ИСП РАН

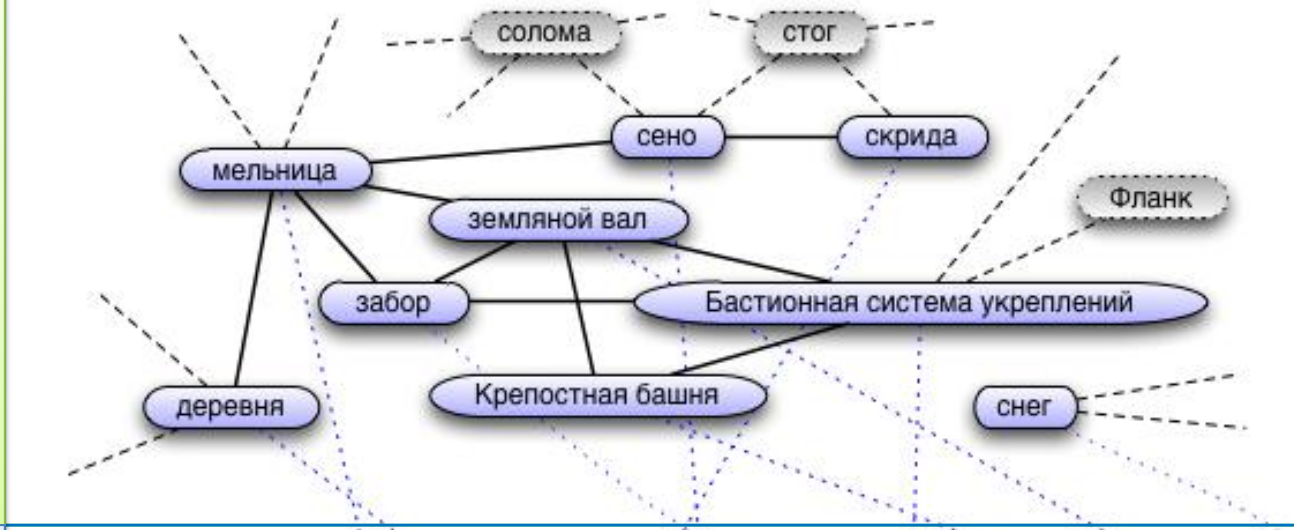
Масштабируемый сервис для анализа текстов на основе системы Texterra

Турдаков Денис
turdakov@ispras.ru

ИСП РАН

Texterra

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

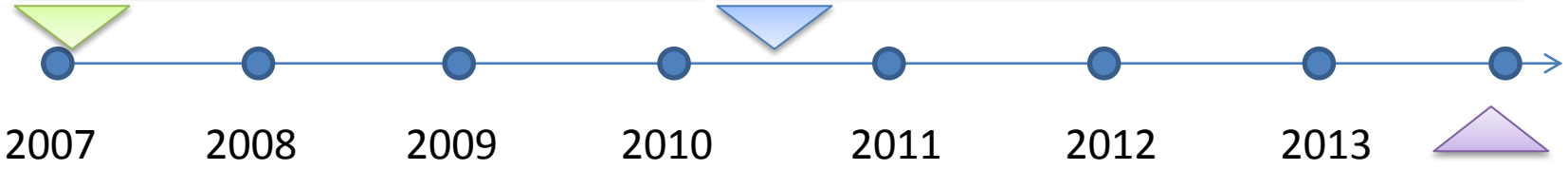
База знаний	
Инструменты обработки текстов	Я глядел во все стороны, ожидая увидеть грозные бастионы, башни и вал; но ничего не видал, кроме деревушки, окруженной бревенчатым забором. С одной стороны стояли три или четыре скирда сена, полузанесенные снегом; с другой — скривившаяся мельница, с лубочными крыльями, лениво опущенными.

