



# Оценка черной магии

**Александр Коваленко**  
Руководитель группы

Я.Субботник, Санкт-Петербург,  
26 февраля 2011 года

# Оглавление

- **З**адача оценки
- **П**рототип
- **У**ниверсальная система
- **Э**пилог

# **З**адача оценки

**Я**ндекс  
Найдётся всё

[Поиск](#) [Карты](#) [Маркет](#) [Новости](#) [Словари](#)

Например, качество снега

кафе

в найденном  в Санкт-Петербурге



[Маркет](#)

Поиск по 17 150 160  
предложениям от 5 536  
магазинов



[Авто](#)

Поиск по 930 160  
объявлениям с 324 сайтов



[Недвижимость](#)

Поиск по 623 265  
объявлениям с 62 сайтов



[Картинки](#)

Поиск по 2 544 067 279  
картинкам и фотографиям

# Белая магия

```
public String left(String str, int len) {  
    if (str == null) {  
        return null;  
    }  
    if (len < 0) {  
        return EMPTY;  
    }  
    if (str.length() <= len) {  
        return str;  
    }  
    return str.substring(0, len);  
}
```

# Черная магия

Было: Санкт-Петербург, Московский пр., 182а

Стало: индекс: 196105  
страна: Россия  
регион: Санкт-Петербург  
город: Санкт-Петербург  
улица: Московский проспект  
дом: 182  
литер: А



# Пример: до улучшения

Название: Ресторан Черемша

Адрес: Санкт-Петербург, поселок Солнечное, 2-я Боровая ул., 16, ...

Название: Солнечное небо

Адрес: Москва, улица Земляной Вал, ...

# Пример: после улучшения

Название: Ресторан Черемша

Адрес: Санкт-Петербург, поселок Солнечное, ...

Название: Солнечное небо

Адрес: Москва, поселок Солнечное, Земляной Вал, ...



Но если...



# **Прототип**

Найти

Пользователи ищут: [драка в таганроге](#) ▼

[расширенный поиск](#)

[Главные новости](#)

[Взрыв в Домодедово](#)

[Политика](#)

[В мире](#)

[Общество](#)

[Экономика](#)

[Спорт](#)

[Происшествия](#)

[Культура](#)

[Наука](#)

[Здоровье](#)

Автоматически обработано 3623 источника, обновлено в 13:40 мск

Выпуск: [Россия](#) ▼

## Общество

[Религия](#)



### [Эпидемия гриппа и ОРВИ в Смоленской области пошла на спад](#) (508 сообщений) 411 мнений

Семь школ закрылись на карантин из-за высокого уровня заболеваемости гриппом и ОРВИ в Находке, сообщила РИА Новости в четверг главный специалист территориального отдела управления Роспотребнадзора по Приморскому краю в городе Людмила Череванина.



### [Прокуратура в Приморье выясняет, кто снял ролик о таможенниках](#) (154) 416

Приморская транспортная прокуратура начала проверку по факту появления в интернете видеоролика с участием сотрудников Владивостокской таможни, которые развлекаются и пьют шампанское в служебном кабинете, сообщил РИА Новости в четверг заместитель прокурора Сергей Селенцев.



### [Фальшивые приставы обманывают москвичей по почте](#) (279) 219

Судебные приставы Колымы, сотрудники ГИБДД и налоговой инспекции с начала 2011 года выявили несколько должников по 14 исполнительным производствам на сумму задолженности почти в 200 тысяч рублей.



### [А. Фурсенко обсудил с депутатами Госдумы реформу образования](#) (209) 316

Напомним, проект федеральных образовательных стандартов вызвал общественный протест из-за предложения оставить в 10-11 классах обязательными лишь физкультуру, ОБЖ, пока не созданный курс "Россия в мире" и подготовку личного проекта.

