НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Лаборатория интернет-исследований

ПРЕПРОЦЕССИНГ РУССКОЯЗЫЧНЫХ ТЕКСТОВ В

Кольцов С.Н

Санкт-Петербург, 2018

ПРЕПРОЦЕССИНГ

Препроцессинг состоит из следующих этапов

1. **Загрузка данных из источника данных**, например csv файл.
2. **Процедура лематизации**. Данная процедура заключается в процессе токинизации текстов.
3. **Удаление стоп слов**. Некоторые слова являются общими и при этом часто встречаются, поэтому их необходимо удалять.
4. **Визуализация текстовых данных в виде облака слов**. Наиболее частотные слова можно представить в виде облака слов, что упрощает некоторое понимание того какие темы внутри коллекции документов.

В дальнейшем нам понадобятся следующие пакеты, которые необходимо установить в Rstudio: **tm, wordcloud, stringr**

ЗАГРУЗКА ДАННЫХ ИЗ CSV ФАЙЛА

Установим пакет tm.

library(tm)

Далее загрузим документы из файла при помощи команды:

text = readLines(file.choose())

Внимание, файл должен быть в кодировке ANSII.

В появившемся окне нужно выбрать файл. В качестве такого файла будем использовать файл etnic3.csv. Данный файл содержит 500 текстов про различные этничности.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	СНОУБОРД Алена Заварзина стала третьей на австрийском этапе Кубка мира в параллельном слаломе. Заварзина в малом финале опередила норвежку Хильде Катри																	
2	БИАТЛОН Кайса Мякяряйнен за евала золото в гонке пресле ва я на 1 0 км на пер м этапе Кубка мира по биатло в шведском Эстерсунде, россиянка Екатерина Глазырина з																	
3	ТУПЫЕ РУСАКИНЕ ЗНАЮТ, ЧТО ЭТО ФЛАГ ЧЕЧ И НАЗЫВАЮТ ЕГО ТУРКМЕНСКИМ/КЫРГЫЗСКИМ/ТАТАРСКИМ ФЛАГОМ #Многонациональный_бугурт s upload.wikimedia.org/w																	
4	? ? Булга?ры, болга?ры (лат. Bulgares, греч.?о?????о?, болг. прабългари, протобългари, тат. Bol?arlar, чуваш.П?лха?рсем) — племена, населявшие cIV века степи Северного При																	
5	бляди хуйбдулкеримовские... и чуркамерзопакостная, если бы его еба ю же шку судейку сб , дали бы 2 0 лет строгача с конфискацией имущества...																	
6	Вкусна жидаиз Т рога Овсянки Банана Какао ды #спортпит #ЗОЖ #правильное пита е #мир лжензнатьчтояем #Instafood #alexfitness_kzn																	
7	Грёбаный стыд. Хохлысварган ответку на ролик Я - русский оккупант . , как сделали. Обосрались по традиции. Хотя бы в том, что зачем-то вытащ на божий свет давно разоблач																	
8	Дагестанцыне хотят саморазвиваться и самопросвещаться - совсем не согласна. Хотят и нас много) зачем за всех Даговго рить?																	
9	Дерьмо кратию придумали ры жиды,чтоб брать и красть что хочется и у кого хочется ,пример Ливия ,Ирак ,Сирия ,Украина ,Россия .После енного пере рота в России были по																	
10	Ёбаные хохлы заебали о по ящику показывают, здесь обсуждают скоро сне буду видеть бляди																	
11	Ебать у хохловприпекает. Четко о с москалями борются от име жителей Чеч . Из пяти косарей писчиков 9 человек из Грозного. Остальные од хохлы Лол, а чего не на АТО?																	
12	Лучшая из украинокзаняла 1 2 -е место на этапе Кубка мира по биатло Невероят ю гонку выдали девушки в спринте Поклюки!																	

ЗАГРУЗКА ДАННЫХ ИЗ CSV ФАЙЛА

После этого конвертируем набор текстов в Corpus. Corpus представляет собой контейнер для хранения текстов.
articles = Corpus(VectorSource(text))

Для того что бы просмотреть конкретный текст нужно набрать команду:
print(articles$content[13])
Результат:
[1] " Лучше нашего Дагестанского (даргинца комментатора постав бы (кто то поня л о ком я). ?? ?? тогда этот бой был бы зрелищным.“

Все наши тексты сохранены в формате: с использованием падежей, пунктуации, строчны х и прописных букв, и прочих частей речи. Все эти элементы
текста несут смысл в контексте, но мешают при построения облака слов.
Библиотека tm содержит инструменты по удалению лишних пробелов, приведению всех букв к строчному виду, удалению цифр, знаков пунктуации и
стоп-слов. Однако удаление стоп слов плохо работает (по крайней мере в Windows).
Поэтому лучше использовать другой софт для удаления стоп слов.

