# Data Mining

Кольцов С.Н.

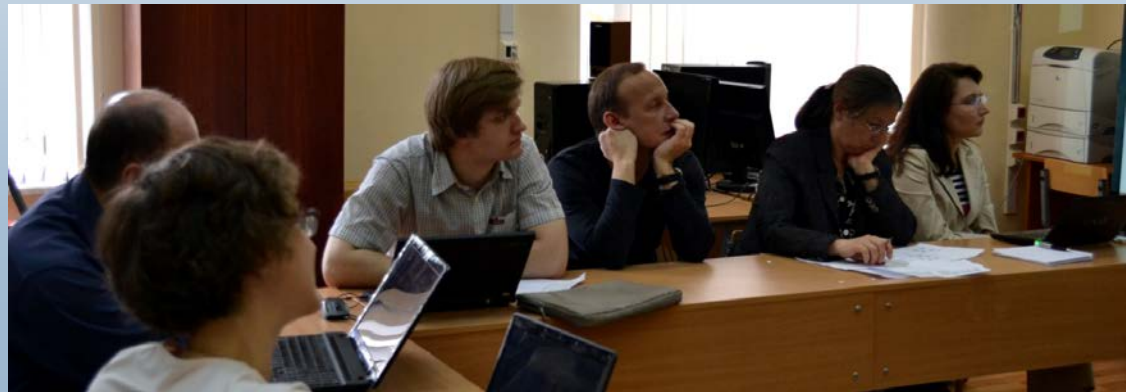
# Междисциплинарная лаборатория интернет-исследований НИУ ВШЭ.

- Big data - серия подходов, инструментов и методов обработки структурированных и неструктурированных данных огромных объёмов.

Методы анализа данных.

1. Тематическое моделирование (topic modeling).
2. Методы кластеризации.
3. Регрессионный анализ.
4. Нейронные сети.
5. Анализ тональности текста (sentiment analysis).
6. Исследование социальных сетей

при помощи теории графов (Network analysis).



# ЧТО ТАКОЕ DATA MINING?

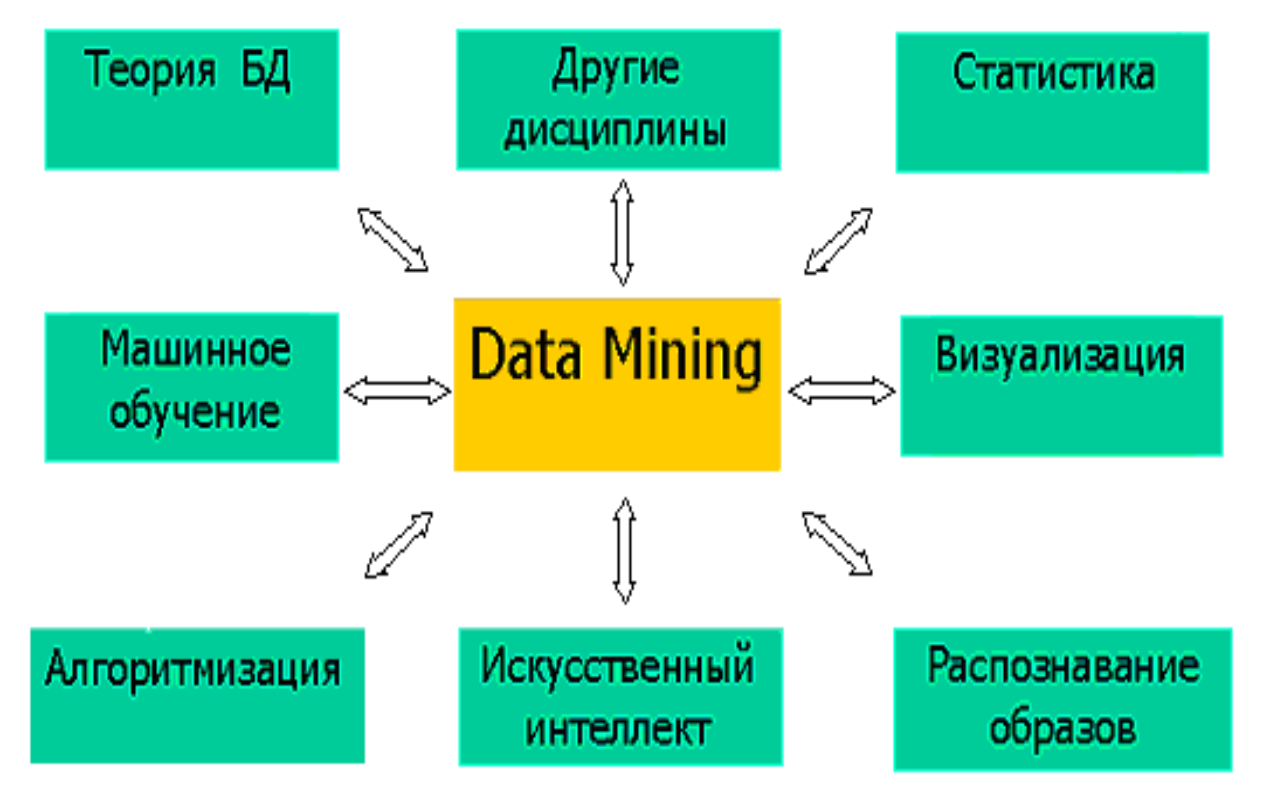
**Data Mining** - мультидисциплинарная область, возникшая и развивающаяся на базе таких наук как прикладная статистика, распознавание образов, искусственный интеллект, теория баз данных и др..

Термин **Data Mining** часто переводится как добыча данных, извлечение информации, раскопка данных, интеллектуальный анализ данных, средства поиска закономерностей, извлечение знаний, анализ шаблонов, "извлечение зерен знаний из гор данных", раскопка знаний в базах данных, информационная проходка данных, "промывание" данных.

**Data Mining** - это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Data Mining находится все еще на ранней стадии развития.
Многие IT-команды увлеклись мифом о том, что средства Data Mining просты в использовании. Предполагается, что достаточно запустить такой инструмент на терабайтной базе данных, и моментально появится полезная информация. На самом деле, успешный Data Mining - проект требует понимания сути деятельности, знания данных и инструментов, а также процесса анализа данных".

# ЧТО ТАКОЕ DATA MINING?




1. Data Mining технология не может заменить аналитика.
2. Технология не может дать ответы на те вопросы, которые не были заданы.
3. Извлечение полезных сведений невозможно без хорошего понимания сути данных

Необходим тщательный выбор модели и интерпретация зависимостей или шаблонов, которые обнаружены в данных.

# КРИТИЧЕСКИЙ ВЗГЛЯД НА **BIG DATA**

1. **Технология**: максимально эффективное использование вычислительных мощностей и точности алгоритмов для сбора, анализа, связи и сравнения больших массивов данных.

2. **Анализ**: опора на большие массивы данных, которые позволят создать образцы, на их основе можно будет делать экономические, социальные, технические и правовые утверждения.

3. **Мифология**: распространённое убеждение, что большие массивы данных приводят к более высокому уровню понимания и новым знаниям, а благодаря этому станет возможно получить ценную информацию, недоступную ранее. Это, в свою очередь, создаёт впечатление правдивости, объективности и точности.

***Big Data** – информационные ресурсы, характеризующиеся большим объемом, высокой скоростью передачи и разнообразием, которые требуют экономичных, инновационных форм переработки информации для улучшения её качества и оптимизации процесса принятия решений.*

# МЕТОДЫ DATA MINING

Особенностью Data Mining является сочетание широкого математического инструментария (от классического статистического анализа до новых кибернетических методов) и последних достижений в сфере информационных технологий. В технологии Data Mining гармонично объединились строго формализованные методы и методы неформального анализа, т.е. количественный и качественный анализ данных.

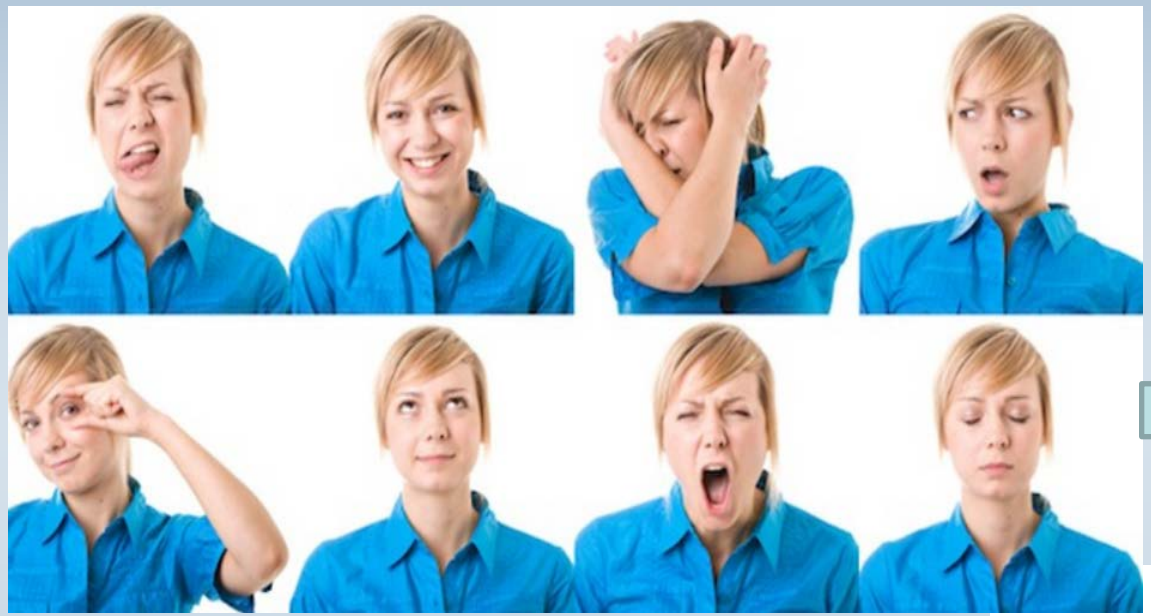
**К методам и алгоритмам Data Mining относятся** следующие: искусственные нейронные сети, деревья решений, символьные правила, методы ближайшего соседа и k-ближайшего соседа, метод опорных векторов, байесовские сети, линейная регрессия, корреляционно-регрессионный анализ; иерархические методы кластерного анализа, неиерархические методы кластерного анализа, в том числе алгоритмы k-средних и k-медианы; методы поиска ассоциативных правил, эволюционное программирование и генетические алгоритмы, различные алгоритмы на основе теории графов, также разнообразные методы визуализации данных.

