

# Масштабируемый сервис для анализа текстов на основе системы Texterra

Турдаков Денис  
[turdakov@ispras.ru](mailto:turdakov@ispras.ru)

# Texterra

Масштабируемое решение для анализа текстов с открытым API, основанное на использовании баз знаний, извлекаемых из Веб-ресурсов



# Texterra: история проекта



## Текущее состояние:

- Скорость обработки (полный разбор текста): 82 Кб/с
- Многоязычность: Английский, Русский, Корейский
- Поддержка нескольких баз знаний
- Облачный сервис

# Базы знаний



WIKIPEDIA  
The Free Encyclopedia

Википедия – основной источник



Любые ресурсы MediaWiki



Linked Open Data

In gemeinsamer Arbeit mit dem Typographen und Schriftkünstler Oldrich Menhart haben wir eine neue Type geschaffen. Es erfüllt uns mit Stolz, eine Schrift nach den Entwürfen dieses anerkannten Künstlers herausgegeben zu haben und zwar mit einem Erfolg, der nicht nur bei den heimischen Fachleuten, sondern auch bei

Text Collections (Astrakhantsev et.al.  
Dialog'14)

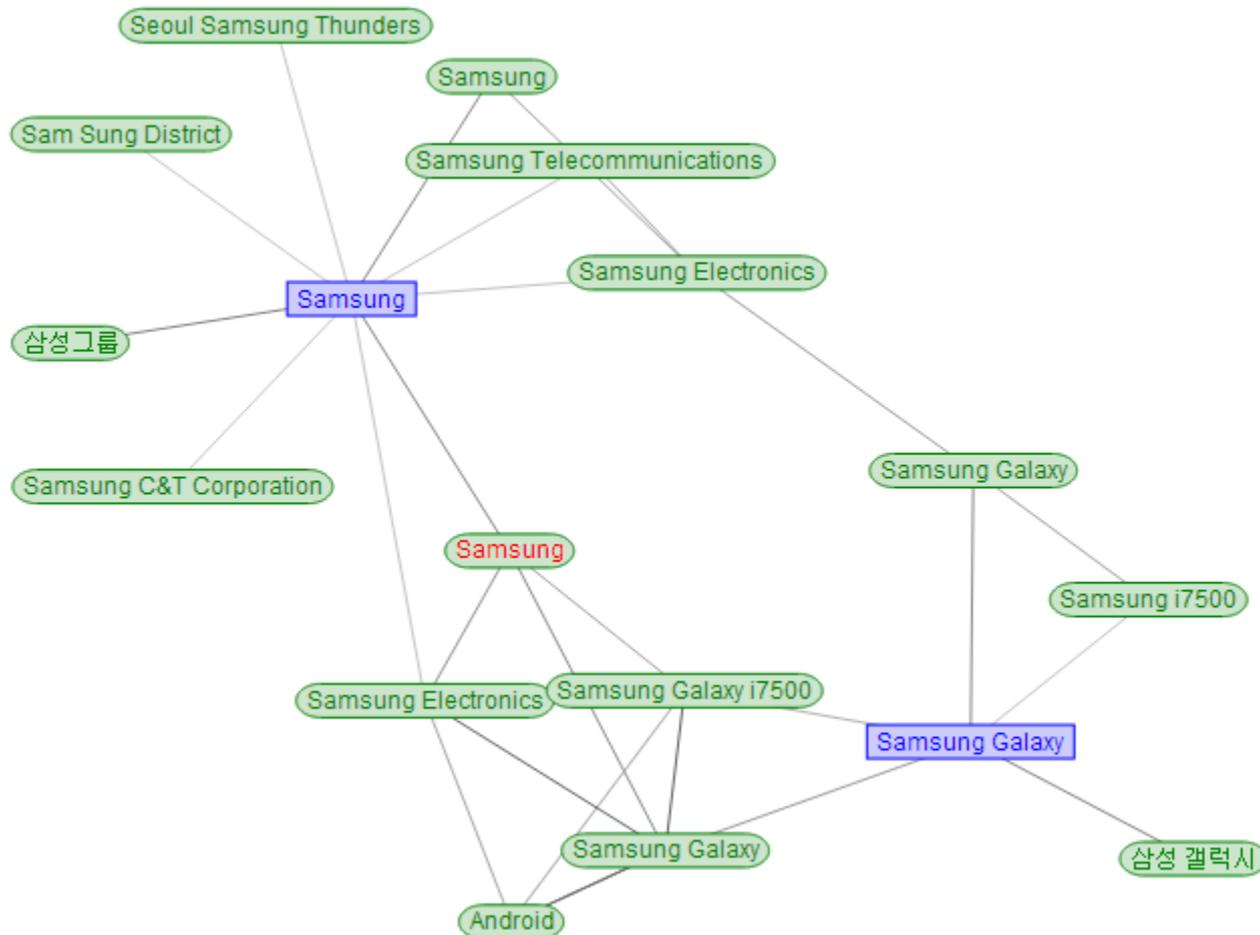
# Работа с базой знаний

Семантически  
близкие понятия

Получение  
синонимов

Проход по  
категориям

Обход  
семантического  
графа



# Инструменты для анализа текстов

Определение  
языка

Определение  
границ  
предложений

Разбиение на  
лексемы

Лемматизация

Определение  
частей речи

Исправление  
ошибок в  
правописании

