Tomita parser

Крылова Ирина, Яндекс

ВШЭ, 20.04.2015

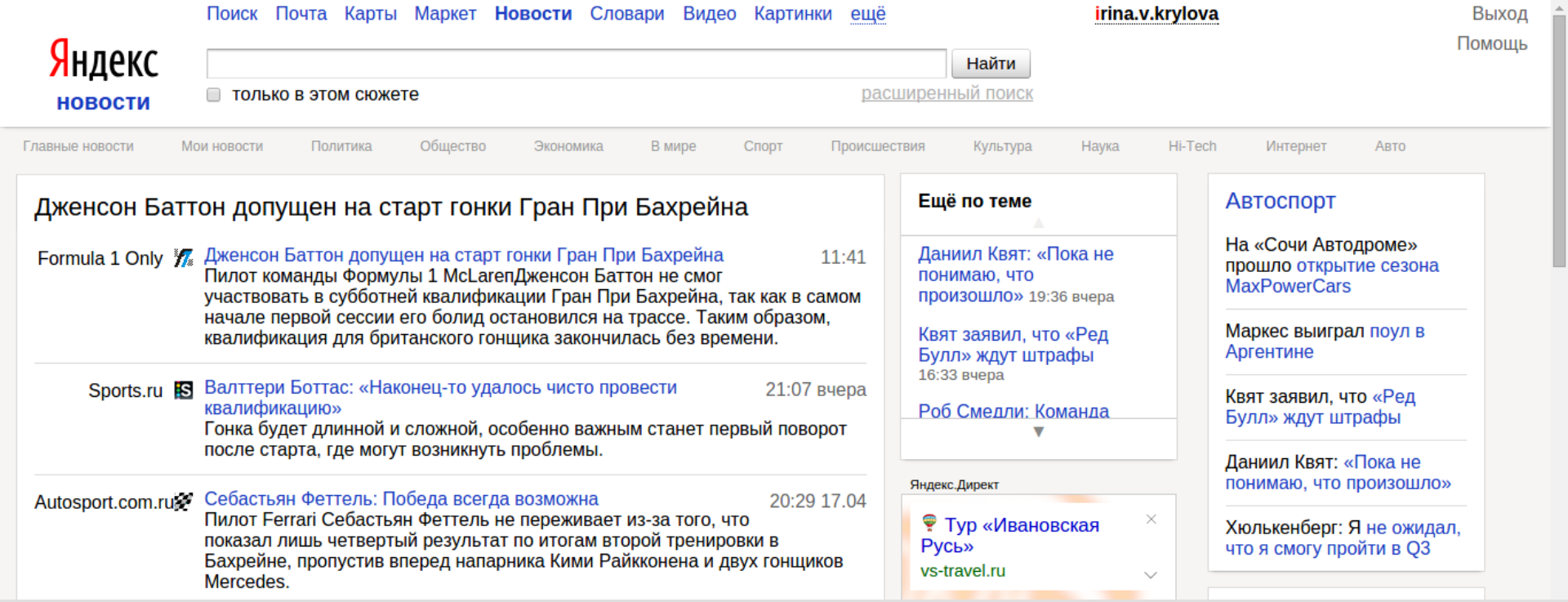
1. Что такое Томита-парсер
2. Примеры использования в Яндексе
3. Примеры использования вне Яндекса
4. Основные принципы работы
5. Некоторые интересные случаи

Томита-парсер

- Инструмент для извлечения структурированных данных из текста
- Контекстно-свободные грамматики
- Поддержка русского языка
- Свободно доступен с исходниками
- Документация

Томита-парсер: аналоги

- JAPE (Java Annotation Patterns Engine)
- AGFL (Affix Grammars Over a Finite Lattice)
- LSPL (LexicoSyntactic Pattern Language)
- AIRE (Artificial Intelligence Information Retrieval Engine)



- Извлечение адресов
- Геопривязка сюжета
- Выделение компаний и персон

Яндекс
работа

Найти
искать в тексте всего объявления

Вакансии в отрасли «IT, интернет, связь, телеком»

17 профессий

программист, менеджер, инженер
специалист, администратор, оператор, дизайнер, аналитик, seo, монтажник, руководитель, оптимизатор, начальник отдела, маркетолог ...

