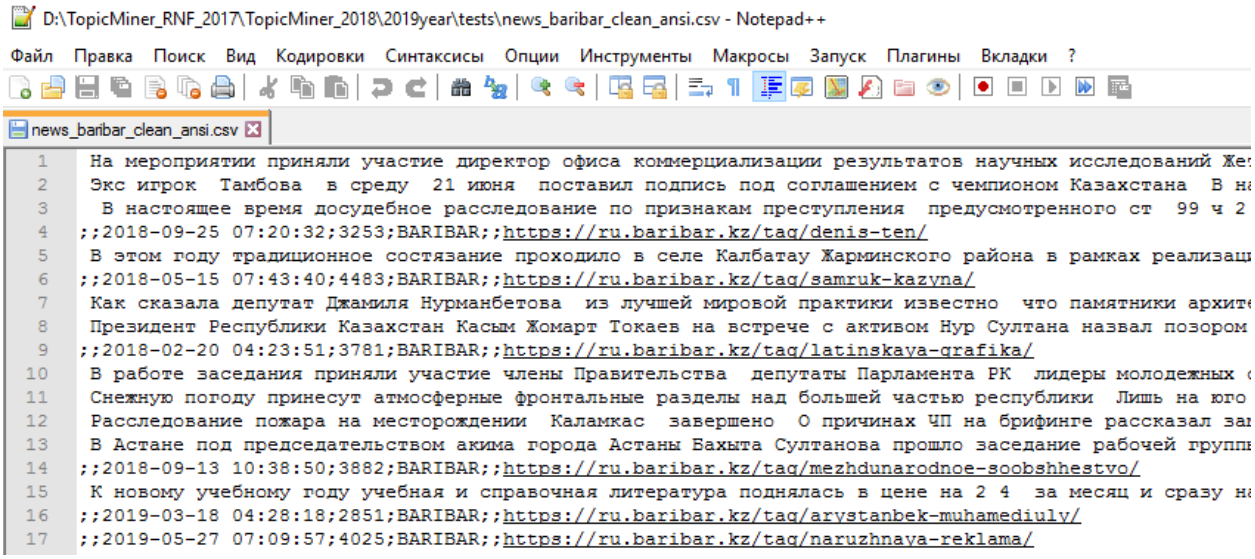
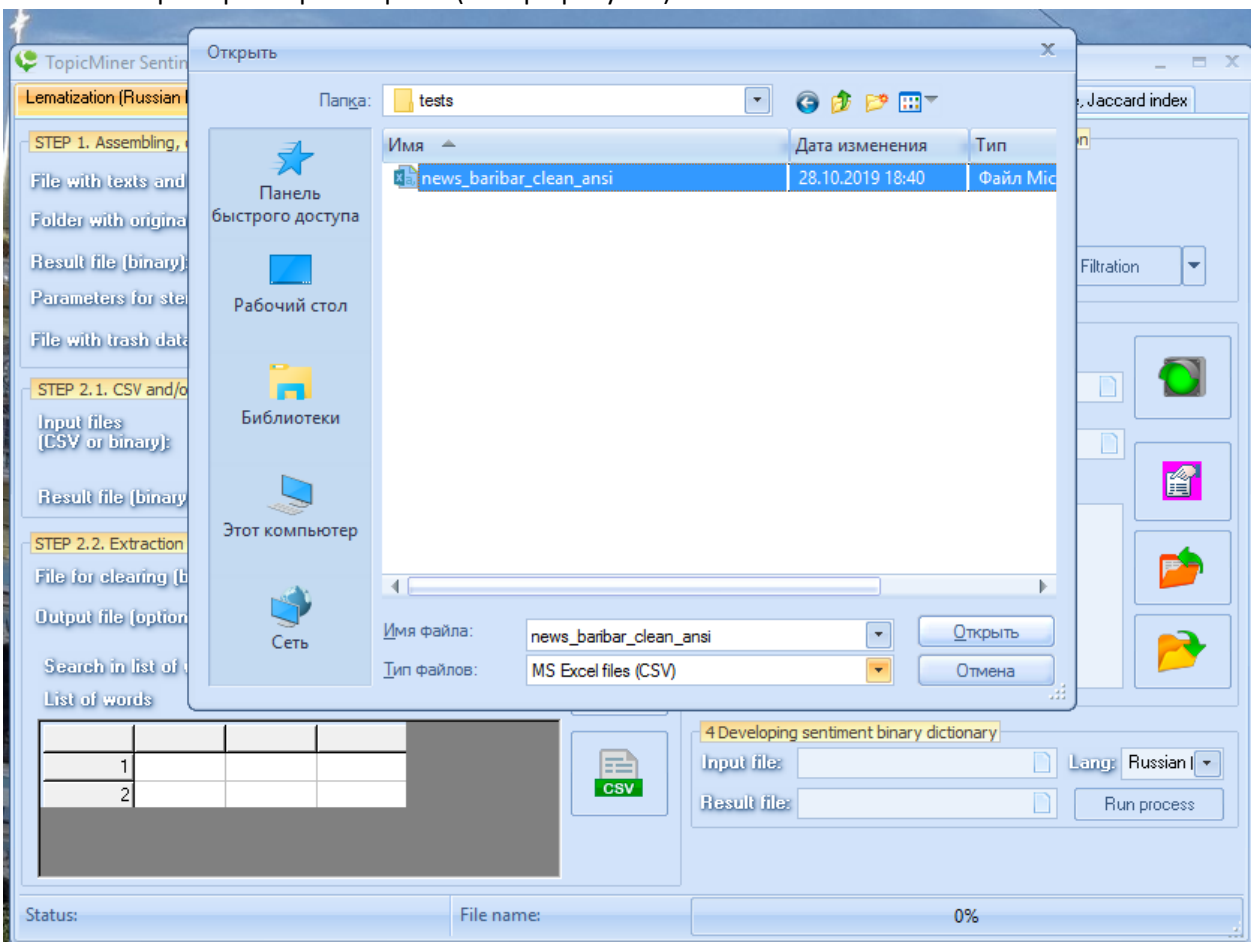
Краткая инструкция

1. **Загрузка и препроцессинг данных.**

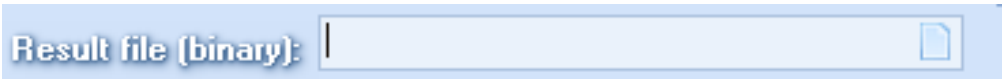
   Для того что бы загрузить данные нужно, например, иметь файлик с данными, в котором одна строка это один документ.



