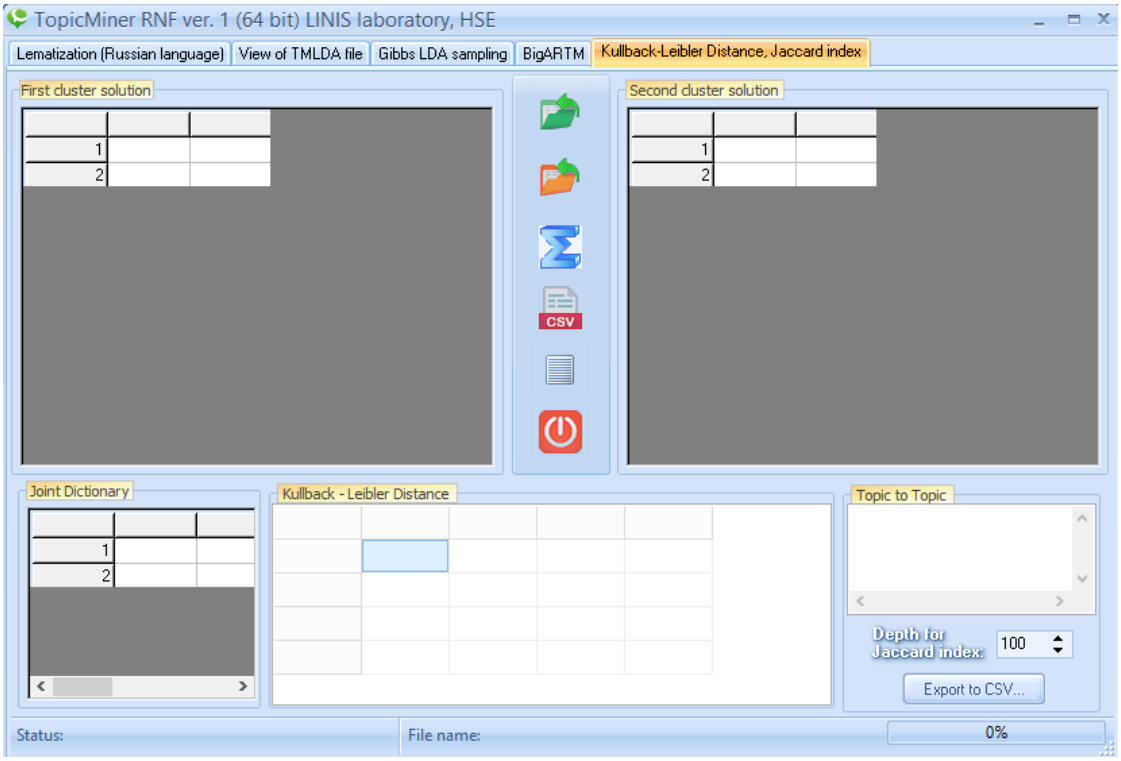


Рис. 5.1. Пример интерфейса для сравнения двух тематических решений.

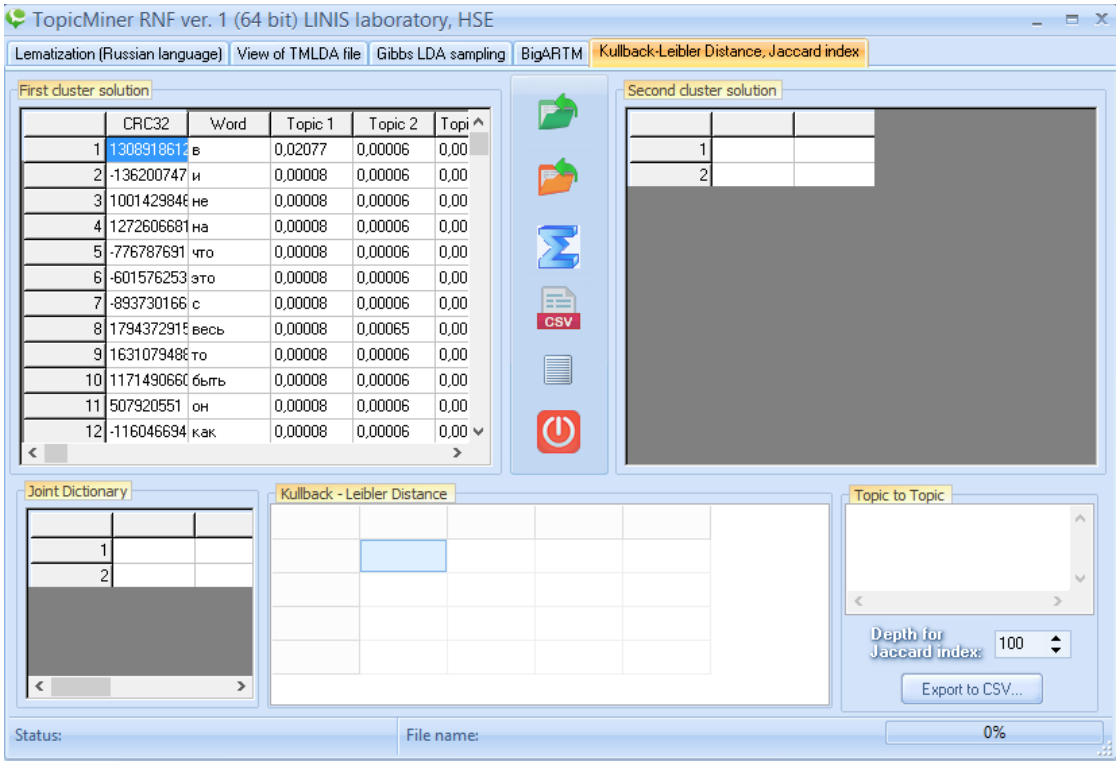


Рис. 5.2. Пример загрузки первого тематического решения.

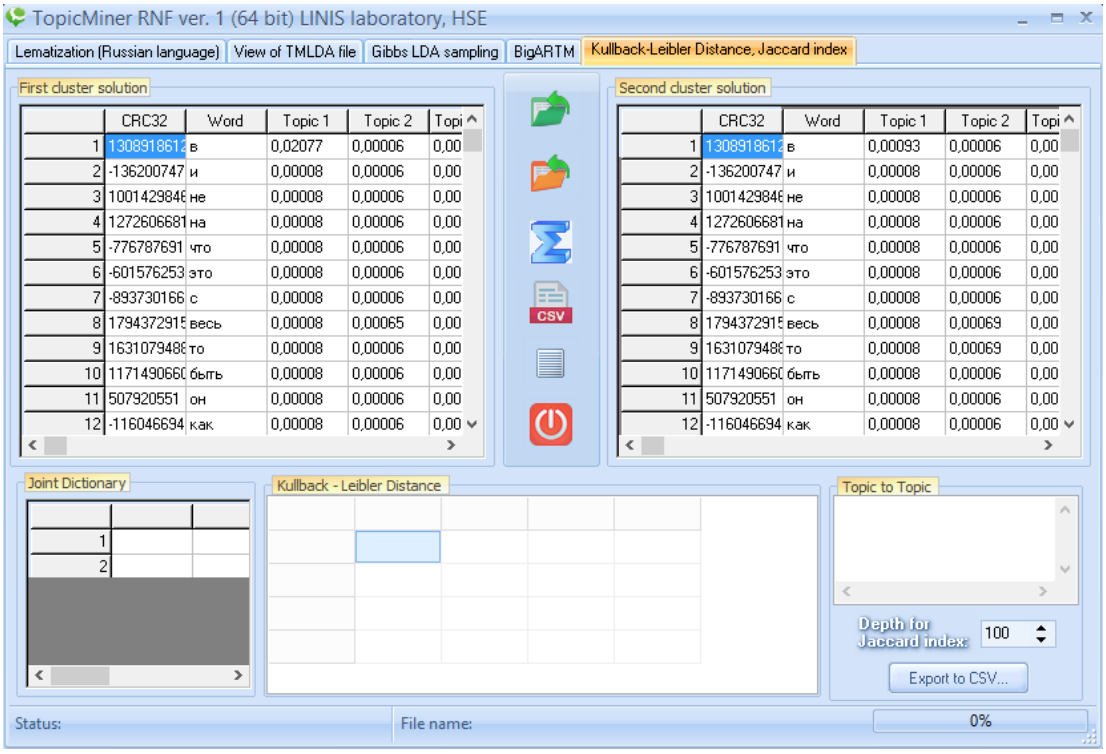


Рис. 5.3. Пример загрузки двух тематических решений.

5.2. Сравнение тематических решений.

Чтобы запустить процедуру попарного сравнения (topic1 vs topic2) двух тематических решений, нужно нажать на кнопку  . После этого запустится процедура сравнения, в которой каждая тема из первого решения будет сравниваться с каждой темой из второго решения. Пример приведен на рисунке 5.4.

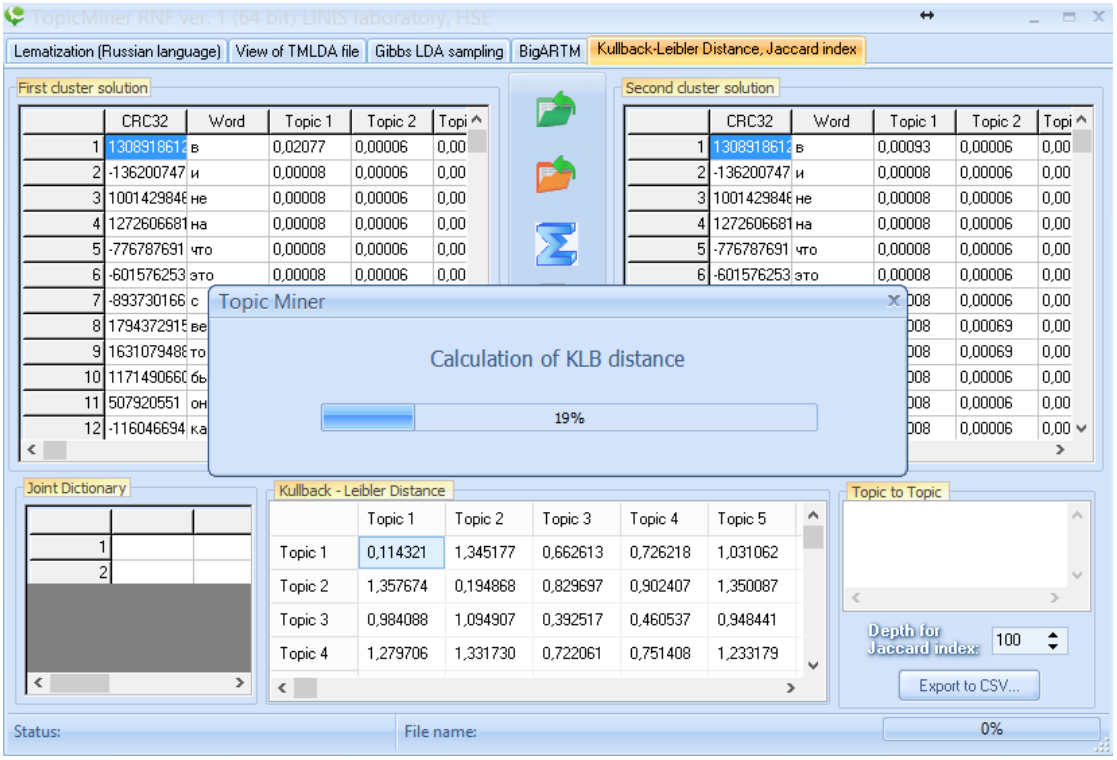
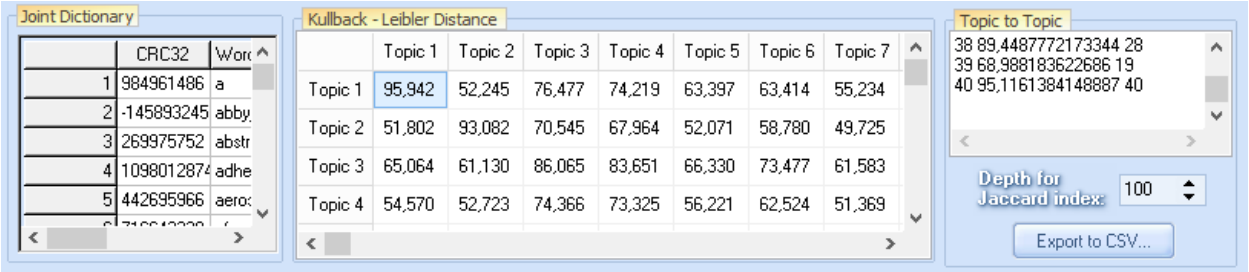


Рис. 5.4. Пример загрузки двух тематических решений.

В результате будут заполнены матрицы ‘Joint Dictionary’, ‘Kullback - Leibler distance’.



‘Joint Dictionary’ – представляет собой список уникальных слов, собранный из двух тематических решений. ‘Kullback - Leibler distance’ – матрица, где в каждой ячейке находится процент сходства между двумя темами. 100% соответствует максимальному сходству.

