Software Engineering Conference Russia
October 2017, St. Petersburg

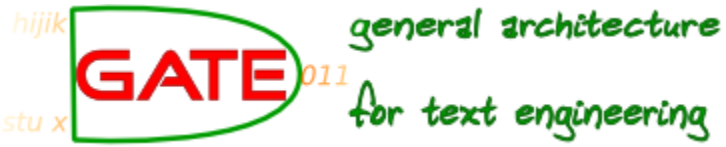
Диспетчеризация задач в комплексе инструментов автоматизированного анализа текста

Екатерина Полицына, МАИ
Сергей Полицын, МАИ
Валерий Шилов, ВШЭ

ЗАДАЧИ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

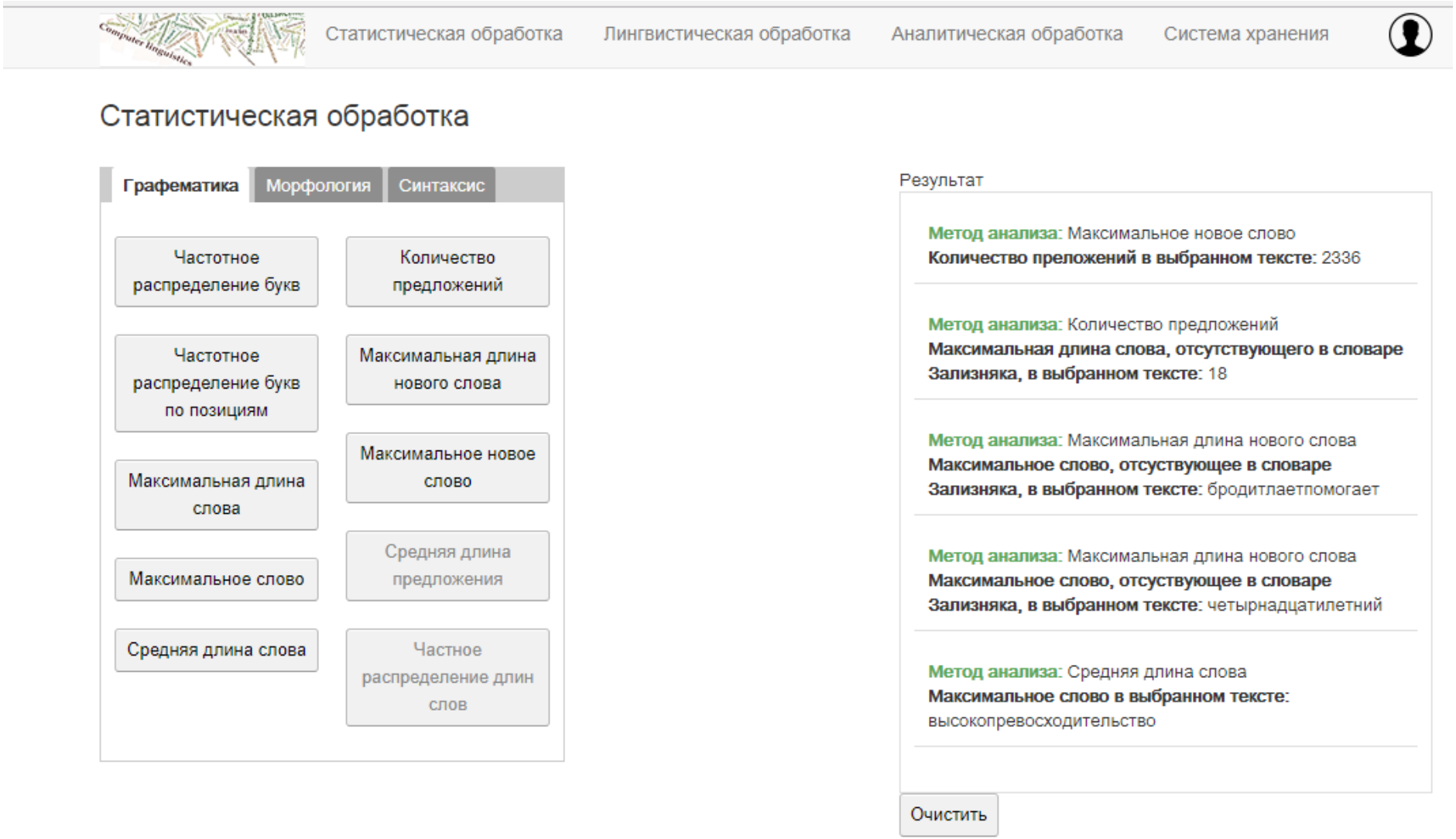
- Автоматизация составления и лингвистической обработки машинных словарей и обработки ошибок правописания
- Автоматическая классификация и реферирование документов
- Упрощение поиска информации
- Перевод текстов с одних естественных языков на другие
- «Общение» пользователей с машинами на естественном языке
- Автоматическое извлечение полезной информации из неформализованных текстов