**В рамках данного курса будут рассмотрены следующие направления.**

1. Анализ тональности текста (Sentiment analysis или Opinion mining)
2. Кластерный анализ (K-Means Cluster, Hierarchical clustering)
3. Тематическое моделирование (Topic modeling)
4. Анализ сетей (network analysis)
5. Регрессионный анализ (regression analysis)
6. Нейронные сети
7. Модели анализа прогнозирования

# DATA MINING (ЭЛЕМЕНТЫ)

# Анализ тональности текста



Сантимент-анализ — это раздел дата майнинга, система автоматического извлечения субъективных мнений из текста, дисциплина на стыке поиска информации и вычислительной лингвистики, которая исследует не столько содержание текста, сколько его тональность. При этом тональность выражается в числовом виде, соответственно можно строить различные функциональные зависимости

# Что такое тональность текста и зачем она нужна?

Под определением тональности текста подразумевается выделение фрагментов текста, выражающих позитивную или негативную эмоциональность по отношению к объекту этой эмоциональной оценки (объекту тональности).

Объектом может выступать как имя собственное, так название продукта (например наименование бренда), организации (например Газпром), услуги, профессии и т. п. по отношению к которому анализируется текст.

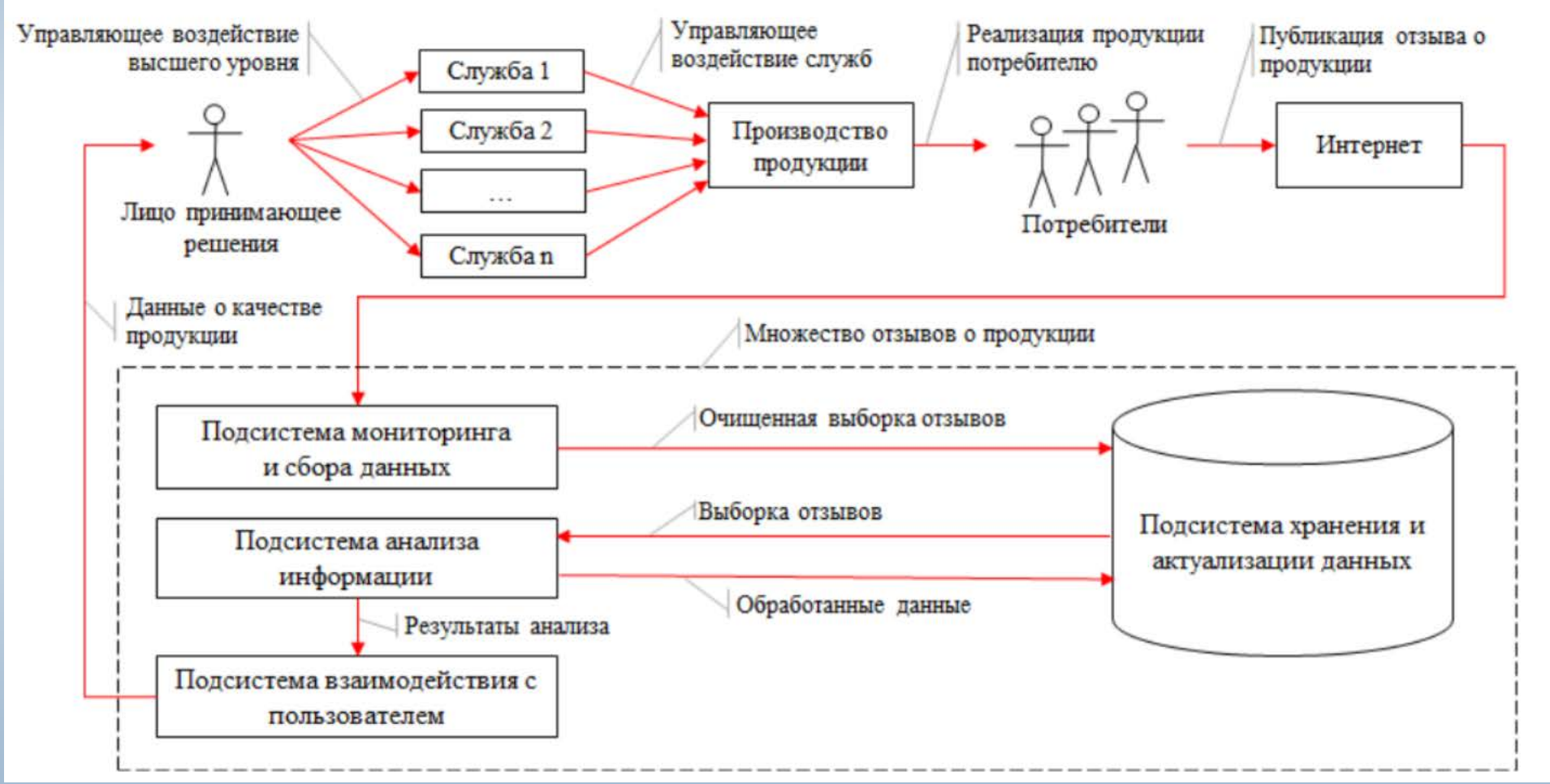
Таким образом, тональность текста определяется такими тремя факторами:

1. **Субъект тональности** – это автор текста, рецензии, комментария и т.п.
2. **Тональная оценка** (позитив/нейтрально/негатив) – это эмоциональное отношение автора к объекту тональности, выраженное в использовании специфической лексики при написании текста.
3. **Объект тональности** — продукт, событие, персона или явления, о которых высказывается автор

Оценка тональности текстов предназначена для получения обратной связи между потребителями услуги и производителями услуги. Например:

1. Оценка деятельности госучреждения в СМИ и в социальных сетях…
2. Возможность предсказания котировок (в качестве дополнительного индекса) на фондовых рынках.
3. Мониторинг экстремистских высказываний в СМИ, интернете.

# Пример системы с обратной связью



# Обзор моделей определения тональности текста

**Метод автоматического анализа текста, с машинным обучением;**

1) на основе правил с использованием шаблонов (rule-based with patterns) . Подход состоит в создании правил, на основе которых будет осуществляться определение тональности текста. Для этого текст разбивается на слова или n-граммы. Далее полученные данные используют для того, чтобы выделить часто встречающиеся шаблоны, которым присваивается положительная или отрицательная оценка.

2) машинное обучение с учителем (supervised learning). В этом подходе требуется наличие обучающей коллекции текстов, заранее размеченных в рамках эмотивного пространства. На основании этой коллекции строится статистический или вероятностный классификатор* (например, байесовский);

3) машинное обучение без учителя (unsupervised learning). Данный подход основан на идее, что наибольший вес в тексте имеют термины, которые чаще встречаются в этом тексте и в то же время присутствуют в небольшом количестве текстов всей коллекции. Выделив данные термины и определив их тональность, можно сделать вывод о тональности всего текста;