СОЗДАНИЕ ОБЛАКА СЛОВ

Очистка текста в контейнере Corpus, реализована при помощи **tm_map**.

articles <- tm_map(articles, stripWhitespace)
articles <- tm_map(articles, tolower)
articles <- tm_map(articles, removeNumbers)
articles <- tm_map(articles, removePunctuation)

Далее загружаем библиотеку **wordcloud** при помощи команды:

library(wordcloud)

Теперь можно построить облако наиболее частотных слов в коллекции документов.

wordcloud(articles, random.order=F, max.words=60, colors=brewer.pal(8, "Dark2"))

Результат:

Видно что в облаке присутствуют похожие слова ‘русский’ и ‘русские’, а также стоп слова вроде ‘для’ или ‘https’.

ПРОЦЕДУРА ЛЕМАТИЗАЦИИ

Пакет tm позволяет удалять стоп слова и производить процесс стеминга, однако все это хорошо работает только для англоязычных текстов.
Для того что бы выполнить процедуру лематизации для русскоязычных текстов то можно воспользоваться программой ‘mystem.exe’. Для того что бы подключить данный лематизатор и использовать для контейнера документов, нужно использовать вот такой код. Прежде всего нужно загрузить библиотеку **stringr.** Данная библиотека позволяет эффективно работать со строками.

library("stringr")

Теперь нужно указать путь к программе mystem и параметры запуска. Так как наша коллекция находится в кодировке ANSII, то для mystem необходимо передавать следующий набор параметров: **-c -wl**
Итоговый параметр:

Myfield= 'D:/Lection_R_Orange_Python/Working_russian_language/mystem.exe -c -wl'

Зеленым цветом выделен путь, а красным параметр лематизации.

ПРОЦЕДУРА ЛЕМАТИЗАЦИИ

Пакет tm позволяет удалять стоп слова и производить процесс стеминга, однако все это хорошо работает только для англоязычных текстов.
Для того что бы выполнить процедуру лематизации для русскоязычных текстов то можно воспользоваться программой ‘mystem.exe’. Для того что бы подключить данный лематизатор и использовать для контейнера документов, нужно использовать вот такой код. Прежде всего нужно загрузить библиотеку **stringr.** Данная библиотека позволяет эффективно работать со строками.

library("stringr")

Теперь нужно указать путь к программе mystem и параметры запуска. Так как наша коллекция находится в кодировке ANSII, то для mystem необходимо передавать следующий набор параметров: **-c -wl**
Итоговый параметр:

Myfield= 'D:/Lection_R_Orange_Python/Working_russian_language/mystem.exe -c -wl'

Зеленым цветом выделен путь, а красным параметр лематизации.

ПРОЦЕДУРА ЛЕМАТИЗАЦИИ

Теперь нужно реализовать процедуру запуска лематизатора:

```
mystem = function(doc) {
    sdoc = system(Myfield, intern=T, input=doc)
    sdoc <- gsub("[{}]", "", sdoc)
    sdoc <- gsub("(\\|[^ ]+)", "", sdoc)
    sdoc <- gsub("\\?", "", sdoc)
    sdoc <- gsub("\\s+", " ", sdoc)
    sdoc = paste(sdoc, collapse=" ")
    attributes(sdoc) <- attributes(doc)
    sdoc
}
```

Красным выделена процедура запуска mystem, зеленым выделен код, связанный с удалением спец символов, синим цветом код передачи обратно результата лематизации. На вход подается не лематизированный корпус документов, на выходе получаем лематизированный корпус документов.

ПРОЦЕДУРА ПРЕПРОЦЕССИНГА В TOPICMINER

Есть путь лучше и быстрее. Для этого нужно использовать пакет TopicMiner. В цело данный пакет предназначен для работы с тематическим моделированием. Но, первая часть этого пакета предназначена для быстрой процедуры препроцессинга русскоязычных и англоязычных текстов.

Препроцессинг данных в TopicMiner состоит из трех этапов.