Найти

Пользователи ищут: [драка в таганроге](#) ▼

[расширенный поиск](#)

[Главные новости](#)

[Взрыв в Домодедово](#)

[Политика](#)

[В мире](#)

[Общество](#)

[Экономика](#)

[Спорт](#)

[Происшествия](#)

[Культура](#)

[Наука](#)

[Здоровье](#)

Автоматически обработано 3623 источника, обновлено в 13:40 мск

Выпуск: [Россия](#) ▼

## Общество

[Религия](#)



### [Эпидемия гриппа и ОРВИ в Смоленской области пошла на спад](#) (508 сообщений) 411 мнений

Семь школ закрылись на карантин из-за высокого уровня заболеваемости гриппом и ОРВИ в Находке, сообщила РИА Новости в четверг главный специалист территориального отдела управления Роспотребнадзора по Приморскому краю в городе Людмила Череванина.



### [Прокуратура в Приморье выясняет, кто снял ролик о таможенниках](#) (154) 416

Приморская транспортная прокуратура начала проверку по факту появления в интернете видеоролика с участием сотрудников Владивостокской таможни, которые развлекаются и пьют шампанское в служебном кабинете, сообщил РИА Новости в четверг заместитель прокурора Сергей Селенцев.



### [Фальшивые приставы обманывают москвичей по почте](#) (279) 219

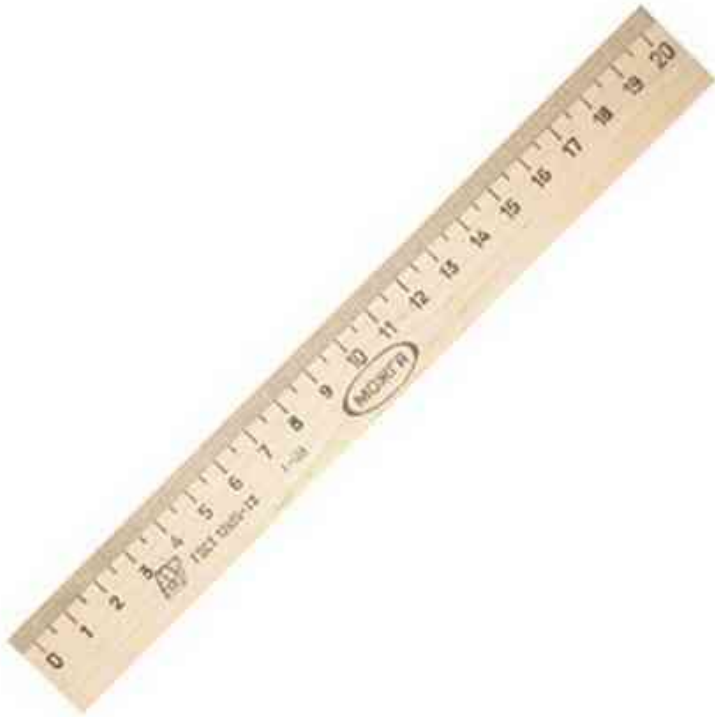
Судебные приставы Колымы, сотрудники ГИБДД и налоговой инспекции с начала 2011 года выявили несколько должников по 14 исполнительным производствам на сумму задолженности почти в 200 тысяч рублей.



### [А. Фурсенко обсудил с депутатами Госдумы реформу образования](#) (209) 316

Напомним, проект федеральных образовательных стандартов вызвал общественный протест из-за предложения оставить в 10-11 классах обязательными лишь физкультуру, ОБЖ, пока не созданный курс "Россия в мире" и подготовку личного проекта.