**Поиск эмотивной лексики в тексте по заранее составленным тональным словарям**

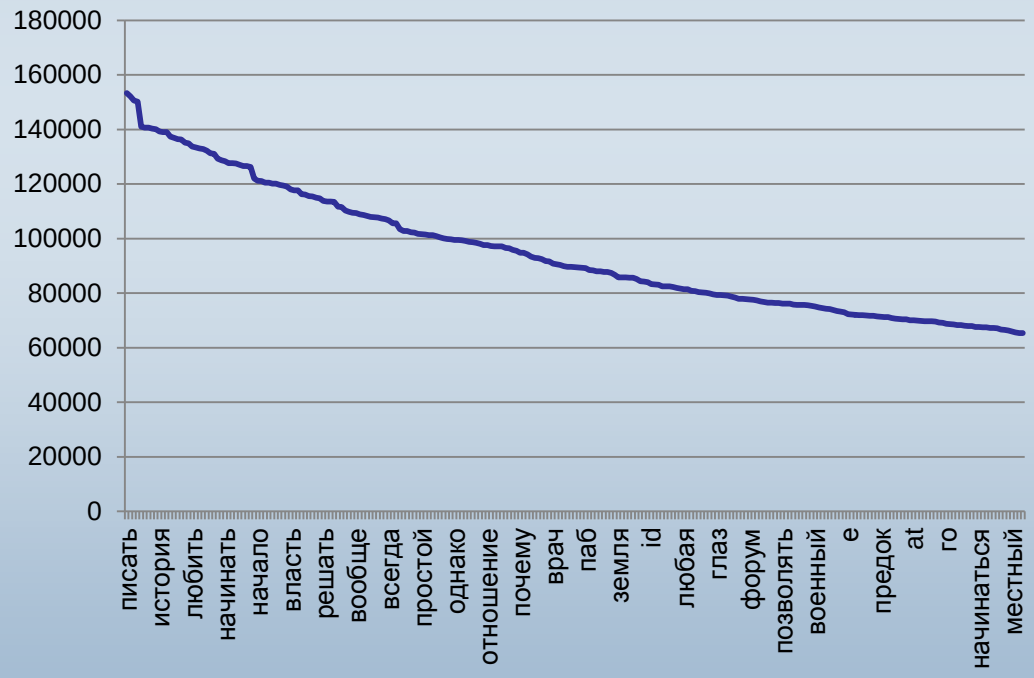
Словарь - это набор слов-маркеров, на присутствие которых в тексте будет реагировать счетчик слов. Если в списке у слова будет положительная оценка, то и текст получит положительную оценку и наоборот. Слова в словаре соотносятся с определенной шкалой эмоциональности.

# Векторная модель текста

В векторной модели осуществляется представление каждого документа коллекции в качестве вектора в векторном пространстве

Построение вектора происходит следующим образом: текст представляется в виде множества слов, каждому из которых соответствует вес (частота встречаемости этого слова в тексте).

| | A | B | C |
|---|---|---|---|
| 70 | хотеть | 290391 | |
| 71 | становить | 277137 | |
| 72 | раз | 276884 | |
| 73 | лицо | 273050 | |
| 74 | тот | 268222 | |
| 75 | страна | 257995 | |
| 76 | пример | 257285 | |
| 77 | больной | 256870 | |
| 78 | день | 255909 | |
| 79 | после | 253643 | |
| 80 | тех | 248496 | |
| 81 | чем | 247507 | |
| 82 | пост | 244468 | |
| 83 | говорить | 240850 | |
| 84 | самый | 236324 | |
| 85 | новое | 232908 | |
| 86 | мало | 231773 | |



# Векторная модель текста

Коллекцию документов предоставляется в виде матрицы термин-документ. Строки будут обозначать отдельные документы (тексты), а колонки – словарь выборки, заключающий в себе все слова в коллекции документов.

Например

- *‘Jane likes coffee and tea’*
- *‘Jane also likes cookies’*

Таблица: матрица термин-документ

| | ***Jane*** | ***likes*** | ***coffee*** | ***and*** | ***tea*** | ***also*** | ***cookies*** |
|---|---|---|---|---|---|---|---|
| **1ый текст** | **1** | **1** | **1** | **1** | **1** | **0** | **0** |
| **2ой текст** | **1** | **1** | **0** | **0** | **0** | **1** | **1** |

Таким образом, получаются бинарные вектора:
$d_1$ = [1,1,1,1,1,0,0] и $d_2$ = [1,1,0,0,0,1,1].

# Препроцессинг текста

Для того, чтобы векторная модель документа наиболее точно отображала тематику текста, необходимо каждое слово в тексте преобразовывать в уникальный терм и только потом считать статистику употребления терма.

**Терм** – это значимая часть слова, очищенная от незначащих аффиксов (т.е. словообразовательных морфем, которые присоединяются к корню слова). Для этого к текстам применяются процессы стемминга и лемматизации.

**Лемматизация** — процесс привода словоформы к лемме – её словарной форме.

**Стемминг** – это процесс нахождения основы слова для заданного исходного слова, то есть отсечение от слова окончаний и суффиксов. При преобразовании текста в векторную модель может также происходить удаление незначимых символов (например, знаков препинания).

После составления списка термов документа производится подсчет статистики употребления слов. Также для экономии времени вычислений в цикле подсчета статистики происходит проверка, входит ли терм в список стоп-слов.

**Стоп-слова** – слова, не несущие в себе смысла и не влияющие на его тематику. Такими словами чаще всего являются предлоги, союзы и другие части речи, предназначенные для связи слов в предложениях.

# Обзор математических моделей определения тональности текста

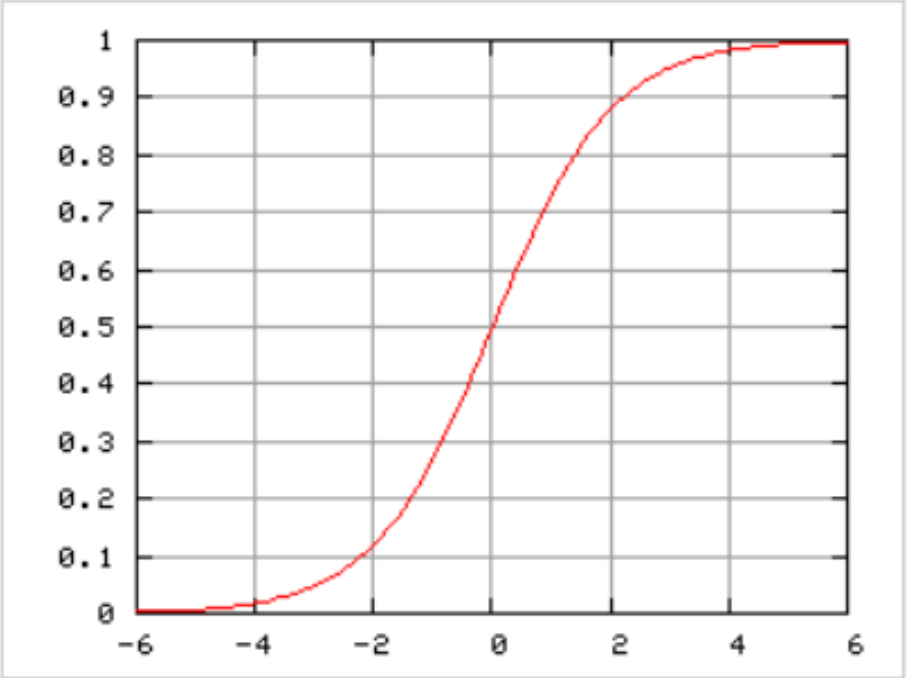
**Логистическая Регрессия** (Logistic Regression)

Логит-регрессия (логистическая регрессия) – это статистическая модель, используемая для предсказания вероятности некоторого события по значениям множества признаков. Для этого вводится так называемая зависимая переменная, принимающая два значения: 0 (текст окрашен негативно) или 1 (текст окрашен положительно). Также вводится множество независимых переменных (отдельные слова из отзыва), на основе которых вычисляется вероятность принятия того или иного значения зависимой переменной. Делается предположение, что вероятность наступления события:

$$\Pr\{y = 1|x\} = f(z)$$

где

$$f(z) = \frac{1}{1 + e^{-z}}$$




Логистическая функция: $f(x) = \frac{1}{1 + e^{-x}}$.

Вероятность рассчитывается так:

$$\ln\left(\frac{p}{1-p}\right) = a * word_1 + b * word_2 + \cdots$$

Если $\Pr\{y = 1|x\} > 0.5$,то документ считается положительно окрашенным, иначе – негативно.

# Обзор математических моделей определения тональности текста

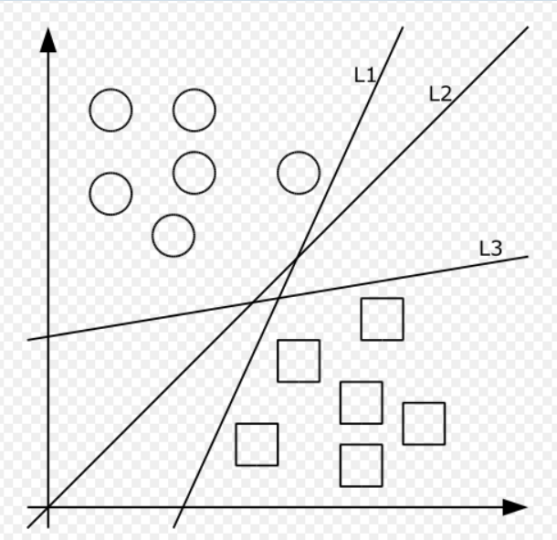
**Метод опорных векторов**

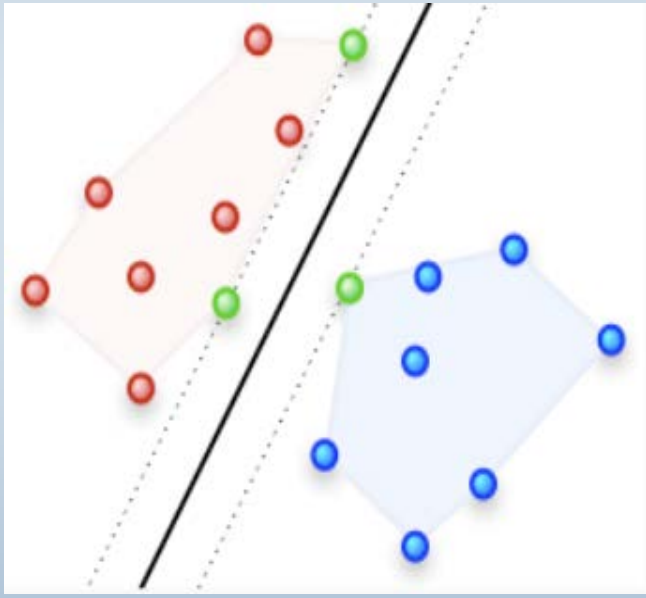
Каждый объект данных (документ) представлен как вектор в **P** мерном пространстве (последовательность чисел). Каждая из этих точек принадлежит только одному из двух классов. Точка (документ) представлена в виде: $\{(\mathbf{x}_1, c_1), (\mathbf{x}_2, c_2), \ldots, (\mathbf{x}_n, c_n)\}$

где константы с равны либо 1 либо -1 (позитив - негатив)

$\mathbf{w} \cdot \mathbf{x} - b = 0.$ - уравнение разделяющей плоскости

Вектор **w** — перпендикуляр к разделяющей гиперплоскости.