4805 вакансий по дате ▾ по компаниям Подписаться

200 000 – 300 000 руб.

Директор по управлению проектами АСУ ТП (нефтегаз)

в агентство Penny Lane Personnel

Требования: Опыт работы: от 6 лет. Высшее техническое образование, предпочтительно в нефтегазовой отрасли; Дополнительная подготовка в сфере управления проектами будет преимуществом Опыт работы на аналогичной ...

Обязанности: В стабильную российскую инжиниринговую компанию, оказывающую весь комплекс проектных услуг от предпроектного обследования, разработки технических заданий до управления проектом, поставки ...

Условия: Уровень заработной платы высокий обсуждается с успешными кандидатами Оформление по ТК РФ График работы 5/2, не нормированный Командировки по России более 50% рабочего времени Офис в центре Москвы

Москва

сегодня с superjob.ru

➢ Для пополнения фильтров

Зарплата
от до тыс. руб. ▾
также без указания зарплаты

Форма занятости
Постоянная Стажировка
Временная Частичная

Выбрать метро

График работы
Гибкий Сменный
На дому Вахтовый
Полный день

Тип вакансии
Прямая Агентство

Отрасль
IT, интернет, связь, телеком ▾

Опыт работы
Не имеет значения ▾

ещё фото

Борис Николаевич Ельцин

Россия, первый президент

Дата рождения — 01.02.1931

Дата смерти — 23.04.2007

- Даты рождения и смерти
- Свободные определения
- Место работы и должность

Кто это

президент (8333 упоминания в СМИ)
В этом случае, Россия из сильной президентской республики, которой она стала при *президенте Б. Ельцине*, превратится в парламентскую.
29.03.05 РБК

политик (198 упоминаний в СМИ)
Как *политик Ельцин* не мог не понимать, что он перестает "возглавлять процесс".
31.01.06 Московский комсомолец

Работа

Россия, первый президент (28798 упоминаний в СМИ)
Дореволюционная традиция ставить елку в новогодние праздники на Соборной площади Кремля возродилась в декабре 1996 года по инициативе *первого президента России Бориса Ельцина*.
28.11.12 РИА Новости

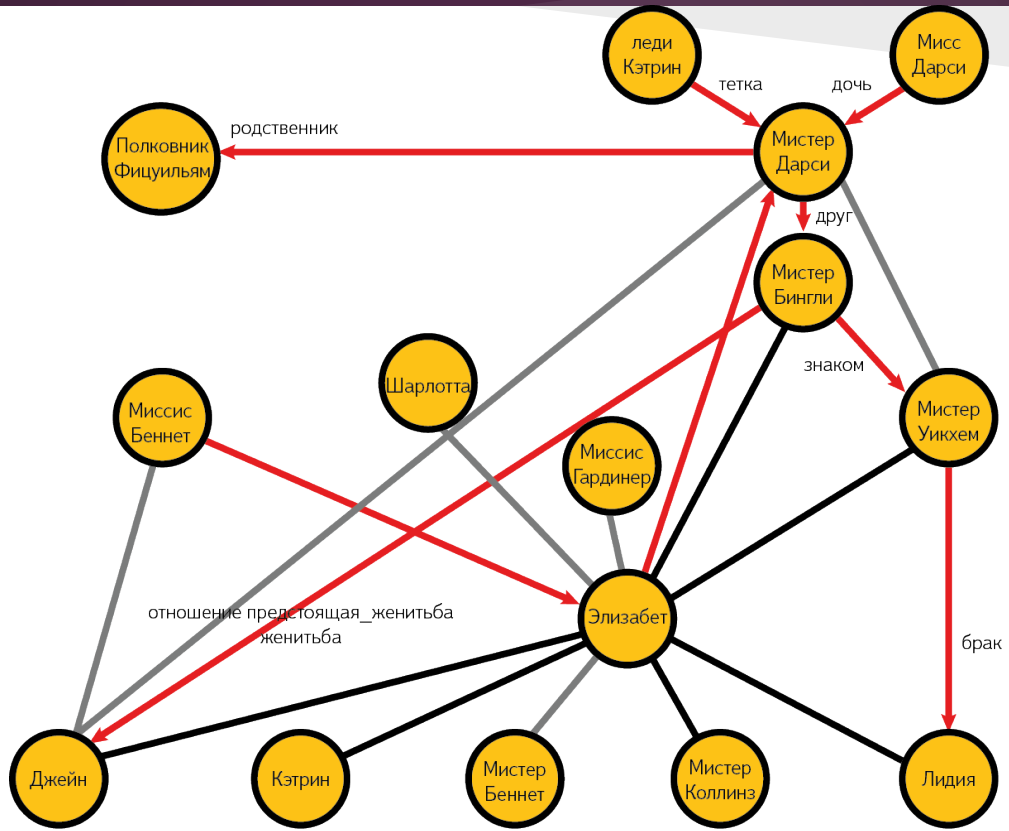
Верховный Совет РСФСР, председатель (878 упоминаний в СМИ)
29 мая 1990 года на I съезде народных депутатов РСФСР *Ельцин* был избран *Председателем Верховного Совета РСФСР* при активной поддержке блока "Демократическая Россия".
01.02.08 РИА Новости