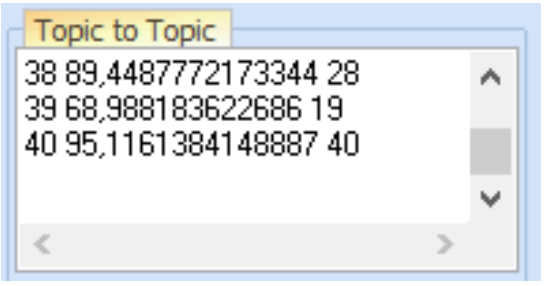
5.2.1. Матрица ‘Kullback - Leibler distance’.

Матрицу ‘Kullback - Leibler distance’ можно выгрузить в формате csv нажатием кнопки  . В появившемся окне нужно указать имя файла. Пример выгрузки показан на рисунке 5.5.

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

Рис. 5.5. Пример выгрузки результатов сравнения по ‘Kullback - Leibler distance’.

Сопоставление максимальных значений ‘Kullback - Leibler distance’ по всем темам выводятся в окне:



В данном примере тема под номером 38 из первого решения похожа на тему 28 из второго решения на 89.44%. Выгрузка результатов сопоставления осуществляется при помощи кнопки  .

5.2.2. Сопоставление тем из разных решений.

Программа может сопоставить (поместить рядом) наиболее похожие темы из двух разных тематических решений, а также рассчитать меру Жаккара. В отличие от ‘Kullback - Leibler distance’, которая считается по всему списку уникальных слов, для меры Жакара нужно указать глубину по словам, то есть количество слов, по которым можно рассчитать меру. Эту глубину можно указать в следующей опции: 

. Типичная величина 100 наиболее вероятностных слов. Чтобы выгрузить таблицу сопоставления похожих тем, в виде совокупности слов нужно нажать на кнопку 

. В появившемся окне нужно указать имя файла. Пример подобной выгрузки показан на рисунке 5.6.

	A	B	C	D	E	F	G	H
1	1 - 95,941	1 - 0,4599	2 - 93,082	2 - 0,4706	3 - 91,222	8 - 0,0363	4 - 88,533	8 - 0,0000
2	в	февраль	кукла	кукла	машин	великобр	северный	великобр
3	февраль	кабул	семья	семья	ji	почувствс	немало	почувствс
4	кабул	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	отчетливі	быстрый	отчетливі
9	демонстр	тысяча	девушка	девушка	интересо	религия	смеяться	религия
10	полицейс	полицейск	родители	религиозі	род	такать	стрит	такать
11	тысяча	демонстра	религиозі	младенец	сильный	источникт	дурак	источникт
12	сша	афганский	младенец	ортодокса	замешате	действо	создавать	действо
13	демонстр	getty	дитя	живой	сантьяго	погон	источник	погон
14	афганский	нато	реборн	реборн	креативні	празднов	интерпре	празднов
15	военный	ap	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	закрыват	объезжат	закрыват
20	оскорбля	marai	также	будущее	диктатор	логотип	подписыв	логотип
21	ap	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	знакомст	отличите.	пингвин	туризм	пингвин

Рис. 5.6. Пример выгрузки результатов сравнения по ‘Kullback - Leibler distance’ и мере Жаккара и выгрузка по словам .

Рассмотрим, что показывает данный пример. В первой паре столбцов, ‘A’ и ‘B’, приведены две темы из двух разных решений, оказавшиеся наиболее похожими. В данном случае это тема №1 из первого решения и тема №1 из второго решения; совпадение их номеров случайно. Это указано в заголовках двух столбцов. В заголовке столбца ‘A’

приведено значение ‘Kullback - Leibler distance’, в данном случае оно 95.941%, а в заголовке столбца ‘B’ приведена мера Жаккара, а данном случае это 0.4599. В ячейках столбцов ‘A’ и ‘B’ приведены наиболее вероятностные слова в этих двух темах. В последующих парах стоблцов, например ‘C’ и ‘D’ приведена следующая пара наиболее сходных тем. Количество пар столбцов равно количеству тем в решении.

Глава 6. Визуализация результатов тематического моделирования на карте Российской федерации.

6.1.Расчет распределений документов по регионам.

Визуализация результатов тематического моделирования реализована при помощи бесплатной картографической системы Quantum GIS (скачать картографическую систему можно по адресу: http://www.qgis.org/ru/site/forusers/download.html).

Внимание: в данном проекте пока не представлены регионы ‘Крым’ и ‘Севастополь’. Эти регионы будут добавлены в следующей версии. Частью этого проекта является файл с расширением dfb, содержащий перечень регионов в картографическом проекте и столбец ‘Topic’, который автоматически заполняется в программе ‘TopicMiner’. Картографический проект находится в каталоге ‘RNF_RF_visualisation’. Главный файл проекта ‘full_project.ggs’.

Прежде чем визуализировать результаты тематического моделирования в картографической системе, нужно рассчитать сумму вероятностей заданной темы по всем регионам. Для этого нужно загрузить в TopicMiner проект со сделанным тематическим моделированием или провести тематическое моделирование. Например, откройте готовый проект из каталога ‘Vk_data_example’.

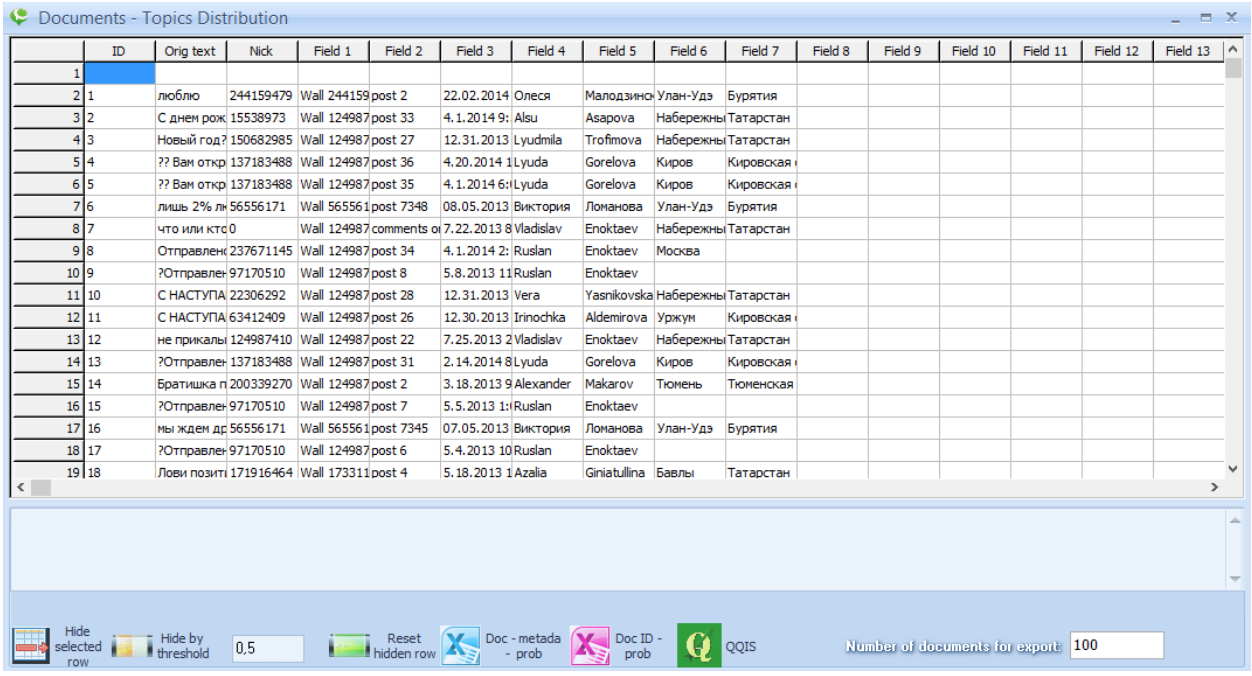


Рис. 6.1. Пример визуализации распределения документов по темам с учетом метаданных.

После загрузки проекта нужно нажать на кнопку  . В появившемся окне (см. рис. 6.1), нужно нажать на кнопку  . В результате появится следующее окно (см. рис. 6.2).

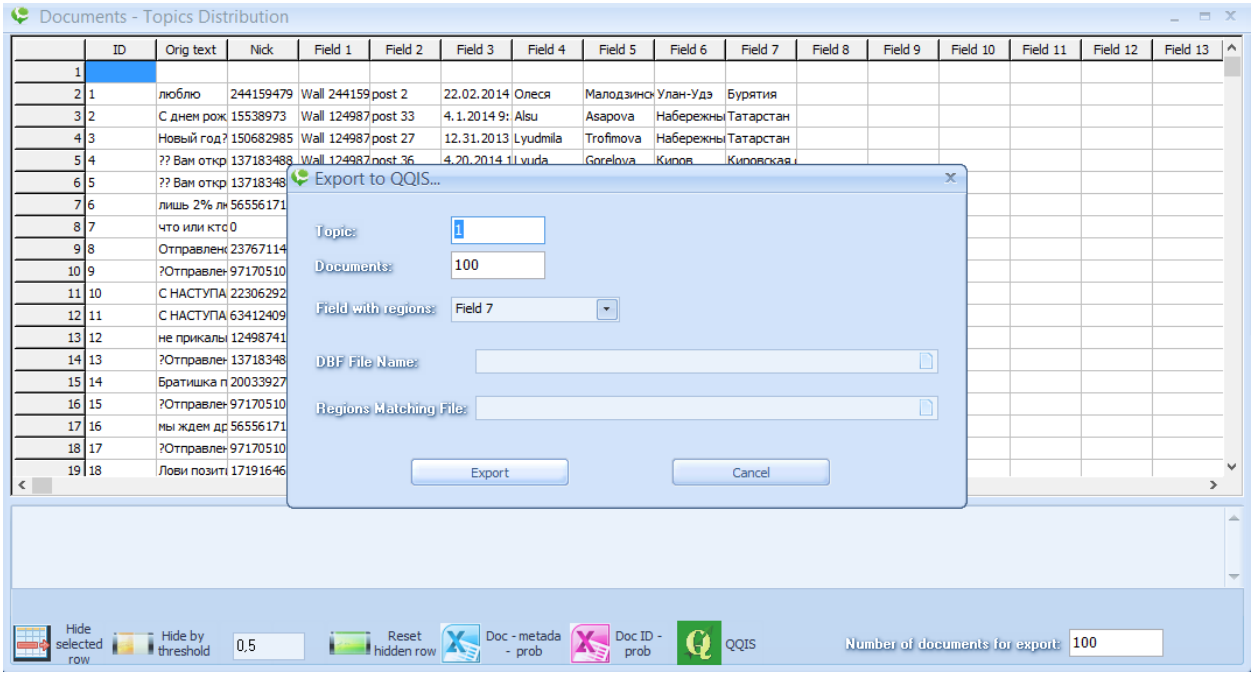


Рис. 6.2. Пример экспорта данных в картографическую систему Quantum GIS.

Для расчета темы по регионам необходимо задать следующие параметры:

1. ‘**Topic**’. Номер темы. Например, задайте номер темы №1, как это показано на рисунке 6.2
2. ‘**Documents**’. Число документов, геотеги которых будут задействованы в расчете. Например, укажите 100 документов, как это показано на рисунке 6.2 Программа выберет 100 наиболее вероятных для заданной темы документов и для каждого региона рассчитает сумму вероятностей всех документов, принадлежащих данному региону. Принадлежность документа региону определяется по геотэгу его автора.
3. ‘**Field with regions**’. В данной опции нужно указать номер столбца, в котором будут находиться наименования регионов. Например, в тестовой коллекции из ВКонтакте наименования регионов находятся в столбце №7 (см. рис. 6.2)
4. ‘DBF file name’. В данной опции необходимо указать имя файла из картографического проекта. Например, файл ‘regions2010_sib_5.dbf’. В данном файле содержатся наименования регионов, выбранных для визуализации, и соответствующие им суммы вероятностей выбранной темы. В этом столбце каждому региону присваивается цвет согласно выраженности выбранной темы в данном регионе. Этим цветом Quantum GIS раскрашивает этот регион на карте Российской Федерации.
5. ‘Regions Matching File’. Поскольку наименования регионов в картографическом проекте и в метаданных из разных социальных сетей могут различаться, необходимо формировать файл, сопоставляющий эти наименования. В данной опции нужно указать имя этого файла. **Внимание: в данной версии мониторинговой системы сформирован файл, в котором наименования регионов из картографического проекта сопоставлены с наименованиями из социальной сети ВКонтакте. Имя данного файла: vktm.dbf.**

После того, как заполнены все поля (см. рис. 6.3), нужно нажать на кнопку ‘Export’.