Процесс расчета заключается в поиске двух опорных векторов (b =1 или -1) при этом ищется максимум расстояния между опорными векторами. Расстояние задано в виде некой функции, например

$$\rho(x, x') = \sqrt{\sum_i^n (x_i - x_i')^2}$$

# Обзор математических моделей определения тональности текста

подход
со словарем

**Подходы, основанные на словарях**, используют так называемые тональные словари (affective lexicons) для анализа текста. В простом виде тональный словарь представляет из себя список слов со значением тональности для каждого слова.

| слово | валентность (1-9) |
|---|---|
| счастливый | 8.21 |
| хороший | 7.47 |
| скучный | 2.95 |
| сердитый | 2.85 |
| грустный | 1.61 |

Чтобы проанализировать текст, можно воспользоваться следующим алгоритмом: сначала каждому слову в тексте присвоить его значением тональности из словаря (если оно присутствует в словаре), а затем вычислить общую тональность всего текста.

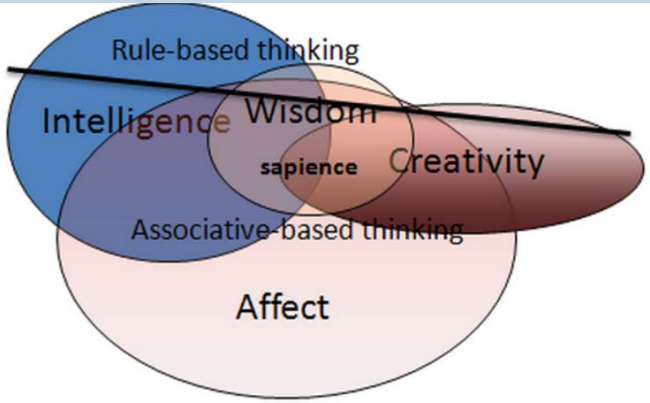
Для каждого текста из обучающей коллекции подсчитывается его вес, равный среднему весу входящих в него оценочных слов:

$$W_T = \frac{\sum_{i=1}^{N} W_i}{N}$$

# Обзор математических моделей определения тональности текста

**Rule-based with patterns** (формирование оценки тональности с учетом правил языка) **:**

Подход состоит в создании правил, на основе которых будет осуществляться определение тональности текста. Для этого текст разбивается на слова или n-граммы. Далее полученные данные используют для того, чтобы выделить часто встречающиеся шаблоны, которым присваивается положительная или отрицательная оценка. Выделенные шаблоны применяются при создании правил вида «ЕСЛИ условие, ТО заключение»;



# Качество разных моделей

| Algorithm | Accuracy | Acc. +/- 1 class | Mean % abs. Error |
|---|---|---|---|
| **SentiStrength** | 60.6% | 96.9% | 22.0% |
| logistic regression | **58.5%** | 96.1% | **23.2%** |
| SVM | **57.6%** | **95.4%** | **24.4%** |
| Classification tree | **55.2%** | **95.9%** | 24.7% |
| Rule-based classifier | **54.3%** | 96.4% | **28.2%** |

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., &Kappas, A. (2010). Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), 2544–2558

# Обзор программных средств.

Test- Download - Java Version - Non-English - Buy! - About

## SentiStrength

SentiStrength estimates the *strength* of positive and negative sentiment in *short texts*, even for informal language. It has human-level accuracy for short social web texts in English, except political texts. SentiStrength reports *two* sentiment strengths:

-1 (not negative) to -5 (extremely negative)

1 (not positive) to 5 (extremely positive)

It can also report binary (positive/negative), trinary (positive/negative/neutral) and single scale (-4 to +4) results. SentiStrength was originally developed for English and optimised for general short social web texts but can be configured for other languages and contexts by changing its input files - some variants are demonstrated below.

**Quick Tests** (English version):

Enter text: I love you but hate the current political climate. Detect Sentiment

Output: Dual, binary, trinary, scale

**Keyword test**:

I love dogs quite a lot but cats I really hate.

Enter keywords (comma-separated list, no spaces): dogs,frogs

Detect Sentiment Around Keywords

**Topic test**:

The movie was very frightening. My phone has a small screen. That is a banging chuuuun. I won the vote but he lied abo

Select domain (broad topic): Films/Movies Detect Sentiment in Domain

**Other languages**: Finnish, German, Dutch Spanish. Russian, Portuguese, French, Arabic, Polish, Persian, Swedish, Greek, Welsh, Italian, Turkish.

**http://sentistrength.wlv.ac.uk/**

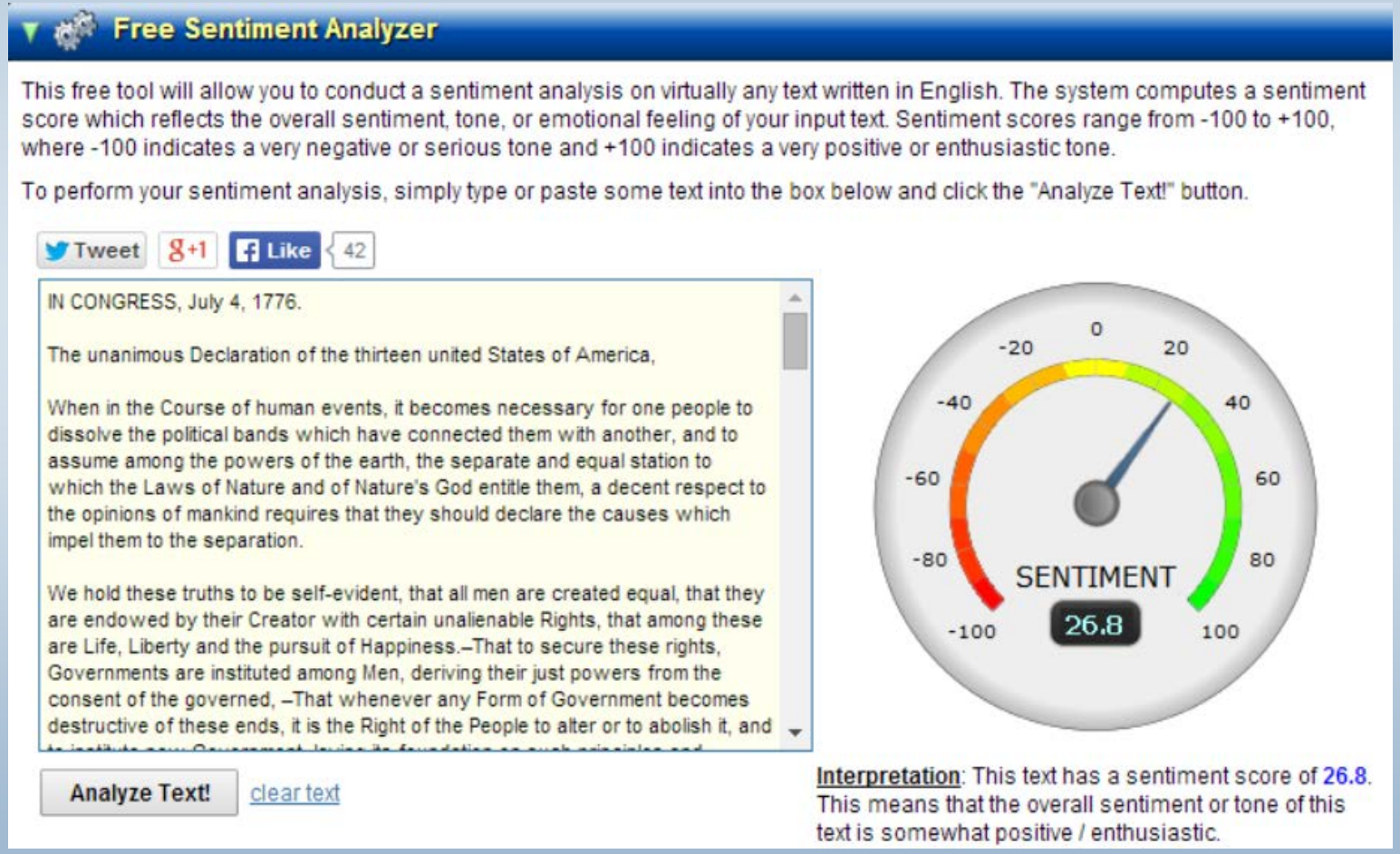
# Обзор программных средств.