1. Процедура лематизации.
2. Подсчет частот и удаление спец символов, которые возникают при использовании mystem.
3. Удаление стоп слов.

Однако, прежде всего нужно преобразовать коллекцию документов в формат входных данных TopicMiner. Такой входной формат представляет собой набор текстовых файлов, каждый из которых содержит один документ. Такое преобразование можно сделать при помощи следующих команд:

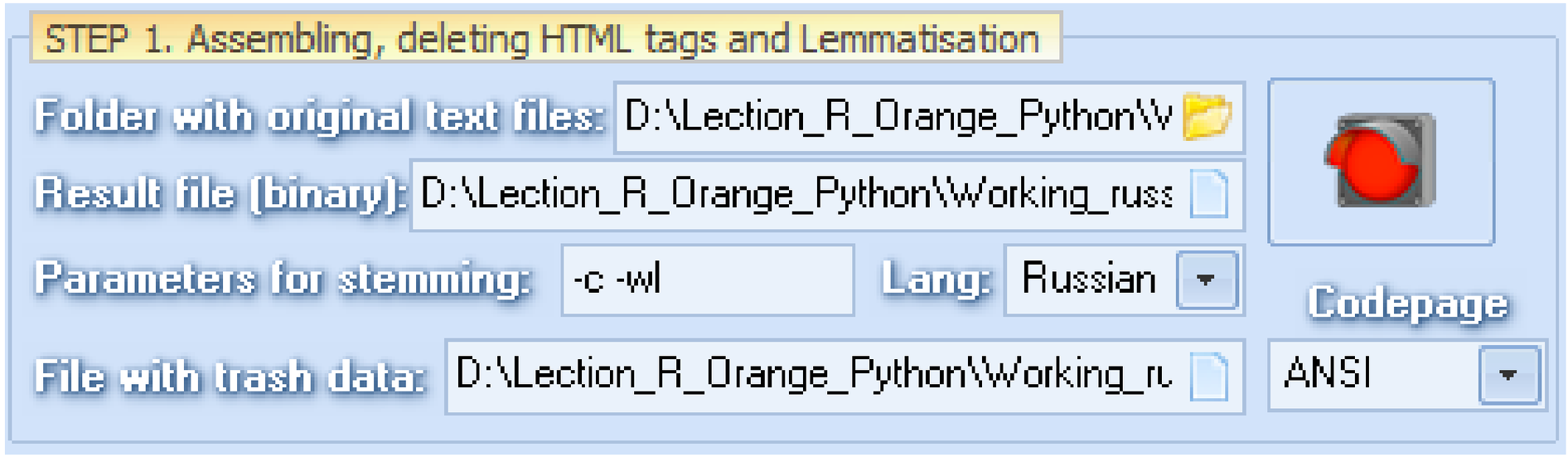
Создаем путь к каталогу для хранения коллекции файлов

path ='D:/Lection_R_Orange_Python/Working_russian_language/test_collection‘

Записываем содержимое коллекции в каталог:

writeCorpus(articles , path, filenames = paste(seq_along(articles), ".txt", sep = ""))