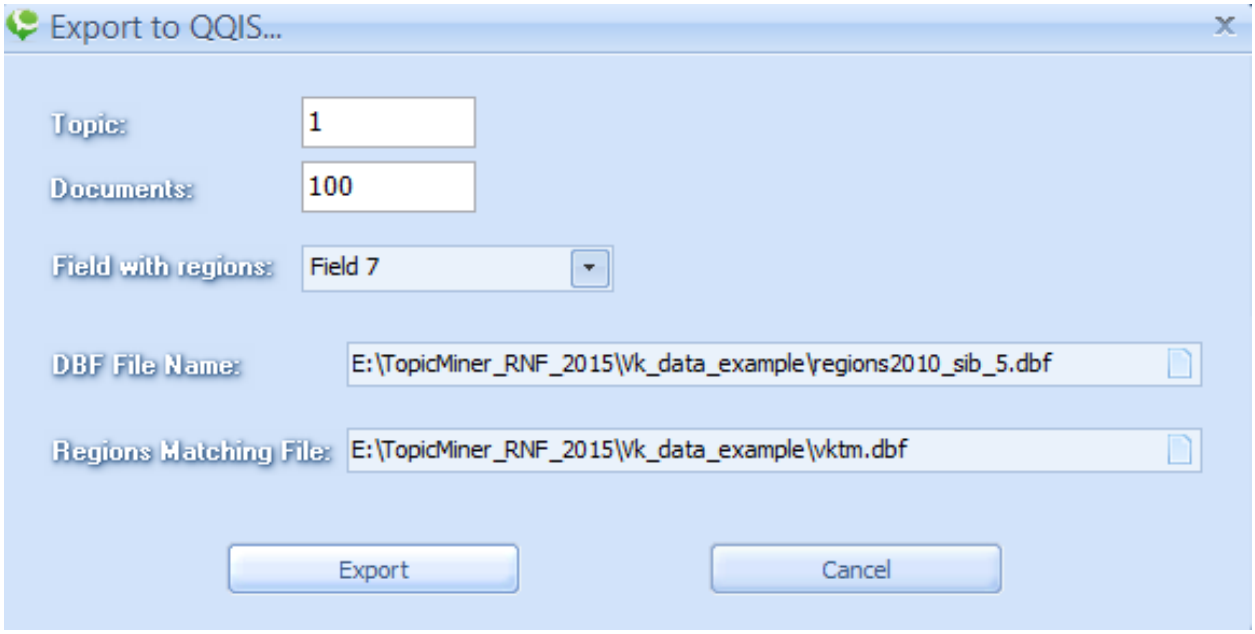
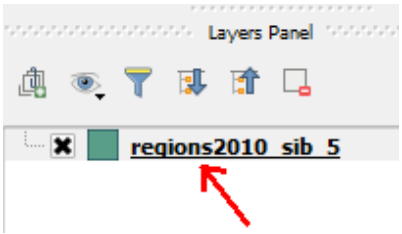


Рис. 6.3. Пример экспорта данных в картографическую систему Quantum GIS.

В ходе расчета программа произведет следующие действия. 1. Определит список регионов, которые присутствуют в заданном количестве отсортированных документов (на основании файла сопоставления). 2. Рассчитает сумму вероятностей документов по каждому региону. 3. Сохранит рассчитанные суммы вероятностей в файл ‘regions2010_sib_5.dbf’ (см. рис. 6.3).

6.2.Визуализация распределения документов в Quantum GIS.

Готовый проект с набором карт регионов Российской Федерации находится в каталоге ‘RNF_RF_visualisation’. Чтобы визуализировать полученные данные, нужно скопировать файл ‘regions2010_sib_5.dbf’ в этот каталог, то есть заменить старый файл с таким же именем на новый файл. После этого нужно кликнуть (дважды) на файле ‘full_project.qgs’. Внимание: картографическая система ‘Quantum GIS’ уже должна быть установлена. В результате запустится ‘Quantum GIS’ и загрузится проект ‘full_project.qgs’ (см. рис. 6.4). Сначала все регионы будут выделены одним цветом. Чтобы раскрасить регионы в цвета в соответствии с суммой вероятностей, нужно изменить стиль рисования. Для изменения стиля нужно дважды кликнуть на наименовании проекта, как показано красной стрелкой на рисунке ниже:



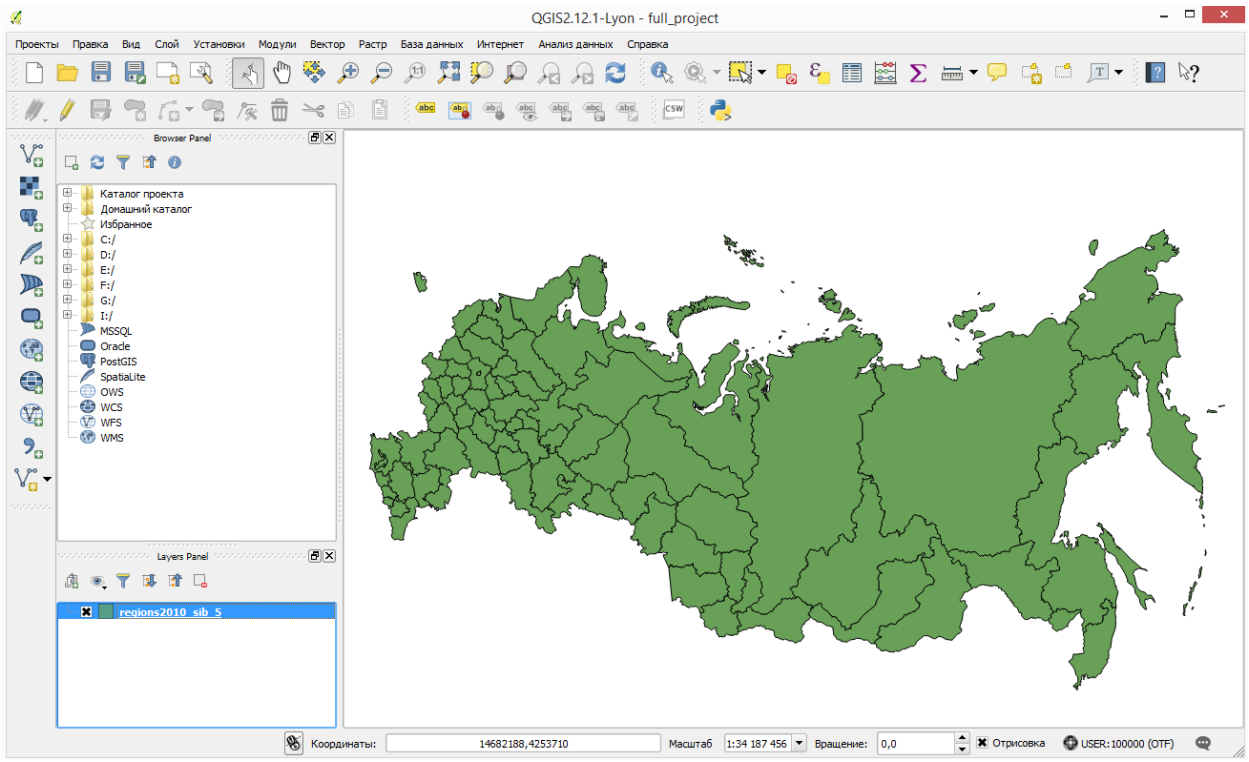


Рис. 6.4. Пример экспорта данных в картографическую систему Quantum GIS.

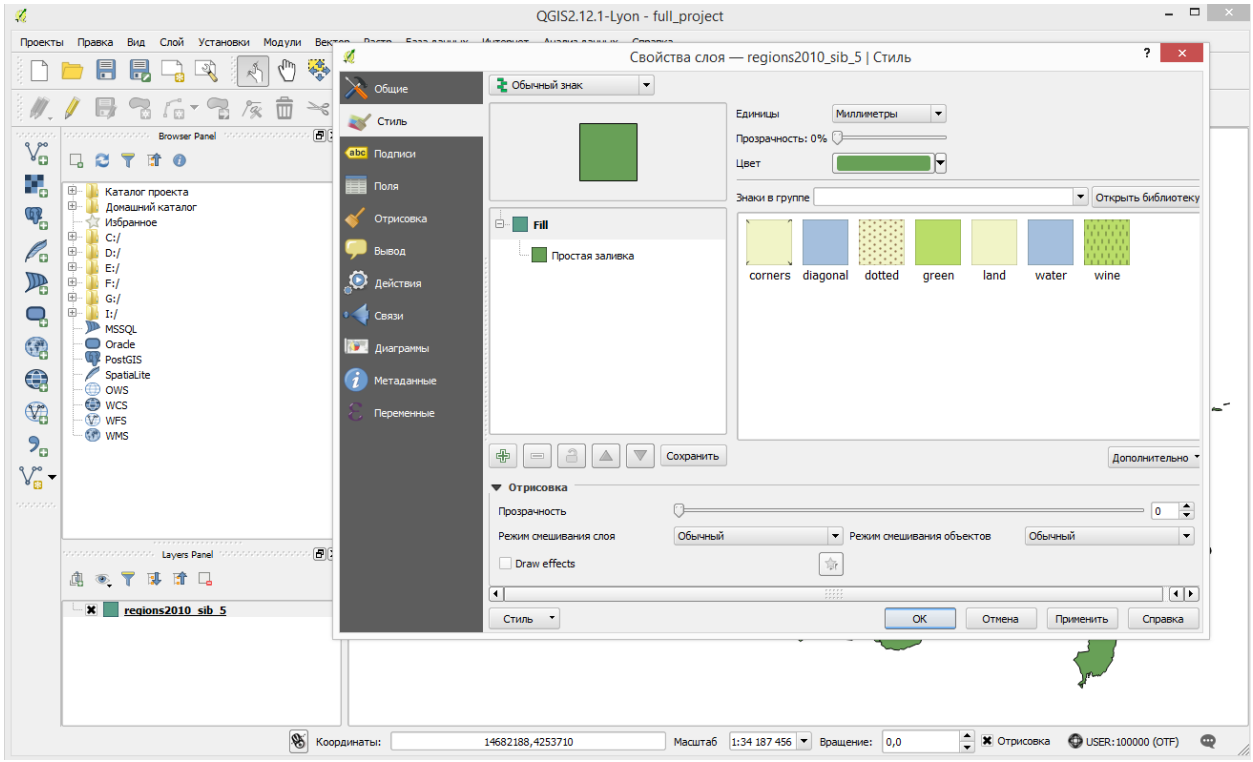


Рис. 6.5. Пример изменения стиля в Quantum GIS.

В результате откроется окно, в котором можно поменять стиль рисования (см. рис. 6.5). Для этого в выпадающем меню, где по умолчанию стоит «Обычный знак», следует выбрать «Уникальные значения», как это показано на рисунке 6.6.

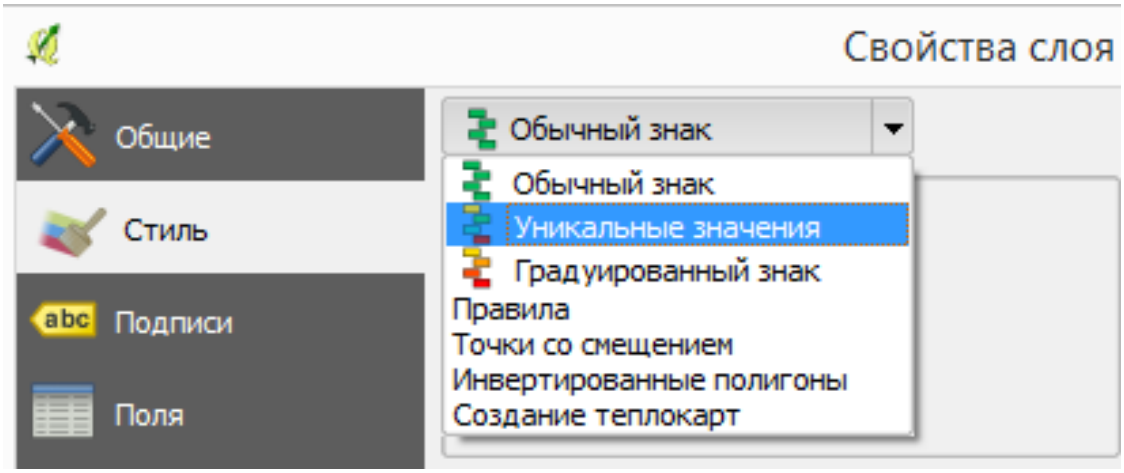


Рис. 6.6. Пример изменения стиля в Quantum GIS.

Затем выбрать поле, по которому нужно рассчитать уникальны значения. В нашем случае это поле ‘Topic’, которое содержит суммы вероятностей по каждому региону. Эти данные берутся из файла ‘regions2010_sib_5.dbf’. После этого нужно нажать на кнопку ‘классифицировать’ (смотри рис. 6.7).