## Twitter Sentiment

Пользователю Twitter Sentiment достаточно ввести слово, и программа проанализирует все последние 100 записей об этом слове. При этом будет построен график соотношения положительных и негативных отзывов.

# Обзор программных средств.

**Free Sentiment Analyzer**




**http://www.danielsoper.com/sentimentanalysis/default.aspx**

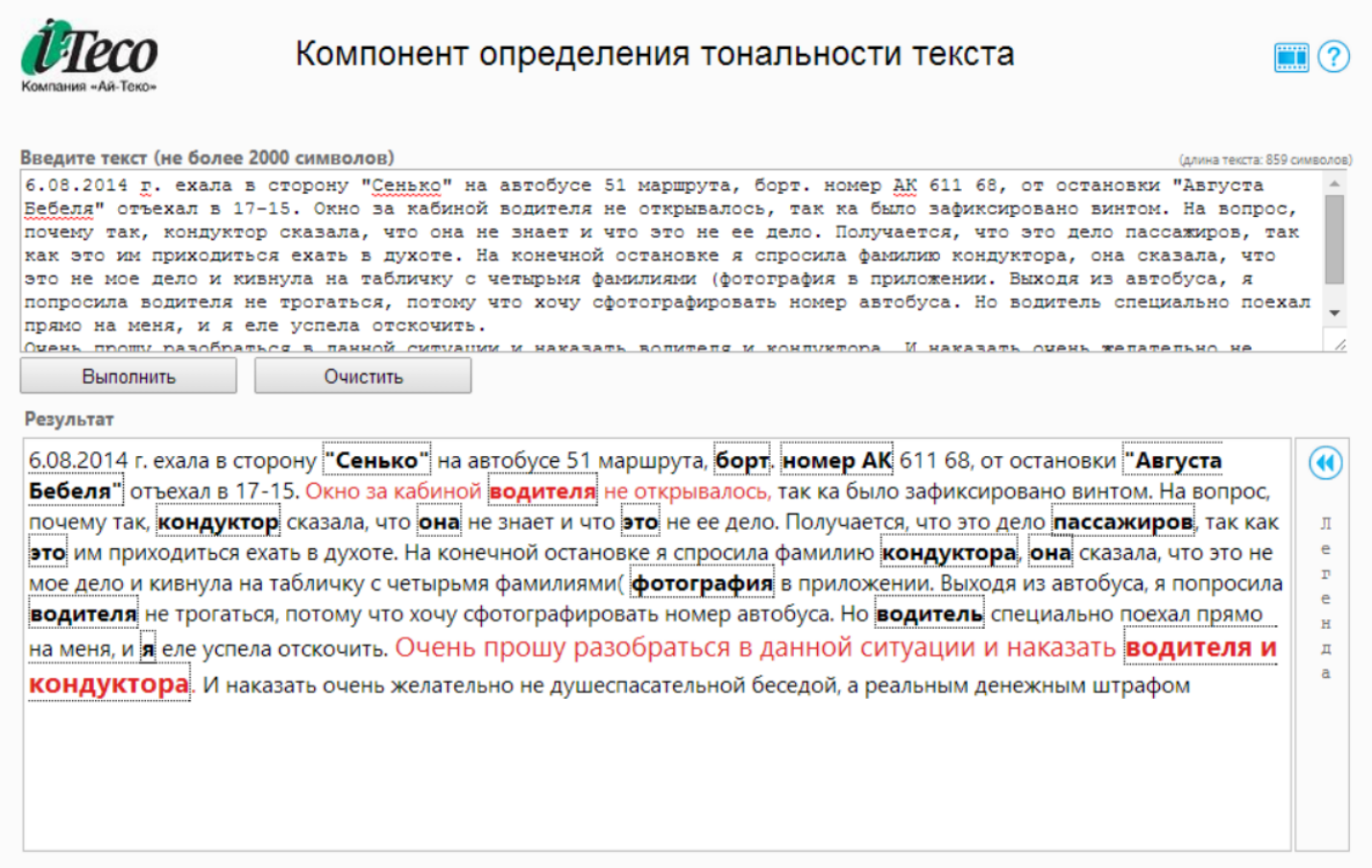
# Обзор программных средств.

**http://x-file.su/tm/Default.aspx**

Компонент определения тональности текста реализует метод, основанный на словарях и правилах. На вход компонента подаётся текст, тональность которого нужно определить

# Обзор программных средств.

**http://x-file.su/tm/Default.aspx**




Данный программный продукт также является веб - сервисом, который позволяет определять эмоциональную окраску текста, введенного пользователем. В нем используются метрики и специальные словари эмоционально окрашенной лексики.

# Работа с TopicMiner и SentStrength

**Препроцессинг данных**

1. Сборка файлов, лематизация, удаление Html тэгов.

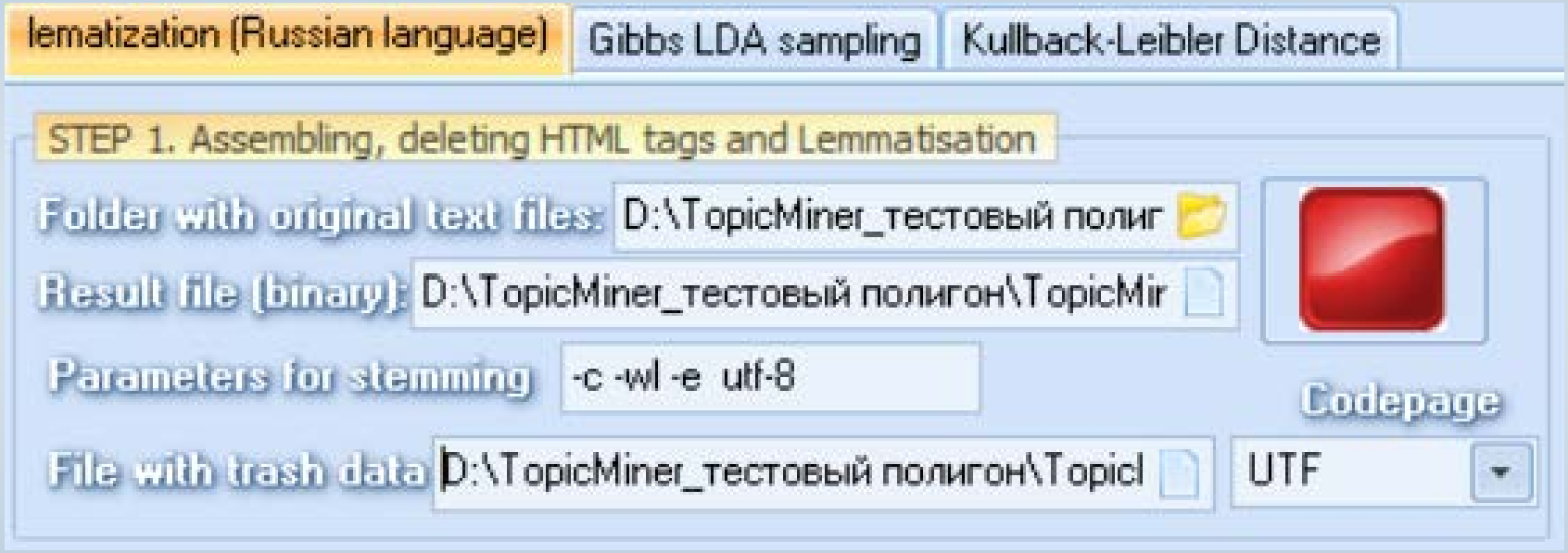
Лематизация производится при помощи программы **mystem.exe** (разработка Yandex)

Михаил Блинков впервые задумались о создании литературно-музыкального проекта. И решили объединить короткие рассказы Марата с музыкальным аккомпанементом Мишы. И получился проект "МыслиЗвуки": Марат читает свою прозу, Михаил наигрывает тихие мелодии. В программе вечера, посвященного двухлетию проекта, - ранее написанные и совершенно новые рассказы.<br В программе вечера ранее написанные и уже знакомые зрителю рассказы, такие как "Бабушкин

{михаил} {блинк|блинков|блинок} {впервые} {задумываться} {о} {создание} {литературный}-{музыкальный} {проект}. {и} {решать} {объединять} {короткий} {рассказ} {марат} {с} {музыкальный} {аккомпанемент} {Мишы??}. {и} {получаться} {проект} &{quot??};{МыслиЗвуки??}&{quot??};: {марат} {читать} {свой} {проза}, {михаил} {наигрывать} {тихий} {мелодия}. {в} {программа} {вечер}, {посвящать} {двухлетие} {проект}, - {ранее} {написать} {и} {совершенно|совершенный} {новое|новый} {рассказ}.<{br??} {в} {программа} {вечер} {ранее} {написать} {и} {уж|уже|узкий|узко} {знакомая|знакомый} {зритель} {рассказ}, {такой} {как} &{quot??};{бабушкин}