Извлечение и  
классификация  
именованных  
сущностей

Анализ  
тональности  
текста

Поиск составных  
терминов

Разрешение  
лексической  
многозначности

Определение  
ключевых  
понятий

# Использование

<https://api.ispras.ru>

Веб формы для  
демонстрации  
отдельных функций

Clear Example Text Example Tweet Example Review

Институт системного программирования РАН имеет более 200 высококвалифицированных постоянных сотрудников и около 80 специалистов, работающих по контрактам. 12 сотрудников - доктора наук и 45 имеют степень кандидата наук. Многие сотрудники преподают в Московском государственном университете и Московском физико-техническом институте.

## Result

Институт системного программирования РАН имеет более 200 высококвалифицированных постоянных сотрудников и около 80 специалистов, работающих по контрактам. 12 сотрудников - доктора наук и 45 имеют степень кандидата наук. Многие сотрудники преподают в Московском государственном университете и Московском физико-техническом институте.

Key Concept

Disambiguation

Lurkification

Sentiment Analysis

Aspect Extraction

Tweet Normalization

## Wikipedia

**Институт системного программирования Российской академии наук (РАН)** был основан 25 января 1994 года на базе бывшего Института проблем кибернетики РАН. ИСП РАН входит в Отделение математических наук РАН.

[See more](#)