   В качестве примера откройте файл (смотри рисунок)



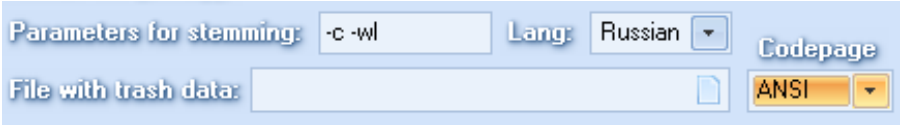
   Далее, нужно указать файл, где будет хранится результаты лематизации



   Например, test1.tmlda

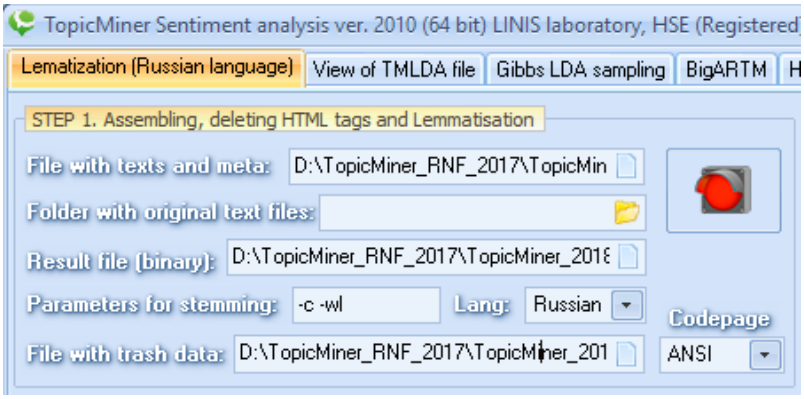
   Далее, нужно указать кодировку текста (в данной версии реализованы два варианта)

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.

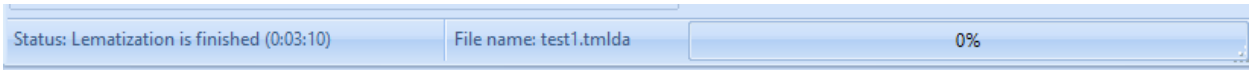
В качестве примеры такого файла смотрите файл trash_data.txt



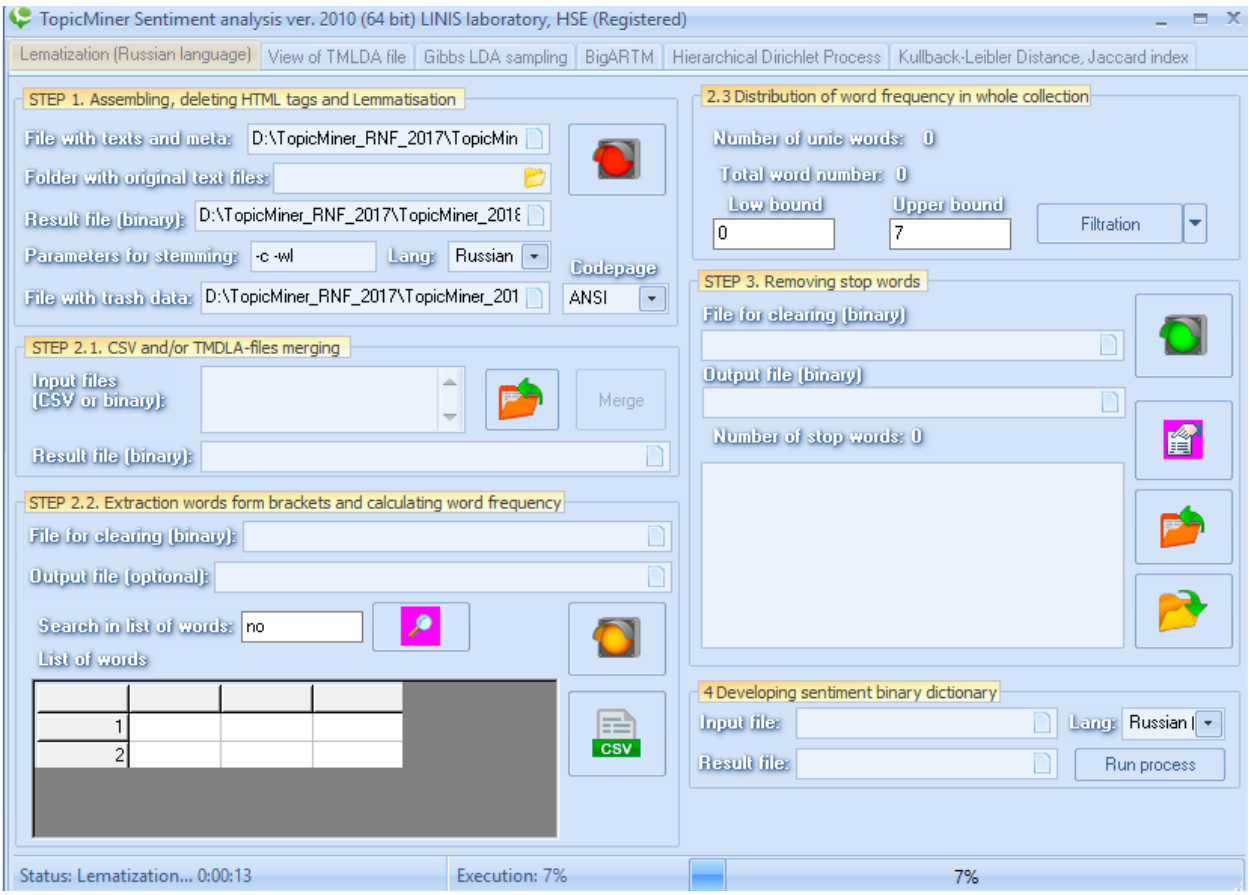
В итоге, когда все что нужно заполнено, нажимайте на красную кнопку.