ИНСТРУМЕНТЫ И ПЛАТФОРМЫ

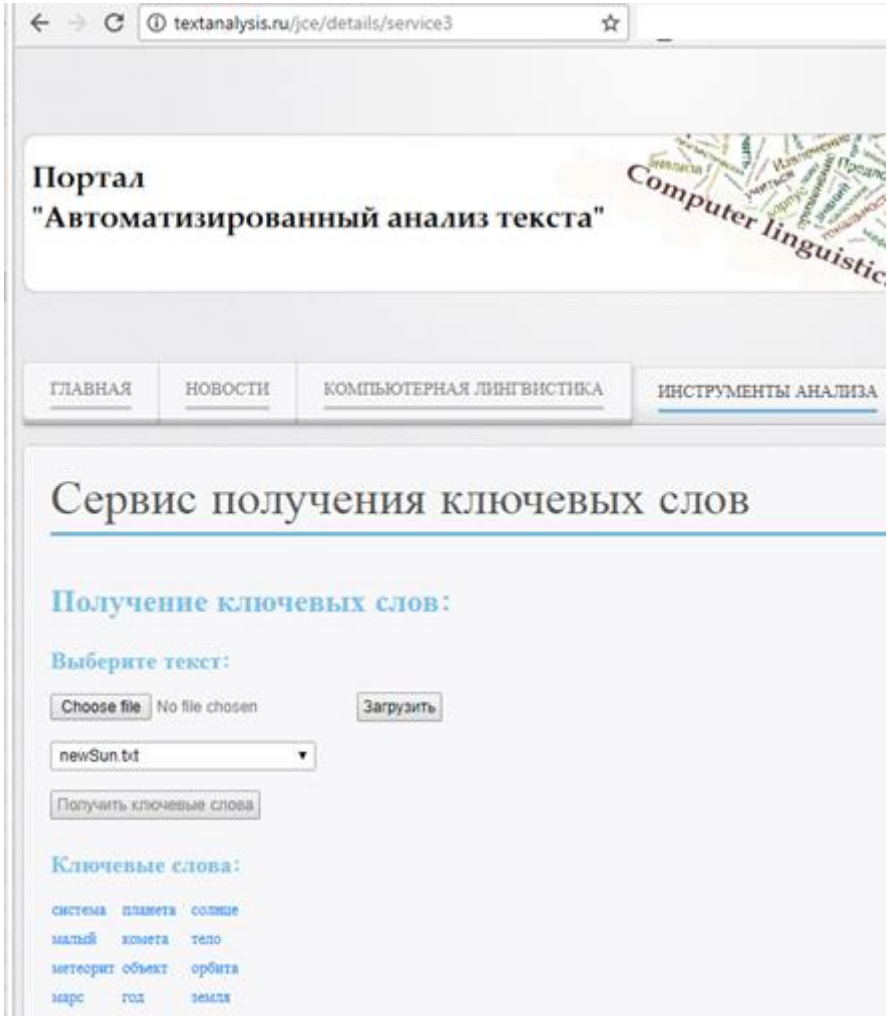


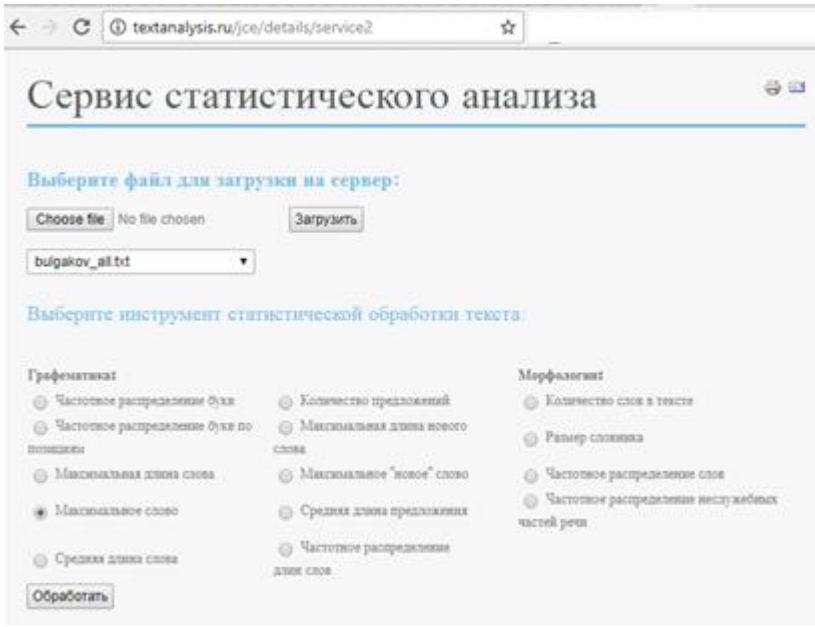
О КОМПЛЕКСЕ ИНСТРУМЕНТОВ

О КОМПЛЕКСЕ ИНСТРУМЕНТОВ

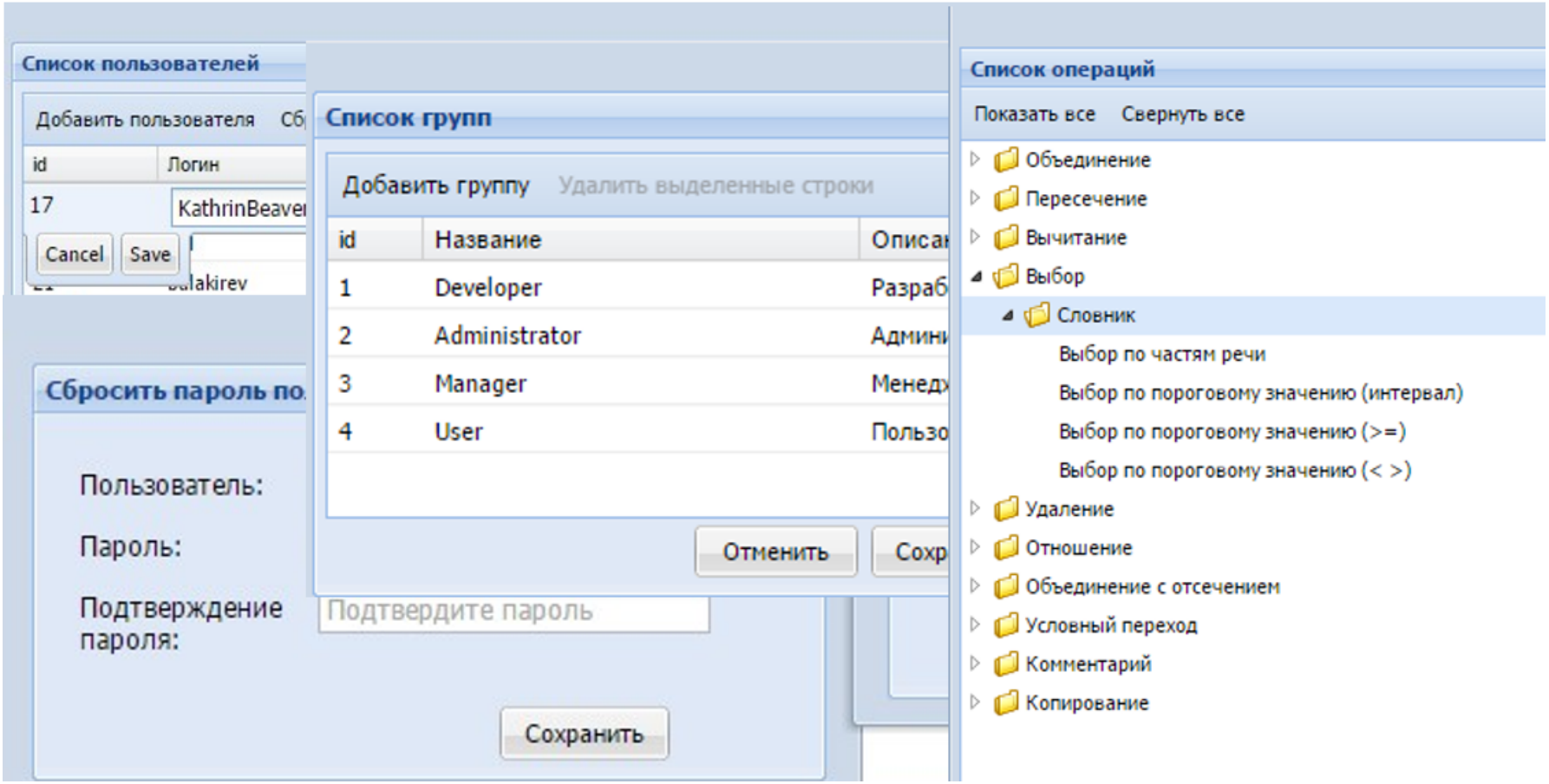


О КОМПЛЕКСЕ ИНСТРУМЕНТОВ



