

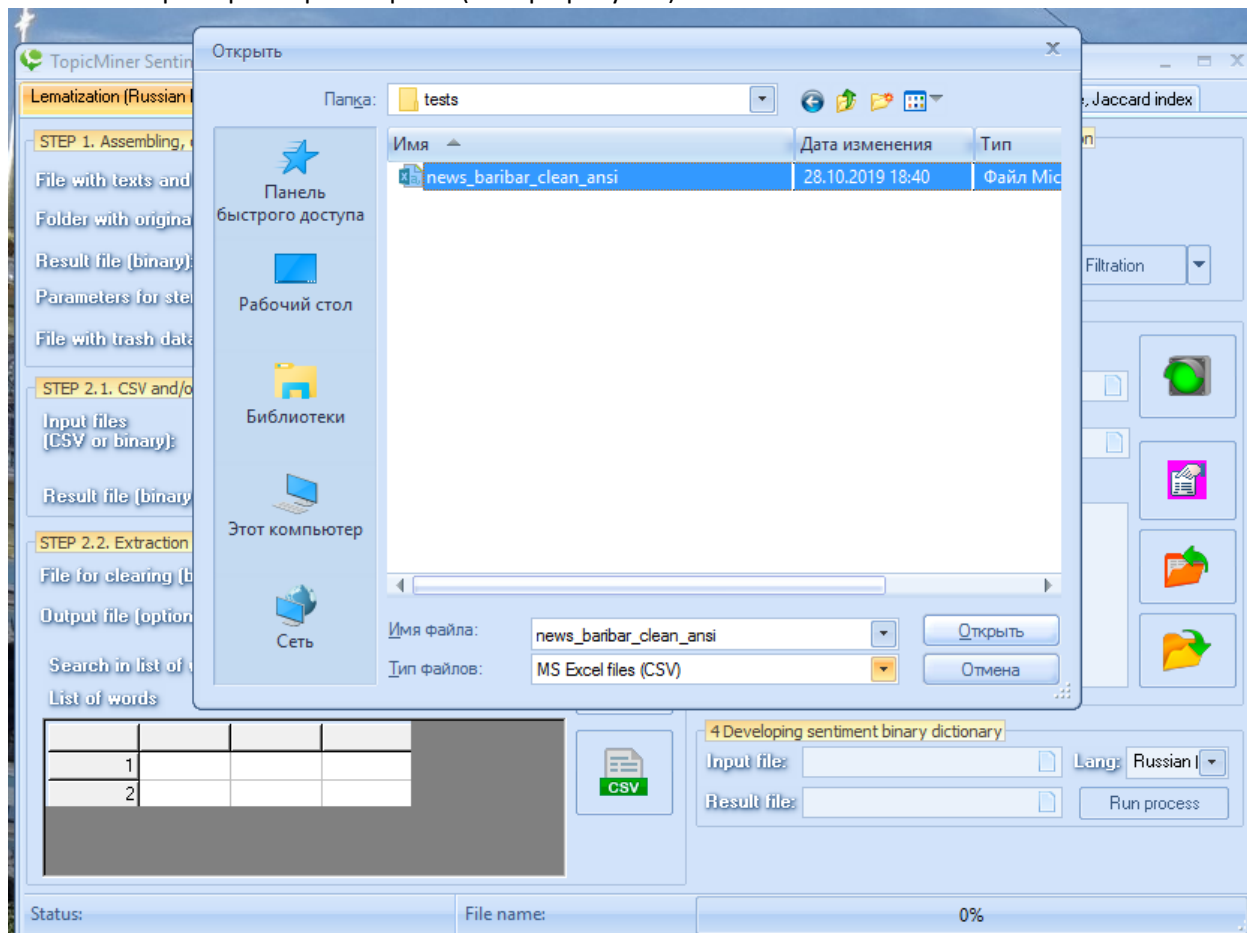
Краткая инструкция

1. Загрузка и препроцессинг данных.

Для того что бы загрузить данные нужно, например, иметь файл с данными, в котором одна строка это один документ.

```
D:\TopicMiner_RNF_2017\TopicMiner_2018\2019year\tests\news_baribar_clean_ansi.csv - Notepad++
Файл Правка Поиск Вид Кодировки Синтаксисы Опции Инструменты Макросы Запуск Плагины Вкладки ?
news_baribar_clean_ansi.csv
1 На мероприятии приняли участие директор офиса коммерциализации результатов научных исследований Же:
2 Экс игрок Тамбова в среду 21 июня поставил подпись под соглашением с чемпионом Казахстана В н:
3 В настоящее время досудебное расследование по признакам преступления предусмотренного ст 99 ч 2
4 ;;2018-09-25 07:20:32;3253;BARIBAR;;https://ru.baribar.kz/tag/denis-ten/
5 В этом году традиционное состязание проходило в селе Калбатау Жарминского района в рамках реализац
6 ;;2018-05-15 07:43:40;4483;BARIBAR;;https://ru.baribar.kz/tag/samruk-kazyna/
7 Как сказала депутат Джамия Нурманбетова из лучшей мировой практики известно что памятники архит:
8 Президент Республики Казахстан Касым Жомарт Токаев на встрече с активом Нур Султана назвал позором
9 ;;2018-02-20 04:23:51;3781;BARIBAR;;https://ru.baribar.kz/tag/latinskaya-grafika/
10 В работе заседания приняли участие члены Правительства депутаты Парламента РК лидеры молодежных (
11 Снежную погоду принесут атмосферные фронтальные разделы над большей частью республики Лишь на юго
12 Расследование пожара на месторождении Каламкас завершено О причинах ЧП на брифинге рассказал за:
13 В Астане под председательством акима города Астаны Бахыта Султанова прошло заседание рабочей групп:
14 ;;2018-09-13 10:38:50;3882;BARIBAR;;https://ru.baribar.kz/tag/mezhdunarodnoe-soobshhestvo/
15 К новому учебному году учебная и справочная литература поднялась в цене на 2 4 за месяц и сразу н:
16 ;;2019-03-18 04:28:18;2851;BARIBAR;;https://ru.baribar.kz/tag/arystanbek-muhamediuly/
17 ;;2019-05-27 07:09:57;4025;BARIBAR;;https://ru.baribar.kz/tag/naruzhnaya-reklama/
```

В качестве примера откройте файл (смотри рисунок)