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

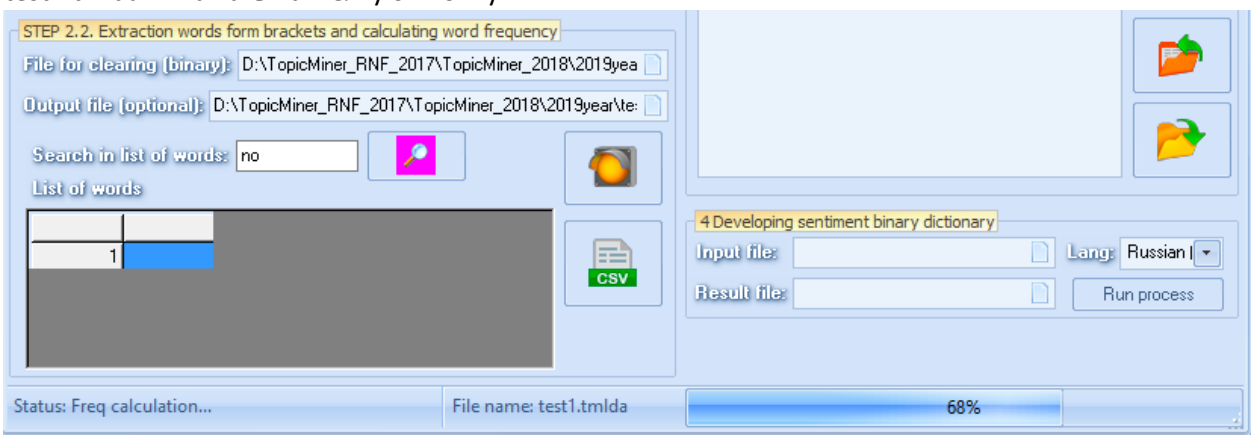
В результате у вас возникнет файл, в котором будут лежать оригинальные и лематизировнные тексты. Когда первый этап препроцессинга закончится (процедура лематизации), у вас будет вот такая надпись



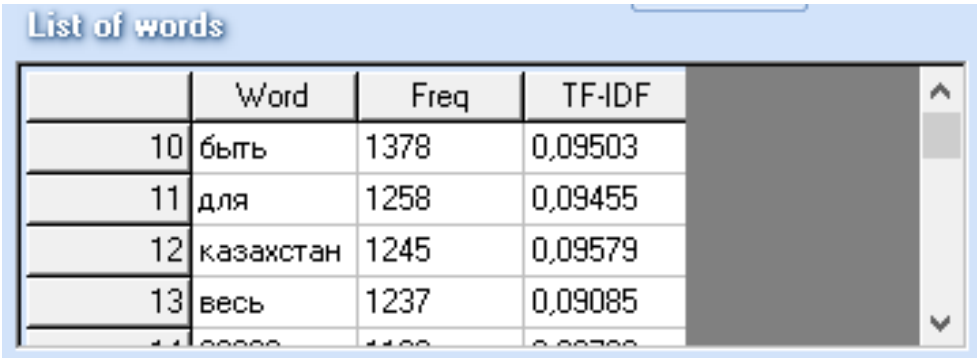
Там же указанно время лематизации.

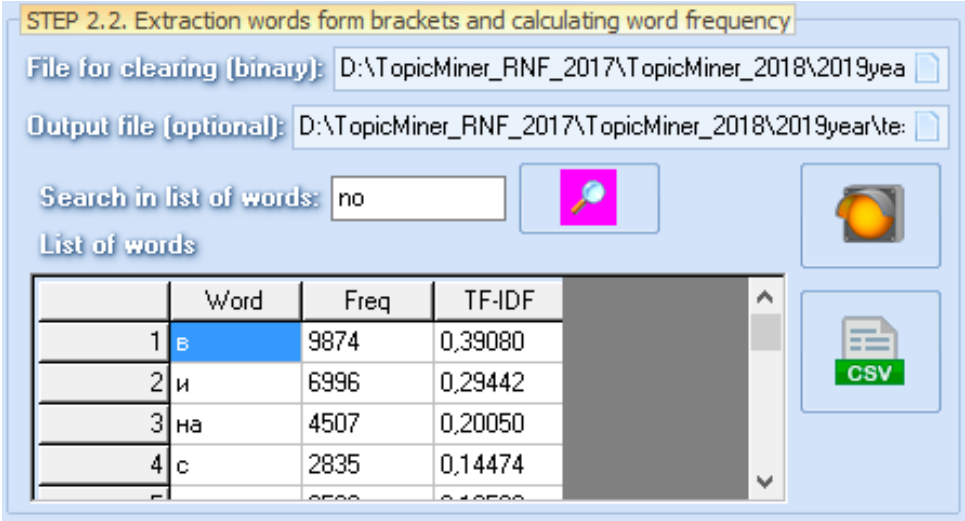


На втором этапе нужно указать в качестве входного файла test1.tmlda, а вкачестве output test2.tmlda. и нажать на желтую кнопку.

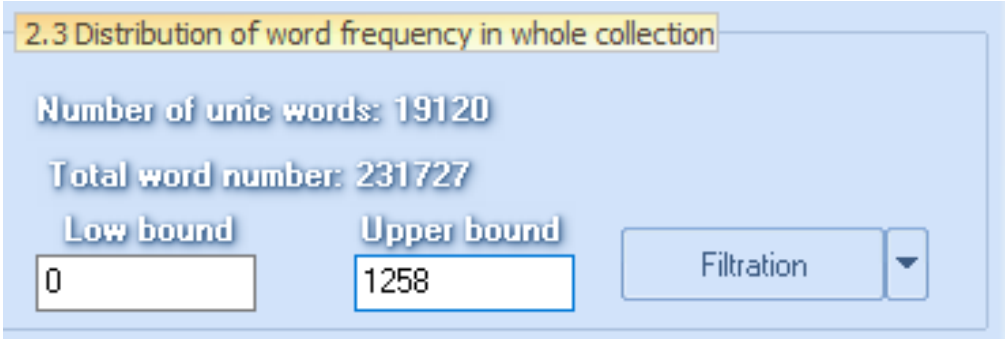


Процесс работы второго этапа (в процентах) также визуализируется. Также появится список слов с частотами. Этот список можно сохранить если на нажать на кнопку ‘csv’. Данная опция полезна для того что бы в далнейшем сформировать список стоп слов. В качестве списке стоп слов можно выбрать например, наиболее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