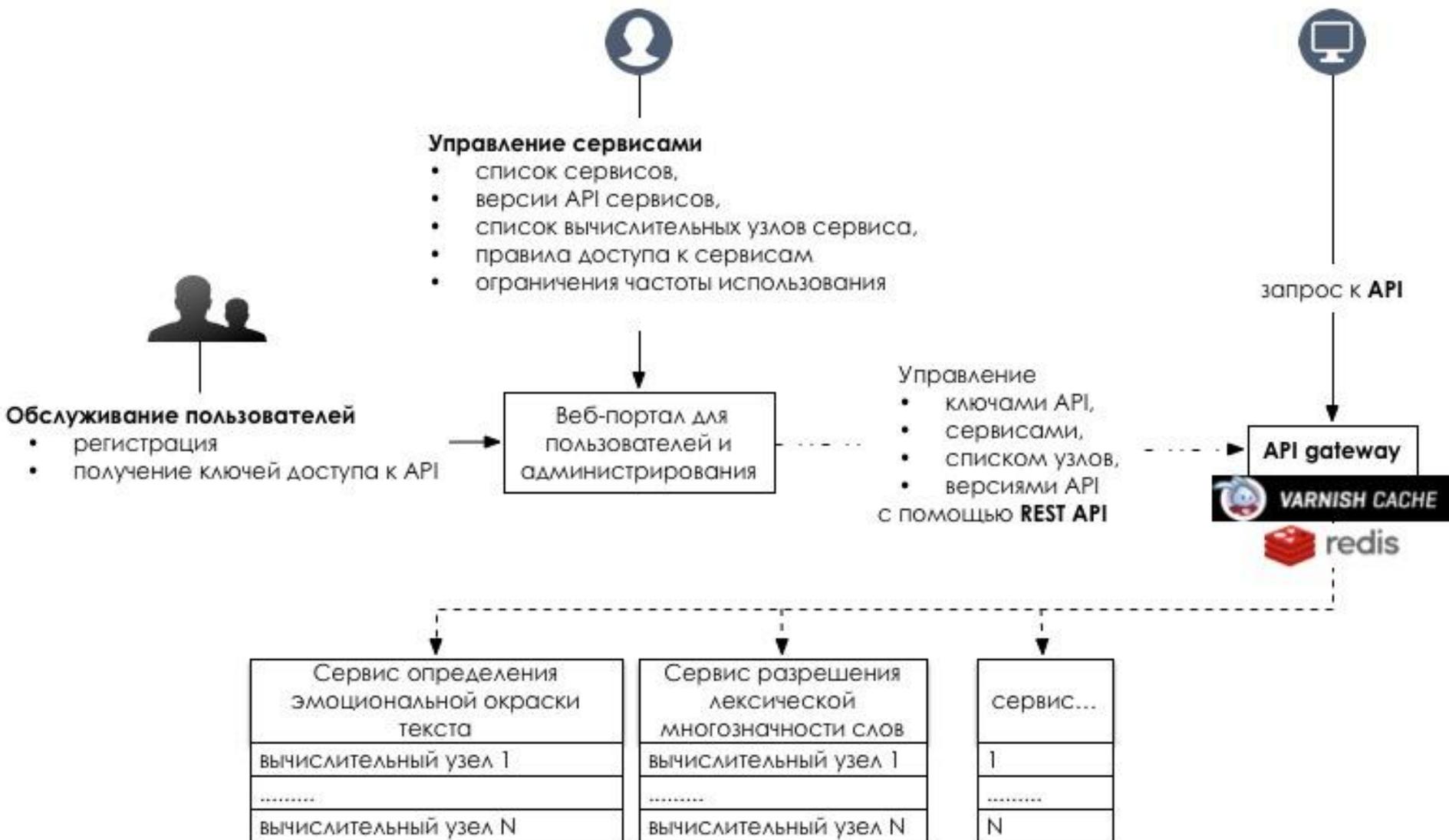
# Использование: REST API

```
<NLP-document>
  <text id="0">The Beatles were more than a great rock band.</text>
  <annotations class="java.util.ArrayList">
    <l-annotation confidence="1.0" class="PennPOSToken">
      <start>0</start>
      <end>3</end>
      <annotated-text reference="0"/>
      <value class="PennPOSTag">DT</value>
    </l-annotation>
    <l-annotation confidence="1.0" class="PennPOSToken">
      <start>4</start>
      <end>11</end>
      <annotated-text reference="0"/>
      <value class="PennPOSTag">NNP</value>
    </l-annotation>
    <l-annotation confidence="1.0" class="PennPOSToken">
      <start>12</start>
      <end>16</end>
      <annotated-text reference="0"/>
      <value class="PennPOSTag">VBD</value>
    </l-annotation>
    ...
  </annotations>
</NLP-document>
```

# Подсистема масштабирования в облаке (входные условия)

- Множество независимых сервисов по обработке текстовых данных при помощи публично доступного API
- Прозрачная поддержка миграции между версиями API
- Вычислительные узлы каждого отдельно взятого сервиса не взаимодействуют друг с другом
- Предоставляемые сервисы расходуют много ресурсов, и вам нужно гибко ограничивать использование вашего API пользователями