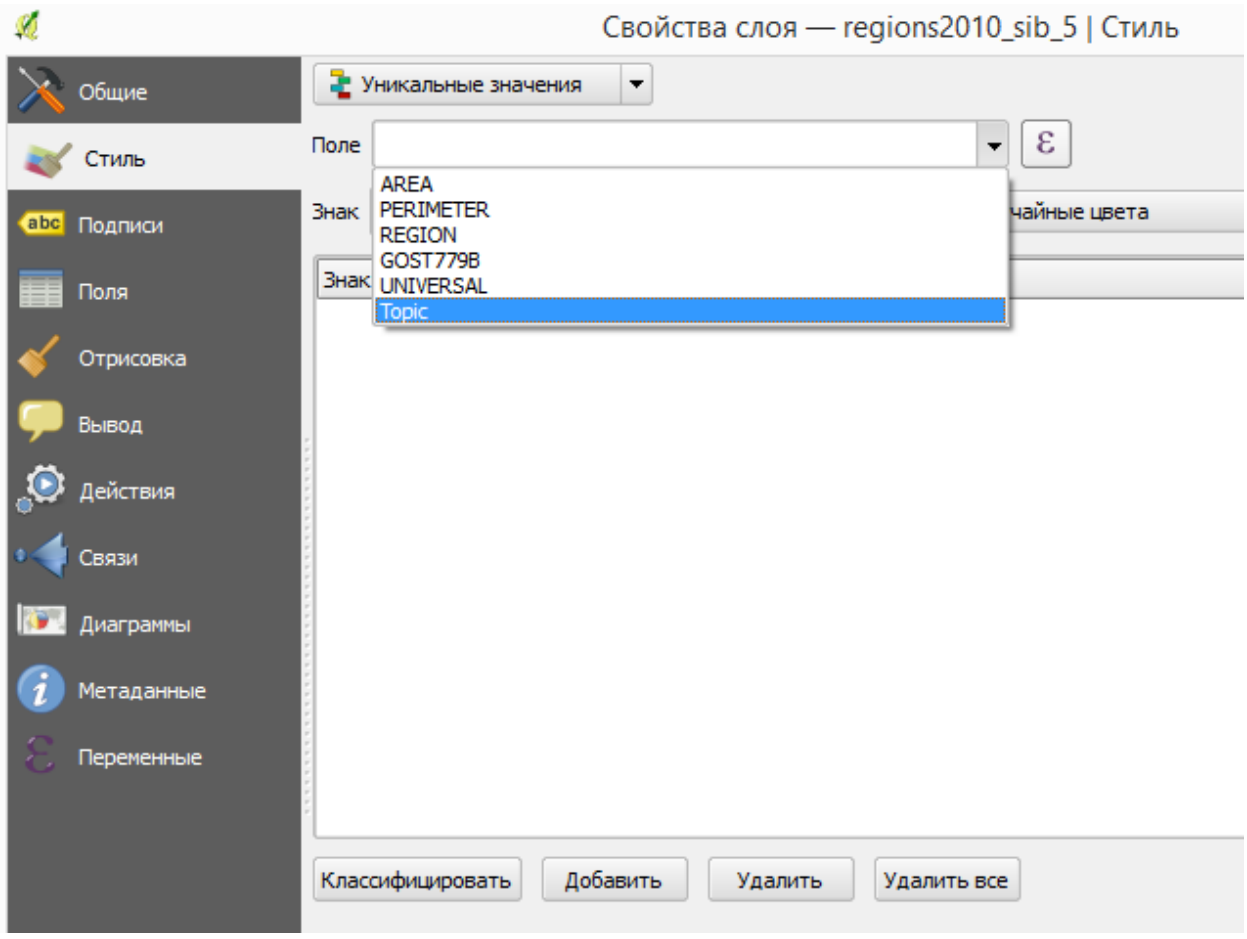


Рис. 6.7. Пример изменения стиля в Quantum GIS.

В результате классификации Quantum GIS определит все уникальные значения (пример см. на рис. 6.8). Теперь нужно указать тип раскраски для найденных значений. Это можно сделать в опции ‘Градиент’ (см. рис. 6.9).

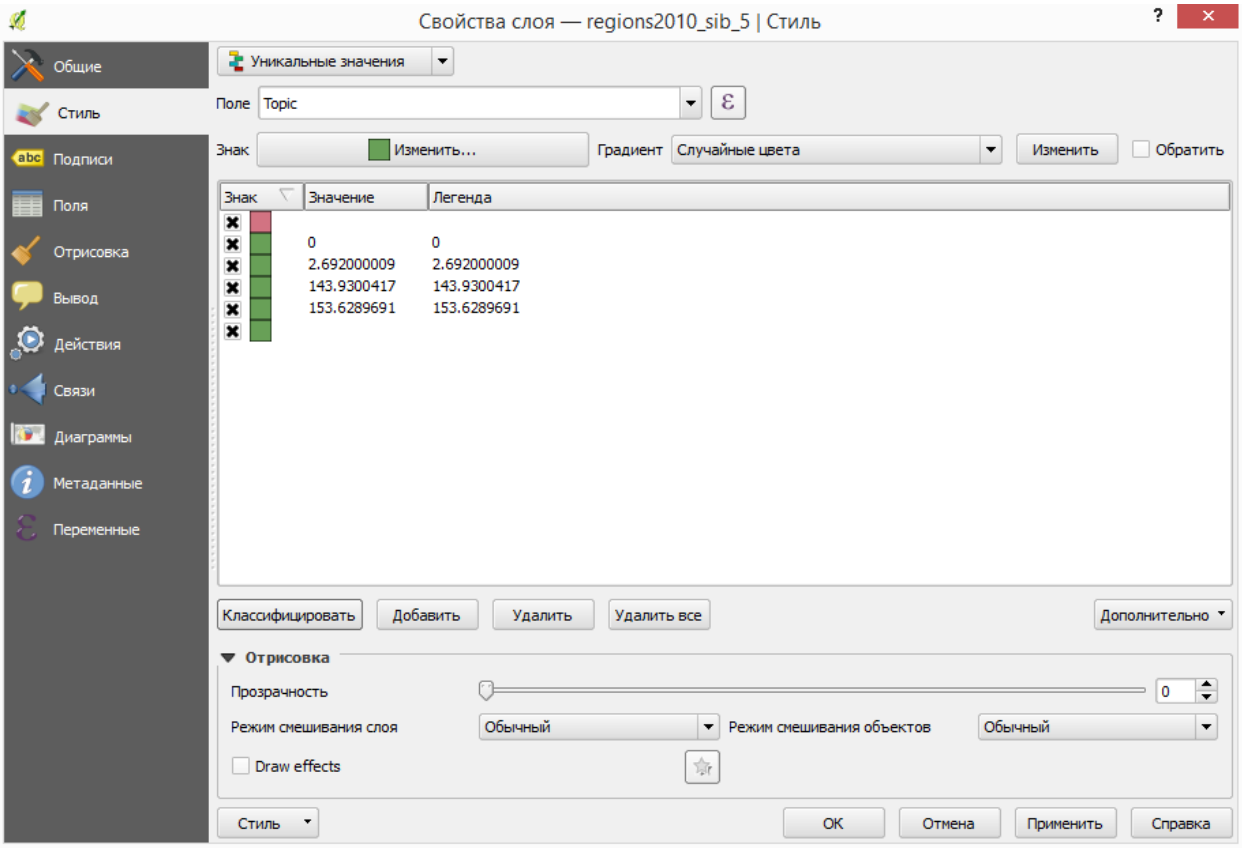


Рис. 6.8. Пример изменения стиля в Quantum GIS.

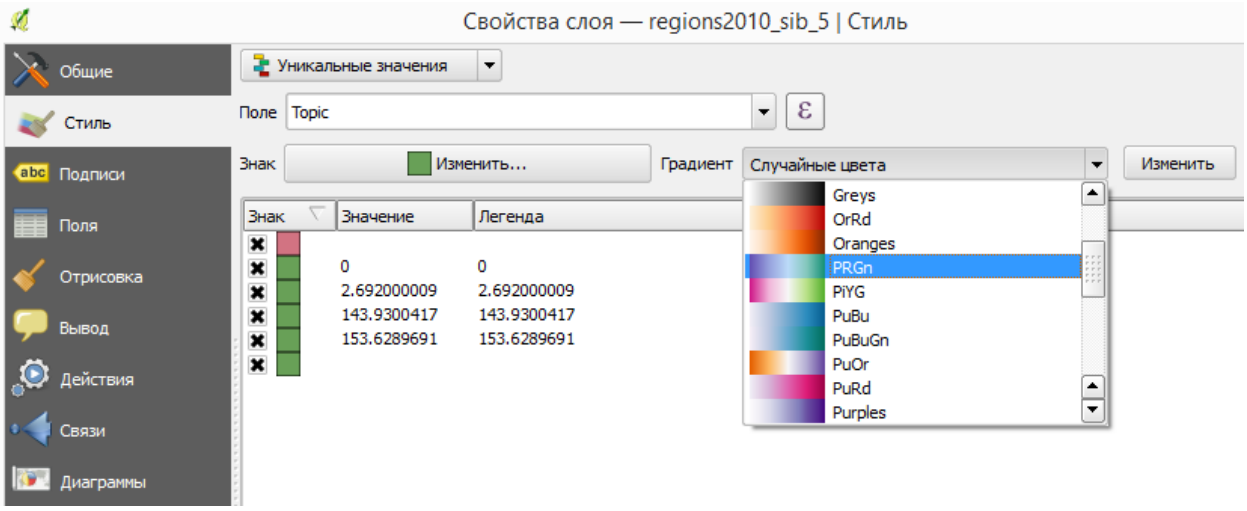


Рис. 6.9. Пример изменения стиля в Quantum GIS.

Чтобы применить цветовую гамму, нужно нажать на кнопку ‘Применить’. Результат по трем найденным уникальным значениям (то есть по трем найденным регионам) показан на рисунке 6.10. Внимание: цветом выделяются только те регионы, для которых в данных нашлись документы, имеющие высокие вероятности по выбранной теме. На рисунке 6.11 приведен пример визуализации темы по регионам на основе 222546 документов в теме №1.

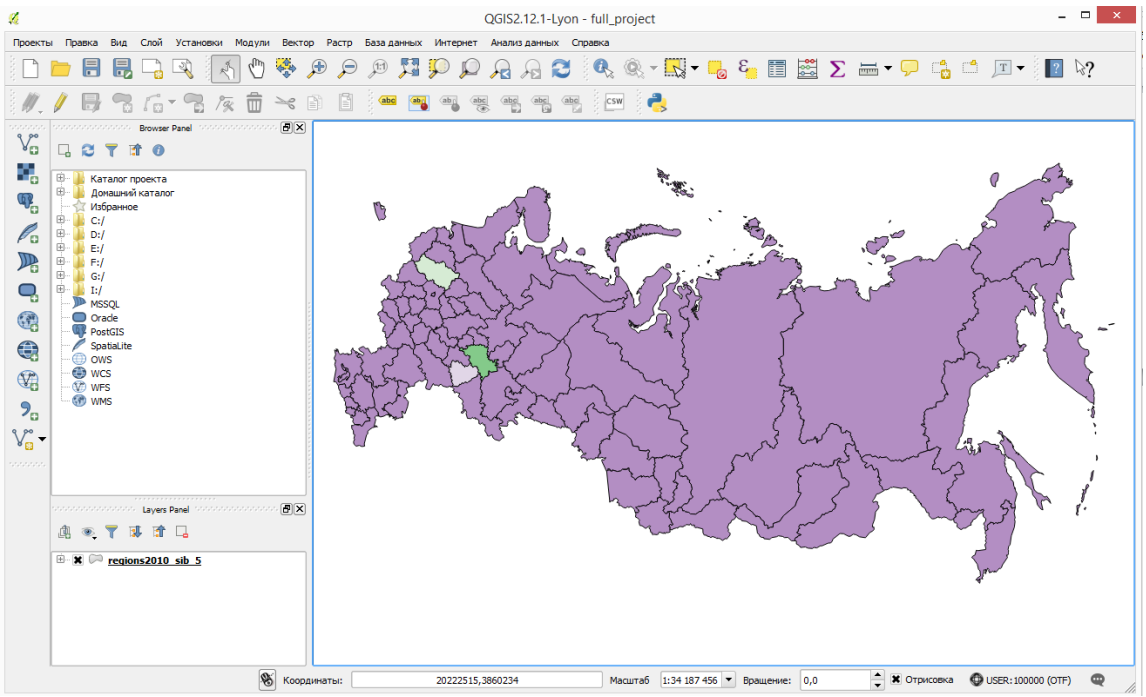


Рис. 6.10. Пример визуализации темы в Quantum GIS по трем регионам.