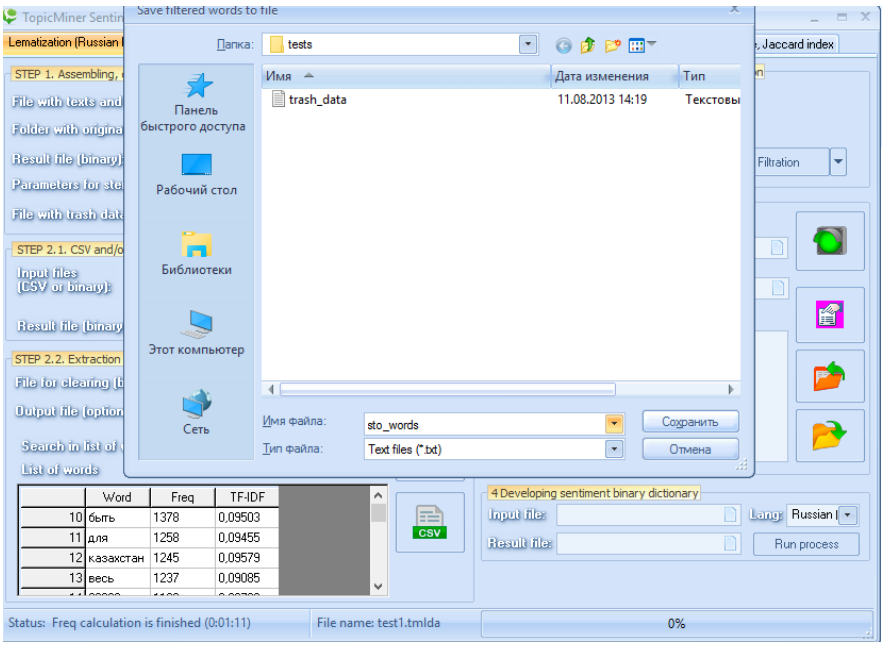




Например, в данном случае, возьмем границу 1258 (до слова казахстан)

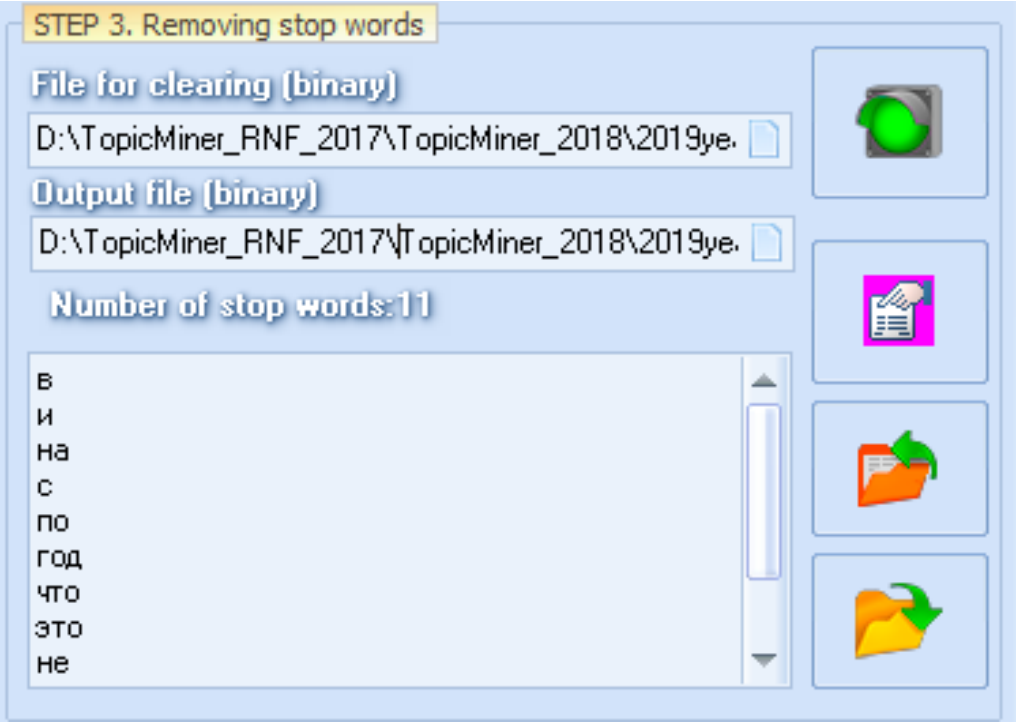


Далее нажимаем на кнопку ‘filtration’ и указываем файл, в которые сохранятся слова, лежащие вне указанного диапазона (то есть все что свыше частоты 1258). Если открыть такой файлик, то увидите список стоп слов (наиболее частотные).



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

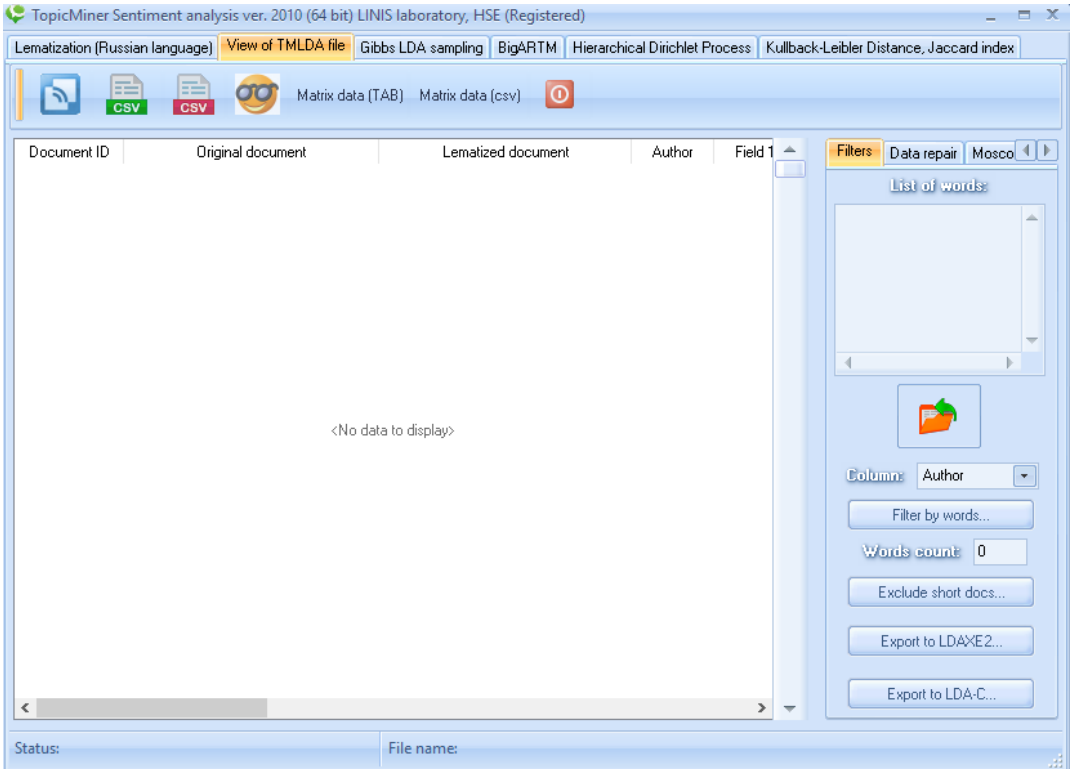
А также загрузить файл со списком стоп слов.