# Что хотели



Простой  
инструмент для  
оценки одного  
из алгоритмов

# Что получилось



Полноценная  
система оценки  
различных  
алгоритмов

**С**истемы оценки нужны!

# **Универсальная система**

Поисково-информационные

- Поиск**  
По всему интернету [RSS](#)
- Видео**  
Поиск роликов, видеохостинг
- Картинки**  
Поиск изображений [RSS](#)
- Музыка**  
Слушать бесплатно и легально
- Карты**  
С тонкими до дома [RSS](#)
- Народная карта**  
Нарисуйте карту своим руками
- Пробки**  
Карта дорожного движения
- Новости**  
Картина дня, созданная автоматически [RSS](#)
- Поиск по блогам и форумам**  
Что происходит в интернете прямо сейчас [RSS](#)
- Каталог**  
Сайты, отобранные вручную [RSS](#)
- Маркет**  
Выбор вещей и поиск товаров [RSS](#)
- Авто**  
Выбор автомобилей и поиск по объявлениям [RSS](#)
- Работа**  
Поиск по вакансиям
- Недвижимость**  
Подбор объявлений о недвижимости
- Услуги**  
Сравнение выходов и кредитов
- Словари**  
Онлайн-словари и справочники, иностранные языки [RSS](#)
- Расписание**  
Поезда и самолеты [RSS](#)
- Адмидз**  
Чем заняться в свободное время [RSS](#)
- Пассажи**  
В России и за рубежом [RSS](#)
- Телеобъявления**  
Центральные, спутниковые, региональные каналы [RSS](#)
- Книги**  
Поиск книг, изданий и авторов
- ЕГЭ 2010**  
Демонстрационная версия
- Время**  
Разницы во времени между городами

Персональные и развлекательные

- Почта**  
Без спама, вирусов и рекламы [RSS](#)
- Фото**  
Много места для ваших фотографий [RSS](#)
- Народ**  
Бесплатный хостинг и хранение файлов
- Мой Конт**  
Сеть профессионалов — найдутся все
- Я.ру**  
Место для приятного общения [RSS](#)
- Дзен**  
Мгновенные платежи в интернете [RSS](#)
- Каталог виджетов**  
Добавьте виджеты на главную страницу
- Стиральня**  
Открыть перед праздником
- Мои вопросы**  
История поисковых запросов
- Закладки**  
Личный архив любимых сайтов
- Почтиски**  
Новости по почте

Мобильные

- Мобильный Яндекс**  
Мобильные сервисы и приложения Яндекса

Вебмастеру

- Метрика**  
Статистика посещаемости сайта
- Рекламная сеть**  
Доход от вашего сайта
- Вебмастер**  
Информация об индексации вашего сайта
- Передача данных о содержимом сайта**  
[Добавить сайт в поиск Яндекса](#)
- Поиск для сайта**  
Поиск Яндекса на вашем сайте
- Почта для сайта**  
Наша почта в вашем адресе
- Партнерские программы**

API Яндекса

- Карты**
- Яндекс.Видео**
- Вид**
- Дзен**
- Поиск по блогам**
- Фото**
- Детектор**
- Счетчик**
- Хостинг JavaScript-библиотеки**

Для бизнеса

- Директ**  
Качественная реклама
- Справочник**  
Добавить организацию на Яндекс.Карты
- Реклама на Яндексе**  
Все рекламные возможности Яндекса
- Статистика**  
Аудитория сервисов Яндекса

Программы для вашего компьютера

- Яндекс.Бар**  
Поиск, Логика, Пробки — одной строкой
- Библиотека Яндекс.Бар**  
Каталог иконок и виджетов
- Скайп**  
Приложение уведомлений, общение в интернете
- Решит Виртуал**  
Автоматический персональный диспетчер клавиатуры
- Mozilla Firefox**  
Официальная версия с поиском Яндекса
- Internet Explorer**  
Официальная версия с поиском Яндекса
- Хром**  
Браузер с поиском Яндекса
- Скринсейвер Яндекс.Фото**  
Зеркало заставки для вашего компьютера

Яндекс.Нано — наши эксперименты

- Ответа**  
Почта ответов на вопросы
- Календарь**  
Поможет вам успеть [RSS](#)
- Лента**  
Истории блогов и RSS-ленток в одной ленте [RSS](#)