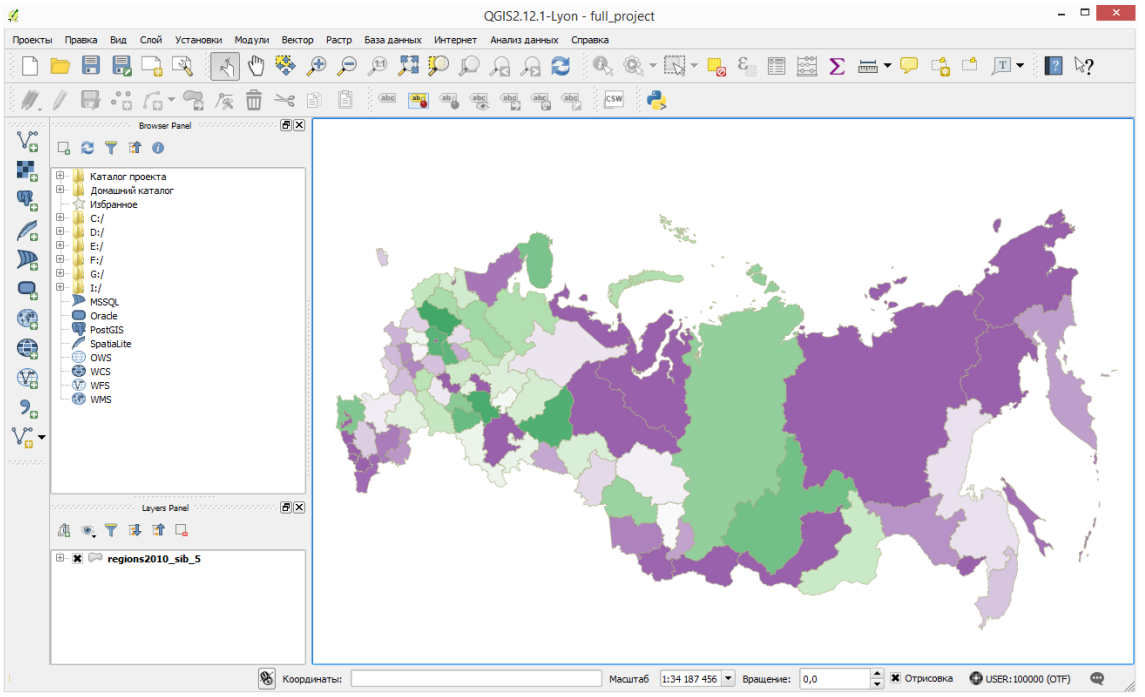


Рис. 6.11. Пример визуализации темы в Quantum GIS по множеству регионов.

Глава 7. Анализ тональности текстов.

7.1. Введение.

Тональный анализ (Сентимент-анализ) или автоматизированный анализ эмоциональной окрашенности текстов (плохо / хорошо, нравится / не нравится и др.) можно отнести к области компьютерной лингвистики, однако, задачи его применения, в основном, лежат за пределами собственно лингвистики. Их можно разделить на две обширные области: маркетинг (в первую очередь — как анализ отзывов на товары и услуги) и социология / политология. Последняя включает, во-первых, анализ текстов СМИ для выявления того, как те или иные социально значимые вопросы преподносятся аудитории и, соответственно, какого отклика можно ожидать на них от публики. Во-вторых, это исследование текстов блогов, социальных сетей, форумов и другого пользовательского контента с целью выявления общественного мнения – точнее, мнения интернет-активной части населения. В рамках данного программного обеспечения реализован подход на основе словаря. В качестве начального словаря использован набор слов, полученный в рамках РГНФ проекта ‘Разработка общедоступной базы данных и краудсорсингового веб-ресурса для создания инструментов сентимент-анализа’, Номер заявки: 14-04-12031. Однако в данном ПО реализована общая технология, подключения любого словаря, который включает в себя этничности, этнофолизмы и так далее.

7.2. Подготовка словаря для сентимент анализа.

В силу того, что результатом тематического моделирования является матрица распределения лематизированных слов по темам, соответственно сентимент оценка должна проводится на основе лематизированного словаря. Подготовка словаря производится следующим образом. На вкладке ‘Lematization’ реализована следующая опция (смотри рис. 7.1)

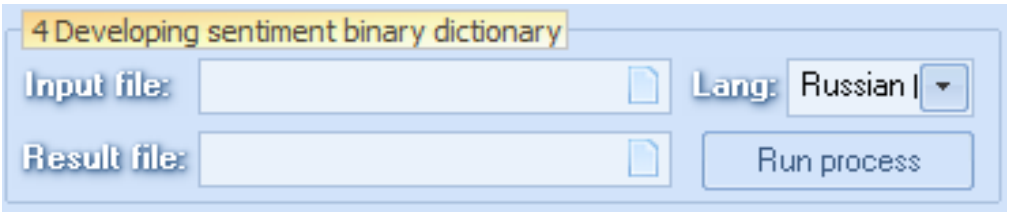


Рис. 7.1. Опция подготовки тонального словаря.

В качестве входного файла должен быть использован файл следующего типа:

Words; average rate

алкаш;-2

апатичный;-2

бедственный;-2

бездарность;-2

безжалостно;-2

безмозглый;-2

безнаказанный;-2

безобразный;-2

безответный;-2

Пример такого словаря входит в состав текущей версии. Для списка слов необходимо указать язык (Русский, Английский) и тип кодировки (UTF, Ansi).

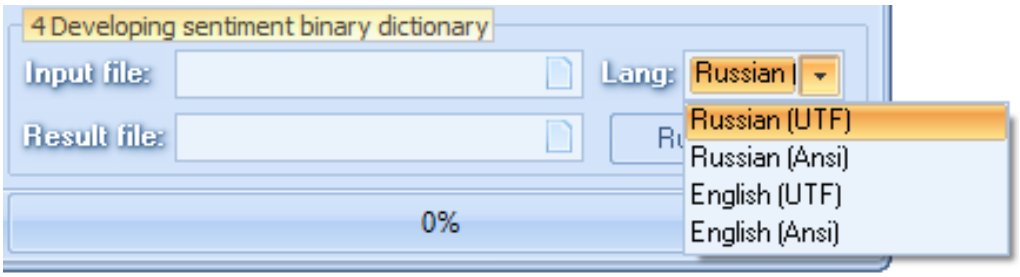
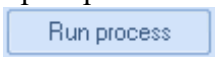


Рис. 7.2. Опция подготовки тонального словаря.

Варианты выбора языка и кодировки приведены на рис. 7.2.

Результатом препроцессинга является бинарный файл, содержащий слова и оценки в бинарном формате. Хранение словаря в бинарном виде позволяет существенно ускорить расчет оценок тональности распределения слов и документов по темам. Для того преобразовать словарь из текстового в бинарный формат нужно нажать на кнопку 

.

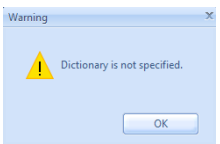
7.2. Подключение словаря к тематической модели.

Расчет тональных оценок производится для уже готового тематического решения. Опция подключения находится на вкладке ‘Gibbs LDA sampling’. Подключение заключается в указании пути и имени словаря в бинарном формате, смотри следующий рисунок:

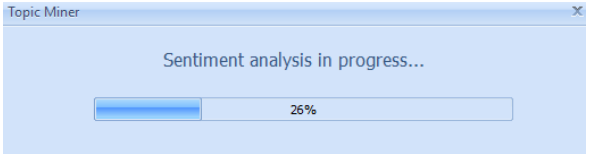
7.3. Тональный расчет распределения слов по темам.

Расчет тональных оценок производиться для готового тематического решения. Это значит, что либо тематическое моделирование должно быть сделано, либо в программу должен быть загружен проект, сделанный ранее. Расчет тональных оценок для распределения слов по темам реализован в окне ‘words with high probability’. Для того что бы данное окно появилось нужно нажать на кнопку 
. В результате появится следующее окно (смотри рис. 7.3). В данном окне визуализируется матрица распределения слов по темам. На данный момент, в каждой ячейке находится слово и вероятность данного слова в теме. Для того, что бы добавить тональные оценки нужно нажать на кнопку 

. В результате запустится расчет тональности. Однако, в случае не подключенного словаря, тогда появится окно предупреждения (смотри ниже):



В случае, когда словарь подключен, тогда показывается процесс расчета (смотри ниже).



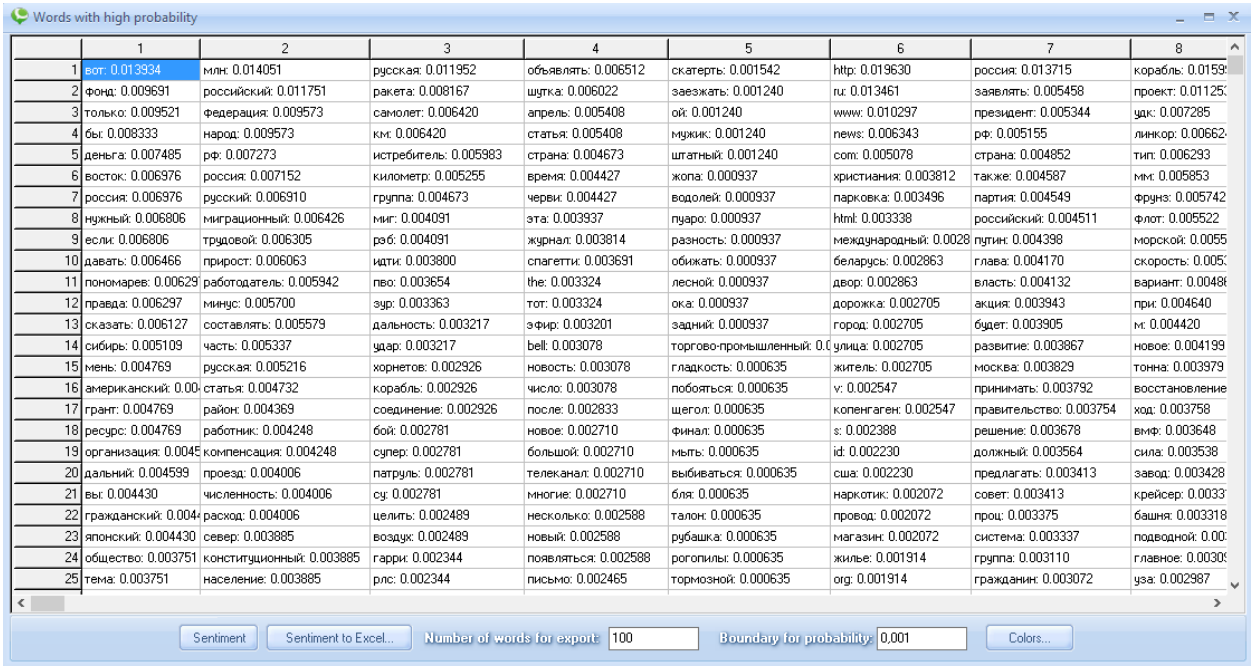


Рис. 7.3. Матрица распределения слов по темам до расчета тональности.

В результате расчета, в каждой ячейке, помимо вероятности, появится целое число, которое характеризует тональность слова (смотри рис. 7.4).

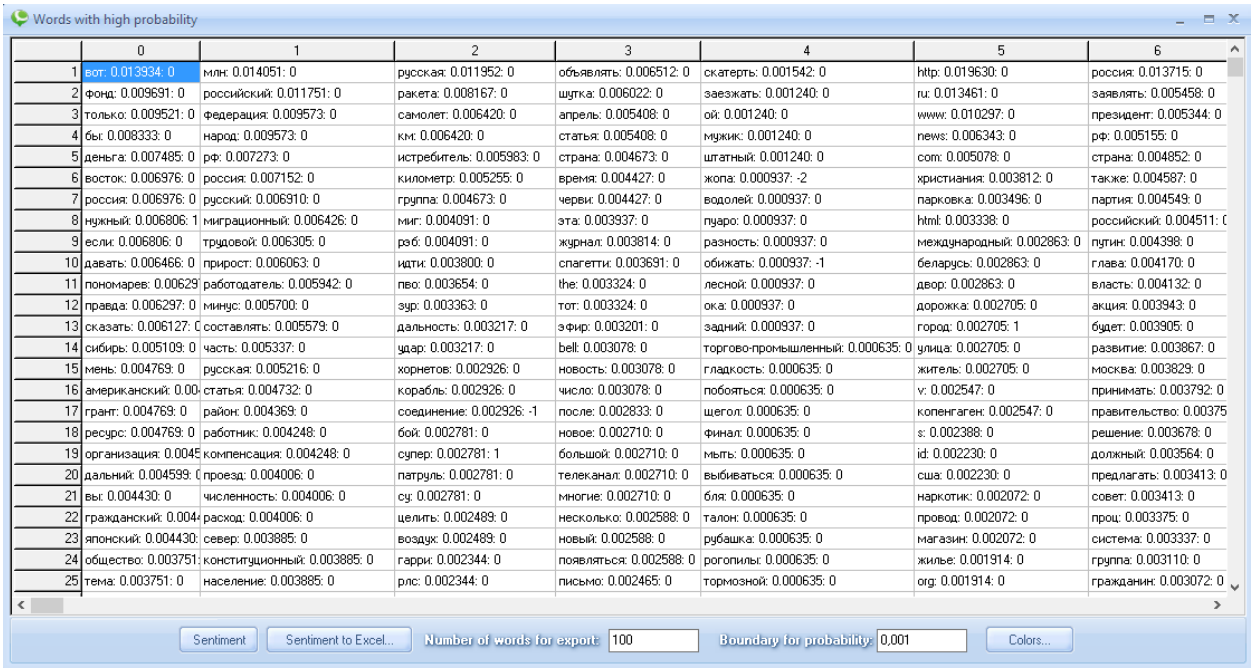
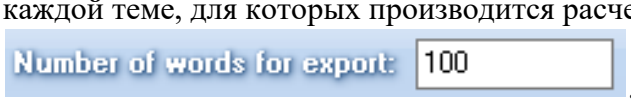


Рис. 7.4. Матрица распределения слов по темам и сентимент анализ.