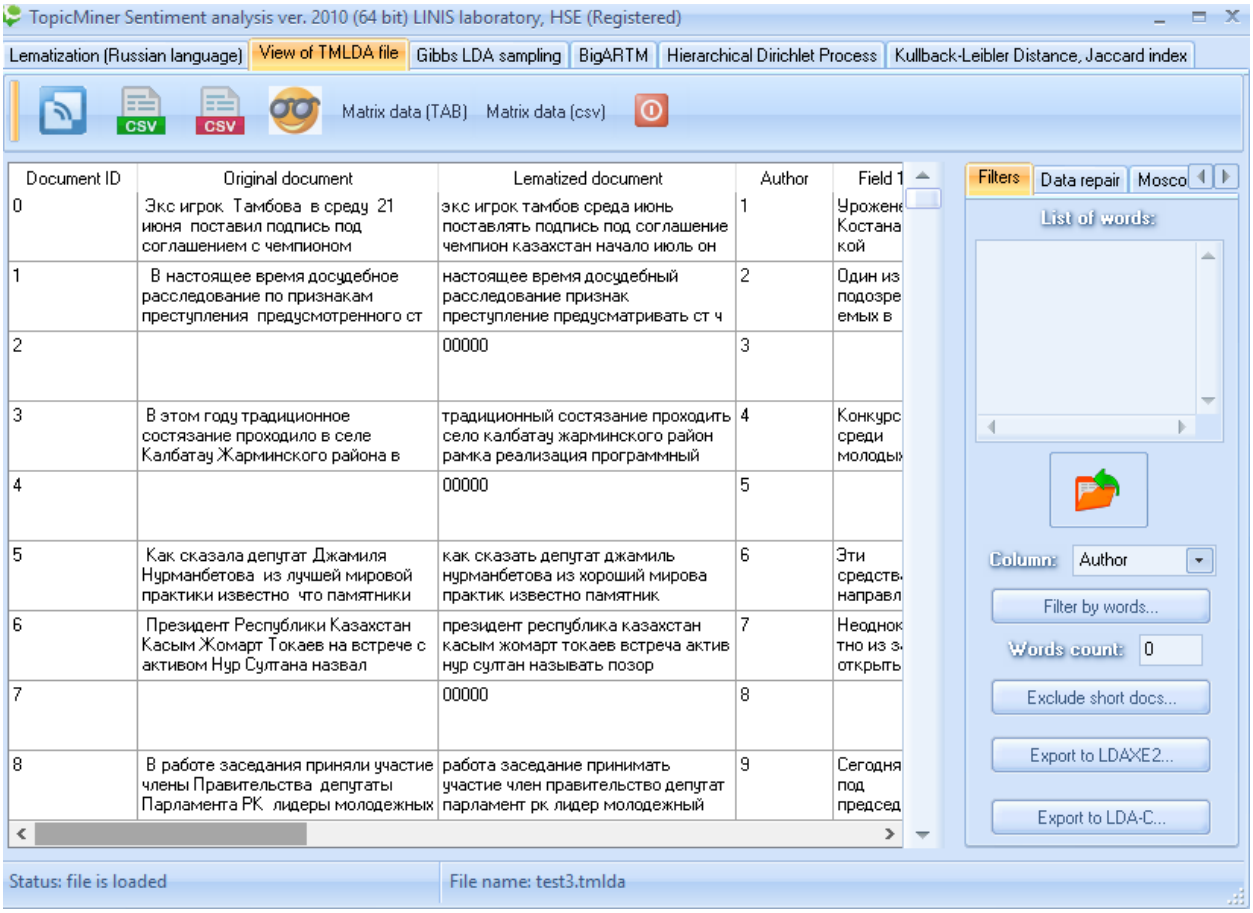
Теперь нажимаете зеленую кнопку.

2. Просмотр результатов препроцессинга

Если нужно посмотреть что подучилось (а также заняться фильтрованием текстов), то нужно перейти на соседнею вкладку. Загрузка файла осуществляется при помощи кнопки  . Укажите файл test3.tmlda.



В результате получится вот такая таблица



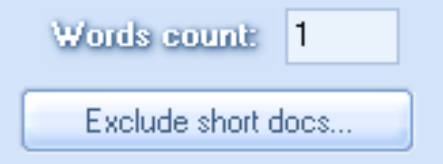
Document ID	Original document	Lematized document	Author	Field 1
0	Экс игрок Тамбова в среду 21 июня поставил подпись под соглашением с чемпионом	экс игрок тамбов среда июнь поставлять подпись под соглашение чемпион казахстан начало июль он	1	Уроженец Костанайской
1	В настоящее время досудебное расследование по признакам преступления предусмотренного ст	настоящее время досудебный расследование признак преступление предусматривать ст ч	2	Один из подозреваемых в
2		00000	3	
3	В этом году традиционное состязание проходило в селе Калбатау Жарминского района в	традиционный состязание проходить село калбатау жарминского район рамка реализация программный	4	Конкурс среди молодых
4		00000	5	
5	Как сказала депутат Джамиля Нурманбетова из лучшей мировой практики известно что памятники	как сказать депутат джамиль нурманбетова из хороший мирова практик известно памятник	6	Эти средства направл
6	Президент Республики Казахстан Касым Жомарт Токаев на встрече с активом Нур Султана назвал	президент республика казахстан касым жомарт токаев встреча актив нур султан называть позор	7	Неоднократно из за открыть
7		00000	8	
8	В работе заседания приняли участие члены Правительства депутаты Парламента РК лидеры молодежных	работа заседание принимать участие член правительство депутат парламент рк лидер молодежный	9	Сегодня под председ