# API gateway: схема компонентов



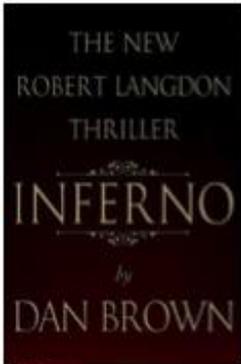
# API Gateway: обзор функций

- Администратор системы может управлять предоставлением сервисов при помощи веб-интерфейса или REST API системы:
  - Задавать имя и версию каждого сервиса.
  - Указывать список вычислительных узлов, относящихся к данной версии API сервиса.
  - Создавать группы доступа к сервисам и распределять пользователей по группам.
  - Задавать количество доступных ключей группам пользователей и частоту использования ключей для каждого предоставляемого сервиса.
- После регистрации на сайте системы пользователь:
  - Получает один или несколько ключей для использования разрешенных ему сервисов.
  - Использует единую точку входа для всех сервисов, не подозревая о том, что за этой точкой входа могут располагаться сотни вычислительных узлов, предоставляющих разные сервисы.
- Система следит за
  - Частотой использования каждого из пользовательских ключей.
  - Равномерной балансировкой поступающих запросов между вычислительными узлами.
  - Жизнеспособностью вычислительных узлов.

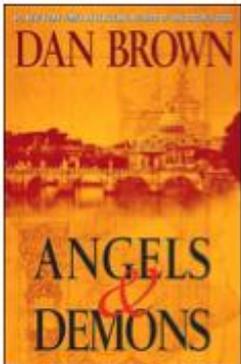
# Примеры использования



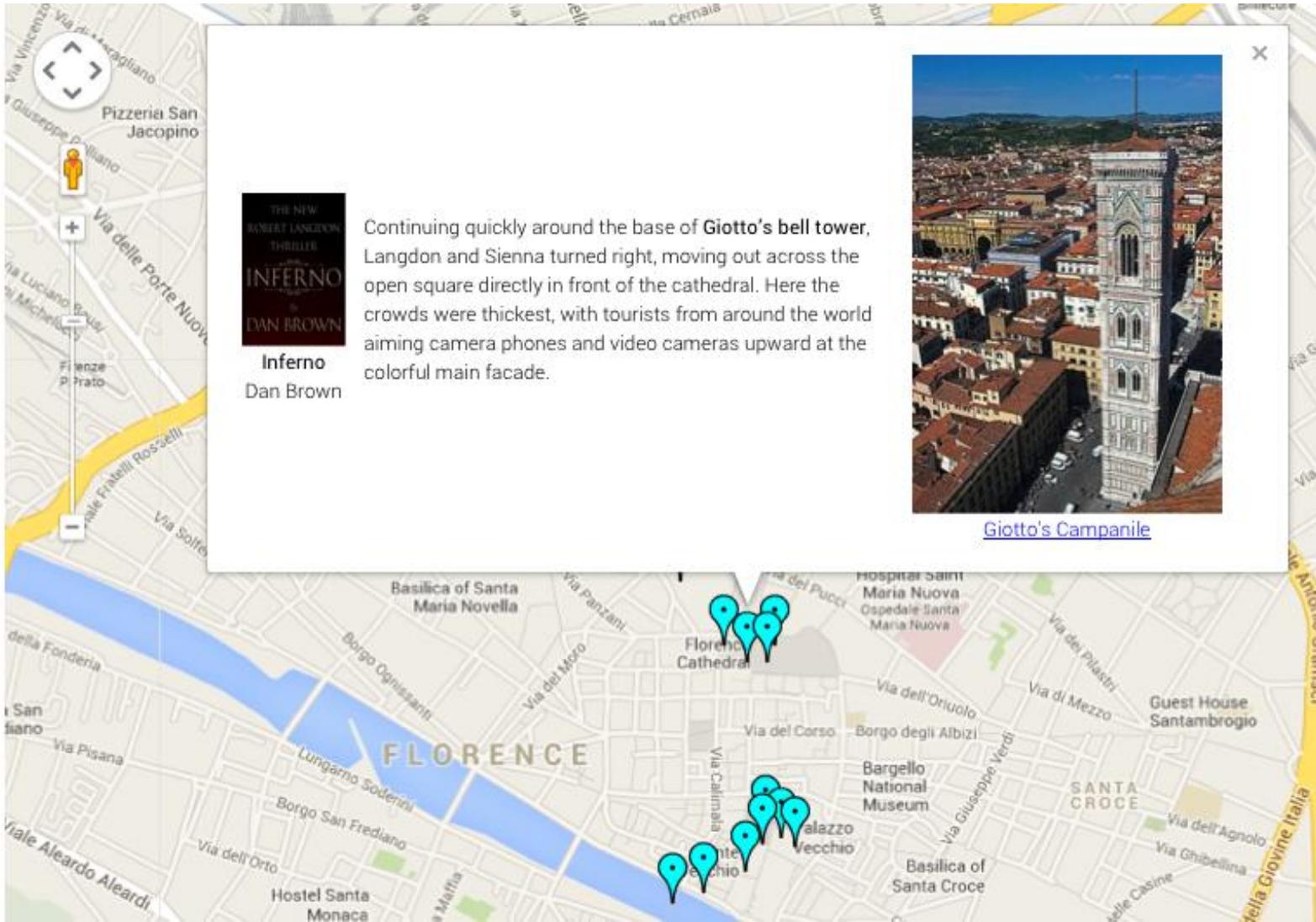
# Книги на карте



Inferno  
Dan Brown



Angels & Demons  
Dan Brown



# Информационный поиск

Расширение и  
уточнение запросов

Динамическая  
фасетная навигация

**BlogNoon**

Search posts  Search blogs

Your query: [Jaguar](#)

Sorted by **relevance** | [date](#)

## Exactly what can swim such

from [ethelexperience.com](#) - 09.12.2011

Church Calgary They're usually present in groups of ten to twenty. They are effortless prey for your **jaguar** and the anacondas who find them to become tasty. Also, they are victim to man searcher who look all of them for their meats, and also their particular pelts, even though nearly all of their