# Препроцессинг в TopicMiner

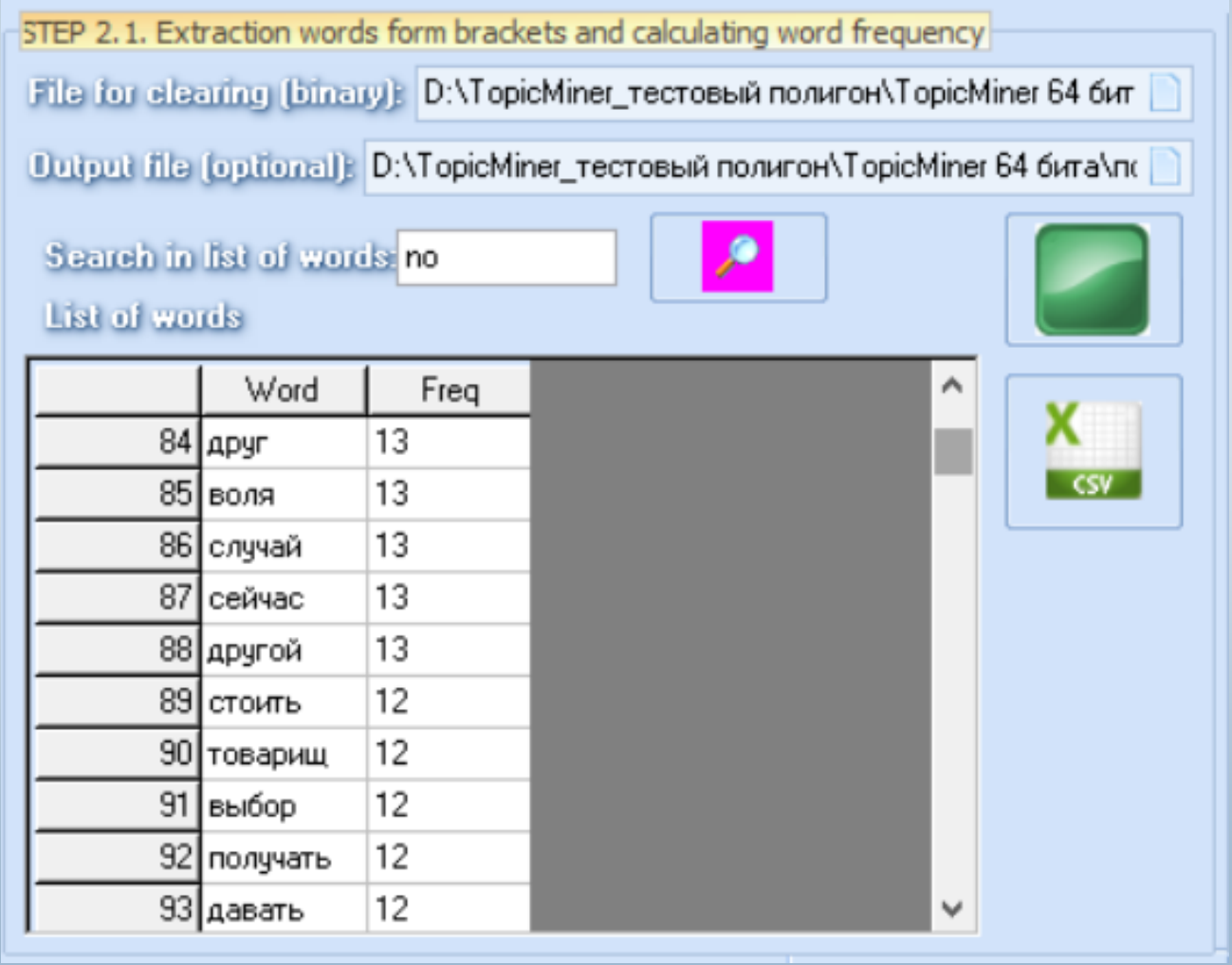
1. Сборка файлов, лематизация, удаление Html тэгов.




В результате работы данной опции появится файл с расширением tmld, в котором хранятся все тексте, как оригинальные так и лематизированные, кроме того лематизированные тексты почищены от Html разметки.

# Препроцессинг в TopicMiner

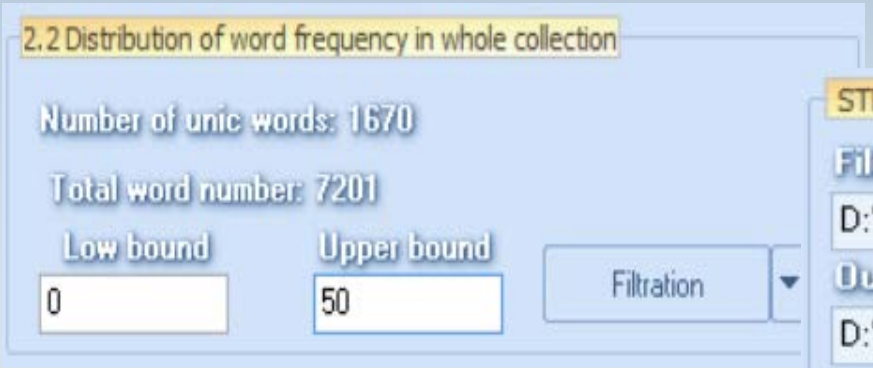
1. Процедура выделения лем и расчет частот.

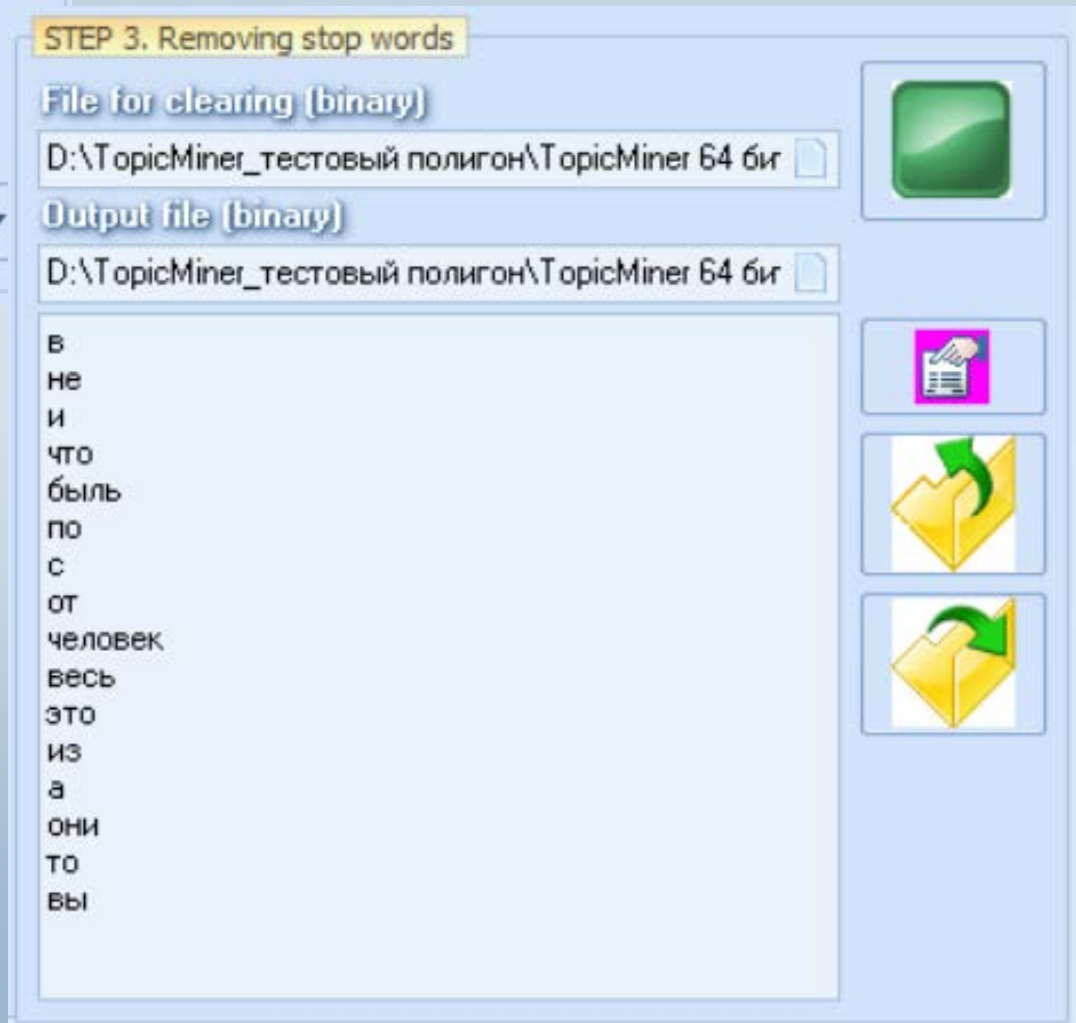



| | Word | Freq |
|---|---|---|
| 84 | друг | 13 |
| 85 | воля | 13 |
| 86 | случай | 13 |
| 87 | сейчас | 13 |
| 88 | другой | 13 |
| 89 | стоить | 12 |
| 90 | товарищ | 12 |
| 91 | выбор | 12 |
| 92 | получать | 12 |
| 93 | давать | 12 |