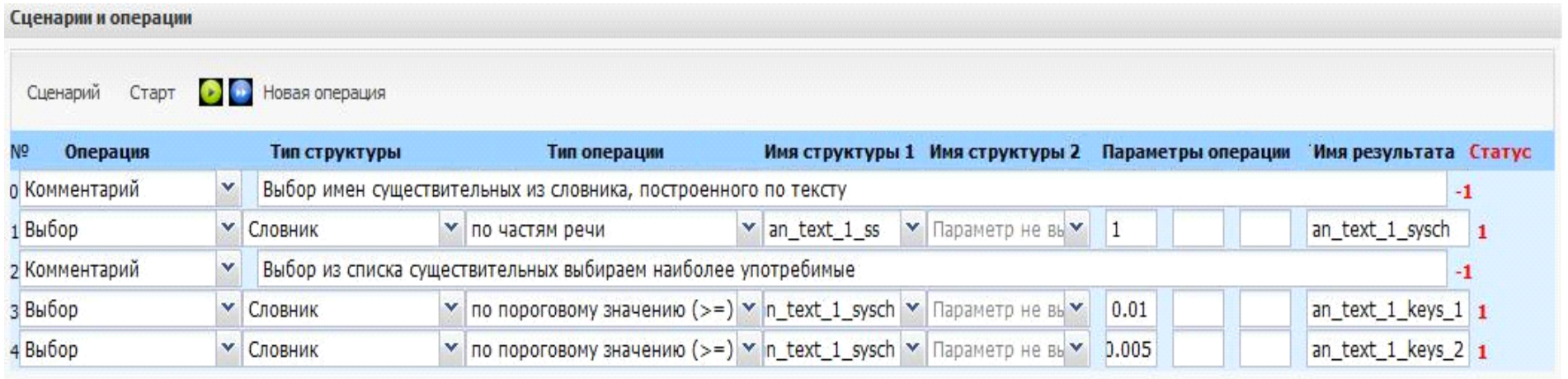


О КОМПЛЕКСЕ ИНСТРУМЕНТОВ



ОБРАБОТКА ТЕКСТА В СИСТЕМЕ АНАЛИЗА

- Базовая обработка
  - Статистическая обработка
  - Лингвистическая обработка
- Аналитическая обработка