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

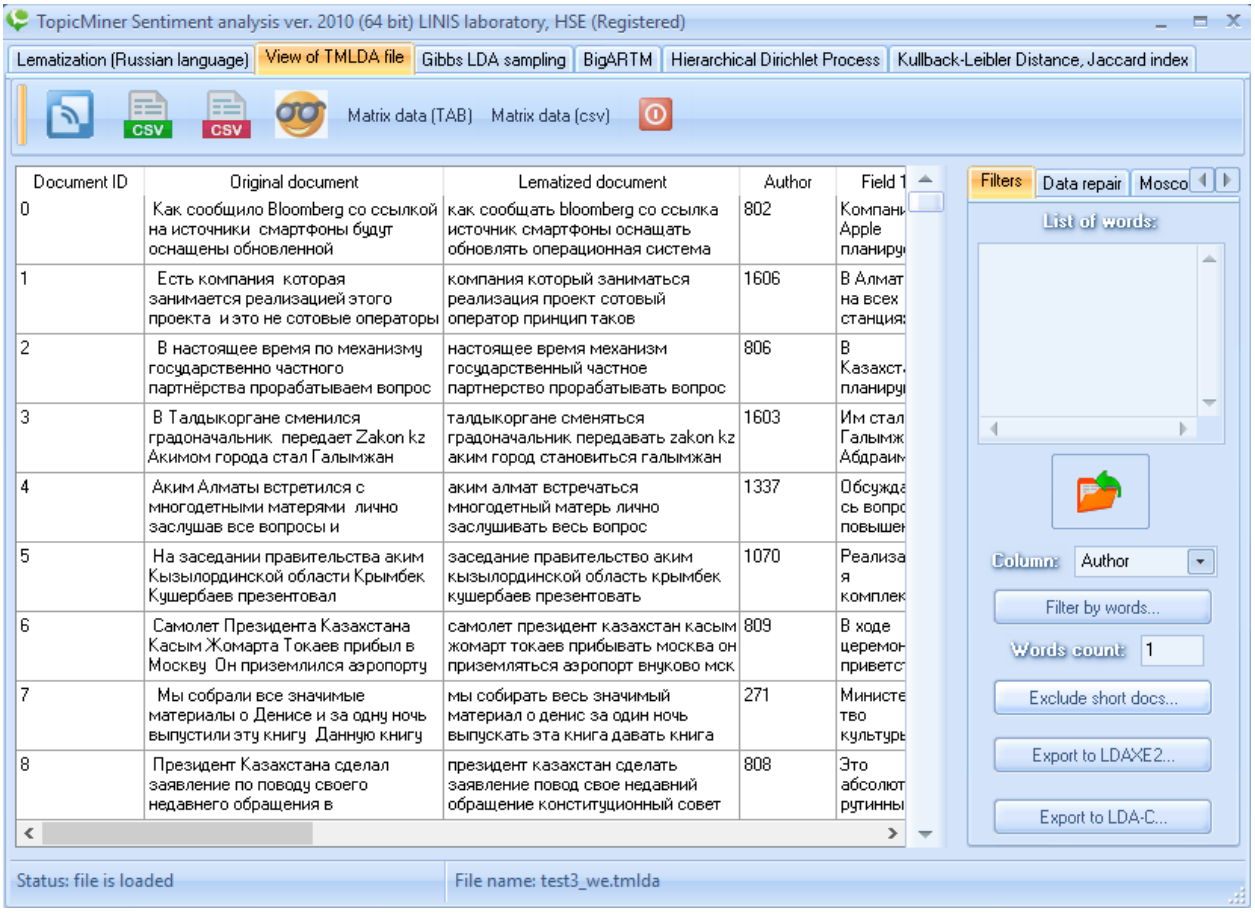
В результате препроцессинга могут получатся пустые документы (например, по причине не указанные концов строки или слова в текстах были удалены при помощи стоп слов).