А.С. Пушкин. Капитанская дочка

Texterra: история проекта

- Начало проекта
- Первая публикация: Turdakov, Velikhov, SYRCoDIS 2008
- Реализация базовой функциональности

- Коммерческие заказчики
- Оптимизация расхода вычислительных ресурсов
- Улучшение удобства использования



Текущее состояние:

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

Базы знаний

- Википедия – основной источник
- Любые ресурсы MediaWiki
- Linked Open Data
- Text Collections (Astrakhantsev et.al. Dialog’14)

ИСП РАН

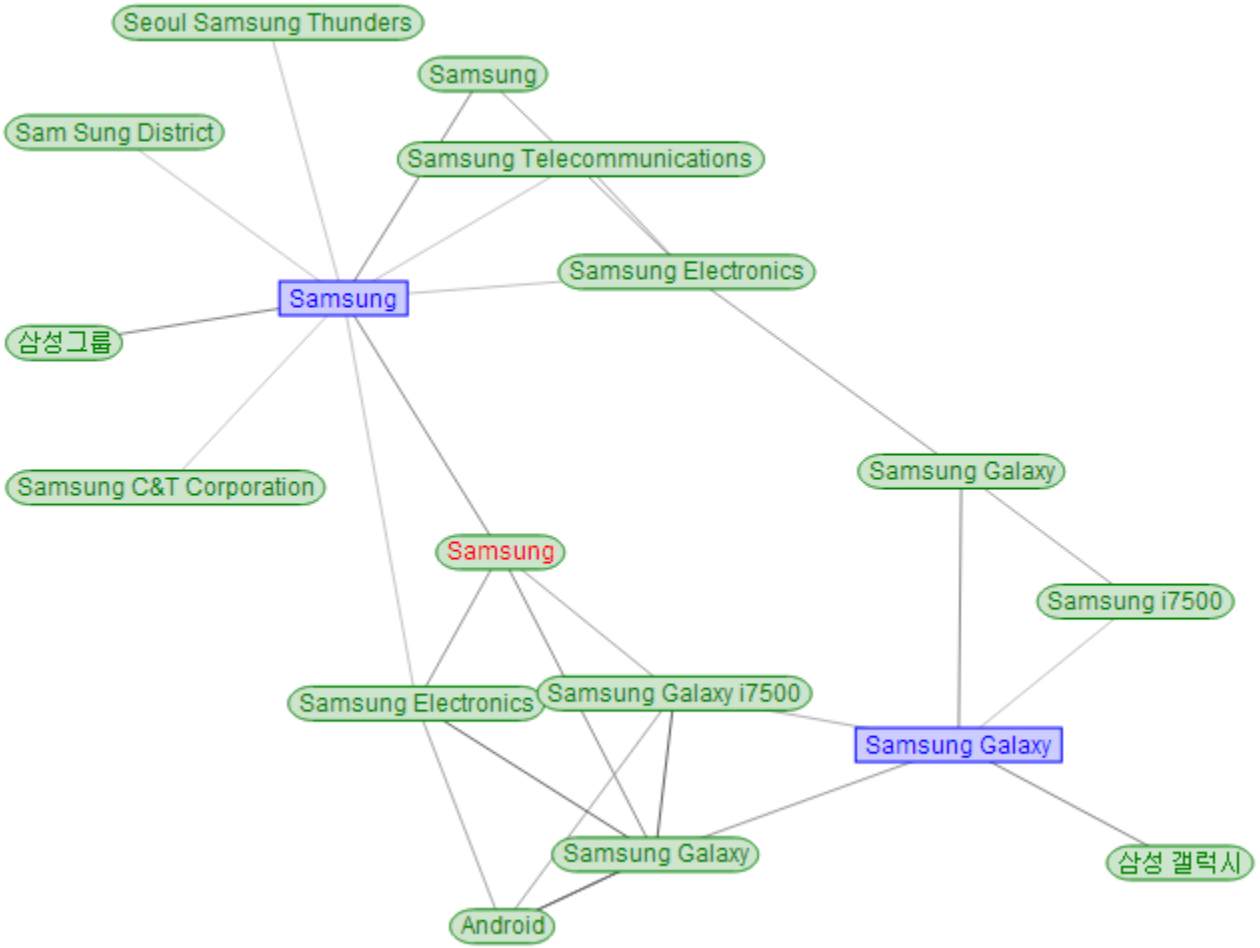
Работа с базой знаний

Семантически близкие понятия

Получение синонимов

Проход по категориям

Обход семантического графа



ИСП РАН

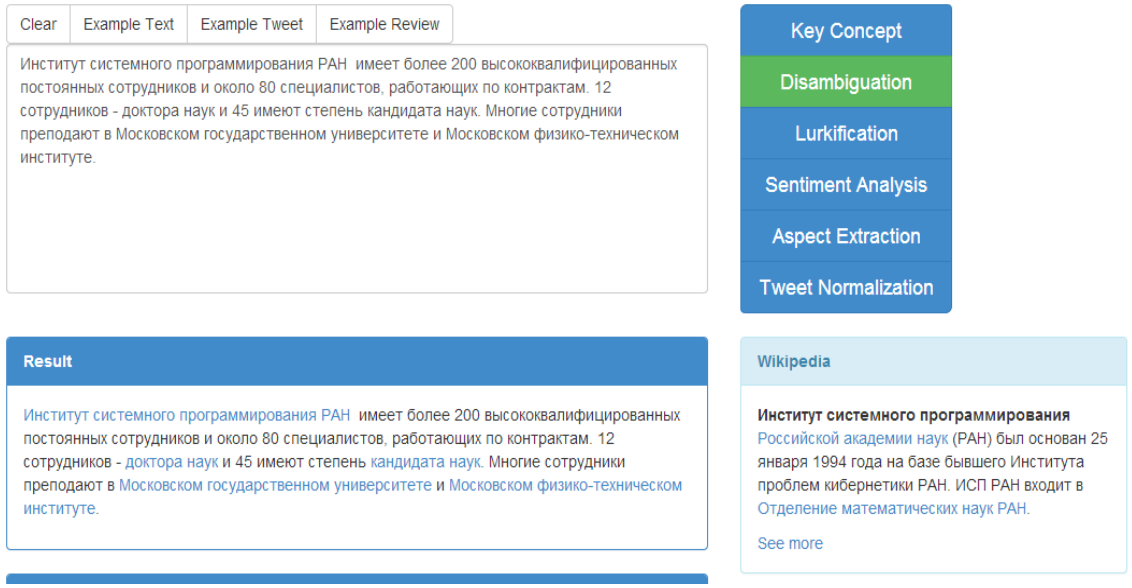
Инструменты для анализа текстов

- 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