# Препроцессинг в TopicMiner

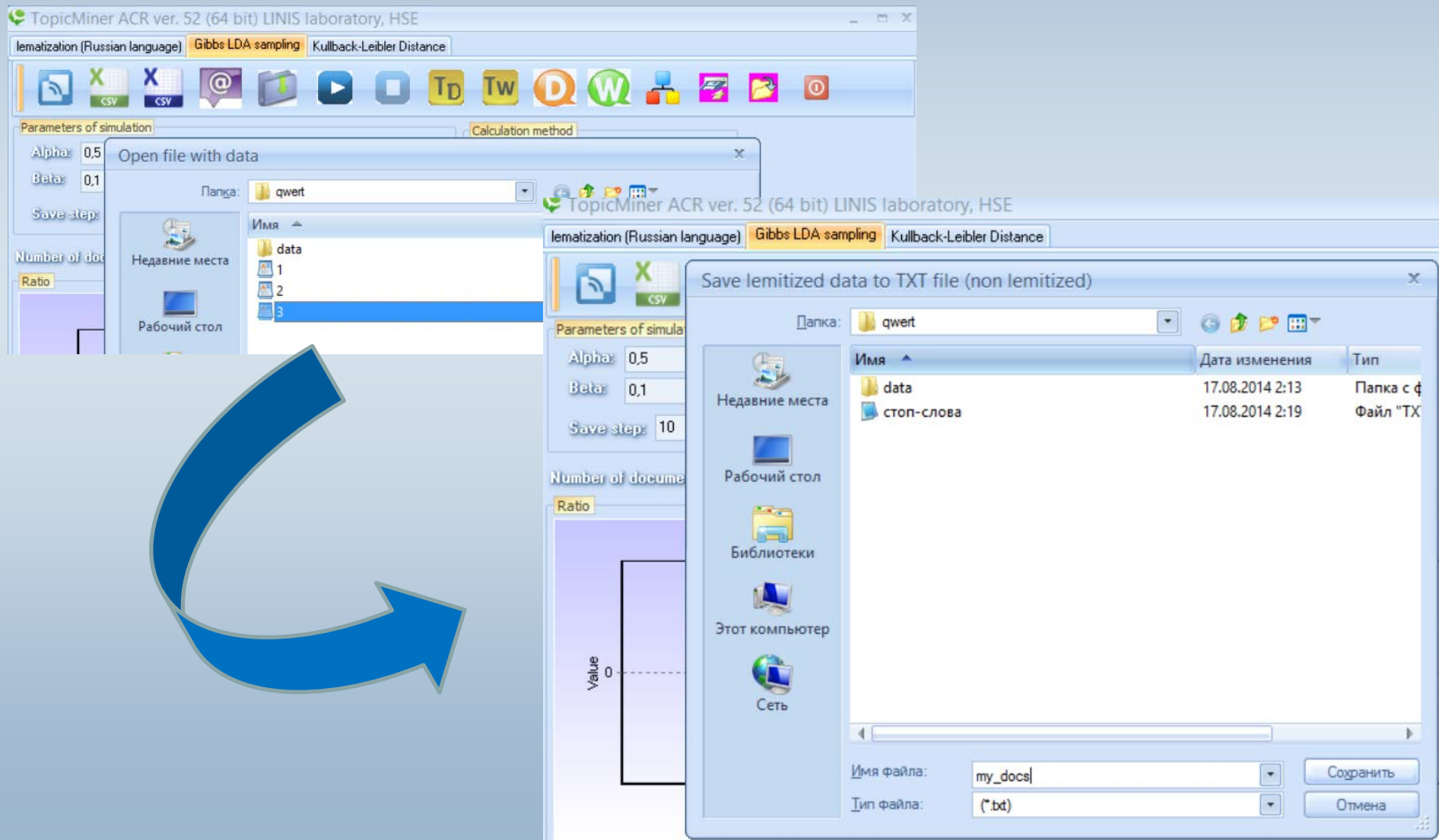
1. Выделение и удаление стоп – слов.



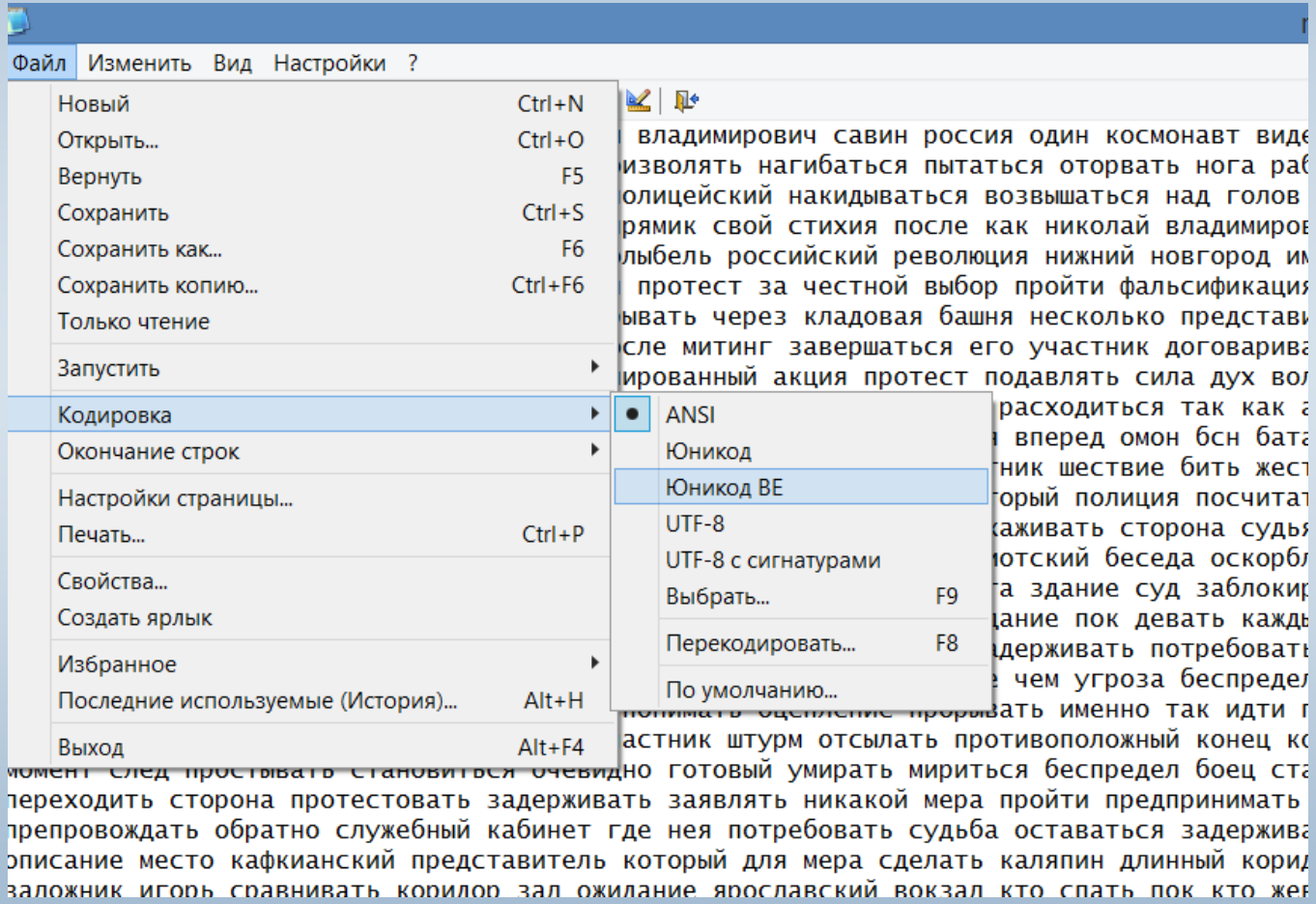




В результате третьего этапа производится удаление стоп слов из лематизированных текстов.