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

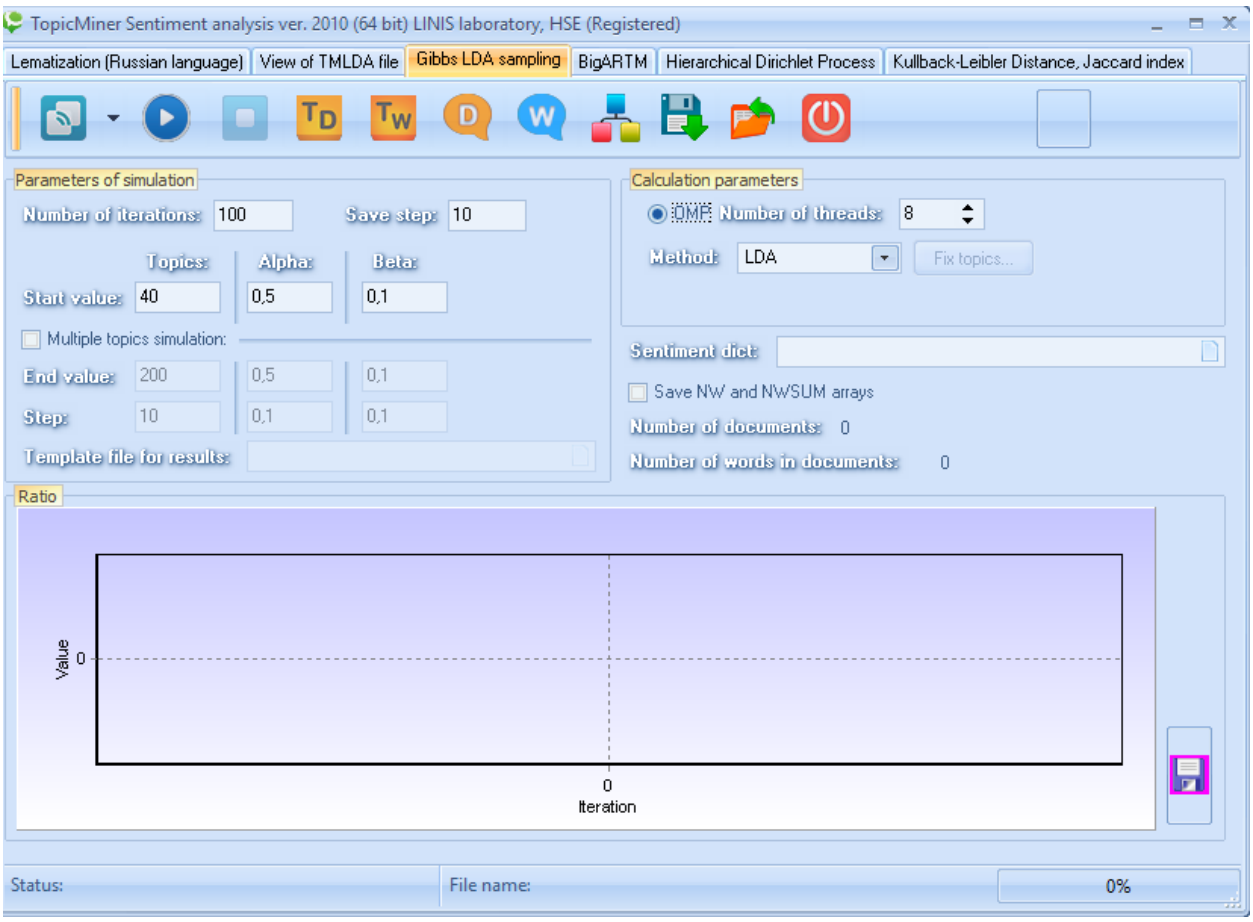


И нажмем на кнопку Exclude short dosc. В результате создаться файл tmlda (test3_we.tmlda), в котором уже не будет этих документов. Например, загрузите файл test3_we.tmlda и увидите