Вокруг интернета

- Интересы**  
Рейтинги на основе поисковых запросов
- Незабудки**  
Сборник спецпроектов
- Я.Интернет**  
Скорость интернет-соединения
- Коллекция**  
Материалы от Яндекса

Сервис 1

Сервис 2

Сервис 3

Сервис 5

Сервис 4

Сервис 1



Сервис 2



Сервис 3



Сервис 5



Сервис 4





Сервис 1



Сервис 2



Сервис 3



Сервис 5



Сервис 4

Сервис 3

Сервис 1

Сервис 2



Сервис 5

Сервис 4

Сервис 1

Сервис 3

Сервис 2



Сервис 5

Сервис 4

# Универсальность по типам алгоритмов

Выделим наиболее общие классы алгоритмов:

- Унификация
- Кластеризация
- Классификация
- Экстракция
- ...

# Унификация

страна: Россия

город: Спб

улица: Московский пр-т.

дом: д.182

литер: а



страна: Россия

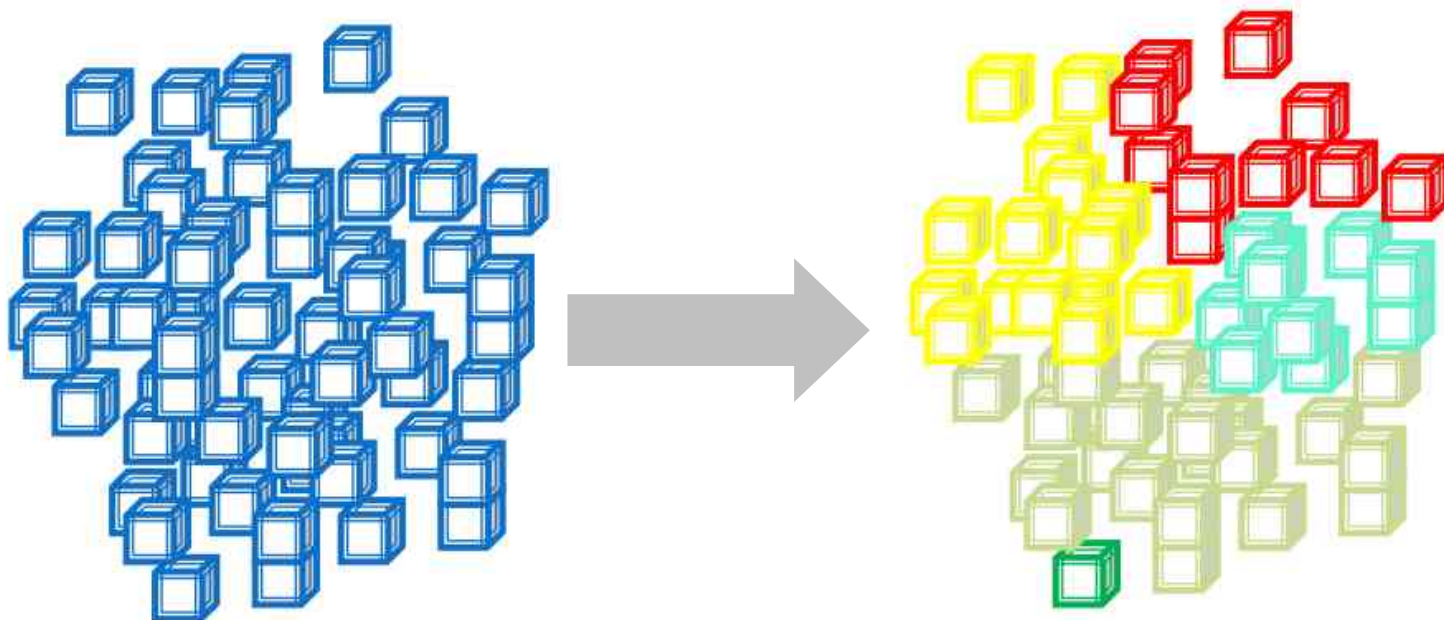
город: Санкт-Петербург

улица: Московский проспект

дом: 182

литер: А

# Кластеризация



**Что нужно для оценки?**

# Данные для опытов



# Эталон

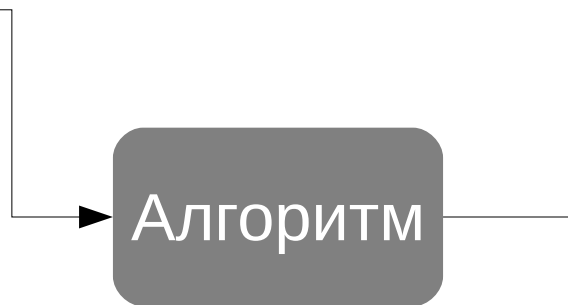


# Инструменты



**Как это работает?**

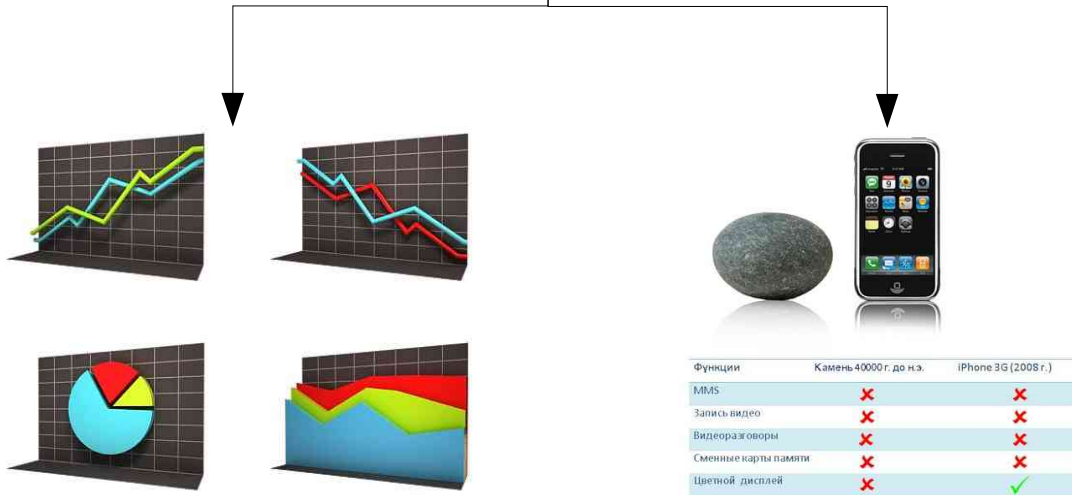
# Запускаем алгоритм



# Сравниваем с эталоном

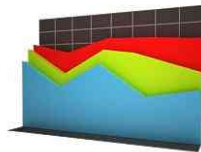


# Показываем результаты





Алгоритм



Функции	Камень 40000 г. до н.э.	iPhone 3G (2008 г.)
MMS	✗	✗
Запись видео	✗	✗
Видеооружия	✗	✗
Сменные карты памяти	✗	✗
Цветной дисплей	✗	✓