Внимание, расчет тональности производится для фиксированного числа наиболее вероятностных слов. Число слов в каждой теме, для которых производится расчет определяется в следующей опции: 

. По умолчанию стоит 100 слов.

7.3.1. Выгрузка матрицы слова - темы с тональными оценками.

Выгрузка матрицы распределений слов по темам вместе с тональными оценками производится при нажатии на кнопку 

(нужно указать имя файла). В результате выгружается все темы. Глубина выгрузки по словам определяется параметрами 

и 

. В результате выгрузки появится файл с заданным именем и содержащий помимо вероятностей, также тональные оценки слов.

7.3.2. Подсказка тем.

При реальных расчетах, количество тем может составлять от нескольких десятков до нескольких сотен. Поиск нужных тем может занимать продолжительное время. С целью оптимизации работы пользователя информационной системы реализовано цветовая подсветка нужных тем. Подсветка тем реализована следующим образом. Пользователь должен нажать на кнопку 

. При этом пользователь указывает файл. Который содержит список слов, по которым нужно искать рассчитанные темы. После этого, программа читает файл, и рассчитывает сколько слов из списка присутствует в среди топовых слов в каждой теме. Все числа делятся на 4 категории, темы с максимальными числами (максимальное число слов из списка) раскрашивается красным цветом. Темы с минимальными числами подкрашены зеленым цветом. Пример такой подсказки приведен на рисунке 7.5.

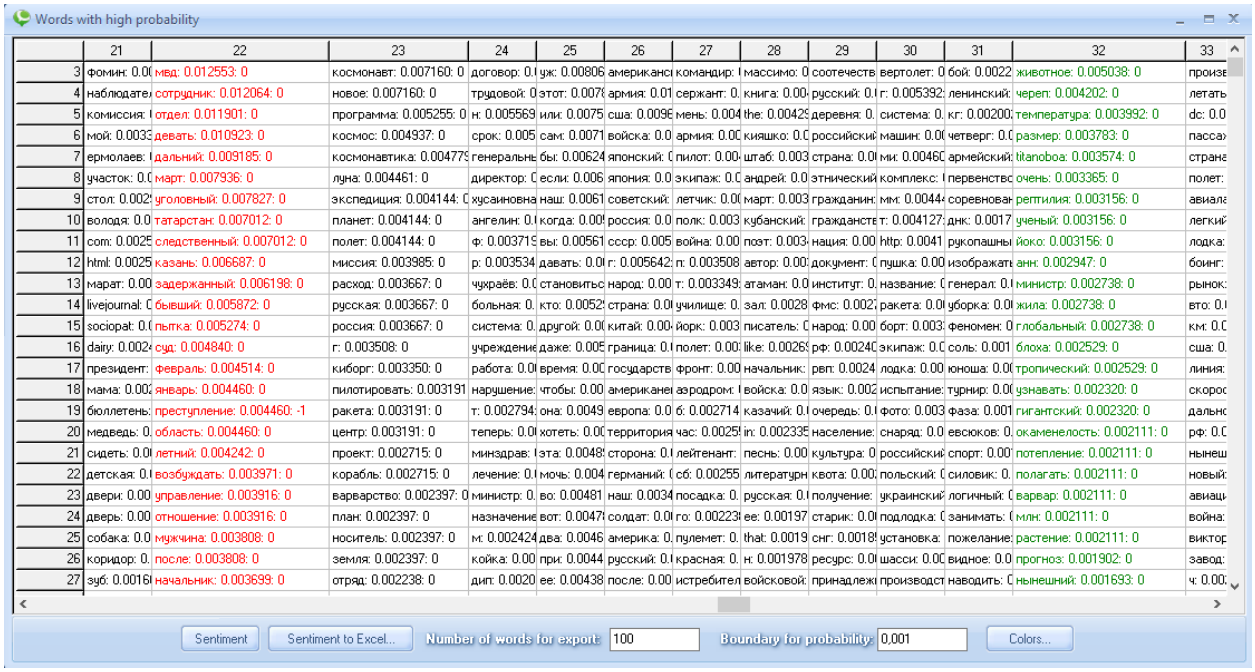


Рис. 7.5. Пример цветовой подсказки в матрице распределения слов по темам.

Как показала практика, такая цветовая дифференциация чрезвычайно удобна.

7.4. Тональный расчет распределения документов по темам.

Расчет тональных оценок для документов производиться для готового тематического решения.

Для того, что бы перейти к сентимент расчету документов нужно нажать на кнопку 
. В результате появится следующее окно (смотри рис. 7.6)

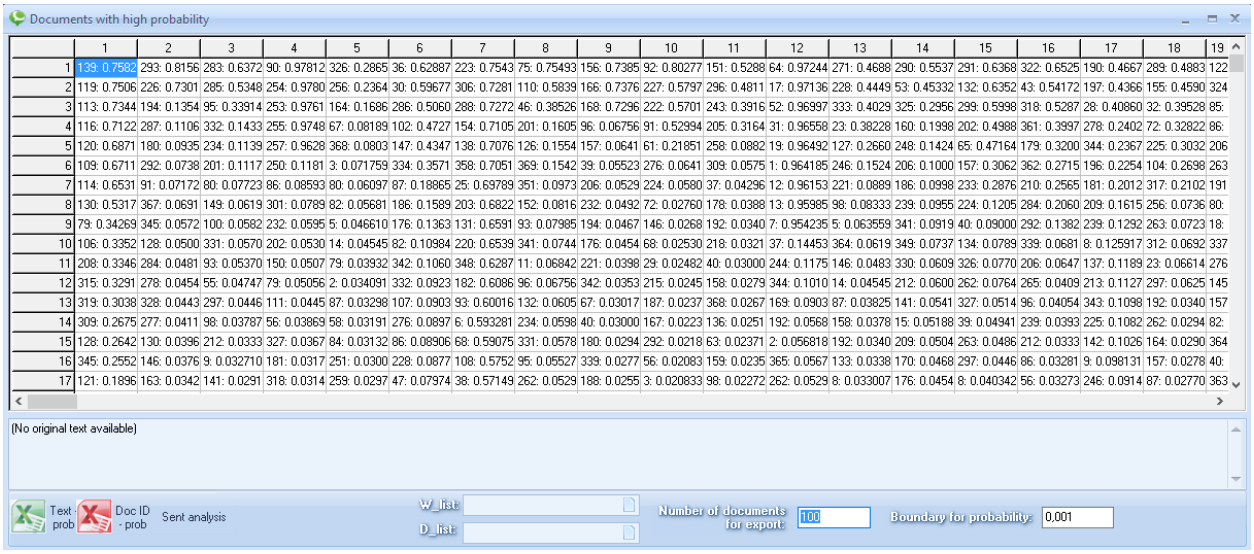
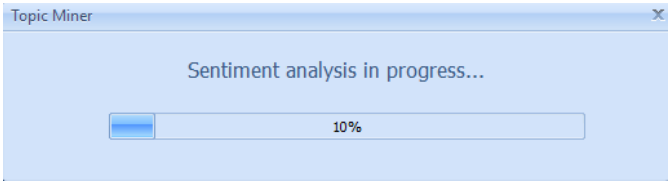


Рис. 7.6. Матрица распределения документов по темам до расчета тональности.

Для того что бы рассчитать тональность документов по темам нужно нажать на кнопку Sent analysis. В результате (если подключен словарь) появится окно (смотри ниже), в котором отражается процесс расчета.



В результате расчета, в каждую ячейку, содержащие номер документа, вероятность документа, добавляется сентимент оценка документа (смотри рис. 7.7).

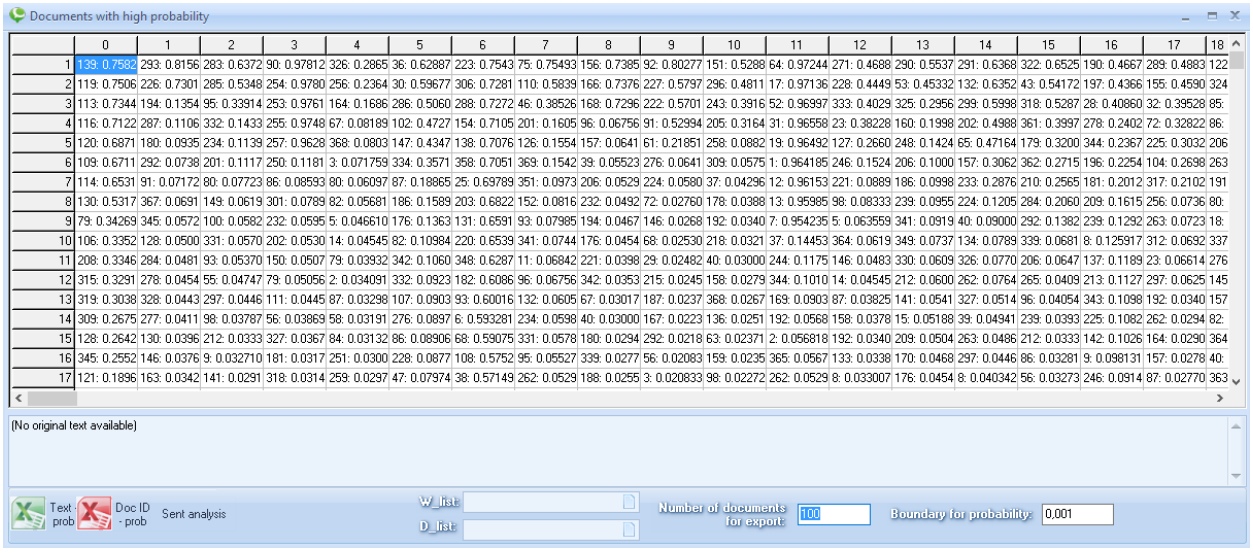
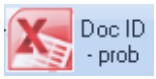


Рис. 7.7. Матрица распределения документов по темам после расчета тональности.

7.4.1. Выгрузка матрицы документы - темы с тональными оценками.

Выгрузка матрицы, в данной версии, реализованы в виде двух функций (две кнопки). Первая выгрузка реализована в виде выгрузки текстов + вероятности + сентимент оценки.

Для того что бы получить такую выгрузку нужно нажать на кнопку 

. В появившимся окне необходимо указать имя файла. Данные сохранятся в формате csv. Вторая выгрузка реализована в виде комбинации ID документов + вероятности + сентимент оценки. Для того что бы выгрузить матрицу в данном виде нужно нажать на кнопку 

. В появившемся окне необходимо указать имя файла. Данные будут сохранены в формате csv.

Число выгруженных документов определяется двумя параметрами: 1. Number of documents for export (как показано на рисунке ниже):

2. Boundary for probability (как показано на рисунке ниже).

В результате выгружаются только те документы (по всем темам), которые удовлетворяют выше указанным условиям.

7.5. Тональный расчет распределения документов по темам для BigArtm.

Расчет тональности модели, рассчитанные в рамках подхода BGARTM, можно произвести следующим образом. Модель рассчитанная в опции BIGARTM нужно сохранить в виде проекта. Затем открыть данный проект на вкладке ‘Gibbs sampling’. Далее провести расчет тональности как это описано в выше. Такой подход обусловлен тем, что формате данных, расчитанных по модели BigArtm и Gibbs sampling идентичны.

Глава 8. Временные тренды в тематических моделях.

В версии 2017 года реализована возможность построения графиков временных трендов. Возможность построения реализована на основе двух вещей. Во-первых, документы должны обладать временными метками. Во-вторых, строить временные графики можно как основе меток у отсортированных документов в заданной теме, так и в мультимодальных тематических моделях, где формируется дополнительная матрица распределения дат по темам.

8.1. Унификация временных дат.

В силу того, что в коллекциях у документов могут встречаться различные форматы представления дат, то в модуле ‘view tmlda’ реализована опция унификация дат. Для того что бы запустить процесс унификации нужно открыть вкладку ‘View of tmlda’, загрузить данные, содержащие временные метки (смотри рис. 8.1).

Далее, нужно открыть опцию‘Data/Time repear’ (смотри рис. 8.2), указать номер поля, который содержит временные метки. В указанном примере это поле №2. Затем нужно нажать на кнопку ‘Save as TMLDA’ и указать имя нового файла.

