



Томи́та-парсер

инструмент для извлечения фактов

Дмитрий Панкратов, **Н**аталья Остапук,
Виктор Бочаров

NLPseminar, Санкт-Петербург, 15 декабря 2012 года

Вступление

Томи́та-парсер

Инструмент для извлечения фактов

- В основе парсера лежит алгоритм GLR – парсинга (<http://ru.wikipedia.org/wiki/GLR-парсер>)
- Автор алгоритма - Масару Томи́та, мы назвали парсер в его честь.
- Извлечение фактов - извлечение структурированных данных из текста на естественном языке.
- Извлечение фактов происходит при помощи контекстно-свободных грамматик и словарей ключевых слов.

Томида-парсер

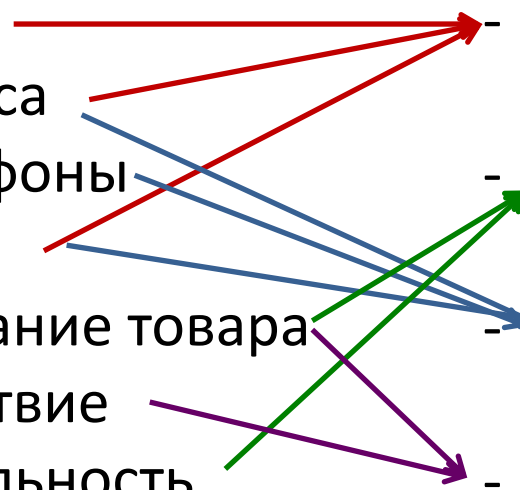
Что можно извлечь?

Объекты в тексте:

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

Связи между этими объектами:

- События
- Мнения и отзывы
- Контактные данные
- Объявления



Объект 1	Объект 2	Тип связи
Яндекс	Аркадий Волож	директор

Извлечение фактов в Яндексе

В проекте Яндекс.Новости

для извлечения адресов

Автоматически обработано 4862 источника, обновлено в 18:53 Москва, статика 2012-12-12 18:55:11

Выпуск: Россия ▼

Новости Москвы



- 1 — Оппозиция не удалось согласовать "Марш свободы"
- 2 — Власти рассказали о состоянии памятников культуры
- 3 — 15 человек задержаны на акции в годовщину событий на Манежной площади

Все новости Москвы >>

Другие регионы >>

Путин поддержал идею возвращения к смешанной системе выборов в Госдуму (28 обработано) 1739 мнений

Президент РФ Владимир Путин согласился с предложением партий вернуться к смешанной системе выборов депутатов в Госдуму - по партийным спискам и одномандатным округам. Такое заявление он сделал, выступая с посланием к Федеральному собранию РФ.

Дело Политковской: обвинение требует посадить Павлюченкова на 12 лет (188) 100

12 лет лишения свободы требует гособвинение для бывшего сотрудника милиции Дмитрия Павлюченкова. Павлюченков - фигурант дела об убийстве обозревателя "Новой газеты" Анны Политковской.

Фигурантка дела «Оборонсервиса» Васильева сменила адвоката (23) 100

Ранее адвокат Ольга Козырева просила суд отменить домашний арест Евгении Васильевой, так как, по ее словам, из-за жестких условий домашнего ареста обвиняемая не получает необходимую медицинскую помощь, не посещает священника и к ней не может прийти домработница.

Найден вертолет Robinson, пропавший в Подмоскovie (246) 1973

По данным LifeNews, частный вертолет бизнесмена Федора Царева, пропавший по дороге из Тверской области в Подмоскovie вечером 8 декабря найден севернее города Солнечногорска.

Оппозиции не удалось согласовать "Марш свободы" (17) 100

Мария Москвы и оппозиция не смогли согласовать маршрут "Марша свободы", который намечен на субботу 15 декабря.