Сценарии и операции

Сценарий | Старт | Новая операция

№	Операция	Тип структуры	Тип операции	Имя структуры 1	Имя структуры 2	Параметры операции			Имя результата	Статус
0	Комментарий	Выбор имен существительных из словника, построенного по тексту								-1
1	Выбор	Словник	по частям речи	an_text_1_ss	Параметр не вы	1			an_text_1_sysch	1
2	Комментарий	Выбор из списка существительных выбираем наиболее употребимые								-1
3	Выбор	Словник	по пороговому значению (>=)	n_text_1_sysch	Параметр не вы	0.01			an_text_1_keys_1	1
4	Выбор	Словник	по пороговому значению (>=)	n_text_1_sysch	Параметр не вы	0.005			an_text_1_keys_2	1

ТОЧКИ ВХОДА ЗАДАЧ

- Точка входа задач статистической обработки
- Точка входа задач лингвистической обработки
- Точка входа задач аналитической обработки
- API системы

ПРОБЛЕМЫ АРХИТЕКТУРЫ

- Отсутствие контроля обращений к ядру системы
- Отсутствие возможности управления задачами и диспетчеризации потока поступающих задач
- Потеря задач после сбоя в системе

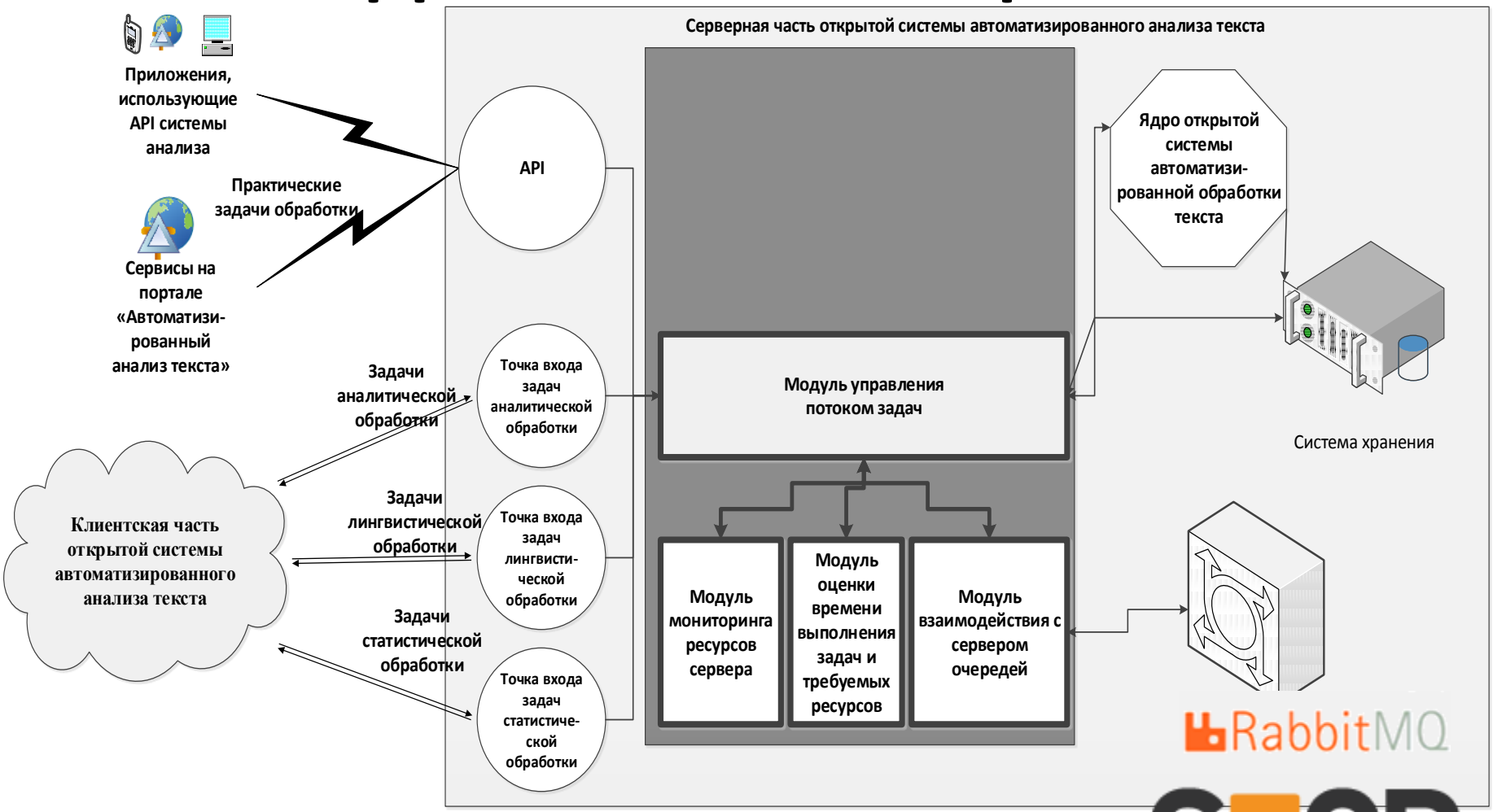
НЕОБХОДИМОСТЬ ДИСПЕТЧЕРИЗАЦИИ ЗАДАЧ

- Среднее DAU (количество активных пользователей в день) портала достигло 74, рост почти в **2 раза.** Количество пользователей системы анализа невелико, но они запускают **очень** «тяжелые» задачи.
- Средние данные по задачам:
  - для линейных задач: 12% CPU load, 3.9 Мб RAM
  - для экспоненциальных: 25% CPU load, 180 Мб RAM
- Скорость обработки текста уменьшается с увеличением размера текста:
  - 0.5 Мб/мин для файлов < 0.5 Мб
  - 0.3 Мб/мин для файлов < 1.5 Мб