Проблема:  
 Нет системной  
 оценки сложных  
 алгоритмов

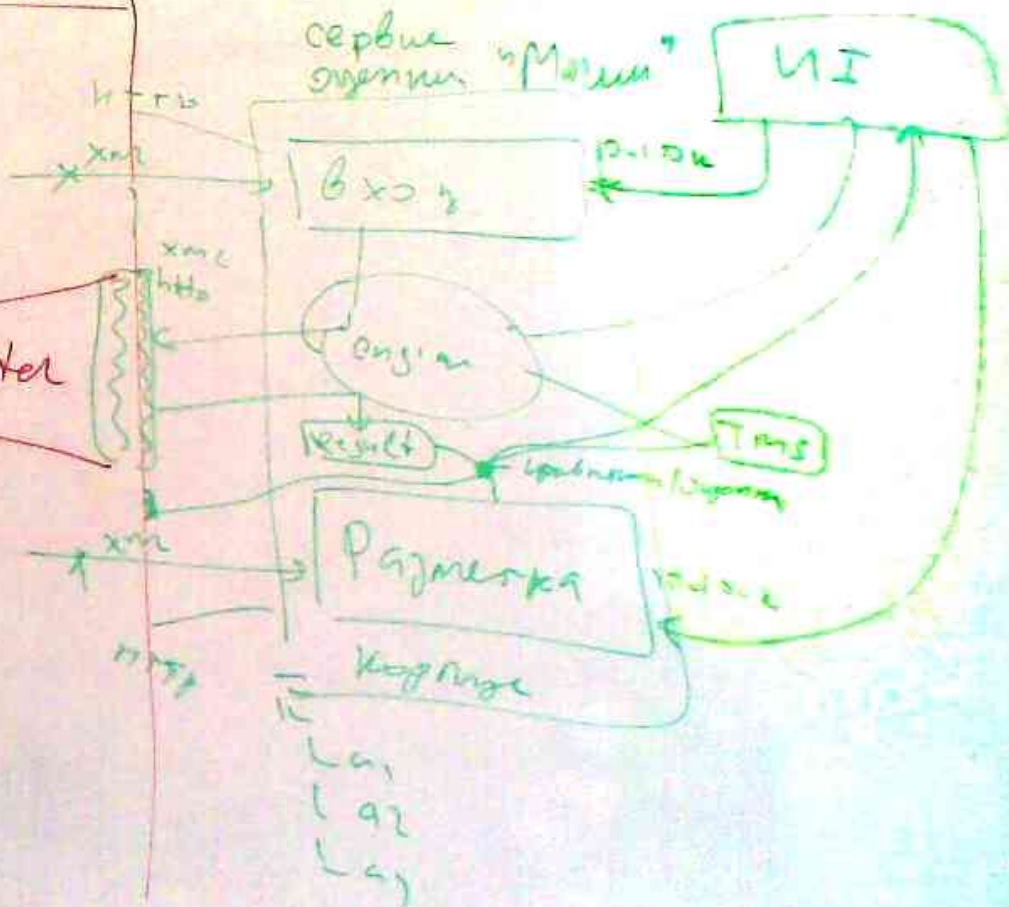
- классификация
- рейтинг
- метрики
- классификация
- экстракция
- кластеризация
- ранжирование

Документ  
 изменения  
 KPI

Any Service



Adapter



+2-3M

- L01
- L02
- L03

# **Н**емного подробностей

# Данные Корпус



# Данные Сущность



индекс: 196105  
страна: Россия  
регион: Санкт-Петербург  
город: Санкт-Петербург  
улица: Московский проспект  
дом: 182  
литер: А

# Данные Эталон

Проблема: где взять?

Варианты:

- «Псевдоэталон»
- Использовать готовый («а вдруг завалялся?..»)
- Взять и сделать



# Данные Результат



# Метрики унификация

Комплексные:

— Полнота по сущностям

— Точность

Единичные:

— Полнота по атрибутам

— Корректность значений




# Статистика



# Цифры

Время	Полнота по атрибутам	Корректность	Точность	Полнота по сущностям
31 Jan 2011, 12:01	↓ 0.996522	↑ 0.994732	↓ 0.849624	↑ 1.000000
29 Jan 2011, 12:01	↑ 0.999130	↑ 0.990673	↑ 0.885918	↓ 0.924217
27 Jan 2011, 12:01	↑ 0.996692	↑ 0.988764	↑ 0.794721	↑ 1.000000
26 Jan 2011, 12:01	0.996682	0.887738	0.002941	0.997067
25 Jan 2011, 17:37	↑ 0.996682	↓ 0.887738	↑ 0.002941	↓ 0.997067
25 Jan 2011, 12:01	0.950900	0.942026	0.000000	1.000000
24 Jan 2011, 12:01	0.950900	0.942026	0.000000	1.000000
23 Jan 2011, 12:02	0.950900	0.942026	0.000000	1.000000
22 Jan 2011, 12:01	0.950900	0.942026	0.000000	1.000000
21 Jan 2011, 12:01	↓ 0.950900	↓ 0.942026	↓ 0.000000	1.000000
28 Dec 2010, 12:02	0.991144	0.977404	0.748092	1.000000
27 Dec 2010, 12:02	0.991144	0.977404	0.748092	1.000000
26 Dec 2010, 12:01	0.991144	0.977404	0.748092	1.000000