ПРОЦЕДУРА ПРЕПРОЦЕССИНГА В TOPICMINER

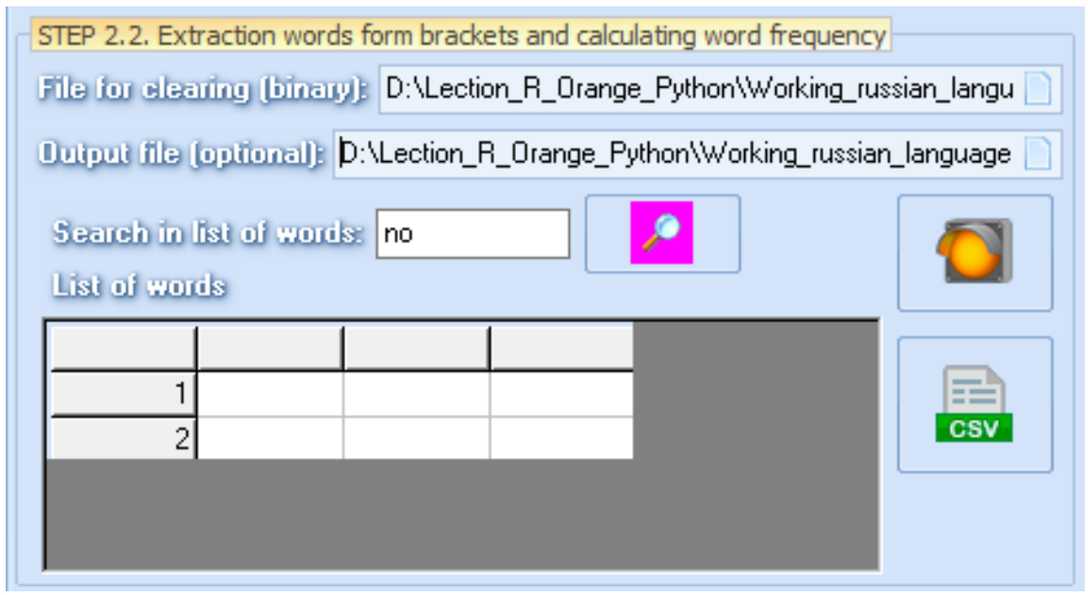


В данной опции нужно установить следующие параметры:

1. Каталог в котором лежат файлы коллекции.
2. Имя файла, в который будет записан результат после первого этапа.
3. Параметры для лематизации
4. Язык (возможны Русский или Английский языки).
5. Текстовый файл с некоторыми словами, которые будут удалятся на первом этапе.
6. Кодировка текстов (для данного примера это ANSI)

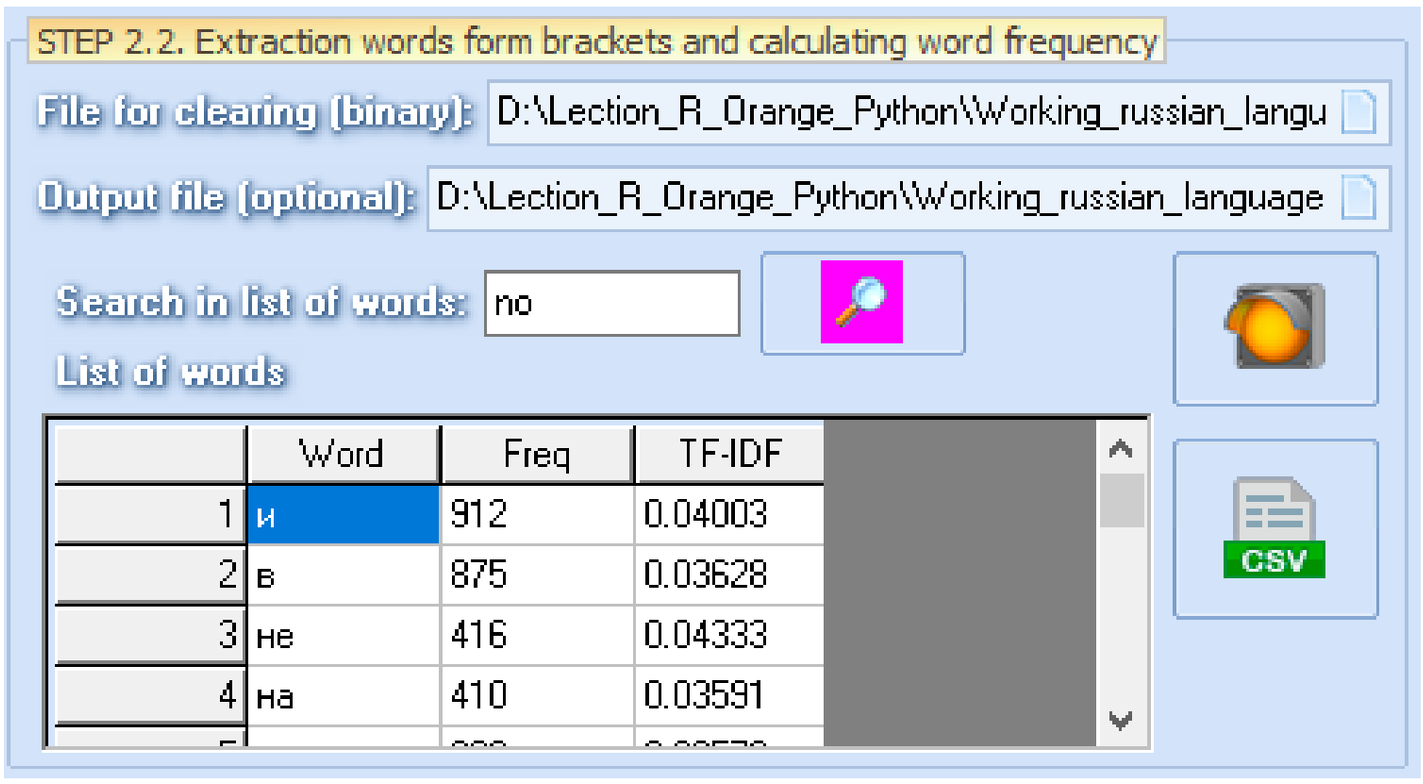
Запуск процедуры первого этапа осуществляется при помощи красной кнопки.