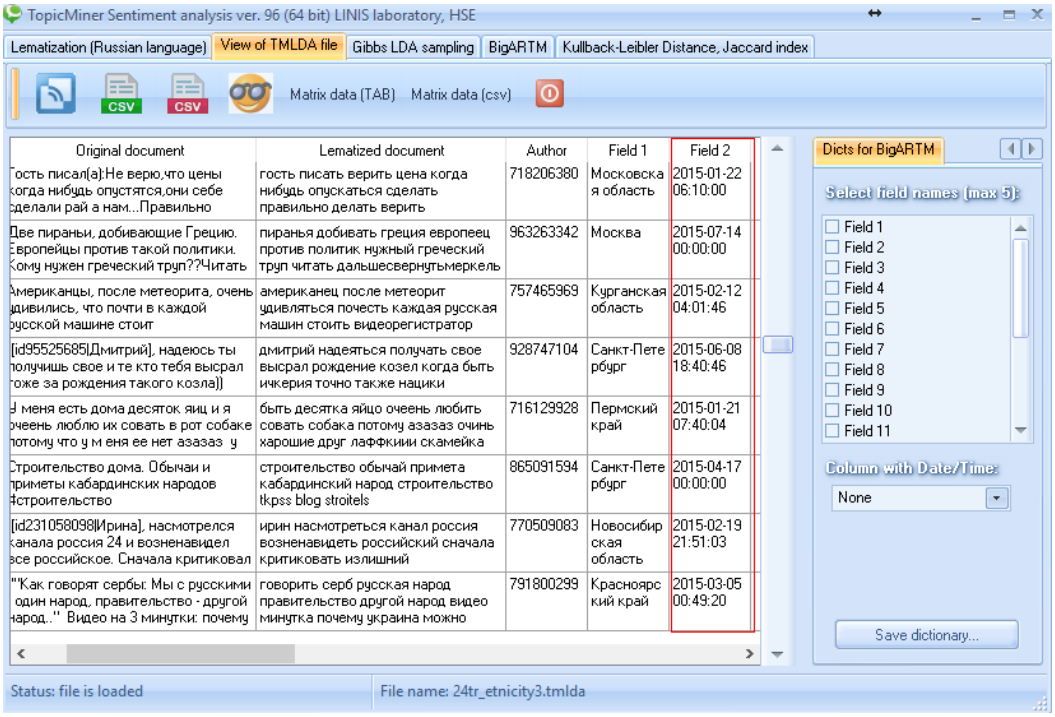


Рис. 8.1. Пример коллекции с датами

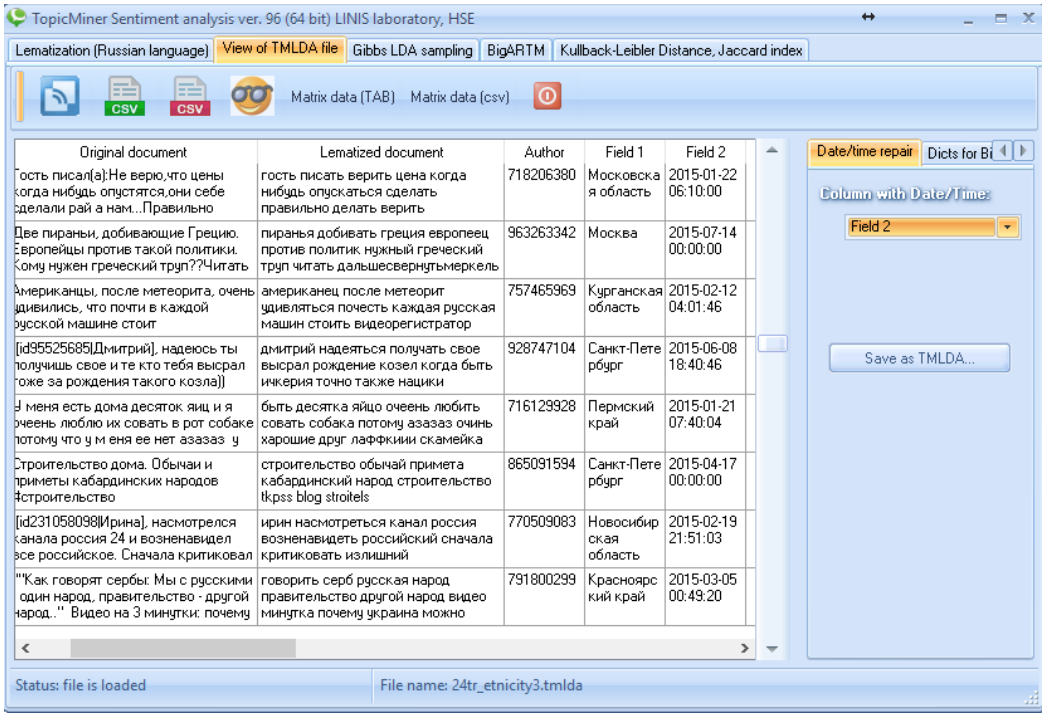


Рис. 8.2. Пример опции ‘Data/Time repair’.

После этого запустится процесс унификации временных данных, и запись результатов в виде нового файла tmlda. Что бы убедится, что все временные метки унифицировались достаточно загрузить новый файл.

Далее нужно сформировать специальный файл tmlda и словарь для мультимодального тематического моделирования. Что бы создать такие файлы нужно перейти на опцию ‘Dics for BigARTM’ (смотри рис. 8.2.1). Далее нужно указать поля, по которым будут строиться дополнительные матрицы при расчете мультимодальных моделей.

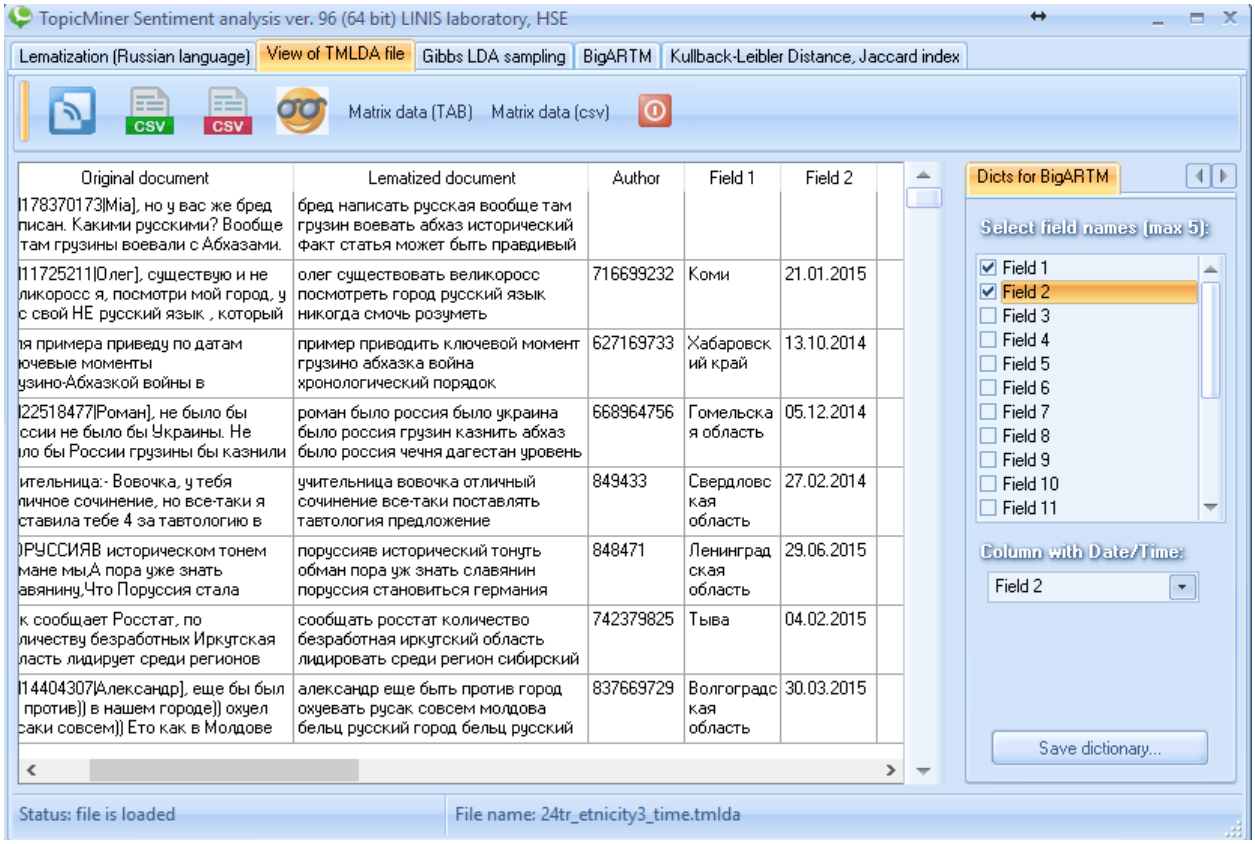


Рис. 8.2.1. Пример создания словаря с метаданными для BigARTM.

Внимание, процесс создания словарь может занимать много времени.

8.2. Построение временных трендов в моделях на основе мультимодального тематического моделирования.

Построение тренда на основе распределения документов по темам.

Построение временных трендов для моделей на мультимодального ТМ возможно лишь для отсортированных матриц распределения документов по темам. Для того что бы построить такой тренд, нужно либо сделать тематический расчет либо загрузить уже сделанные расчеты. В итоге нужно открыть матрицу распределения отсортированных документов (кнопка ) на вкладке ‘BigARTM’. Получится вот такая матрица (смотри рис. 8.3).

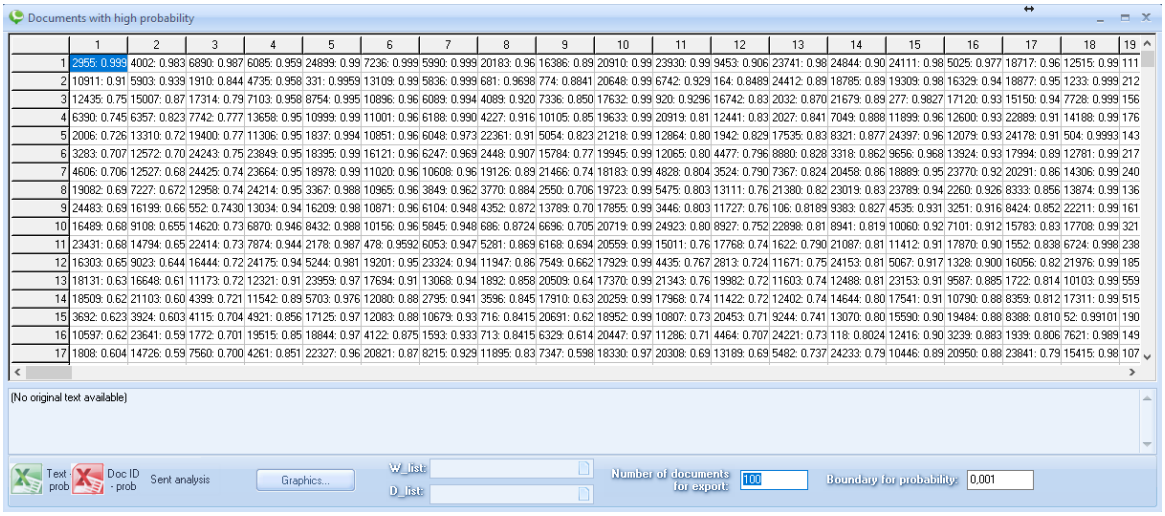


Рис. 8.3. Пример построения временных трендов в матрице распределения документов по темам.

В данном окне присутствует кнопка ‘Graphics’. При нажатии на данную кнопку появится окно графика.

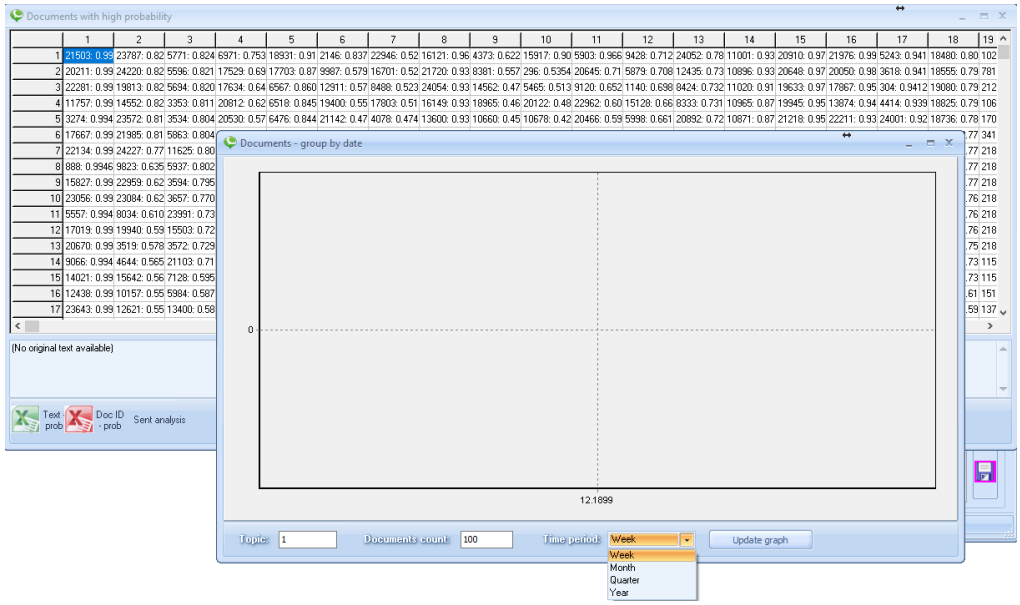


Рис. 8.4. Пример построения временных трендов в матрице распределения документов по темам.