# Цифры

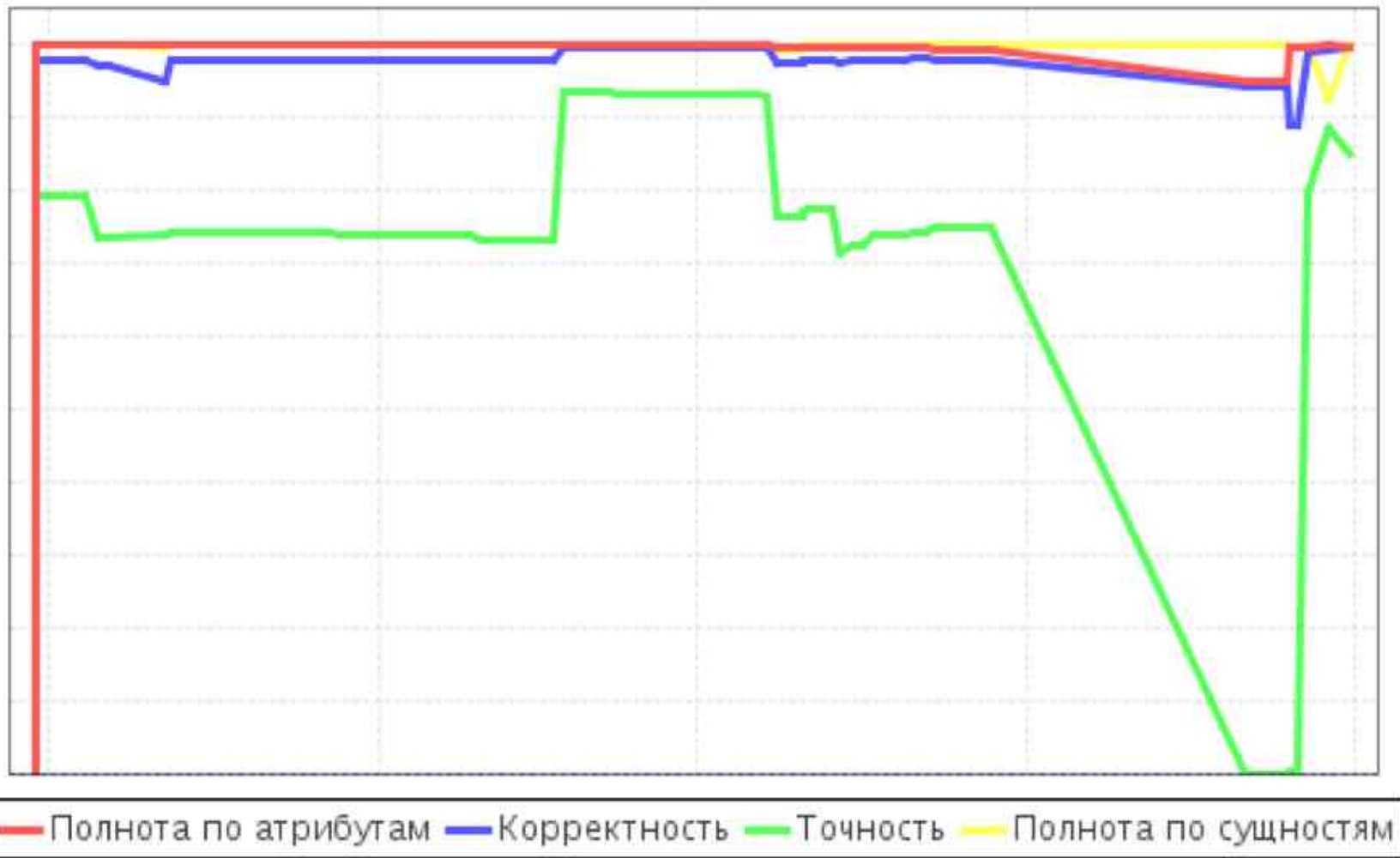


Время	Полнота по атрибутам	Корректность	Точность	Полнота по сущностям
31 Jan 2011, 12:01	↓ 0.996522	↑ 0.994732	↓ 0.849624	↑ 1.000000
29 Jan 2011, 12:01	↑ 0.999130	↑ 0.990673	↑ 0.885918	↓ 0.924217
27 Jan 2011, 12:01	↑ 0.996692	↑ 0.988764	↑ 0.794721	↑ 1.000000
26 Jan 2011, 12:01	0.996682	0.887738	0.002941	0.997067
25 Jan 2011, 17:37	↑ 0.996682	↓ 0.887738	↑ 0.002941	↓ 0.997067
25 Jan 2011, 12:01	0.950900	0.942026	0.000000	1.000000
24 Jan 2011, 12:01	0.950900	0.942026	0.000000	1.000000
23 Jan 2011, 12:02	0.950900	0.942026	0.000000	1.000000
22 Jan 2011, 12:01	0.950900	0.942026	0.000000	1.000000
21 Jan 2011, 12:01	↓ 0.950900	↓ 0.942026	↓ 0.000000	1.000000
28 Dec 2010, 12:02	0.991144	0.977404	0.748092	1.000000
27 Dec 2010, 12:02	0.991144	0.977404	0.748092	1.000000
26 Dec 2010, 12:01	0.991144	0.977404	0.748092	1.000000

# Цифры

Время	Полнота по атрибутам	Корректность	Точность	Полнота по сущностям
31 Jan 2011, 12:01	↓ 0.996522	↑ 0.994732	↓ 0.849624	↑ 1.000000