НЕОБХОДИМОСТЬ ДИСПЕТЧЕРИЗАЦИИ ЗАДАЧ

- Несколько источников задач
- Рост числа пользователей
- Ядро системы – единый центр выполнения задач
- Сложность алгоритмов анализа текста
- Экспоненциальная зависимость времени и требуемых ресурсов обработки текста от его размера
- Рост объема текстовой информации и потребности в оперативной обработке
- Углубление алгоритмов анализа для решения различных задач

СТРУКТУРА ИНСТРУМЕНТА ДИСПЕТЧЕРИЗАЦИИ



МОДУЛЬ ВЗАИМОДЕЙСТВИЯ С СЕРВЕРОМ ОЧЕРЕДЕЙ

Отправка задач в очередь

Получение задач из очереди

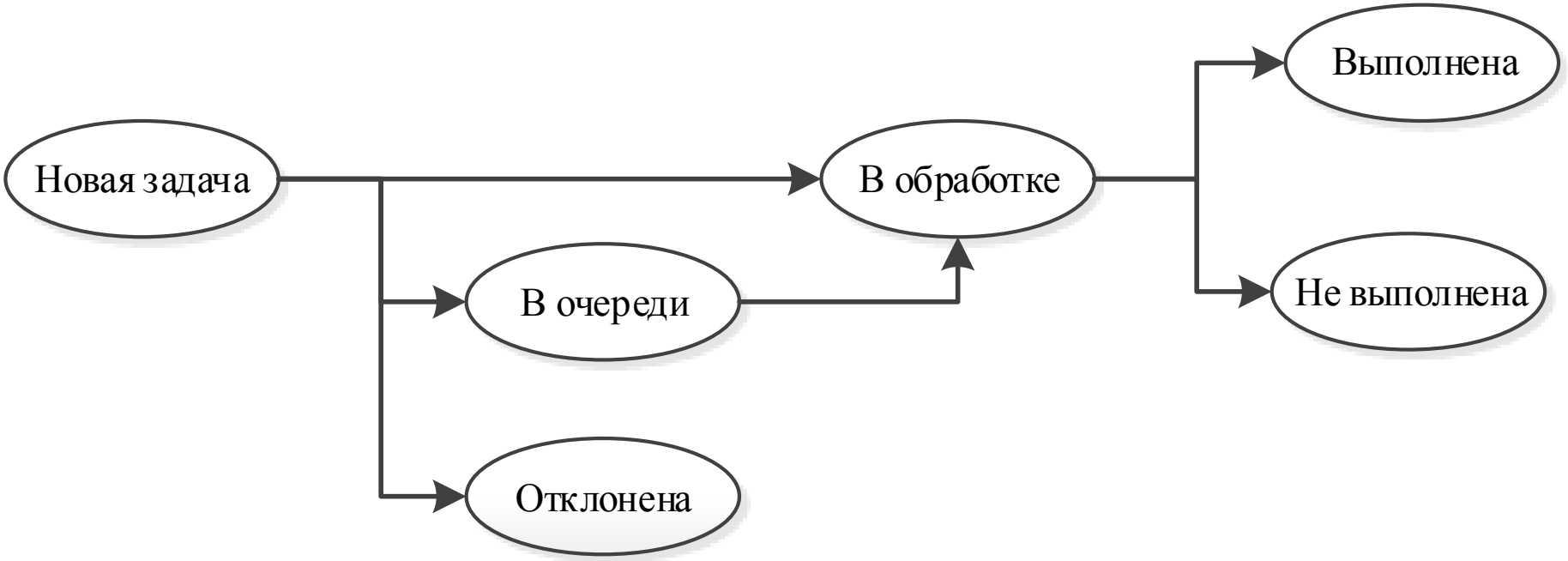
Получение загруженности очередей и выбранной очереди

МОДУЛЬ УПРАВЛЕНИЯ ПОТОКОМ ЗАДАЧ

Основные задачи модуля:

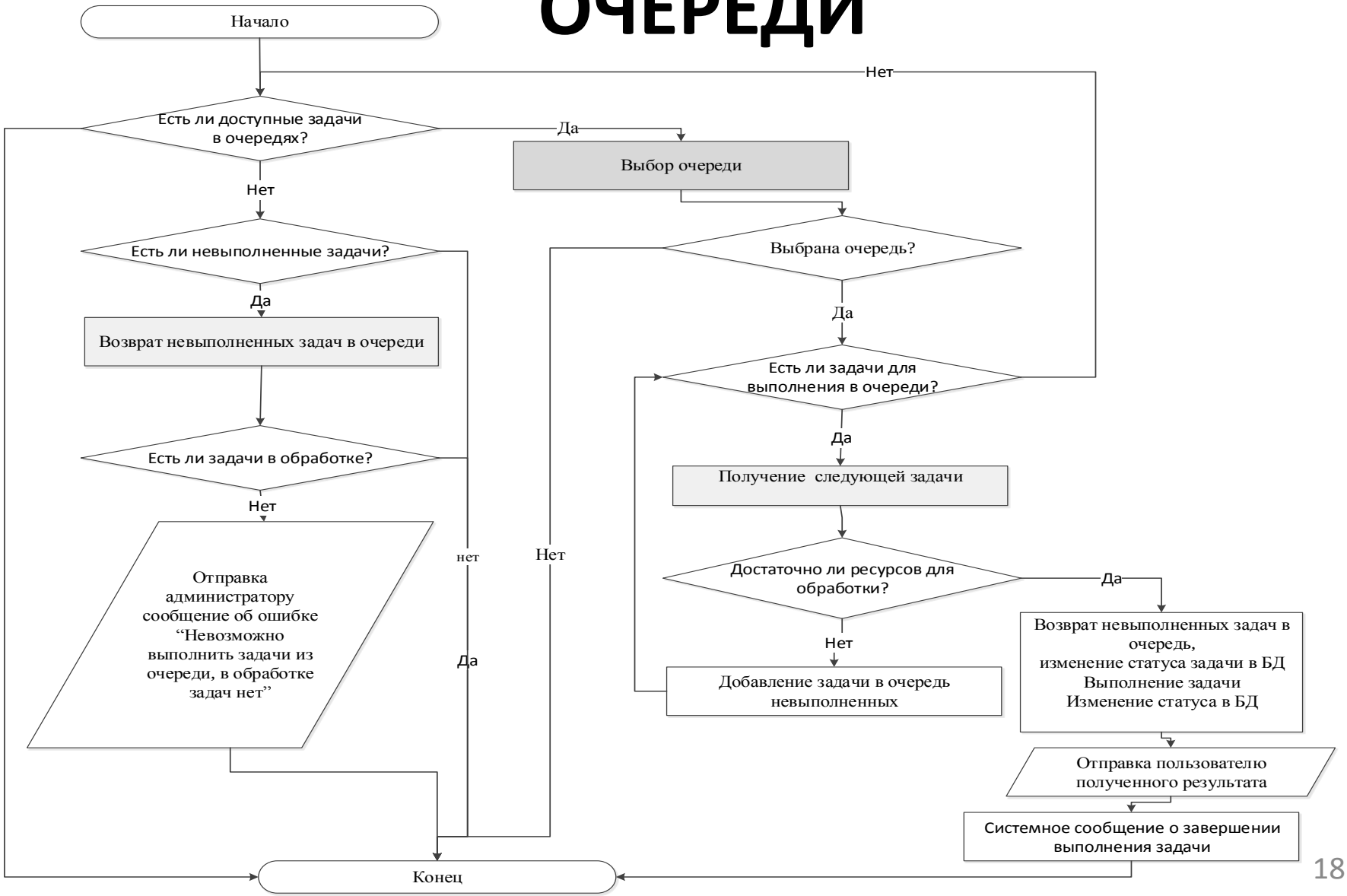
- Оценка «сложности» задачи (может ли задача быть обработана при нулевой нагрузке)
- Проверка занятости системы и очередей
- Отправка задачи в очередь или на исполнение

СТАТУСЫ ЗАДАЧ



АЛГОРИТМ ОБРАБОТКИ НОВОЙ ЗАДАЧИ

АЛГОРИТМ ВЫБОРА ЗАДАЧИ ИЗ ОЧЕРЕДИ



МОДУЛЬ МОНИТОРИНГА РЕСУРСОВ СИСТЕМЫ

- Проверка работоспособности ядра системы
- Получение информации о загрузке процессора
- Получение информации о загрузке оперативной памяти

МОДУЛЬ ОЦЕНКИ ТРЕБУЕМЫХ РЕСУРСОВ

- Данные для анализа:
  - Экспериментальные данные зависимости времени обработки и требуемых ресурсов от размера текста
- Сложность задач:
  - Линейная $y = k * x + b$
  - Экспоненциальная $y = k * e^{x + b}$
- Метод наименьших квадратов:

$$k = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 \left(\sum_{i=1}^{n} x_i^2 \right)^2}$$

$$b = \frac{1}{n} \left(\sum_{i=1}^{n} y_i - k \sum_{i=1}^{n} x_i \right)$$

МОДУЛЬ ОЦЕНКИ ТРЕБУЕМЫХ РЕСУРСОВ

Экспериментальные данные по линейным задачам

	Размер файла, Кб	Оперативная память, Кб	Загрузка процессора, %
1	2	55,000	1,5
2	12	85000	2
3	25	132000	5
4	143	177000	8
5	837	300000	1
6	1128	150000	17
7	2993	196000	25
8	5898	200000	25