В данном окне нужно, во-первых, указать номер темы, количество документов, чьи метки будут использованы для построения графика и период агрегации (смотри рис. 8.4). Для того что бы обновить содержимое графика нужно нажать на кнопку ‘update graph’. Пример такого графика по теме №11 приведен на рис. 8.5

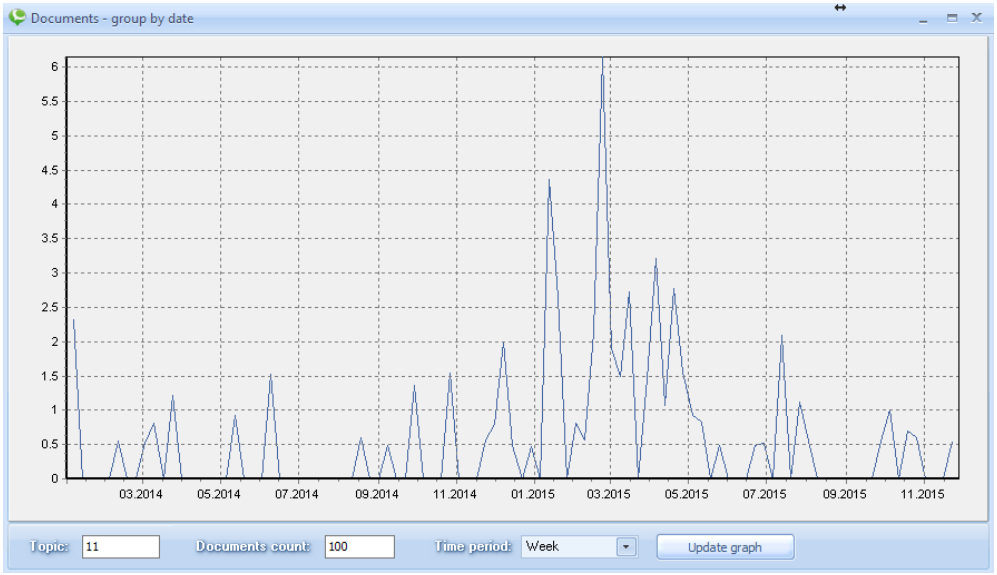


Рис. 8.5. Пример построения временных трендов по неделям.

Распределение временных меток по темам.

В мультимодальных расчетах поле с временными метками можно задействовать непосредственно в расчете. В результат расчета получается дополнительная матрица распределения временных меток по темам. В качестве примера рассмотрим датасет ‘24tr_etnicity3_time_bigartm.tmlda’. Данный датасет входит в состав информационной системы. В этом датасете временные метки находятся в поле №2, соответственно, дополнительная матрица также имеет наименование ‘field2’. Чтобы открыть матрицу распределения дат по темам нужно выбрать кнопку  . Пример выбора матрицы показан на рисунке 8.6.

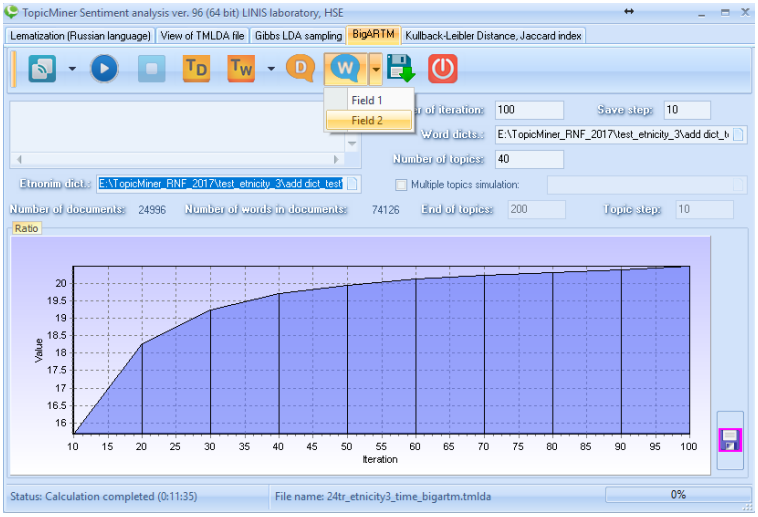


Рис. 8.6. Пример открытия матрицы распределения временных меток по темам.

В результате получится вот такая матрица (смотри рис. 8.7).

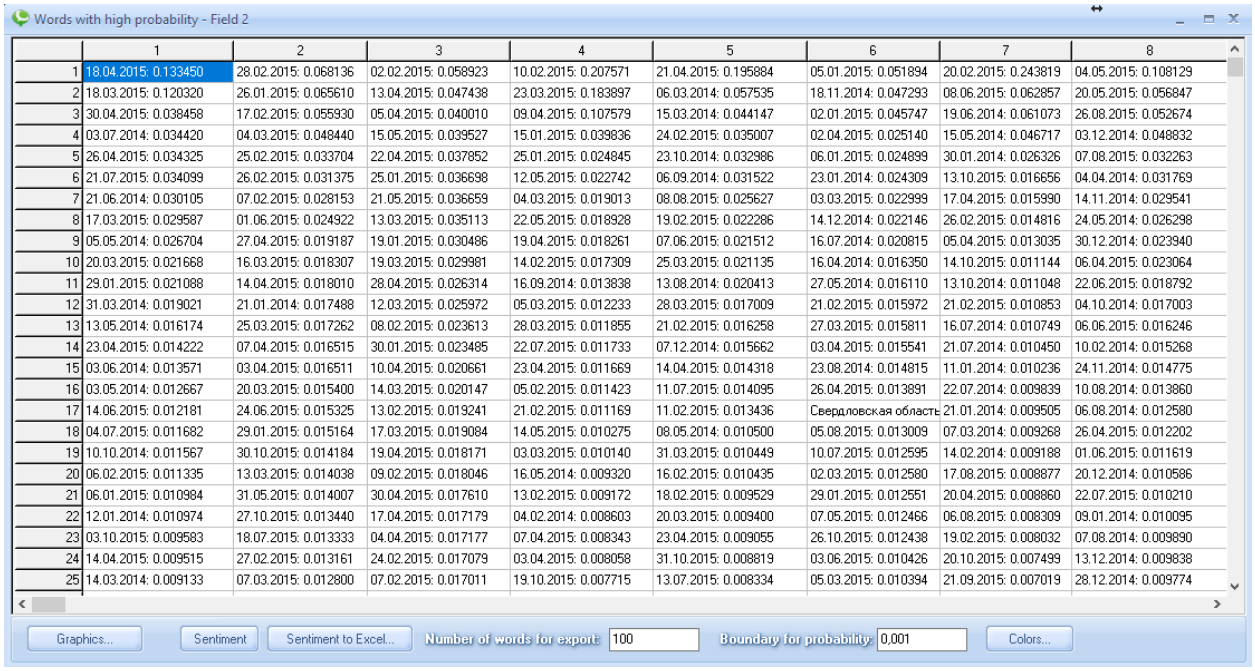


Рис. 8.7. Пример матрицы распределения временных меток по темам.

Каждая ячейка в данной матрице содержит дату и вероятность даты в соответствующей теме.

Для того что бы построить временной график нужно нажать на кнопку ‘graphics’ и в появившемся окне нужно указать номер темы, количество документов, чьи метки будут использованы для построения графика и период агрегации. Пример построения графика приведен на рисунке 8.8.

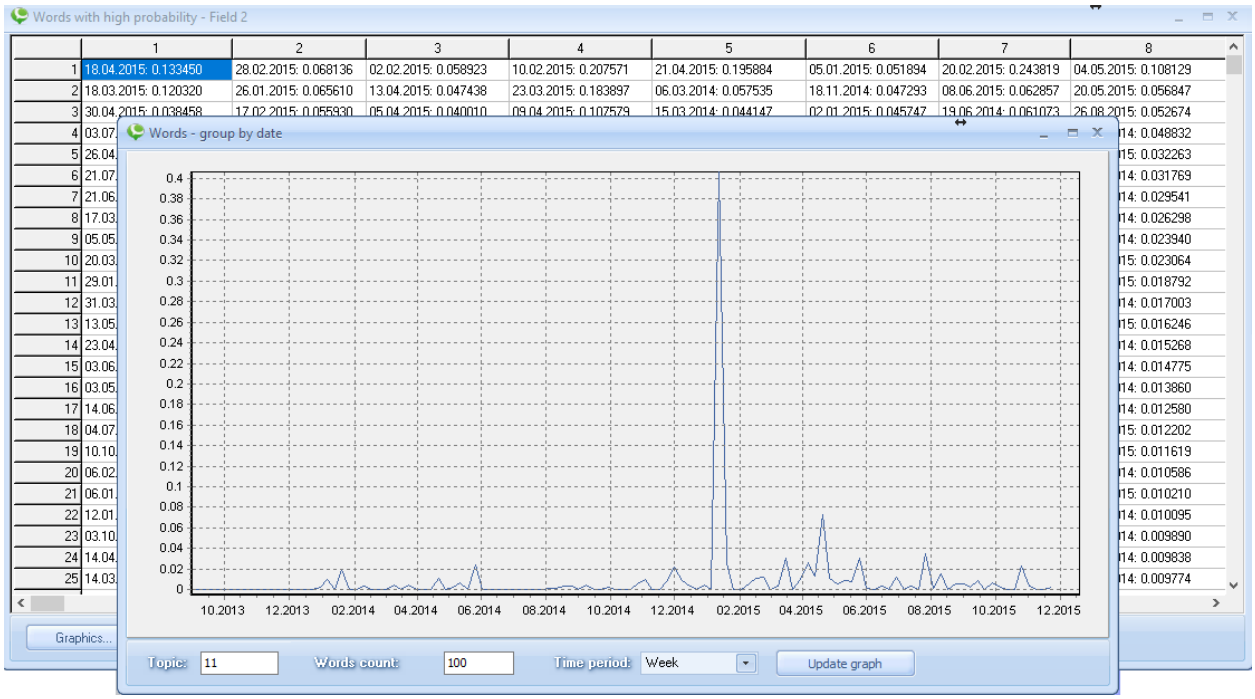


Рис. 8.8. Пример матрицы распределения временных меток по темам.

8.3. Построение временных трендов в моделях на основе сэмплирования Гиббса.
Модели на основе сэмплирвоания Гиббса позволяют построить временной тренд только для матрицы распределения документов по темам. Для этого нужно нажать на кнопку  на вкладке ‘Gibbs LDA sampling’. В появившемся окне нужно жать на кнопку ‘Graphics’. Далее в новом окне указать следующие параметры: 1. Номер темы (для которой нужно строить тренд). 2. Число документов, которые будут использованы для создания тренда. 3. Период агрегации (неделя, месяц..). 4. Номер поля, которое содержит дату конкретного документа. После задания параметров нужно нажать на кнопку ‘update graph’. Пример построения такого графика приведен на рисунке 8.9.

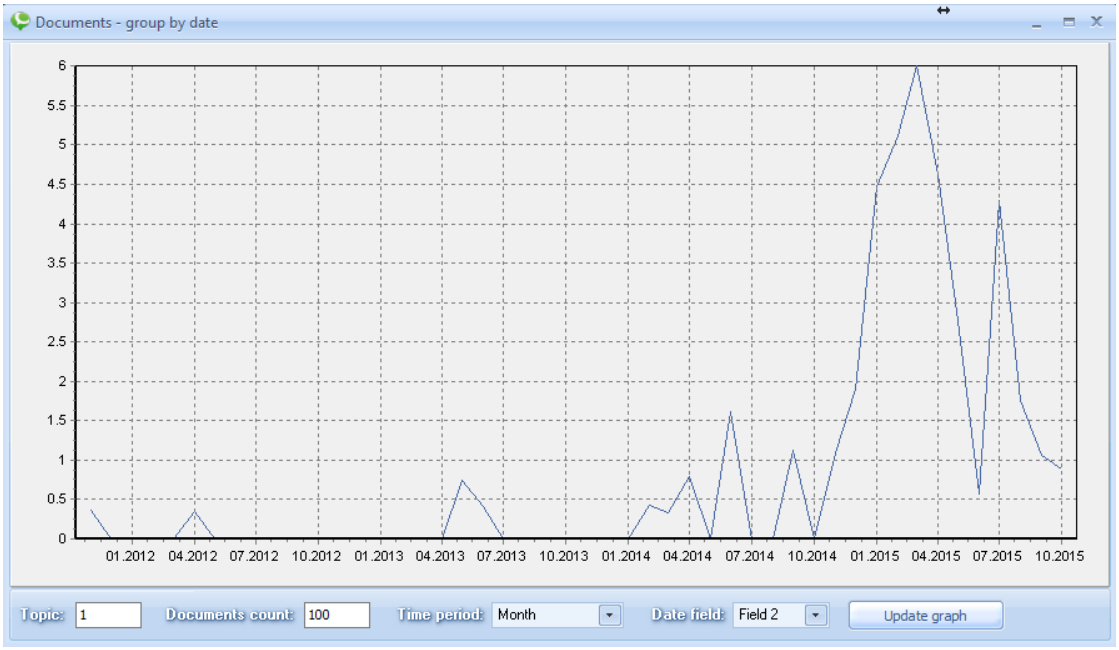


Рис. 8.8. Пример временного тренда для сэмплирования Гиббса.

Заключение.
Все вопросы по применению мониторинговой системы ‘TopicMiner’ и ‘Web TopicMiner’ просьба направлять в лабораторию интернет-исследований Кольцову С.Н (skoltsov@hse.ru)