ПРОЦЕДУРА ПРЕПРОЦЕССИНГА В TOPICMINER



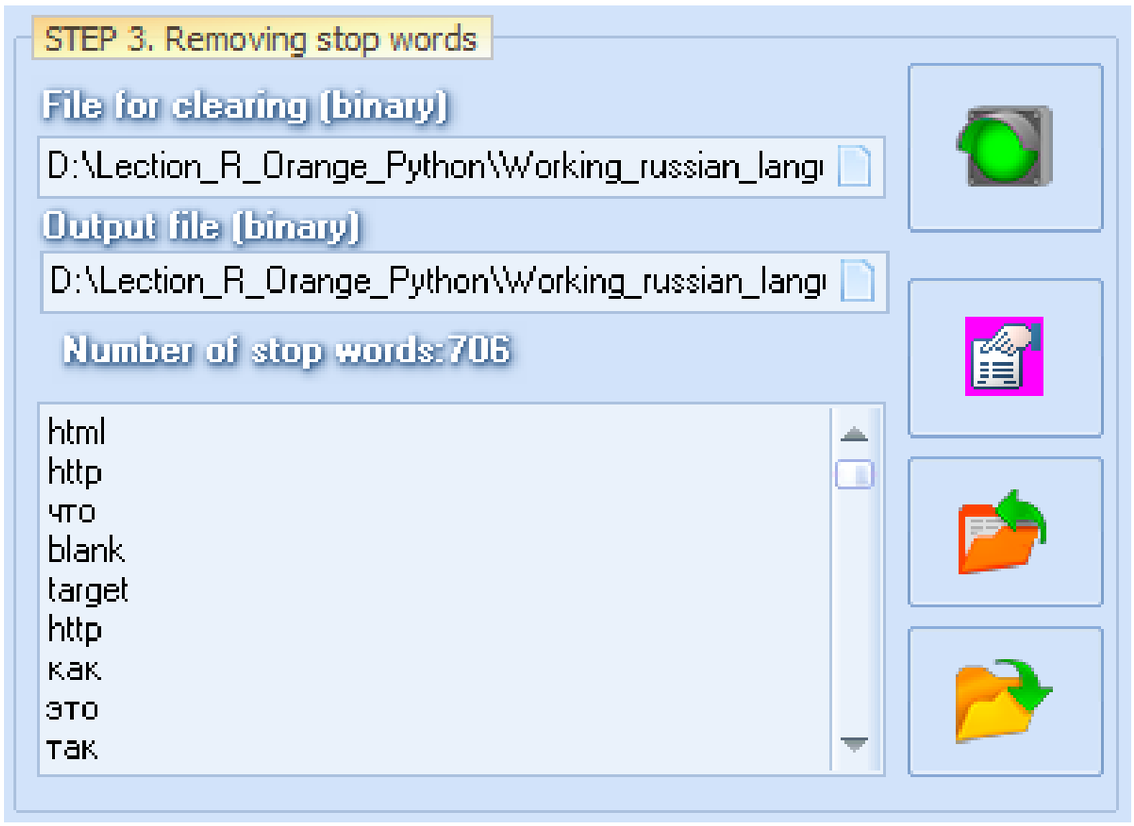
На втором этапе нужно указать файл, который получился после 1 этапа и файл, в который будут записаны результаты второго этапа.

Запуск второго этапа осуществляется при нажатии на большую желтую кнопку.



Результатам является лематизированная коллекция и список частот, которые можно сохранить в файл при помощи кнопки csv.