Все места работы (3)

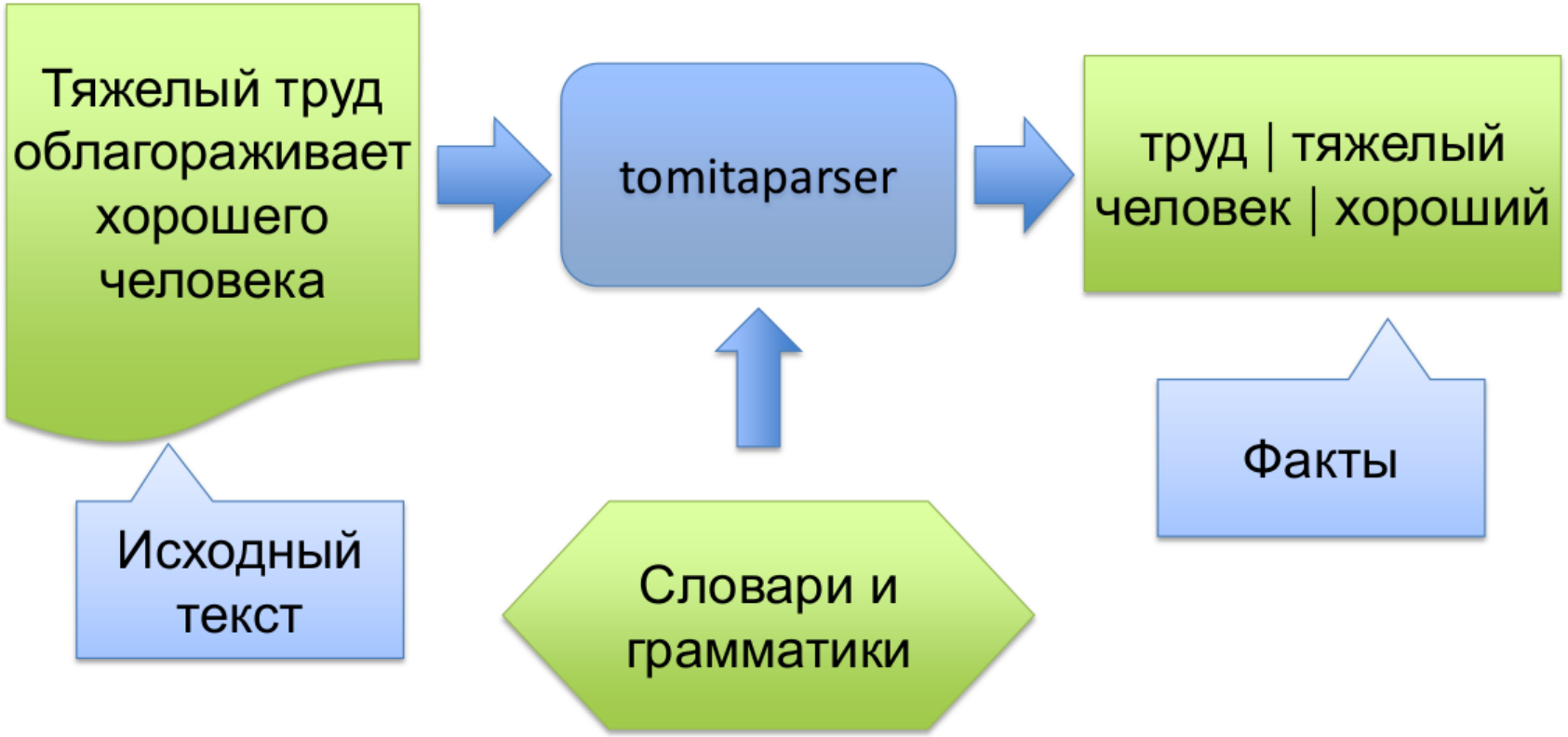
Томита-парсер вне Яндекса

- Законодательство РФ
  - Изменения в текстах законов (Apps4Russia 2013)
  - Извлечение определений (AINL 2014)
- Художественная литература
  - Отношения между персонажами (Диалог 2014)
- Описания товаров
- Отзывы

“Гордость и предубеждение” Джейн Остин



Основные принципы работы



Грамматики. Правила

NounPhrase -> ‘хороший’ Noun;

- Нетерминалы
  - строятся из терминалов
- Терминалы
  - ‘лемма’
  - Noun, Verb, Adj
  - Comma, Punct, Hyphen
  - Percent, Dollar

Интерпретация факта

S -> Word<gram= “persn”> interp (FIO.Name::not_norm);

Себастьян Феттель : Победа всегда возможна .

```
FIO
{
   Name = Себастьян
}
```

Пример грамматики (ne_base.cxx)

```
#encoding "utf-8"    // кодировка
#GRAMMAR_ROOT S