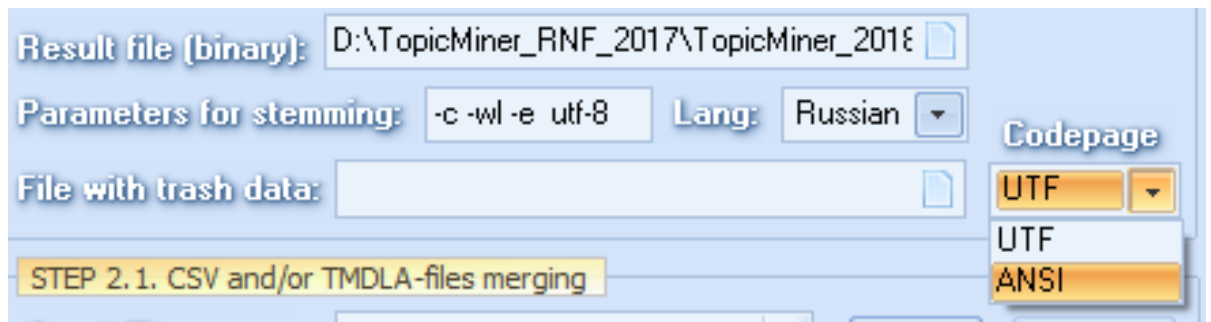


Далее, нужно указать файл, где будут храниться результаты лематизации

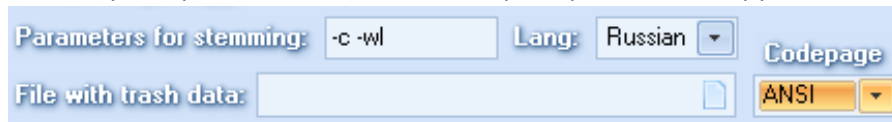
Result file (binary):

Например, test1.tmla

Далее, нужно указать кодировку текста (в данной версии реализованы два варианта)

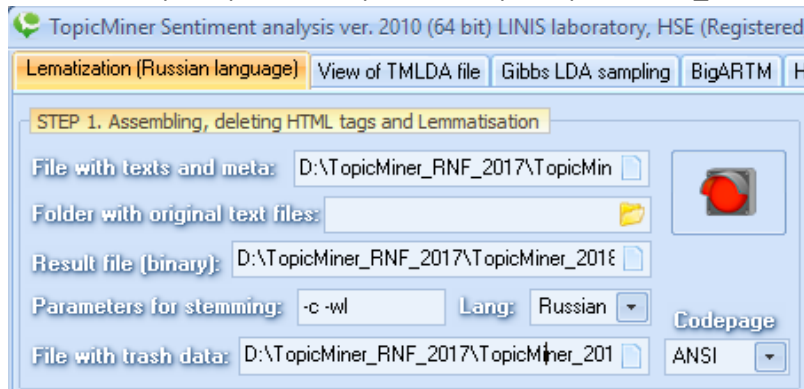


Также нужно указать язык (в данной версии реализованы русский и английский)



Последнее, еще нужно задать файл, который содержит некоторые предварительные слова, которые также можно удалить.

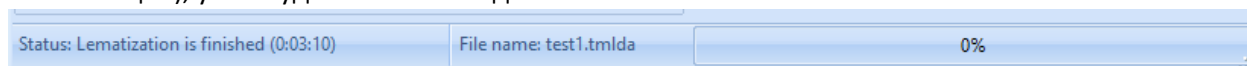
В качестве примеров такого файла смотрите файл trash_data.txt



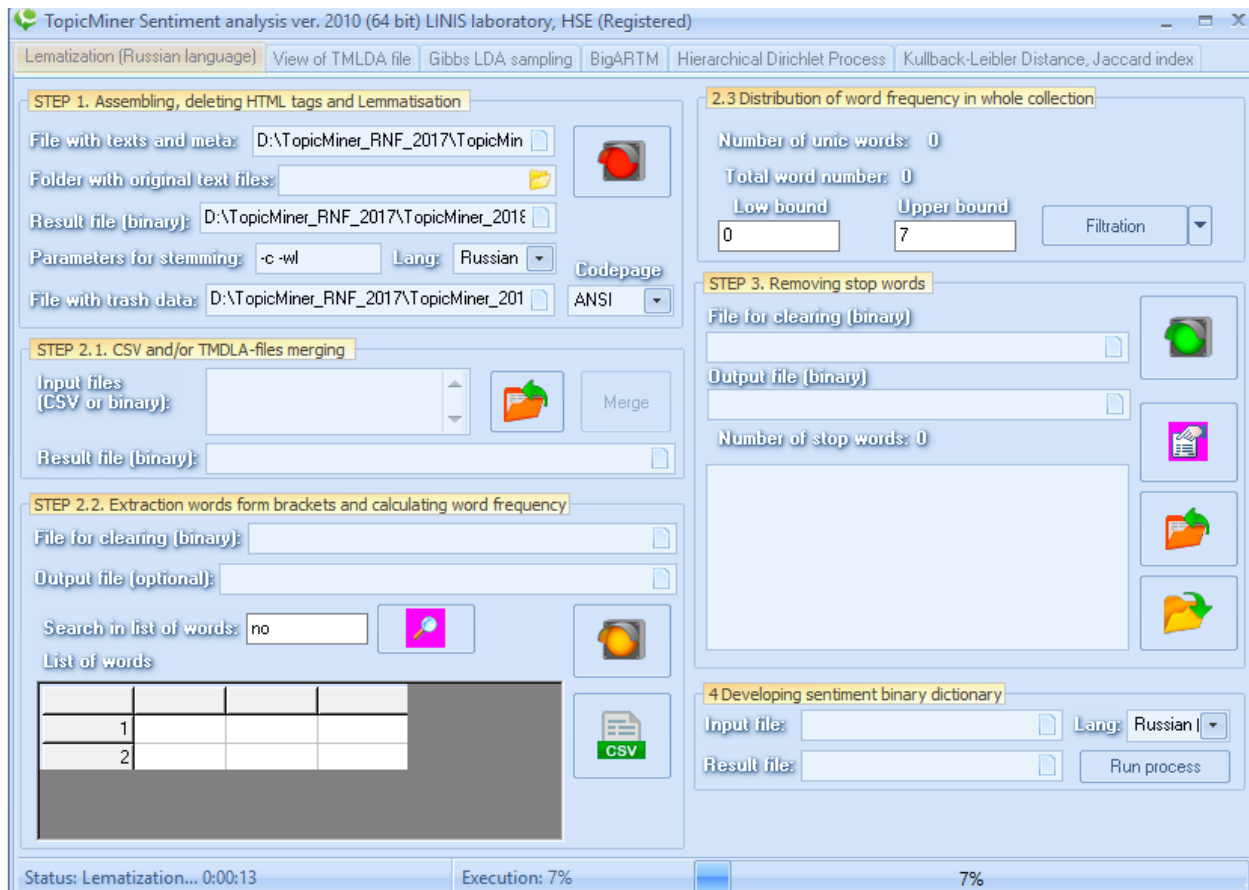
В итоге, когда все что нужно заполнено, нажимайте на красную кнопку.

У вас запустится процесс лематизации (процент выполнения смотрите внизу экрана).