ИСП РАН

Использование: REST API

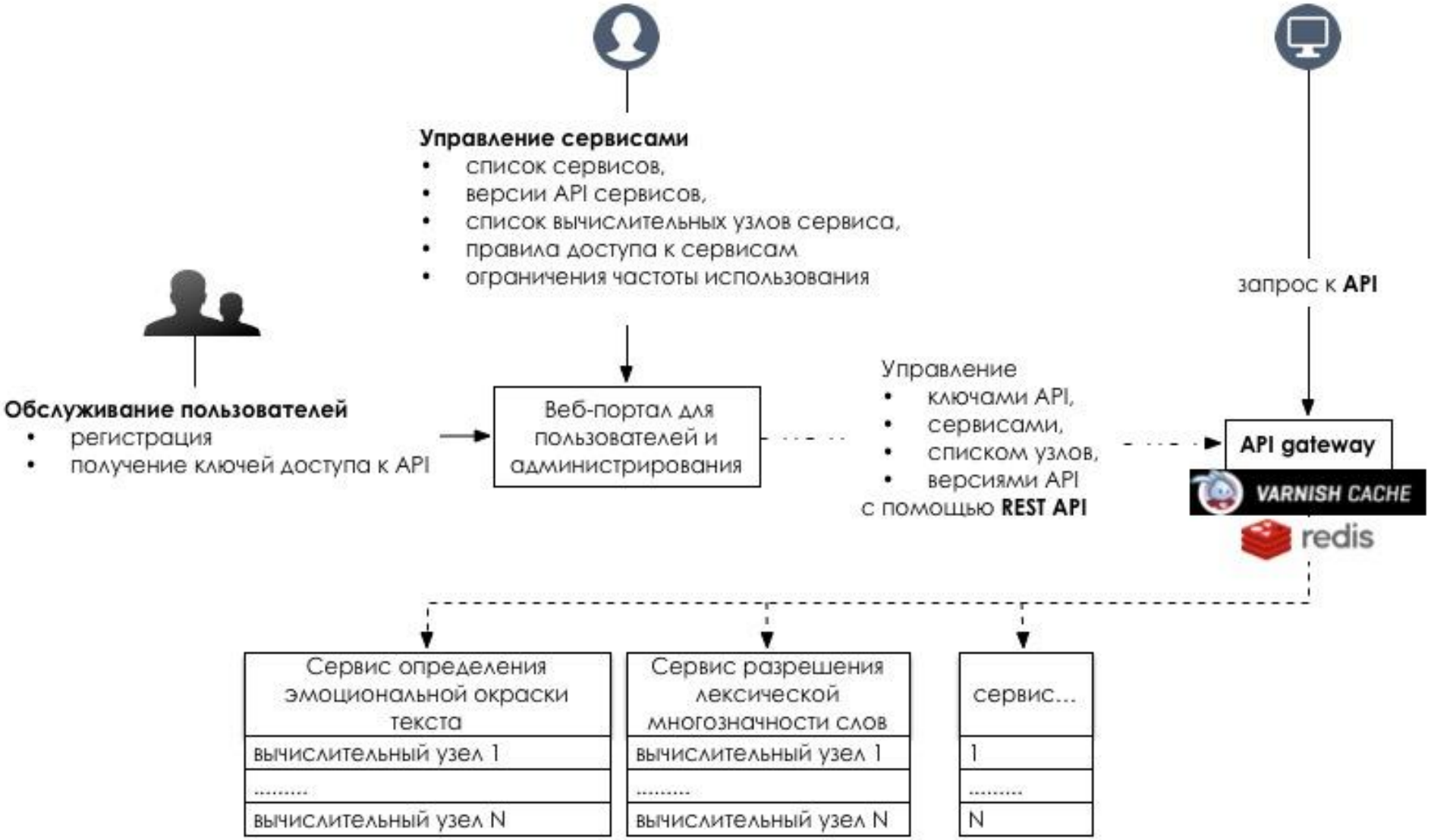
```
<NLP-document>
  <text id="0">The Beatles were more than a great rock band.</text>
  <annotations class="java.util.ArrayList">
    <l-annotation confidence="1.0" class="PennPOSToken">
      <start>0</start>
      <end>3</end>
      <annotated-text reference="0"/>
      <value class="PennPOSTag">DT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