S -> Word<gram="persn"> interp (FIO.Name::not_norm);
```

Факты (facttypes.proto)

```
import "base.proto";          // описание protobuf-типов
import "facttypes_base.proto";  // описание protobuf-типа NFactType.TFact

message FIO : NFactType.TFact
{
  required string Name = 1;
  optional string Surn = 2;
  optional string Patron = 3;
}
```

Словари (maindic.gzt)

```
encoding "utf8";
import "base.proto";
import "articles_base.proto";
import "facttypes.proto";

TAuxDicArticle "именные_группы"
{
    key = { "tomita:ne_base.cxx" type=CUSTOM }
}
```

Словари

- написать свой (key, lemma)
- создать из txt файла

```
key = { "complex_prep.txt" type=FILE };
```

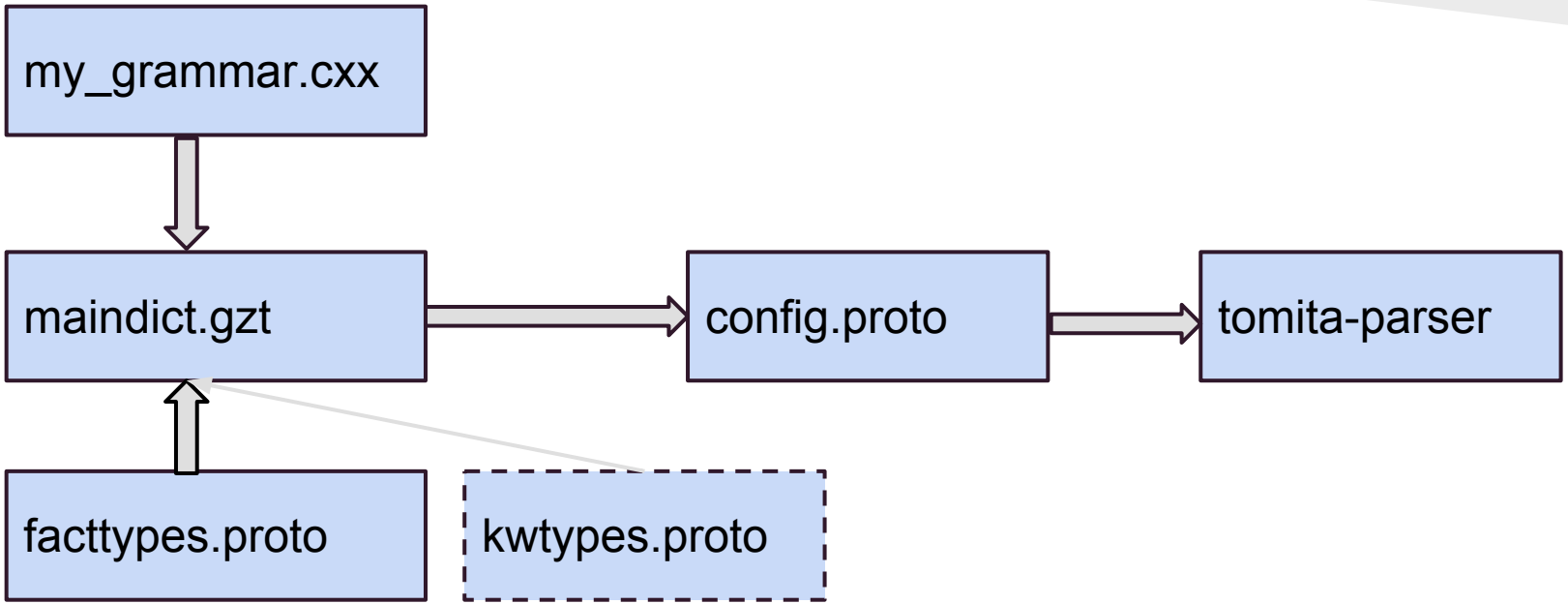
- создать из другой грамматики

```
key = { "tomita:ne_base.cxx" type=CUSTOM }
```