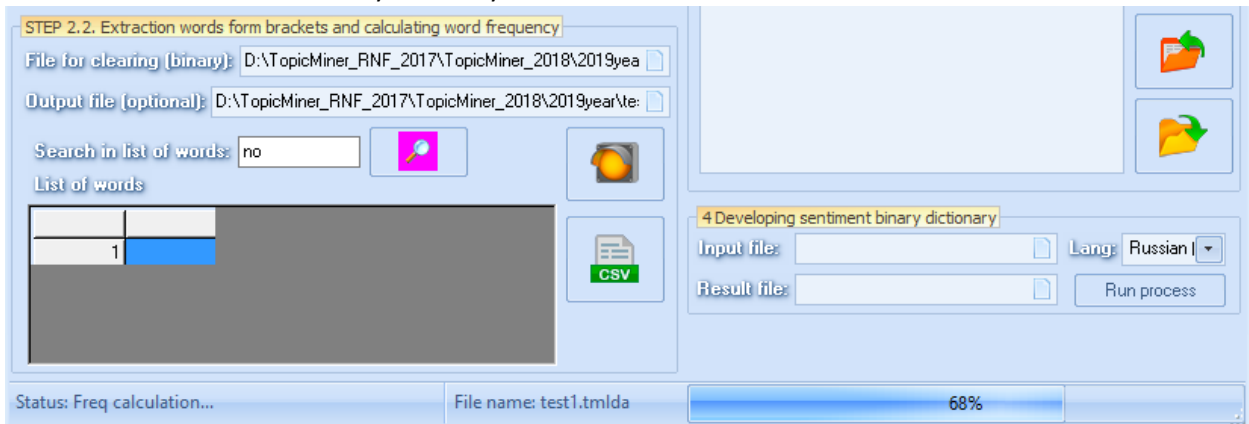
В результате у вас возникнет файл, в котором будут лежать оригинальные и лематизированные тексты. Когда первый этап препроцессинга закончится (процедура лематизации), у вас будет вот такая надпись



Там же указано время лематизации.



На втором этапе нужно указать в качестве входного файла test1.tmla, а в качестве output test2.tmla. и нажать на желтую кнопку.



Процесс работы второго этапа (в процентах) также визуализируется. Также появится список слов с частотами. Этот список можно сохранить если на нажать на кнопку 'csv'. Данная опция полезна для того что бы в дальнейшем сформировать список стоп слов. В качестве списке стоп слов можно выбрать например, наиболее частотные и поставить фильтр, который сохранить в тестовый файл эти слова.


List of words

	Word	Freq	TF-IDF
10	бьтъ	1378	0,09503
11	для	1258	0,09455
12	казахстан	1245	0,09579
13	весь	1237	0,09085

STEP 2.2. Extraction words form brackets and calculating word frequency



File for clearing (binary): D:\TopicMiner_RNF_2017\TopicMiner_2018\2019yea

Output file (optional): D:\TopicMiner_RNF_2017\TopicMiner_2018\2019year\te:

Search in list of words: 

List of words

	Word	Freq	TF-IDF
1	в	9874	0,39080
2	и	6996	0,29442
3	на	4507	0,20050
4	с	2835	0,14474

Например, в данном случае, возьмем границу 1258 (до слова казахстан)


2.3 Distribution of word frequency in whole collection

Number of unic words: 19120

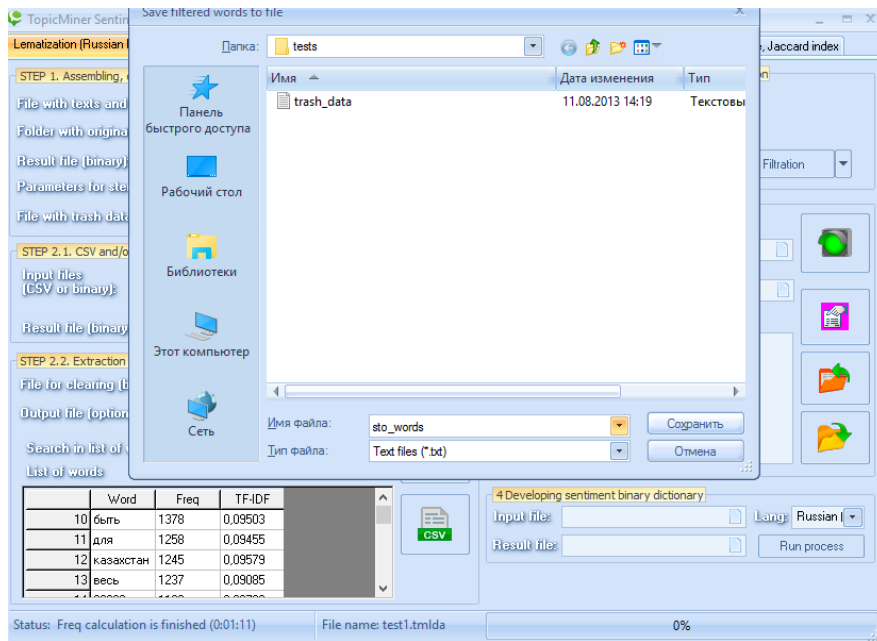
Total word number: 231727

Low bound:

Upper bound:

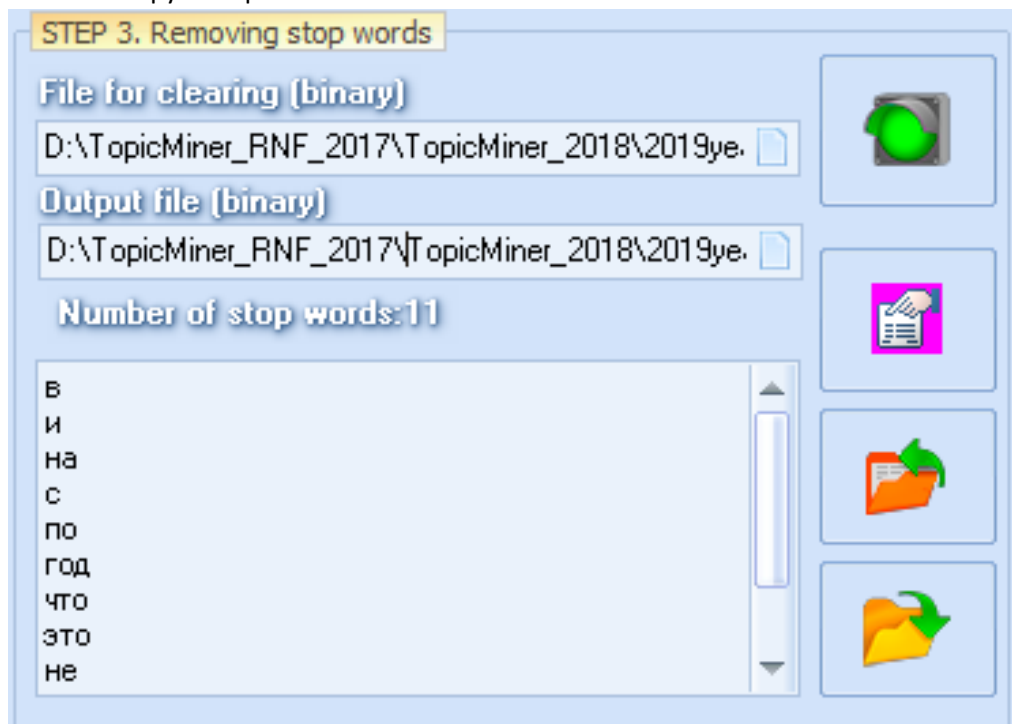


Далее нажимаем на кнопку 'filtration' и указываем файл, в которые сохраняются слова, лежащие вне указанного диапазона (то есть все что выше частоты 1258). Если открыть такой файл, то увидите список стоп слов (наиболее частотные).