ПРОЦЕДУРА ПРЕПРОЦЕССИНГА В TOPICMINER



На третьем этапе препроцессинга происходит удаление стоп слов.
Для того что бы удалить стоп слова из лематизированной коллкции нужно:

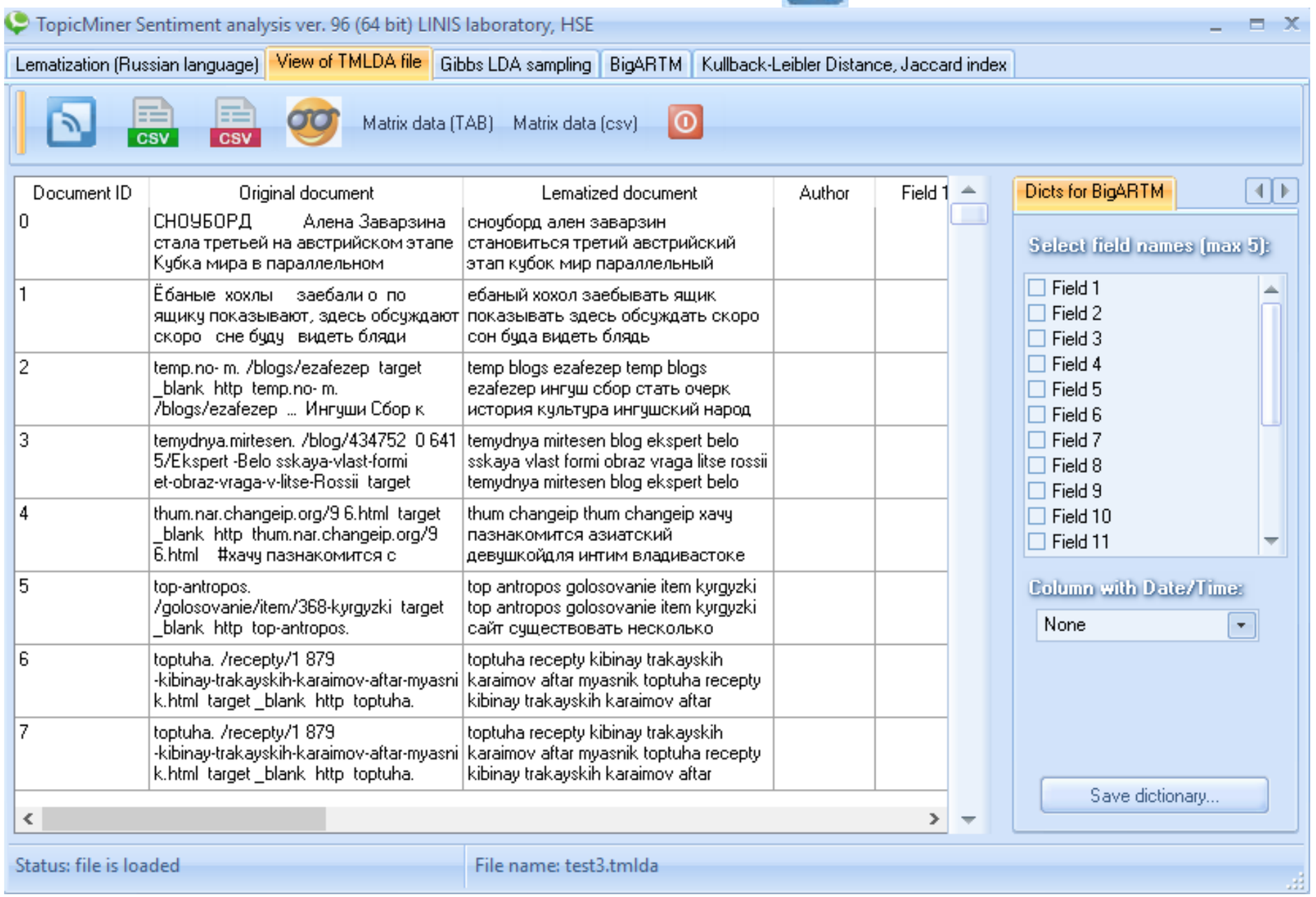
1. Указать файл после второго этапа.
2. Указать имя файла, в который сохранятся результаты работы третьего этапа.
3. Загрузить список стоп слов из внешнего текстового файла.

Процедура удаления запускается при помощи зеленой кнопки.

В результате в последнем файле будут хранится несколько вариантов коллекций (файл в формате **tmlda**): 1. Исходная коллекция. 2. Лематизированная версия коллекции. 3. Цифровая версия коллекции. В данной версии все слова заменены на crc32 коды слов.

