БАЗОВЫЙ СЕМАНТИЧЕСКИЙ АНАЛИЗАТОР TEXTERRA

Texterra — масштабируемая платформа для извлечения семантики из текста. Базовый комплекс технологий для создания многофункциональных прикладных приложений. Анализирует тексты с помощью выделения концептов. Включена в Единый реестр российского ПО (№4048).

ИСП РАН

ОСОБЕННОСТИ И ПРЕИМУЩЕСТВА

Texterra осуществляет уникальный анализ русскоязычных текстов на основе выделения концептов, а не только слов. Отличается от иностранных аналогов преимущественным вниманием к русскому языку. Базируется на результатах фундаментальных исследований и предоставляет возможность интеграции с поисковой системой Elasticsearch, существенно расширяя ее возможности. Удачное сочетание технологий позволяет платформе конкурировать с проектами уровня IBM Watson Natural Language Understanding.

TEXTERRA — ЭТО

- Высокая скорость обработки текста (морфологический анализ — 69 000 слов в секунду, синтаксический — 39 100 слов/сек, разрешение кореферентности — 10 100 слов/сек, полный разбор текста — приблизительно 13 600 слов/сек);
- Максимальное внимание к русскому языку (в отличие от аналогичных проектов spaCy и UDPipe, а также IBM Watson Natural Language Understanding, который не поддерживает анализ эмоций и концептов в русскоязычных текстах);
- Большой объем знаний (более 7 миллионов понятий);
- Построение базы знаний без привлечения экспертов (автоматическое пополнение с помощью Wikipedia, MediaWiki, Linked Open Data и др.);
- Масштабируемость как по скорости обработки текстов, так и по объему знаний (с помощью Apache Ignite и облачной среды Asperitas (ИСП РАН));
- Высокая точность анализа текста благодаря ряду ключевых особенностей:
  - Многоуровневый поиск по смежным понятиям;
  - Адаптивность к сленгу, хэштегам и ошибкам;
  - Анализ эмоциональной окраски (с разделением отношения к объектам и их атрибутам);
  - Определение взаимосвязей людей и компаний (на основе информации в тексте);
  - Определение неявных упоминаний объектов в дискуссиях.
- Высокая скорость разработки индивидуального решения;
- Два варианта использования:
  - в качестве отчуждаемого продукта на локальном сервере заказчика с доступом как по протоколу HTTP (REST-архитектура), так и по протоколу RMI;
  - онлайн на сайте https://texterra.ispras.ru/;
- Простое и быстрое освоение специфичных предметных областей и возможность интеграции новых языков для анализа (благодаря современному подходу к машинному обучению).

ДЛЯ КОГО ПРЕДНАЗНАЧЕНА TEXTERRA?

- Разработчики корпоративного ПО (в частности, чат-ботов);
- Разработчики систем семантического поиска для специфических предметных областей (информационная безопасность, медицина, аудит и т.п.);
- Разработчики прикладных систем обработки текста.

ОПЫТ ВНЕДРЕНИЯ

Texterra доработана до промышленного уровня в рамках сотрудничества с HP и Samsung (цель совместных проектов — получение технологий для анализа корпоративной отчетности и поддержки работы смарт-телевидения). В настоящее время Texterra используется в работе ряда технологий ИСП РАН (в частности, платформы для обработки данных Talisman). Texterra используется также рядом государственных ведомств России.

ПОДДЕРЖИВАЕМЫЕ ЯЗЫКИ

Texterra анализирует тексты на русском и английском языках.

СИСТЕМНЫЕ ТРЕБОВАНИЯ

- Любые платформы, поддерживаемые Java 8;
- Не менее 16 Гб оперативной памяти для каждого из анализируемых языков;
- Рекомендуется применение 64-битной версии ОС.

СХЕМА РАБОТЫ