На последнем этапе препроцессинга. Нужно указать на вход файл test2.tmla, а на выходе test3.tmla.

А также загрузить файл со списком стоп слов.



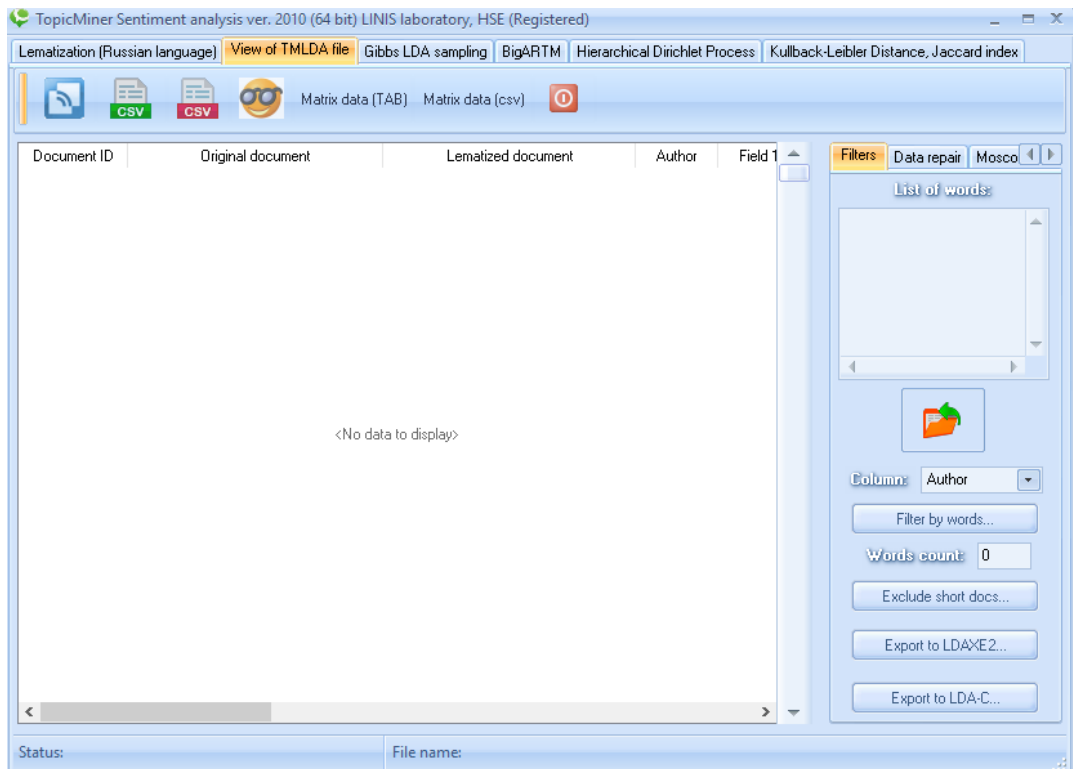
Теперь нажимаете зеленую кнопку.

2. Просмотр результатов препроцессинга

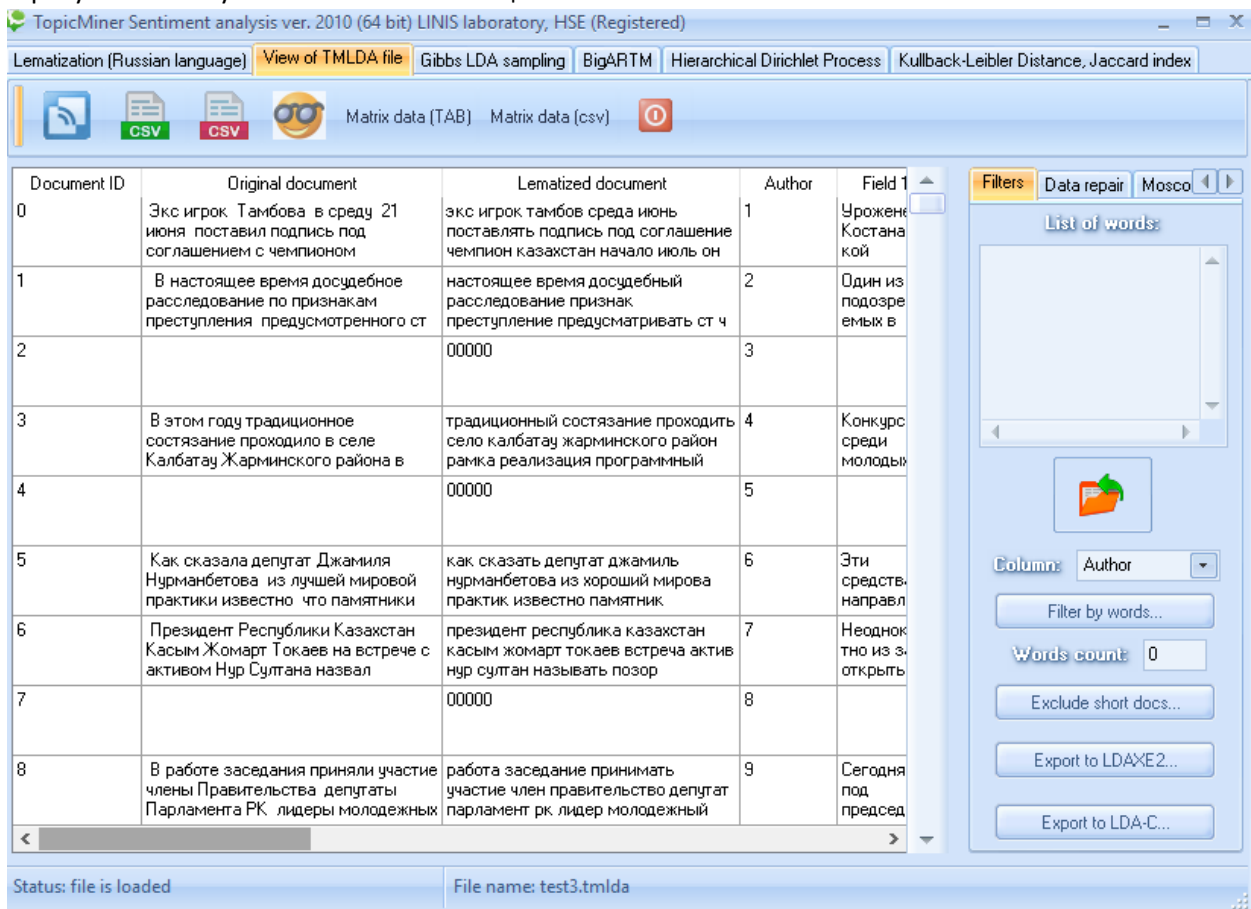
Если нужно посмотреть что подучилось (а также заняться фильтрованием текстов), то нужно перейти на соседнюю вкладку. Загрузка файла осуществляется при помощи кнопки



. Укажите файл test3.tmla.