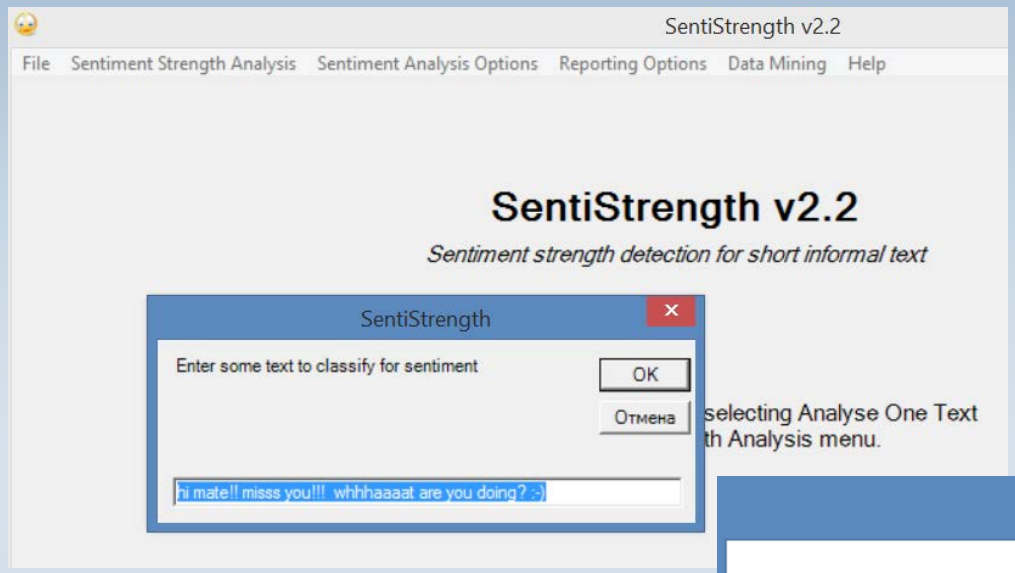
# Выгрузка лематизированных текстов из TopicMiner

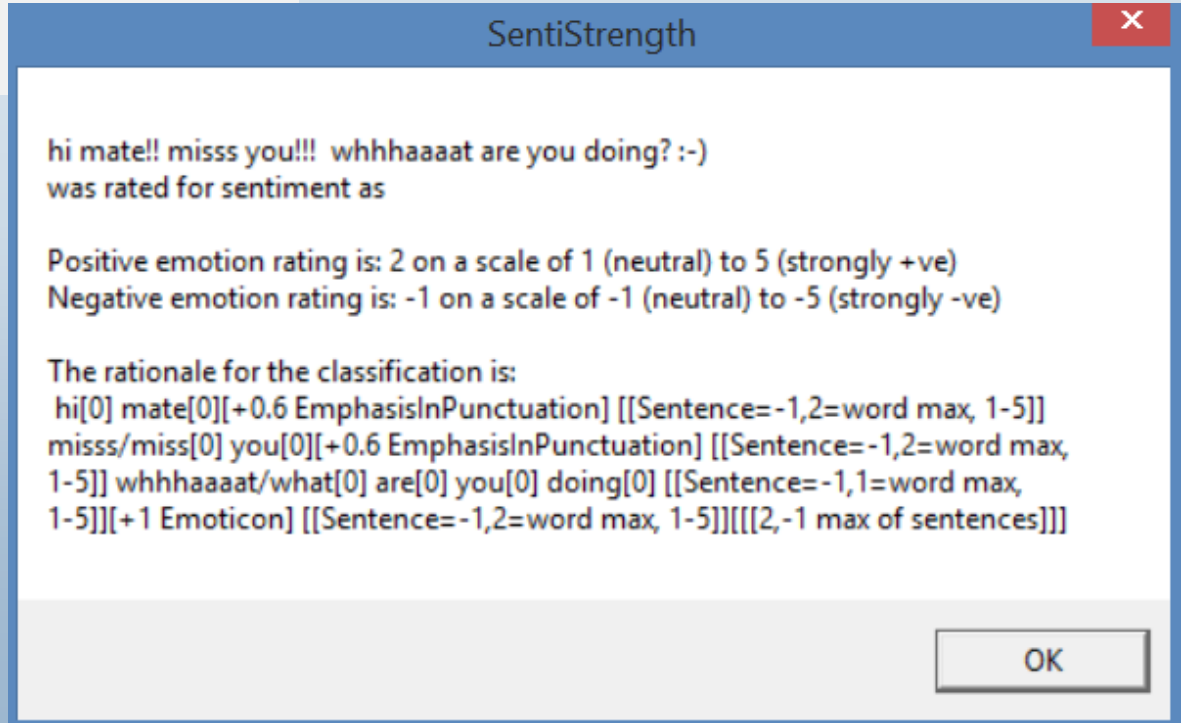


# Конвертация текста в Юникод

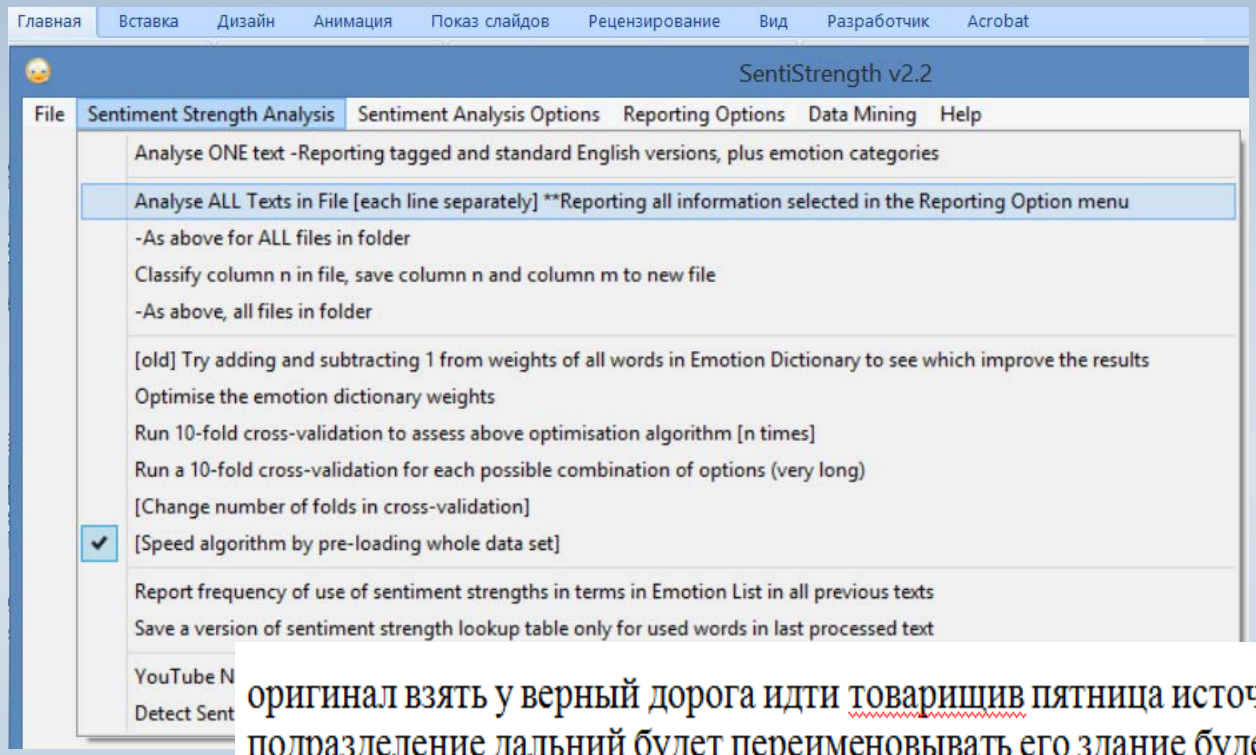


# Работа с SentiStrength



# Работа с SentiStrength




оригинал взять у верный дорога идти товаришив пятница источник мвд сообщать возможно подразделение дальний будет переименовывать его здание будет разрушать заново отстраивать земля закопать надпись написать оригинал взять у верный дорога идти товаришив пятница источник мвд сообщать возможно подразделение дальний будет переименовывать его здание будет разрушать заново отстраивать земля закопать надпись написать **3** **-1** оригинал[0] взять[0] у[0] верный[2] дорога[0] идти[0] товаришив[0] пятница[0] источник[0] мвд[0] сообщать[0] возможно[0] подразделение[0] дальний[0] будет[0] переименовывать[0] его[0] здание[0] будет[0] разрушать[0] заново[0] отстраивать[0] земля[0] закопать[0] надпись[0] написать[0] [[Sentence=-1,3=word max, 1-5]][[[3,-1 max of sentences]]]

INTERNET STUDIES LAB

HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY
SAINT PETERSBURG

# THANK YOU!

www.linis.hse.spb.ru/index.php/glavnaja.html

ВЫСШАЯ ШКОЛА ЭКОНОМИКИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
САНКТ-ПЕТЕРБУРГ

ЛАБОРАТОРИЯ ИНТЕРНЕТ-ИССЛЕДОВАНИЙ

- Главная
- О нас
- События
- Публикации
- Исследования
- Материалы
- Ссылки
- Контакты

ENGLISH

## Новости

### Чем дышит блогосфера? Семинар ЛИнИс в Москве

11:54 27.04.2012

24 апреля команда ЛИнИс провела в НИУ-ВШЭ (Москва) семинар на тему "Чем дышит блогосфера? Методы анализа больших массивов Интернет-данных для социологических задач". Выступление прошло в рамках академического семинара по социологической теории и методологии кафедры анализа социальных институтов (рук. Инна Девятко), который на этот раз проходил совместно с Лабораторией экономико-социологических исследований (рук. В.Радаев).

Подробнее...

### ЛИнИс – участник Балтийского партнерства по новым медиа

## Анонсы

11.05.2012 (Пятница)

**Презентация Лаборатории Интернет-Исследований**

11 мая в 17-00 состоится презентация новой исследовательской площадки Национального исследовательского университета «Высшая школа экономики» (Санкт-Петербург) - Лаборатории интернет-исследований (ЛИнИс).

Подробнее...

26.09.2012 (Среда)

**Новые СМИ: меняющийся медийный ландшафт**

27-28 сентября 2012 года лаборатория интернет-исследований совместно с зарубежными партнерами проводит конференцию "New media: changing media landscapes". Конференция