$$y = 0{,}005695118 * t + 4{,}406082214$$

$$y = 0{,}022325434 * t + 98{,}2274039$$

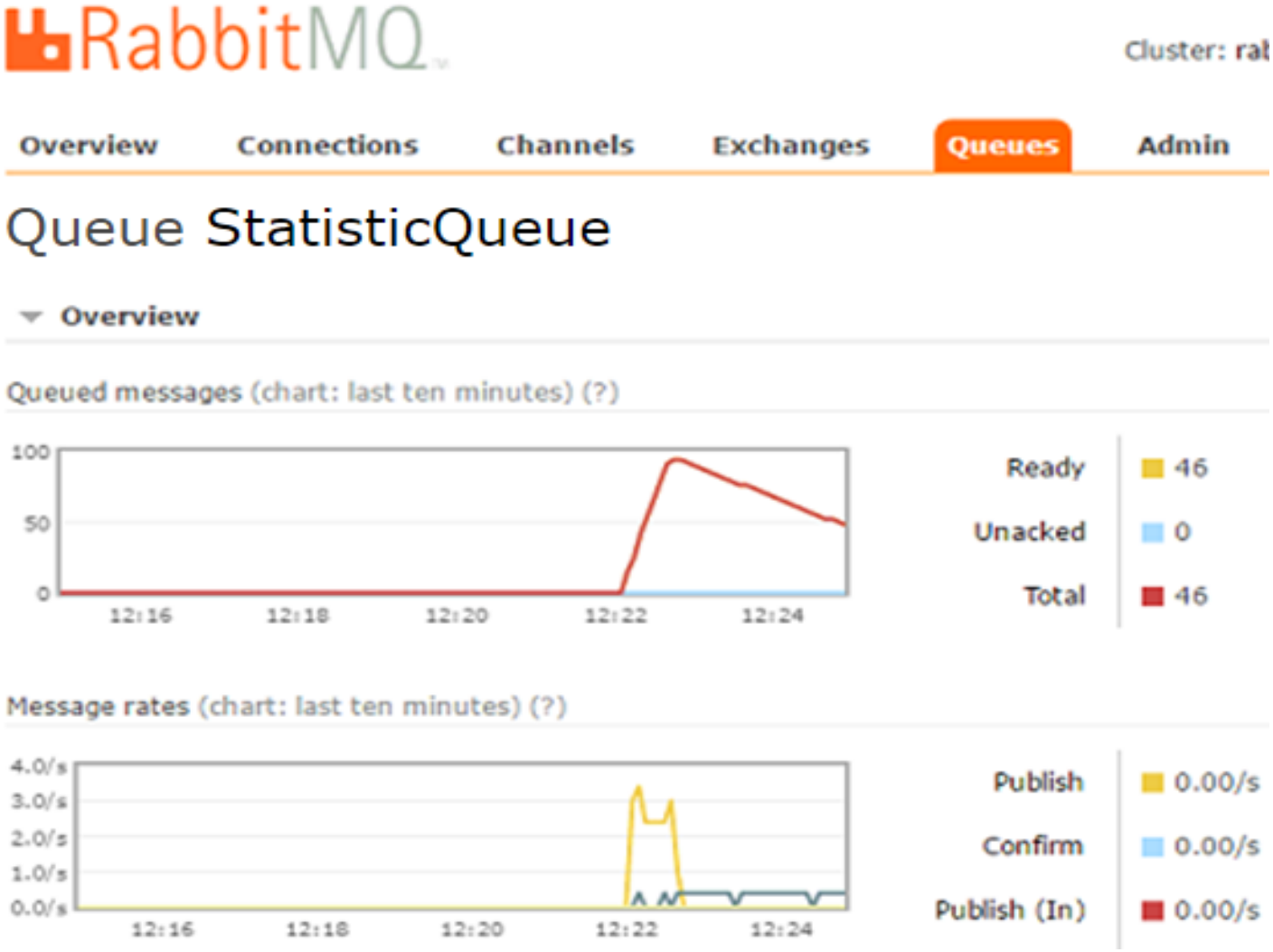
Экспериментальные данные по экспоненциальным задачам

	Размер текста, Кб	Оперативная память, Кб	Загрузка процессора, %
1	2	30720	3
2	12	30720	5
3	25	35840	8
4	143	40960	21
5	837	51200	25
6	1128	61440	25
7	2993	112640	25
8	5898	204800	25