В результате получится вот такая таблица

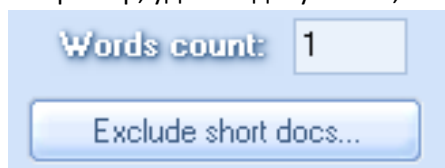


Первый столбик – исходные данные, второй столбик – лематизированные данные. Третий и последующие строки метаданные.

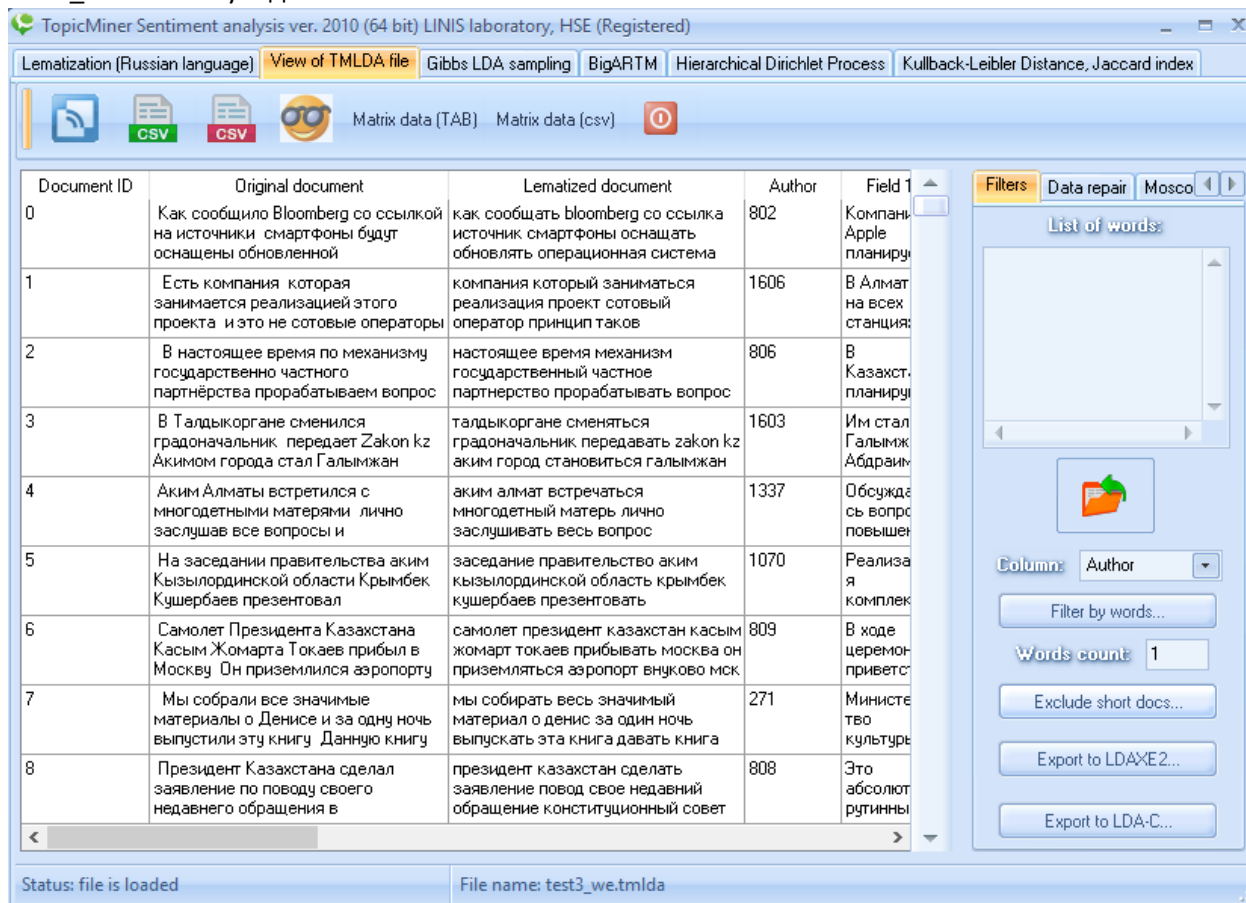
В результате препроцессинга могут получаться пустые документы (например, по причине не указанные концов строки или слова в текстах были удалены при помощи стоп слов).

Конечно такие документы нужно удалять из коллекции. Это можно сделать при помощи фильтрации.

Например, удалим документы, в которые лежит слово 00000. Для этого укажем число слов