ПРОСМОТР ДАННЫХ В TOPICMINER

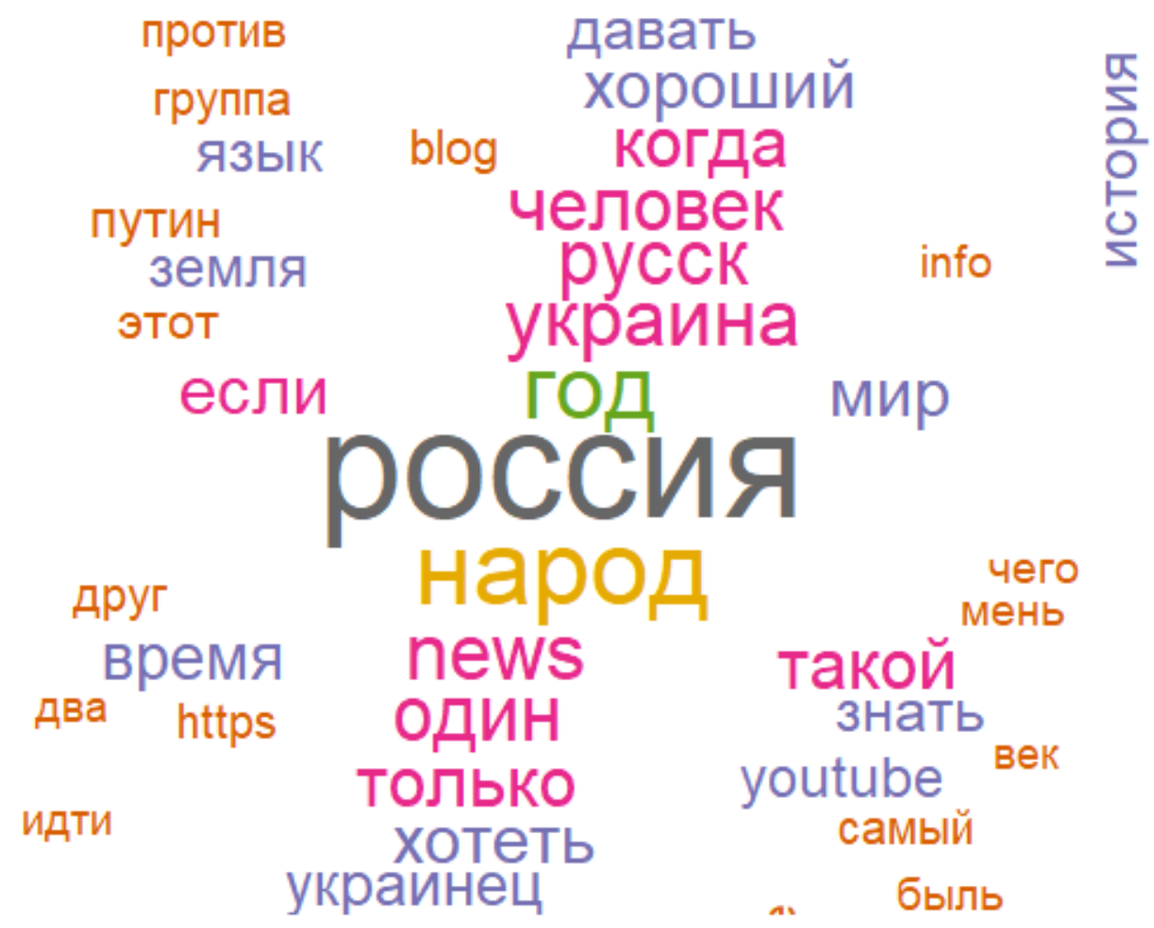
Результаты работы модуля препроцессинга можно посмотреть на вкладке View of TMLDA files. Для этого нужно нажать на кнопку  и загрузить файл после третьего этапа.



Чтобы выгрузить лематизированные документы в файл csv, нужно выбрать красную кнопку csv и указать файд. Полученный файл можно загрузить в Rstudio и построить облако слов. Только нужно потом файл csv преобразовать в кодировку ANSI.

ОБЛАКО ЛЕМАТИЗИРОВАННЫХ СЛОВ

```
text = readLines(file.choose())
articles  = Corpus(VectorSource(text))
wordcloud(articles, random.order=F, max.words=20,  colors=brewer.pal(8, "Dark2"))
```



НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ

https://linis.hse.ru/

Phone: +7 (911) 981 9165

Email: skoltsov@hse.ru