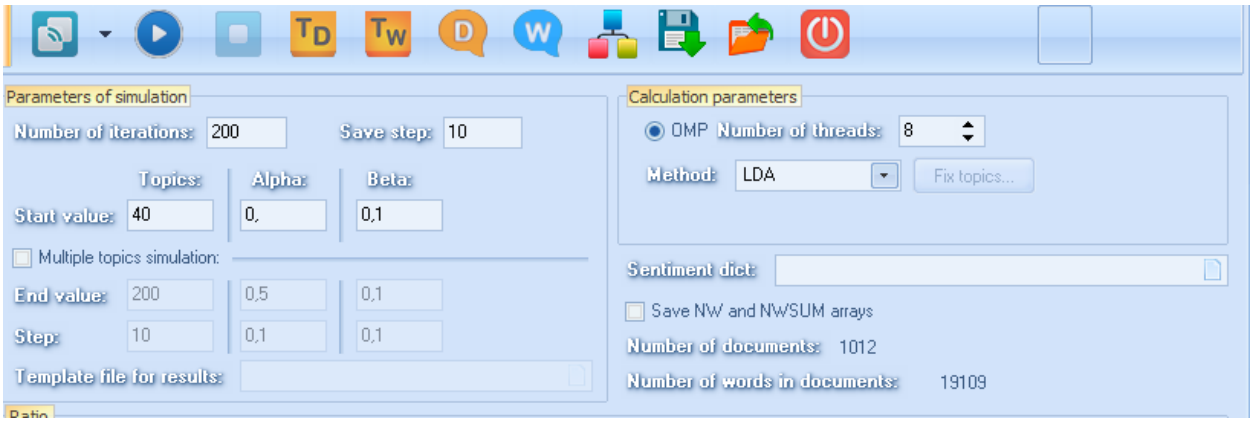


3. **Тематическое моделирование.**

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

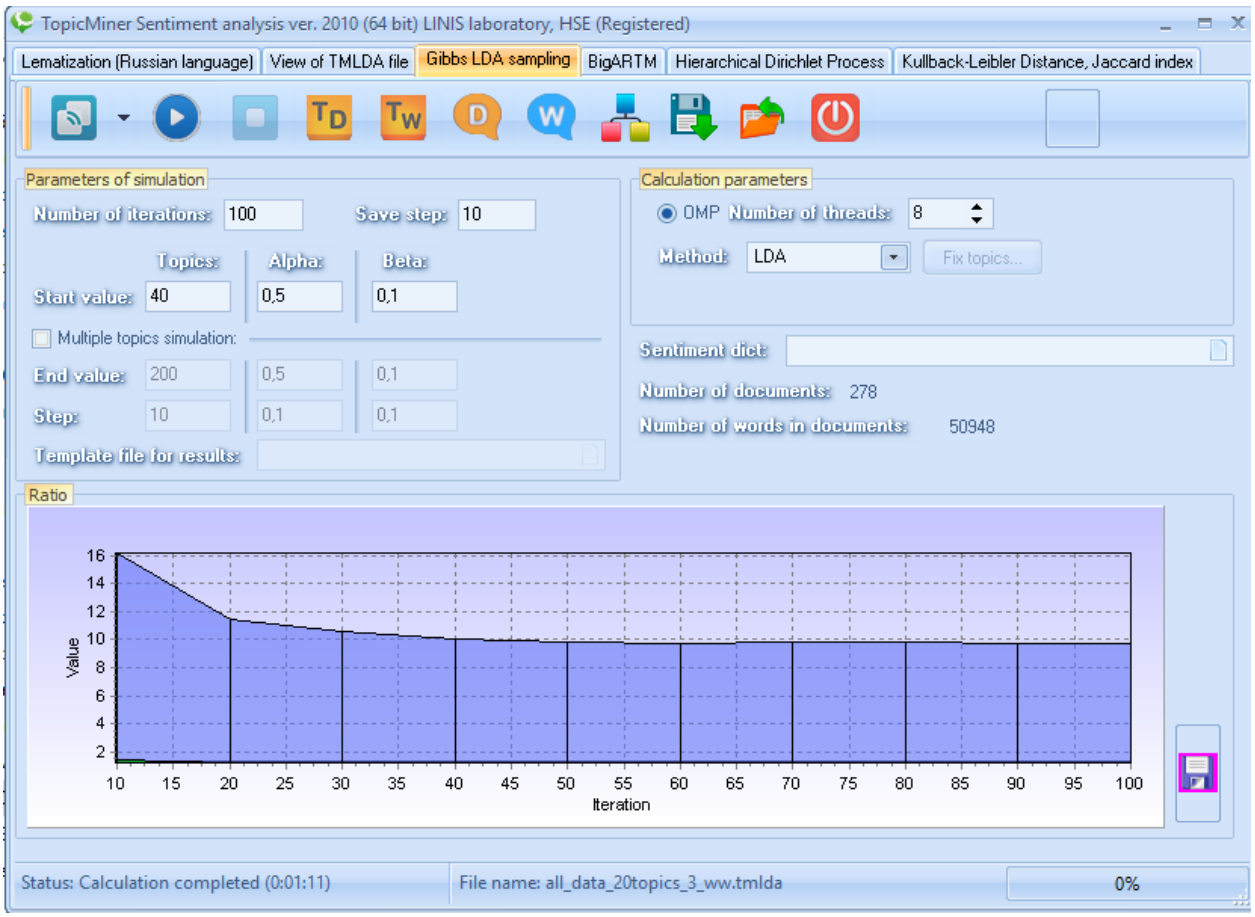


Нужно загрузить файл после препроцессинга и удаления пустых локументов.
Например загрузим файл test3_we.tmlda. Не обращайте внимание на процент загрузки (в текущей версии это просто показатель длины коллекции).
Далее укажите параметры модели (число тем, количество итераций, шаг визаулизации, и параметры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



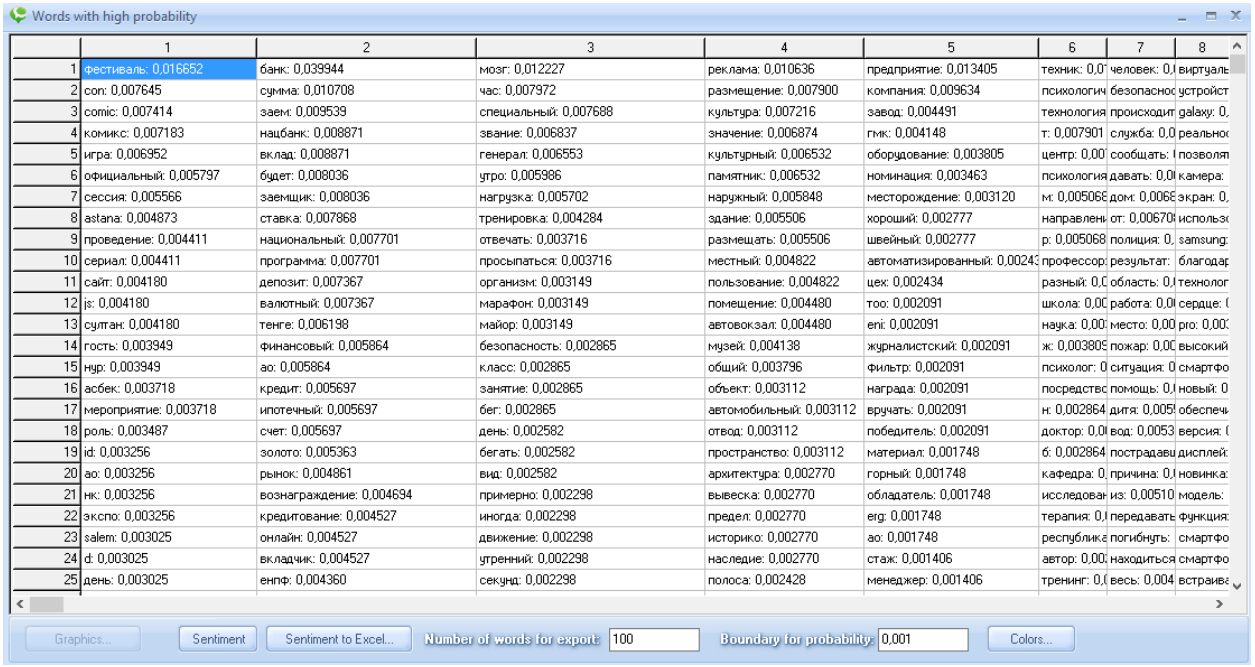
И нужно выбрать модель, например LDA.

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

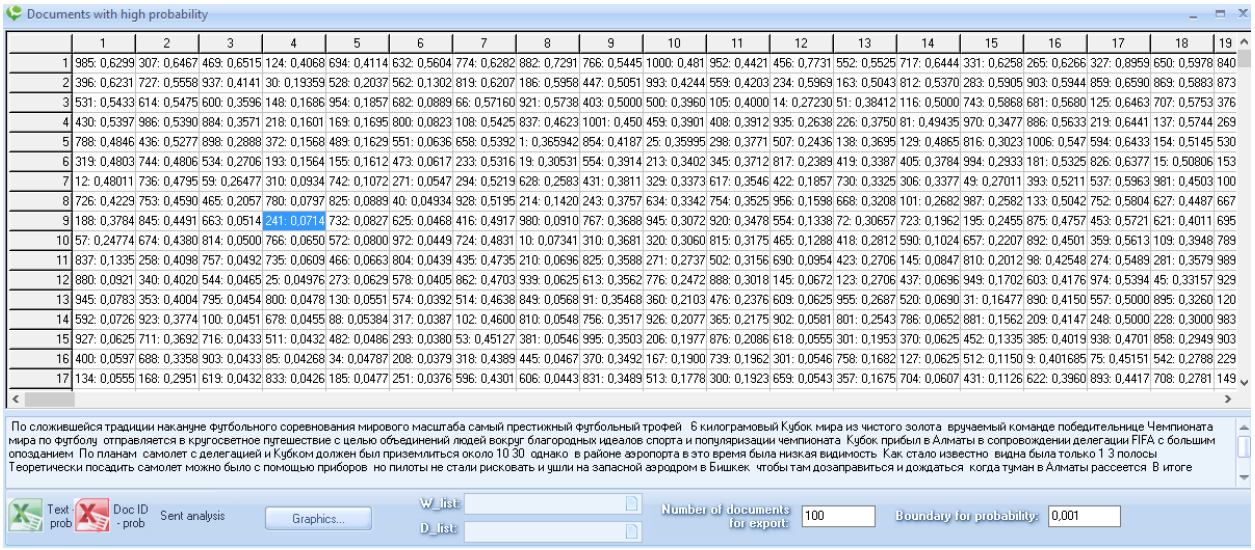


Синяя линия – это процент слов с высокой вероятностью. Из графика видно, что после 100 итерации, процент слов перестает меняться, что значит дальнейшее увеличение итераций не нужно делать. Зеленая линия это процент документов с высокой вероятностью.

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



Если хотите почитать наиболее вероятностные документы в какой либо теме, то можно нажать на кнопку  , и увидите следующее



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  .