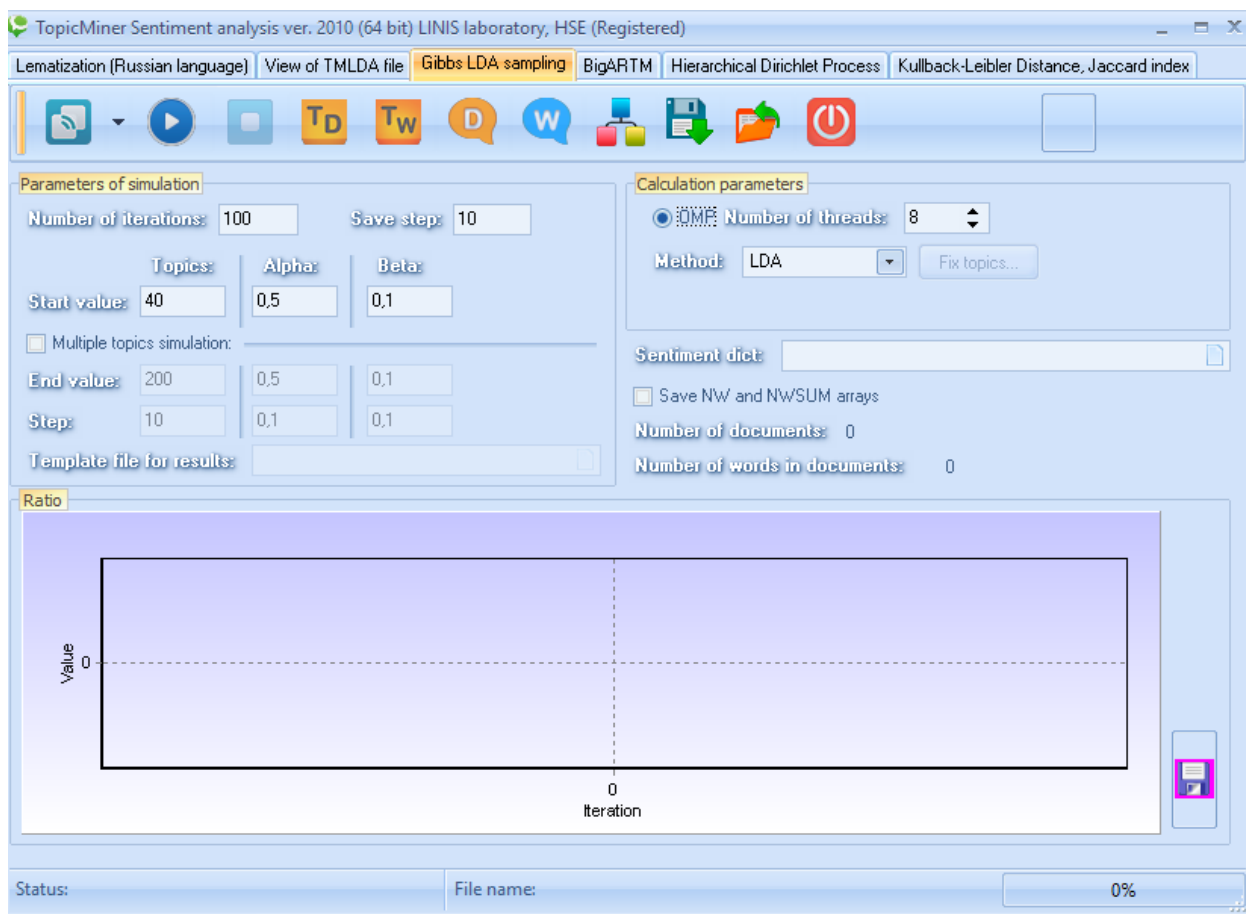


И нажмем на кнопку Exclude short docs. В результате создается файл tmla (test3_we.tmla), в котором уже не будет этих документов. Например, загрузите файл test3_we.tmla и увидите

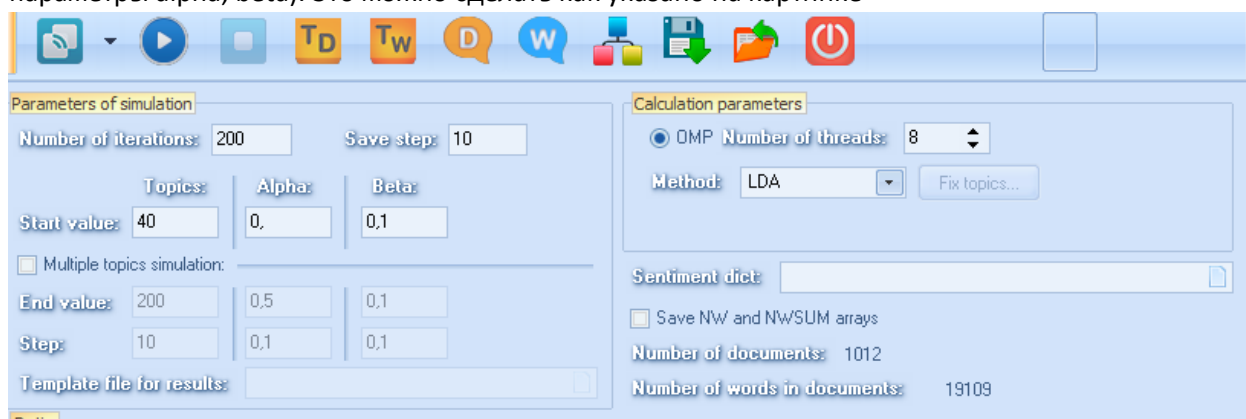


3. Тематическое моделирование.

Для того что бы запустить тематическое моделирование, например, на основе сэмпирования Гиббса (метод монте- карло), нужно перейти на другую вкладку




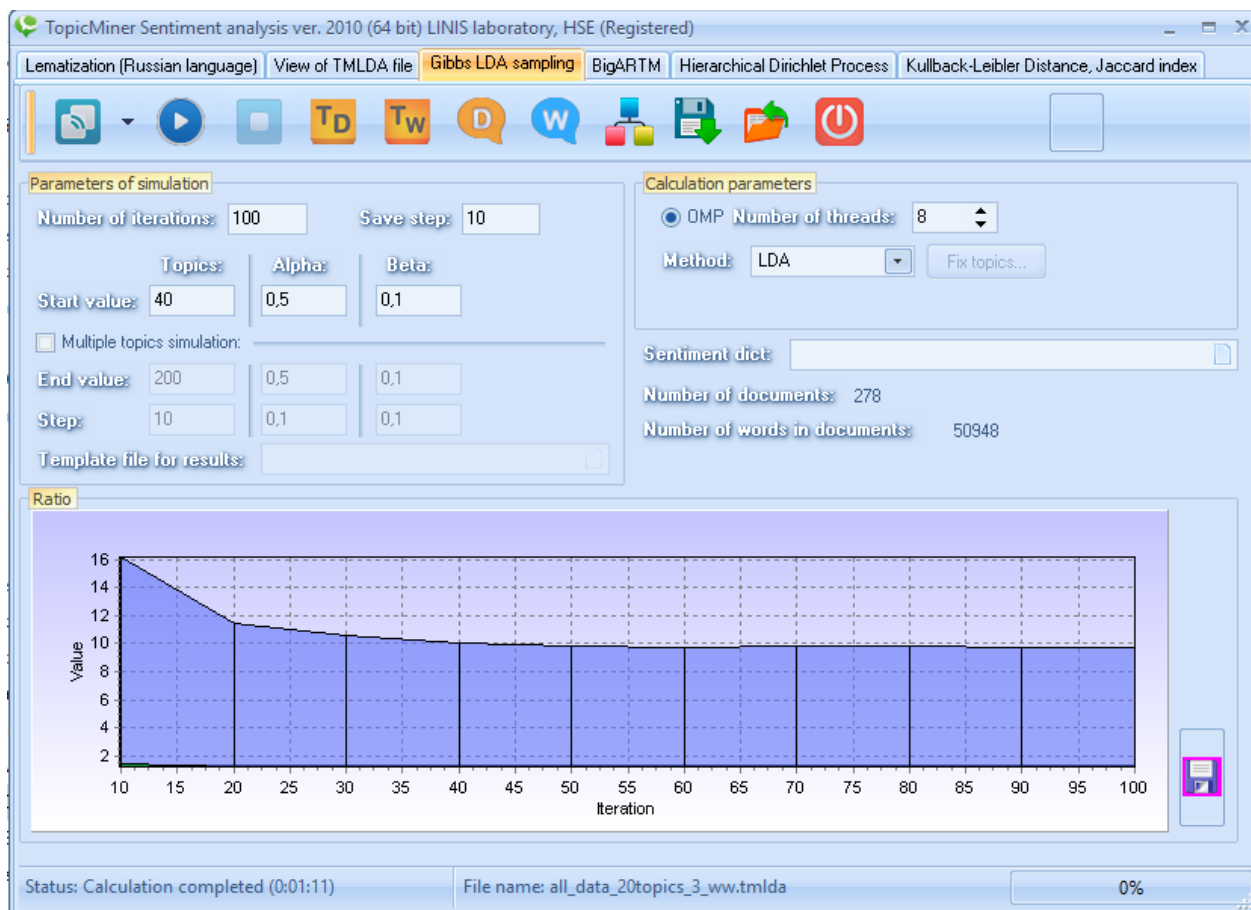
Нужно загрузить файл после препроцессинга и удаления пустых документов.
 Например загрузим файл test3_we.tmla. Не обращайте внимание на процент загрузки (в текущей версии это просто показатель длины коллекции).
 Далее укажите параметры модели (число тем, количество итераций, шаг визуализации, и параметры alpha, beta). Это можно сделать как указано на картинке



И нужно выбрать модель, например LDA.