ИСП РАН

API gateway: схема компонентов



API Gateway: обзор функций

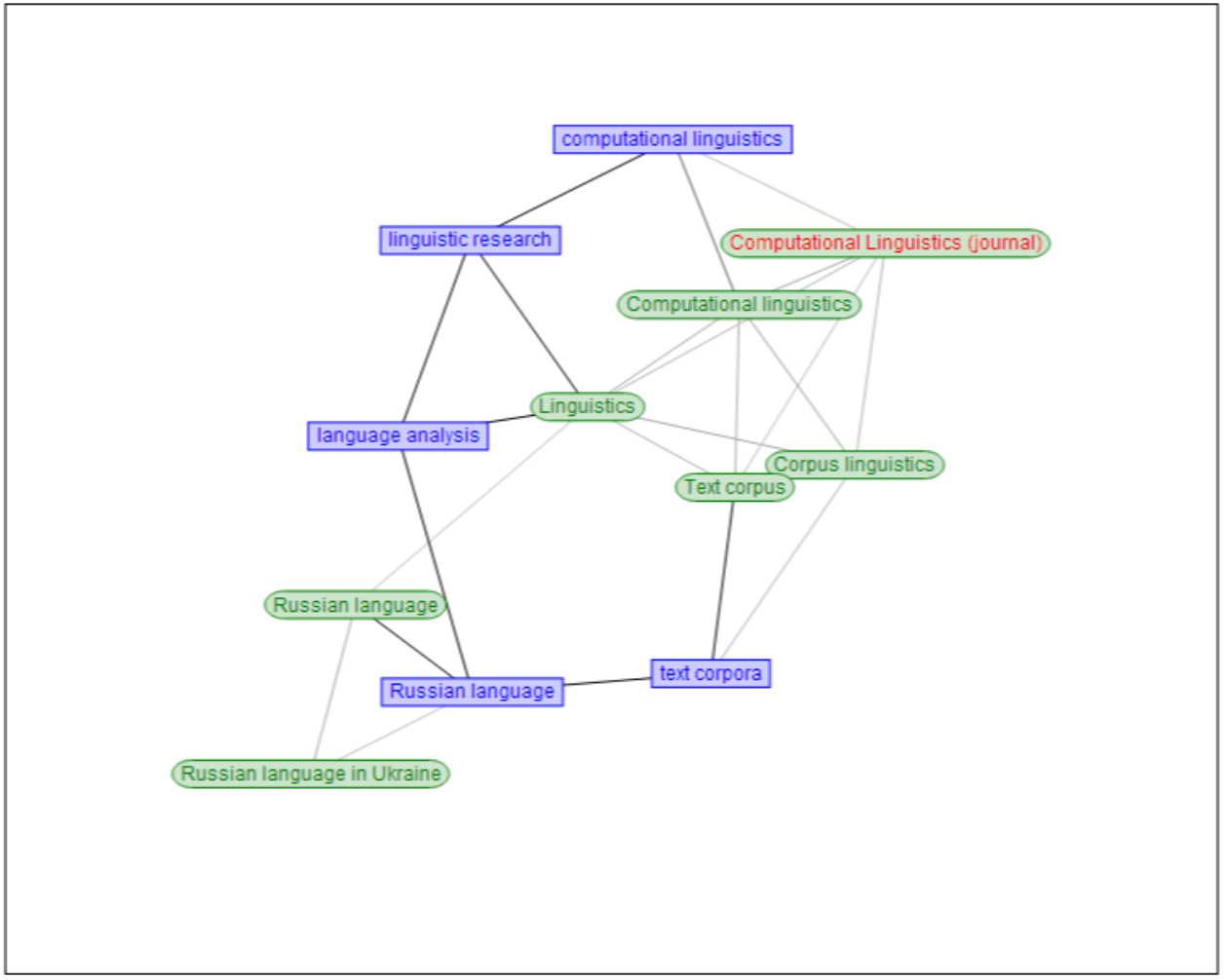
- Администратор системы может управлять предоставлением сервисов при помощи веб-интерфейса или REST API системы:
  - Задавать имя и версию каждого сервиса.
  - Указывать список вычислительных узлов, относящихся к данной версии API сервиса.