29 Jan 2011, 12:01	↑ 0.999130	↑ 0.990673	↑ 0.885918	↓ 0.924217
27 Jan 2011, 12:01	↑ 0.996692	↑ 0.988764	↑ 0.794721	↑ 1.000000
26 Jan 2011, 12:01	0.996682	0.887738	0.002941	0.997067
25 Jan 2011, 17:37	↑ 0.996682	↓ 0.887738	↑ 0.002941	↓ 0.997067
25 Jan 2011, 12:01	0.950900	0.942026	0.000000	1.000000
24 Jan 2011, 12:01	0.950900	0.942026	0.000000	1.000000
23 Jan 2011, 12:02	0.950900	0.942026	0.000000	1.000000
22 Jan 2011, 12:01	0.950900	0.942026	0.000000	1.000000
21 Jan 2011, 12:01	↓ 0.950900	↓ 0.942026	↓ 0.000000	1.000000
28 Dec 2010, 12:02	0.991144	0.977404	0.748092	1.000000
27 Dec 2010, 12:02	0.991144	0.977404	0.748092	1.000000
26 Dec 2010, 12:01	0.991144	0.977404	0.748092	1.000000

# Графики



# Эпилог

Сервис 1

Сервис 2

Сервис 3

Сервис 5

Сервис 4

**VS**

Сервис 1

Сервис 3

Сервис 2

Система  
оценки

Сервис 5

Сервис 4

# Вопросы



## **Александр Коваленко**

Руководитель группы

195027, Россия, Санкт-Петербург,  
Свердловская набережная, д. 44

[alex-kovalenko@yandex-team.ru](mailto:alex-kovalenko@yandex-team.ru)