После чего. Нажмите на кнопку .
 Ход моделирования будет отображаться на графике



Синяя линия – это процент слов с высокой вероятностью. Из графика видно, что после 100 итерации, процент слов перестает меняться, что значит дальнейшее увеличение итераций не нужно делать. Зеленая линия это процент документов с высокой вероятностью.

4. Просмотр полученных результатов и выгрузка во внешний файл. Что бы посмотреть уже отсортированные результаты расчета на нужно нажать на кнопку



. В итоге увидите следующую картинку

Words with high probability

	1	2	3	4	5	6	7	8
1	фестиваль: 0,016652	банк: 0,039944	мозг: 0,012227	реклама: 0,010636	предприятие: 0,013405	техник: 0,0	человек: 0,0	виртуаль
2	соп: 0,007645	сумма: 0,010708	час: 0,007972	размещение: 0,007900	компания: 0,009634	психологич	безопаснос	устройст
3	comic: 0,007414	заем: 0,009539	специальный: 0,007688	культура: 0,007216	завод: 0,004491	технология	происходит	galaxy: 0,
4	комикс: 0,007183	нацбанк: 0,008871	звание: 0,006837	значение: 0,006874	гик: 0,004148	т: 0,007901	служба: 0,0	реально
5	игра: 0,006952	вклад: 0,008871	генерал: 0,006553	культурный: 0,006532	оборудование: 0,003805	центр: 0,00	сообщать: 0	позволт
6	официальный: 0,005797	бюджет: 0,008036	утро: 0,005986	памятник: 0,006532	номинация: 0,003463	психология	давать: 0,0	камера:
7	сессия: 0,005566	заемщик: 0,008036	нагрузка: 0,005702	наружный: 0,005848	месторождение: 0,003120	м: 0,005066	дом: 0,0066	экран: 0,
8	astana: 0,004873	ставка: 0,007868	тренировка: 0,004284	здание: 0,005506	хороший: 0,002777	направленн	от: 0,00670	использ
9	проведение: 0,004411	национальный: 0,007701	ответить: 0,003716	размещать: 0,005506	швейный: 0,002777	р: 0,005068	полиция: 0,	samsung:
10	сериял: 0,004411	программа: 0,007701	просыпаться: 0,003716	местный: 0,004822	автоматизированный: 0,0024	профессор:	результат:	благодар
11	сайт: 0,004180	депозит: 0,007367	организм: 0,003149	пользование: 0,004822	чек: 0,002434	разный: 0,0	область: 0,0	технолог
12	is: 0,004180	валютный: 0,007367	марафон: 0,003149	помещение: 0,004480	тоо: 0,002091	школа: 0,00	работа: 0,0	сердце: 0,
13	султан: 0,004180	тенге: 0,006198	майор: 0,003149	автовокзал: 0,004480	награда: 0,002091	наука: 0,00	место: 0,00	про: 0,00
14	гость: 0,003949	финансовый: 0,005864	безопасность: 0,002865	музей: 0,004138	журналистский: 0,002091	ж: 0,003805	пожар: 0,00	высокий
15	нур: 0,003949	ао: 0,005864	класс: 0,002865	общий: 0,003796	фильм: 0,002091	психолог: 0	ситуация: 0	смартфо
16	асбест: 0,003718	кредит: 0,005697	занятие: 0,002865	объект: 0,003112	награда: 0,002091	посредств	помощь: 0,0	новый: 0,
17	мероприятие: 0,003718	ипотечный: 0,005697	бег: 0,002865	автомобильный: 0,003112	вручать: 0,002091	н: 0,002864	дети: 0,005	обеспеч
18	роль: 0,003487	счет: 0,005697	день: 0,002582	отвод: 0,003112	победитель: 0,002091	доктор: 0,0	вод: 0,0053	версия: 0,
19	id: 0,003256	золото: 0,005363	бегать: 0,002582	пространство: 0,003112	материал: 0,001748	б: 0,002864	пострадав	дисплей:
20	ао: 0,003256	рынок: 0,004861	вид: 0,002582	архитектура: 0,002770	горный: 0,001748	кафедра: 0,	причина: 0,0	новинка:
21	нк: 0,003256	вознаграждение: 0,004694	примерно: 0,002298	вывеска: 0,002770	обладатель: 0,001748	исследова	из: 0,00510	модель:
22	экспо: 0,003256	кредитование: 0,004527	ингода: 0,002298	предел: 0,002770	eng: 0,001748	терапия: 0,0	передавать	функция:
23	salern: 0,003025	онлайн: 0,004527	движение: 0,002298	историко: 0,002770	ао: 0,001748	республик	погибнуть:	смартфо
24	d: 0,003025	вкладчик: 0,004527	утренний: 0,002298	наследие: 0,002770	стаж: 0,001406	автор: 0,00	находится	смартфо
25	день: 0,003025	енпф: 0,004360	секунда: 0,002298	полоса: 0,002428	менеджер: 0,001406	тренинг: 0,0	весь: 0,004	встраи

Graphics... Sentiment Sentiment to Excel... Number of words for export: 100 Boundary for probability: 0,001 Colors...