Конфиг (config.proto)

```
encoding "utf8"; // указываем кодировку, в которой написан конфигурационный файл

TTextMinerConfig {
  Dictionary = "maindic.gzt"; // путь к корневому словарю

  PrettyOutput = "NE.html";

  Articles = [
    { Name = "именные_группы" } // название статьи в корневом словаре,
                                // которая содержит запускаемую грамматику
  ]

  Facts = [
    {
     Name = "FIO";
    }
  ]

  Output = {
  Format = text;
  }

//  PrintRules = "rules.txt";
//  PrintTree = "tree.txt";
}
```

Как работать. Памятка



Чуть сложнее

- пометы-ограничения, операторы, согласование
- регулярные выражения
- тип факта
- веса, главное слово
- способы убрать лишние срабатывания
- объединение фактов из разных предложений текста
- анафора
- отладка

Регулярные выражения

Даты

Date -> **AnyWord**<wff="(19[0-9]{2})|(20[0-1][0-9])">;

URL

Sobst_site -> **AnyWord**<wff="([a-z]{3,10}://)?(www|ввв)?\\.?([А-Яа-я0-9-_]+\\.?){1,4}\\.рф">;

Тип факта. Вес. Главное слово.

```
Sobst -> Noun<gnc-agr[1]> Word<rt,gram="persn", gnc-agr[1]> interp
        (NamedEntity.Type="имя собственное") {weight=0.3};

        NamedEntity
            {
                Type = имя собственное
            }

Sobst_super -> Adj<c-agr[1]> Sobst<c-agr[1]>;
```

Способы убрать лишние срабатывания

1. однословные

   N -> Noun<kwtype=~[bad_noun]>;

2. многословные

   #GRAMMAR_KWSET[“плохие_даты”]

```
Some_type "плохие_даты"
{
    key = "улица 8 !марта"
}
```

```
Some_type "плохие_даты"
{
    key = { "tomita:bad_dates.cxx" type=CUSTOM }
}
```

Объединение фактов. Анафора

1. Правила для извлечения конкретных фактов
   a. именные группы
   b. отношения
2. Пост-процессинг
   a. небольшая статистика - MI
   b. машинное обучение (кластеризация)
   c. всё, что сможете придумать сами

Что можно извлекать

Объекты в тексте

- даты
- адреса
- телефоны
- ФИО
- название товара
- действие
- тональность

Связи между объектами

- события
- мнения и отзывы
- контактные данные
- объявления

Контакты


https://github.com/yandex/tomita-parser


https://tech.yandex.ru/tomita/


tomita@yandex-team.ru

Литература

- Bodrova, Bocharov. Relationship extraction from literary fiction.
- M. Hearst. Automatic Aquisition of Hyponyms from Large Corpora
- Извлечение именованных сущностей (грамматика для именных групп)