  - Создавать группы доступа к сервисам и распределять пользователей 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

ИСП РАН

VisOntia

Russian language | Add | Remove All | 0.01 | Similar Threshold



Rendering | Pause | Resume | Options

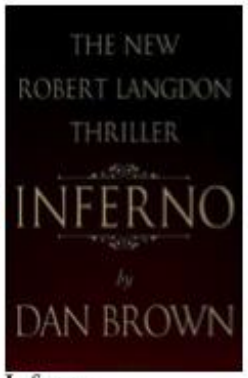
Selection

Name: Computational Linguistics (journal)
Node type: concept
ID: 924945:enwiki

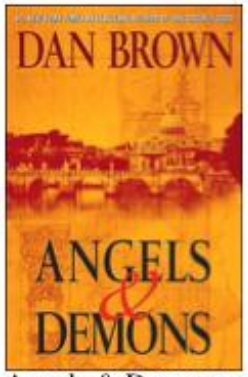
Graph | Wiki | Similar | Remove

ИСП РАН

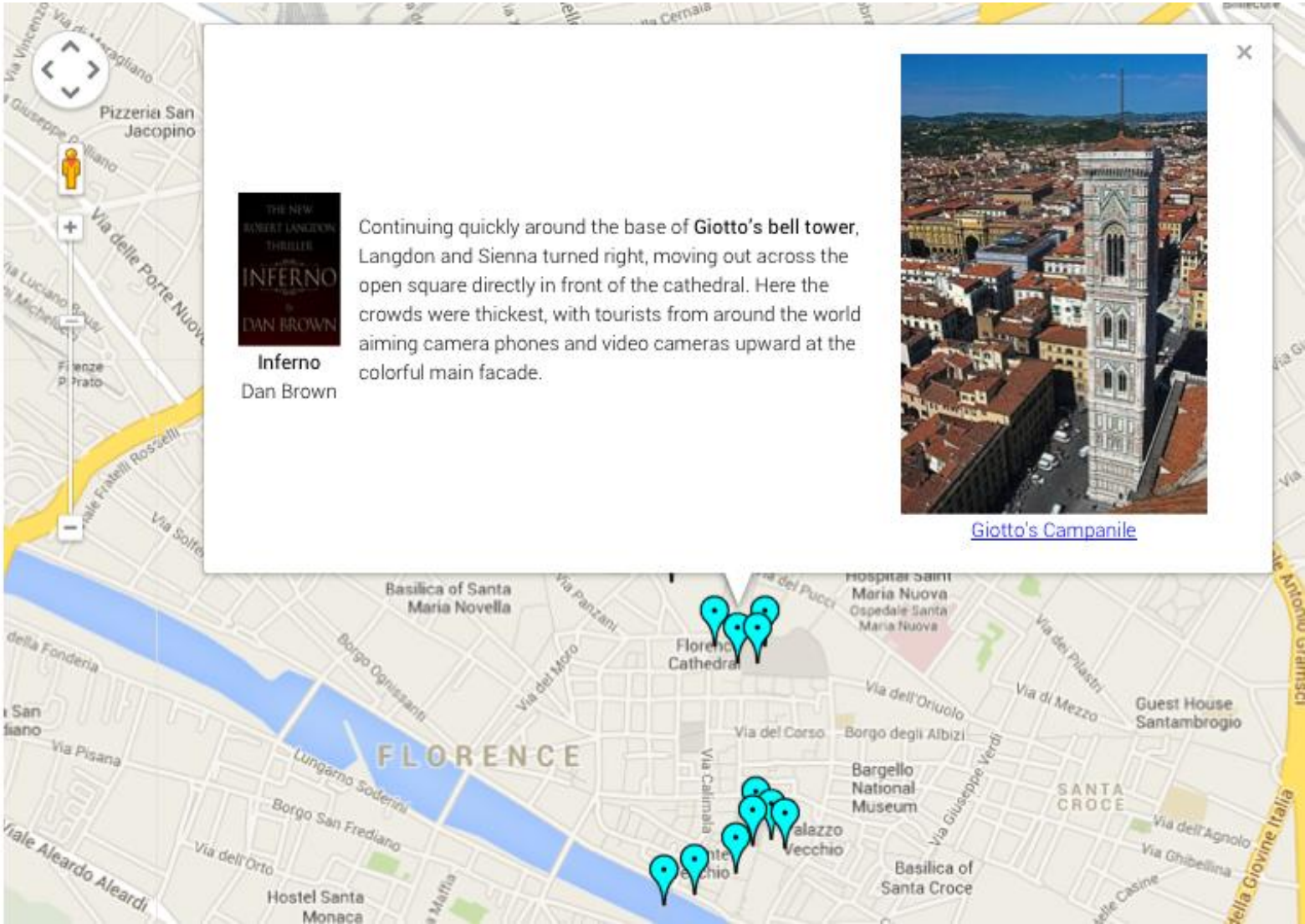
Книги на карте



Inferno
Dan Brown



Angels & Demons
Dan Brown



Inferno
Dan Brown

Continuing quickly around the base of **Giotto's bell tower**, Langdon and Sienna turned right, moving out across the open square directly in front of the cathedral. Here the crowds were thickest, with tourists from around the world aiming camera phones and video cameras upward at the colorful main facade.