[Capybara](#) ; [Aquatic plant](#) ; [South America](#) ; [Pork](#) ; [Habitat](#) ; [Ingroups and outgroups](#) ; [Meat](#)

## Recommended Reads Saturday: Magic Bites by Ilona Andrews

[Book](#) ; [Fantasy](#) ; [Saturday](#) ; [Promenade II](#) ; [Atlanta](#) ; [CARE \(relief](#)

entary from a reader's point of view - 24.03.2012  
another. A lean shadow flickered in the corner of my  
in the jutting block of concrete, an elegant statue  
ca. The killer who

[Promenade II](#) ; [Atlanta](#) ; [CARE \(relief](#)

[v motion](#)

their adult lives. When a sloth reaches six months of age, it's old enough to be left on its own. Before that time, however, if a youngster falls from a tree, the mother will not attempt to rescue it; the risk of attack by a bird or **jaguar** is too great. Young sloths separated from their

[Sloth](#) ; [Two-toed sloth](#) ; [Leafcutter ant](#) ; [Algae](#) ; [Moth](#) ; [Tai chi chuan](#) ; [Harpy Eagle](#)

## Your Next Query

relevance | [bundles](#)

### Felids

[Panthera hybrid](#)

[Panthera](#)

[Big cat](#)

### Mesoamerica

[Olmec](#)

[Mexico](#)

[Mesoamerica](#)

### South America

[Amazon Rainforest](#)

[Beringia](#)

[South America](#)

### Miscellaneous

[Animal](#)

[Predation](#)

[Habitat](#)

[Venezuela](#)

[Sloth](#)

[Jaguar Cars](#)

[Capybara](#)

Поиск по значениям  
многозначных слов

# Другие базы знаний

Clear Example Text Example Tweet Example Review

Бюджет ЦРУ составлял тогда 4 миллиарда долларов

Key Concept

Disambiguation

Lurkification

Sentiment Analysis

Aspect Extraction

Tweet Normalization

Result

Бюджет **ЦРУ** составлял тогда 4 миллиарда долларов

Wikipedia

**Кровавая гэбня** (*кгававая гэбня*) — выражение, употребляемое **поцреотами** для насмешек в адрес **лейбералов** и сторонников **теории заговора**. Используется для обозначения **ФСБ**

<https://lurkmore.to/>

# Вопросы?

<https://api.ispras.ru>

turdakov@ispras.ru