Если хотите почитать наиболее вероятностные документы в какой либо теме, то можно



нажать на кнопку , и увидите следующее

Documents with high probability

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	985: 0.6299	307: 0.6467	469: 0.6515	124: 0.4068	694: 0.4114	632: 0.5604	774: 0.6282	882: 0.7291	766: 0.5445	1000: 0.481	952: 0.4421	456: 0.7731	552: 0.5525	717: 0.6444	331: 0.6258	265: 0.6266	327: 0.8959	650: 0.5978	840
2	396: 0.6231	727: 0.5558	937: 0.4141	30: 0.19359	528: 0.2037	562: 0.1302	819: 0.6207	186: 0.5958	447: 0.5051	993: 0.4244	559: 0.4203	234: 0.5969	163: 0.5043	812: 0.5370	283: 0.5905	903: 0.5944	859: 0.6590	869: 0.5883	873
3	531: 0.5433	614: 0.5475	600: 0.3596	148: 0.1686	954: 0.1857	682: 0.0889	66: 0.57160	921: 0.5738	403: 0.5000	500: 0.3960	105: 0.4000	14: 0.27230	51: 0.38412	116: 0.5000	743: 0.5868	681: 0.5680	125: 0.6463	707: 0.5753	376
4	430: 0.5397	986: 0.5390	884: 0.3571	218: 0.1601	169: 0.1695	800: 0.0823	108: 0.5425	837: 0.4623	1001: 0.450	459: 0.3901	408: 0.3912	935: 0.2638	226: 0.3750	81: 0.49435	970: 0.3477	886: 0.5633	219: 0.6441	137: 0.5744	269
5	788: 0.4846	436: 0.5277	898: 0.2888	372: 0.1568	489: 0.1629	551: 0.0636	658: 0.5392	1: 0.365942	854: 0.4187	25: 0.35995	298: 0.3771	507: 0.2436	138: 0.3695	129: 0.4865	816: 0.3023	1006: 0.547	594: 0.6433	154: 0.5145	530
6	319: 0.4803	744: 0.4806	534: 0.2705	193: 0.1564	155: 0.1612	473: 0.0617	233: 0.5316	19: 0.30531	954: 0.3914	213: 0.3402	345: 0.3712	817: 0.2389	419: 0.3387	405: 0.3784	994: 0.2933	181: 0.5325	826: 0.6377	15: 0.50806	153
7	12: 0.48011	736: 0.4795	59: 0.26477	310: 0.0934	742: 0.1072	271: 0.0547	294: 0.5219	628: 0.2583	431: 0.3811	329: 0.3373	617: 0.3546	422: 0.1857	730: 0.3325	306: 0.3377	49: 0.27011	393: 0.5211	537: 0.5963	981: 0.4503	100
8	726: 0.4229	753: 0.4590	465: 0.2057	780: 0.0797	825: 0.0889	40: 0.04934	928: 0.5195	214: 0.1420	243: 0.3757	634: 0.3342	754: 0.3525	956: 0.1598	668: 0.3208	101: 0.2682	987: 0.2582	133: 0.5042	752: 0.5804	627: 0.4487	667
9	188: 0.3784	845: 0.4491	663: 0.0514	241: 0.0714	732: 0.0827	625: 0.0468	416: 0.4917	980: 0.0910	767: 0.3688	945: 0.3072	920: 0.3478	554: 0.1338	72: 0.30657	723: 0.1962	195: 0.2456	875: 0.4757	453: 0.5721	621: 0.4011	695
10	57: 0.24774	674: 0.4380	814: 0.0500	756: 0.0650	572: 0.0800	972: 0.0449	724: 0.4831	10: 0.07341	310: 0.3681	320: 0.3060	815: 0.3175	465: 0.1288	418: 0.2812	590: 0.1024	657: 0.2207	892: 0.4501	359: 0.5613	109: 0.3948	789
11	837: 0.1335	258: 0.4098	757: 0.0492	735: 0.0609	466: 0.0663	804: 0.0439	435: 0.4735	210: 0.0636	825: 0.3588	271: 0.2737	502: 0.3156	690: 0.0954	423: 0.2706	145: 0.0847	810: 0.2012	98: 0.42548	274: 0.5489	281: 0.3579	989
12	880: 0.0921	340: 0.4020	544: 0.0465	25: 0.04976	273: 0.0629	578: 0.0405	862: 0.4703	939: 0.0625	613: 0.3562	776: 0.2472	888: 0.3018	145: 0.0672	123: 0.2706	437: 0.0636	949: 0.1702	603: 0.4176	974: 0.5394	45: 0.33157	929
13	945: 0.0783	353: 0.4004	795: 0.0454	800: 0.0478	130: 0.0551	574: 0.0392	514: 0.4638	849: 0.0568	91: 0.35488	360: 0.2103	476: 0.2376	609: 0.0625	955: 0.2687	520: 0.0690	31: 0.16477	890: 0.4150	557: 0.5000	895: 0.3260	120
14	592: 0.0726	923: 0.3774	100: 0.0451	678: 0.0455	88: 0.05384	317: 0.0387	102: 0.4600	810: 0.0548	756: 0.3517	926: 0.2077	365: 0.2175	902: 0.0581	801: 0.2543	786: 0.0652	981: 0.1562	209: 0.4147	248: 0.5000	228: 0.3000	983
15	927: 0.0625	711: 0.3692	716: 0.0433	511: 0.0432	482: 0.0486	293: 0.0380	53: 0.45127	381: 0.0546	995: 0.3503	206: 0.1977	876: 0.2086	618: 0.0555	301: 0.1953	370: 0.0625	452: 0.1335	385: 0.4019	938: 0.4701	858: 0.2949	903
16	400: 0.0597	688: 0.3358	903: 0.0433	85: 0.04268	34: 0.04787	208: 0.0379	318: 0.4389	445: 0.0467	370: 0.3492	167: 0.1900	739: 0.1962	301: 0.0546	758: 0.1682	127: 0.0625	512: 0.1150	9: 0.401685	75: 0.45151	542: 0.2788	229
17	134: 0.0555	168: 0.2951	619: 0.0432	833: 0.0426	185: 0.0477	251: 0.0376	596: 0.4301	606: 0.0443	831: 0.3489	513: 0.1778	300: 0.1923	659: 0.0543	357: 0.1675	704: 0.0607	431: 0.1126	622: 0.3960	893: 0.4417	708: 0.2781	149

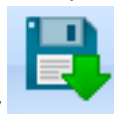
Text prob Doc ID - prob Sent analysis Graphics... W_list D_list Number of documents for export: 100 Boundary for probability: 0,001

По сложившейся традиции накануне футбольного соревнования мирового масштаба самый престижный футбольный трофей – 6 килограммовый Кубок мира из чистого золота вручается команде победительнице Чемпионата мира по футболу, отправляется в кругосветное путешествие с целью объединений людей вокруг благородных идеалов спорта и популяризации чемпионата. Кубок прибыл в Алматы в сопровождении делегации FIFA с большим опозданием. По плану самолет с делегацией и Кубком должен был приземлиться около 10.30, однако в районе аэропорта в это время была низкая видимость. Как стало известно, видна была только 1-3 полосы. Теоретически посадить самолет можно было с помощью приборов, но пилоты не стали рисковать и ушли на запасной аэродром в Бишкек, чтобы там дозвониться и дожидаться, когда туман в Алматы рассеется. В итоге

В каждой ячейке лежит номер документа и его вероятность в теме (то есть в колонке). Что бы посмотреть содержимое документа (оригинальный не лематизированный текст), достаточно ткнуть мышкой в ячейку.

5. Сохранения результатов расчета в виде проекта.

Результаты расчета можно сохранить в виде проекта, соответственно, в следующий раз (на другой день, месяц или год), это проект можно открыть в TopicMiner и продолжить работу.



Что бы сохранить нужно использовать кнопку . А что бы открыть проект



используйте кнопку .