$$y = 35.517e^{0.000316t}$$

$$y = 9.75e^{0.000226t}$$

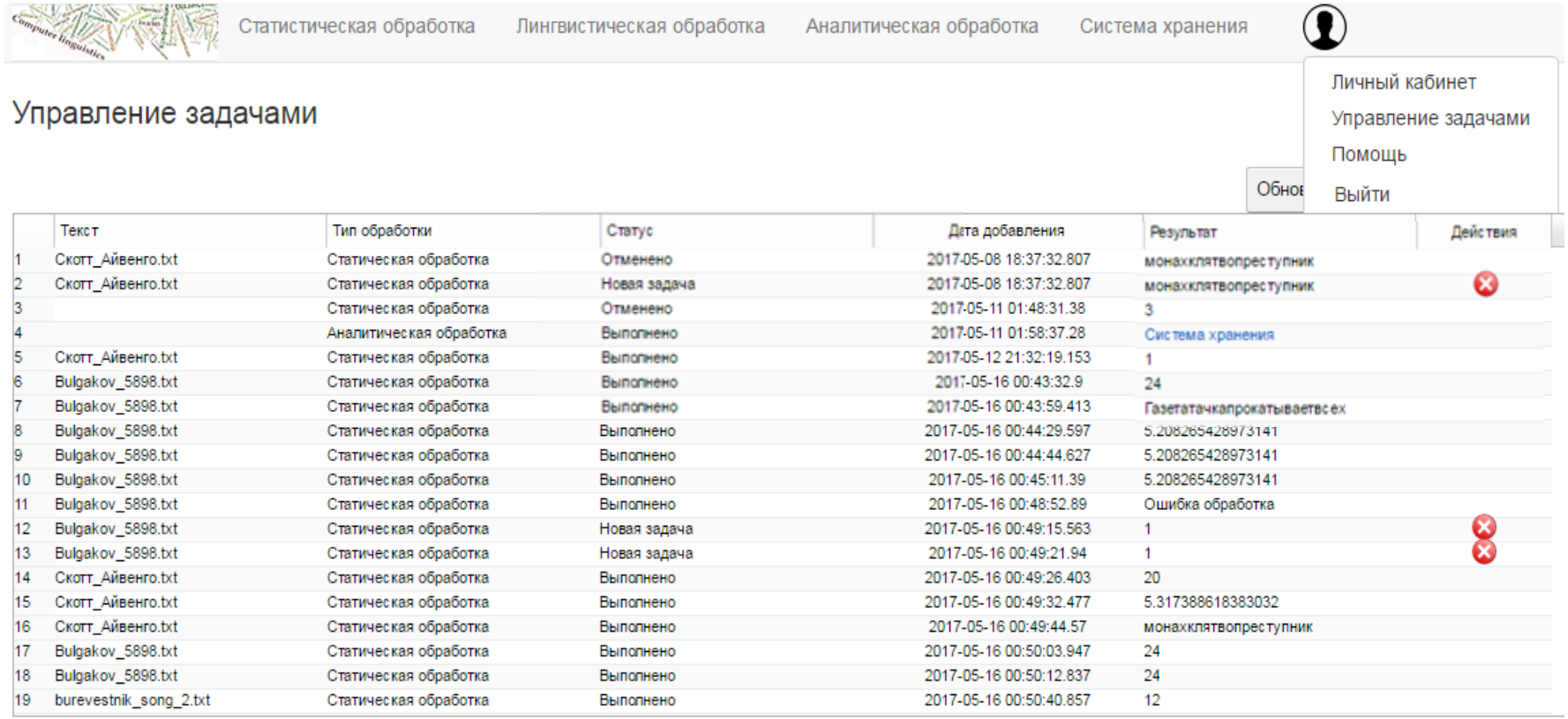
НАГРУЗОЧНОЕ ТЕСТИРОВАНИЕ



НАГРУЗОЧНОЕ ТЕСТИРОВАНИЕ

- При отсутствии возможности обработки задачи, она попадает в очередь, завершение обработки задачи вызывает выполнение новой задачи из очереди
- При отправке одновременно большого количества задач повышается нагрузка на процессор и ядро системы не может принять задачу в обработку, задачи отправляются в очередь
- В случае сбоя в системе задачи добавляются в конец соответствующей очереди и при возобновлении работы системы отправляются на обработку

ИЗМЕНЕНИЯ В ЛИЧНОМ КАБИНЕТЕ ПОЛЬЗОВАТЕЛЯ



	Текст	Тип обработки	Статус	Дата добавления	Результат	Действия
1	Скотт_Айвенго.txt	Статическая обработка	Отменено	2017-05-08 18:37:32.807	монахклятвопреступник	
2	Скотт_Айвенго.txt	Статическая обработка	Новая задача	2017-05-08 18:37:32.807	монахклятвопреступник	
3		Статическая обработка	Отменено	2017-05-11 01:48:31.38	3	
4		Аналитическая обработка	Выполнено	2017-05-11 01:58:37.28	Система хранения	
5	Скотт_Айвенго.txt	Статическая обработка	Выполнено	2017-05-12 21:32:19.153	1	
6	Bulgakov_5898.txt	Статическая обработка	Выполнено	2017-05-16 00:43:32.9	24	
7	Bulgakov_5898.txt	Статическая обработка	Выполнено	2017-05-16 00:43:59.413	Газетатачкапрокатываетвсех	
8	Bulgakov_5898.txt	Статическая обработка	Выполнено	2017-05-16 00:44:29.597	5.208265428973141	
9	Bulgakov_5898.txt	Статическая обработка	Выполнено	2017-05-16 00:44:44.627	5.208265428973141	
10	Bulgakov_5898.txt	Статическая обработка	Выполнено	2017-05-16 00:45:11.39	5.208265428973141	
11	Bulgakov_5898.txt	Статическая обработка	Выполнено	2017-05-16 00:48:52.89	Ошибка обработка	
12	Bulgakov_5898.txt	Статическая обработка	Новая задача	2017-05-16 00:49:15.563	1	
13	Bulgakov_5898.txt	Статическая обработка	Новая задача	2017-05-16 00:49:21.94	1	
14	Скотт_Айвенго.txt	Статическая обработка	Выполнено	2017-05-16 00:49:26.403	20	
15	Скотт_Айвенго.txt	Статическая обработка	Выполнено	2017-05-16 00:49:32.477	5.317388618383032	
16	Скотт_Айвенго.txt	Статическая обработка	Выполнено	2017-05-16 00:49:44.57	монахклятвопреступник	
17	Bulgakov_5898.txt	Статическая обработка	Выполнено	2017-05-16 00:50:03.947	24	
18	Bulgakov_5898.txt	Статическая обработка	Выполнено	2017-05-16 00:50:12.837	24	
19	burevestnik_song_2.txt	Статическая обработка	Выполнено	2017-05-16 00:50:40.857	12	

РЕЗУЛЬТАТЫ

- **Разработка и внедрение инструмента диспетчеризации задач** в комплексе инструментов анализа текста **повышает надежность** работы комплекса и дает **возможность распараллеливания** выполнения задач путем расширения аппаратной части
- Использование **сервера очередей** и реализация системы запуска задач из очереди после сбоя в системе обеспечивает с**охранность задач** пользователей и **гарантирует их обработку**
- На основе разработанного инструмента диспетчеризации задачи и полученного опыта его внедрения в комплекс может быть разработан **самостоятельный программный продукт** для применения его в других системах с набором ресурсоемких задач разных типов и приоритетов

ПОЛЕЗНЫЕ ЗАМЕТКИ

- Однородные ли задачи? Разные ли приоритеты задач? Есть ли типизация задач?
- Сколько они выполняются и могут ли быть выполнены?
- Важно ли их выполнить ASAP?
- Сколько допустимо ждать?
- Какая нагрузка планируется и какая есть на текущий момент?
- Насколько система масштабируема?
- Тестируйте и «ломайте» решения. Анализ поломок позволяет сделать «мир лучше».

СПАСИБО ЗА ВНИМАНИЕ!