Giotto's Campanile

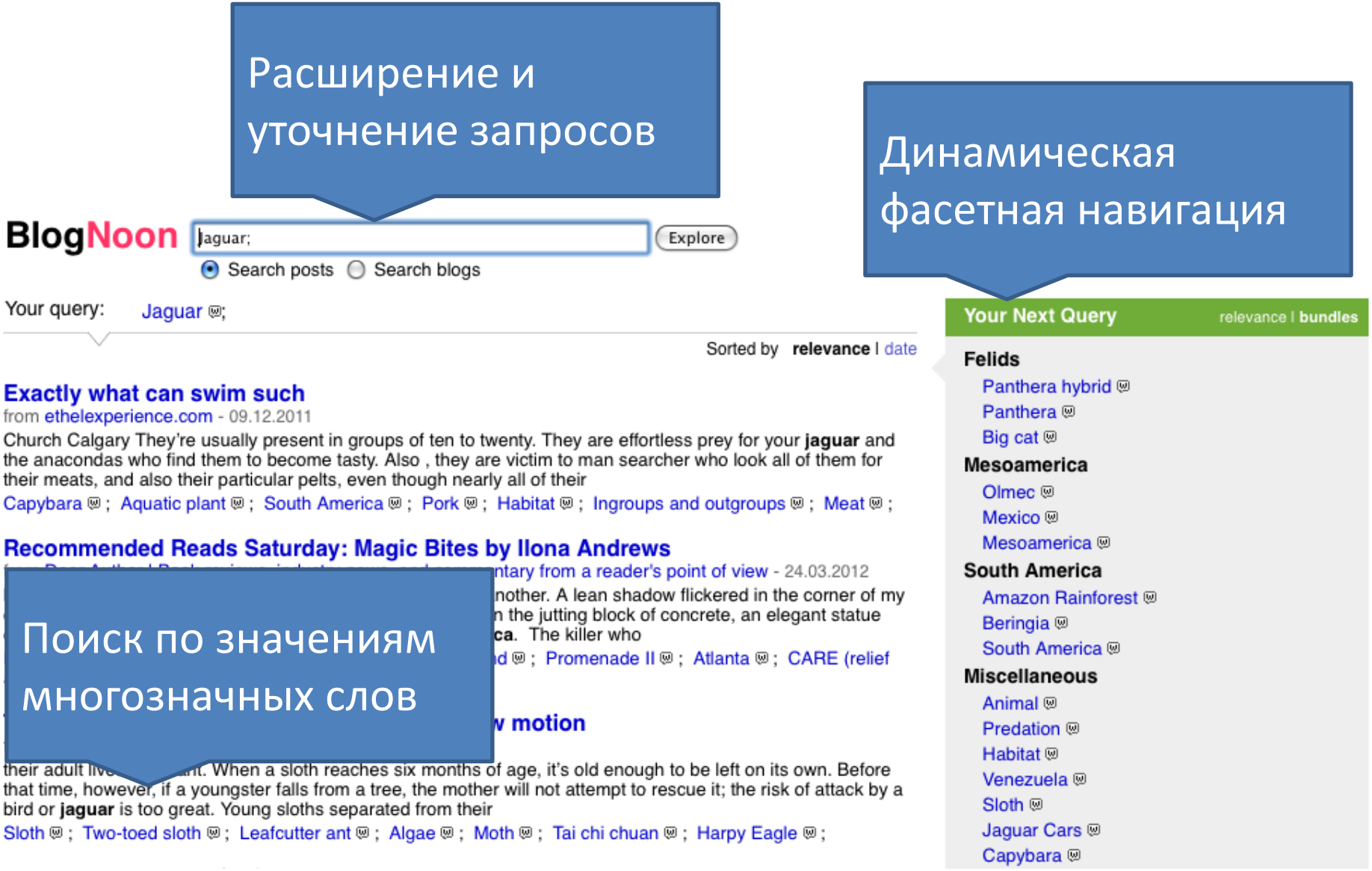
ИСП РАН

Информационный поиск

Расширение и уточнение запросов

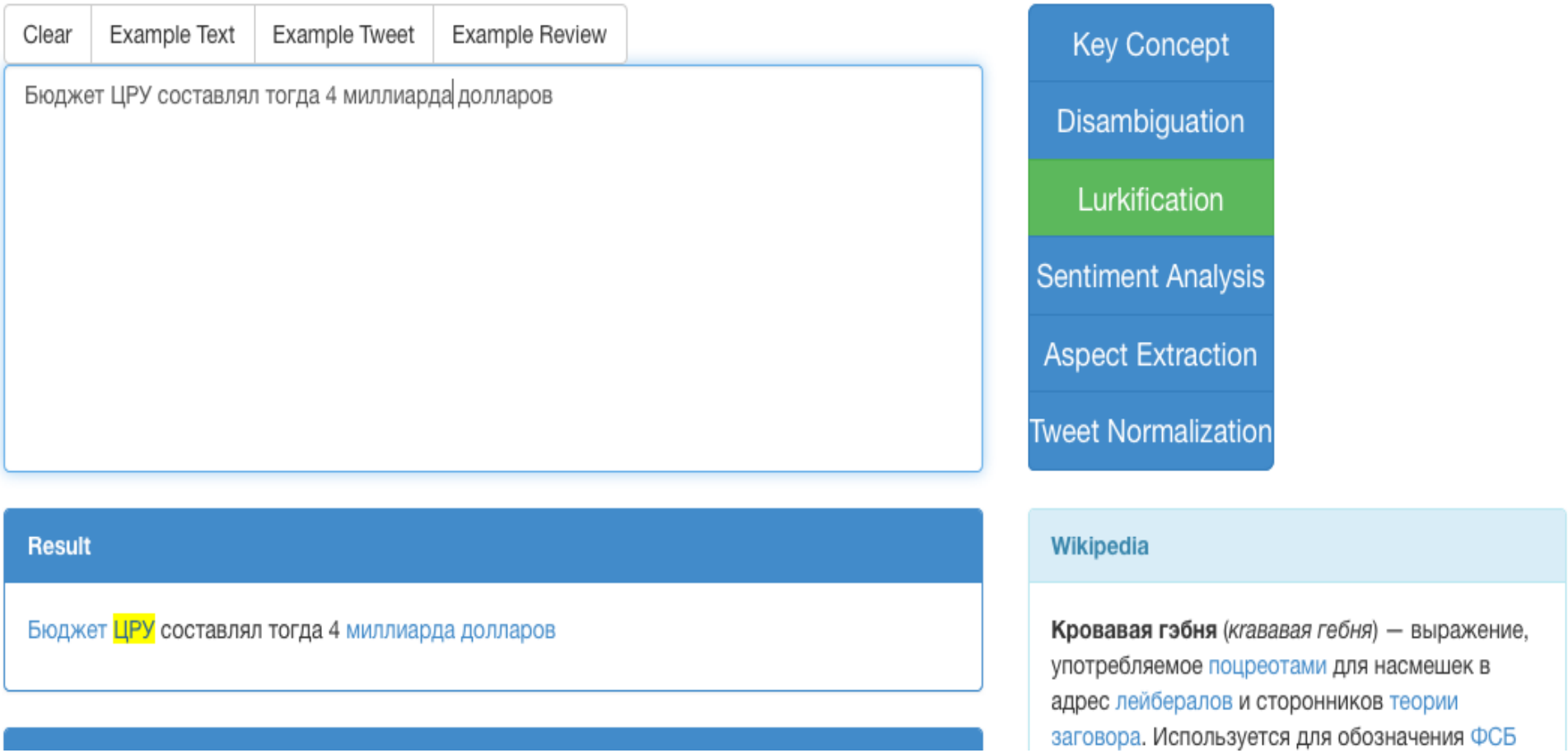
Динамическая фасетная навигация

Поиск по значениям многозначных слов



ИСП РАН

Другие базы знаний



https://lurkmore.to/

ИСП РАН

Вопросы?

https://api.ispras.ru

turdakov@ispras.ru