Срок следствия по делу Развозжаева продлен до 1 апреля (111) 100

Ространснадзор назвал причину аварии сухогруза «Амурская» (16)

"ВКонтакте" удалила песни Лазарева в борьбе за культурные ценности (24)

Оппозиционную коалицию Сирии признали 100 стран мира (130) 120

Режиссера Костомарова снова вызвали на допрос в СК (234) 100

Обыски по делу об уклонении от налогов прошли в офисах Deutsche bank (12)

Долг "Патамерка Бизнес Групп" перед туристами оценили в 2,2 млн руб (26) 100

Актриса Наталья Кустинская вышла из комы (45) 100

Власти Москвы решили сотрудничать с фондом «РосЖКХ» Навального (31) 100

Галину Вишневскую похоронят на Новодевичьем кладбище (24) 100

для геопривязки
сюжетов

для выделения
компаний и персон



Люди

Организации



Владимир Путин



Анна Политковская



Евгения Васильева



Сергей Удальцов

Следственный комитет России, МЧС, Новая газета, Минобороны, Басманный суд Москвы, Центр оперного пения, Левый фронт, Министерство иностранных дел России, Боткинская больница Москвы, компания БИИ, страхование

Новости в блогах

Путин поддержал идею возвращения к смешанной системе выборов в Госдуму (2734)

Путин не считает нужным отказываться от плоской шляпы НДФЛ (2727)

В Стокгольме завершилось награждение лауреатов Нобелевской премии-2012 (2343)

Извлечение фактов в Яндексе В проекте Яндекс.Работа

50 000–100 000 руб.

[Врач стоматолог-терапевт](#)

в компанию [Архидент](#)


Требования Опыт работы: от 3 лет. Владение всеми методиками терапевтической стоматологии. Обязательно умение выполнять несложные удаления. Гражданство Россия или Беларусь

Обязанности Прием пациентов. Терапевтическая помощь, несложные удаления. Реставрация. Эндодонтическое лечение. Снятие зубных отложений. Отбеливание.

Условия Место работы: м. Волоколамская (10 мин. пешком). Зарплата выплачивается без задержек.

Москва, м. Волоколамская

вчера с [Вакансия Профи.ру](#) [похожие](#)

 **Выбрать метро**

Волоколамская ×

График работы

- | | |
|--------------------------------------|-----------------------------------|
| <input type="checkbox"/> Гибкий | <input type="checkbox"/> Сменный |
| <input type="checkbox"/> На дому | <input type="checkbox"/> Вахтовый |
| <input type="checkbox"/> Полный день | |

Тип вакансии

- | | |
|---------------------------------|------------------------------------|
| <input type="checkbox"/> Прямая | <input type="checkbox"/> Агентство |
|---------------------------------|------------------------------------|

Отрасль

Медицина, фармацевтика ▼

Опыт работы

От 3 до 6 лет ▼

[еще параметры](#)

для пополнения фильтров



[ещё фото](#)

Борис Николаевич Ельцин

Россия, первый президент

Дата рождения — 01.02.1931

Дата смерти — 23.04.2007

даты рождения и смерти

Кто это [Работа](#) [Интервью](#) [Связанные пресс-портреты](#) [Новости](#)

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](#)

свободные определения

политик ([198 упоминаний в СМИ](#))

Как *политик Ельцин* не мог не понимать, что он перестает "возглавлять процесс".

31.01.06 [Московский комсомолец](#)

Работа

место работы и должность

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 \(3\)](#)

Я

http://news.yandex.ru/people/el1tsin_boris.html

Томи́та-парсер: Аналоги

JAPE (Java Annotation Patterns Engine)

- Конечный автомат над множеством регулярок
- Часть GATE – фреймворка для лингвистических задач
- <http://gate.ac.uk/>

AGFL (Affix Grammars Over a Finite Lattice)

- Контекстно-свободные грамматики ограничены набором predetermined значений категорий
- <http://www.agfl.cs.ru.nl/>

LSPL (LexicoSyntactic Pattern Language)

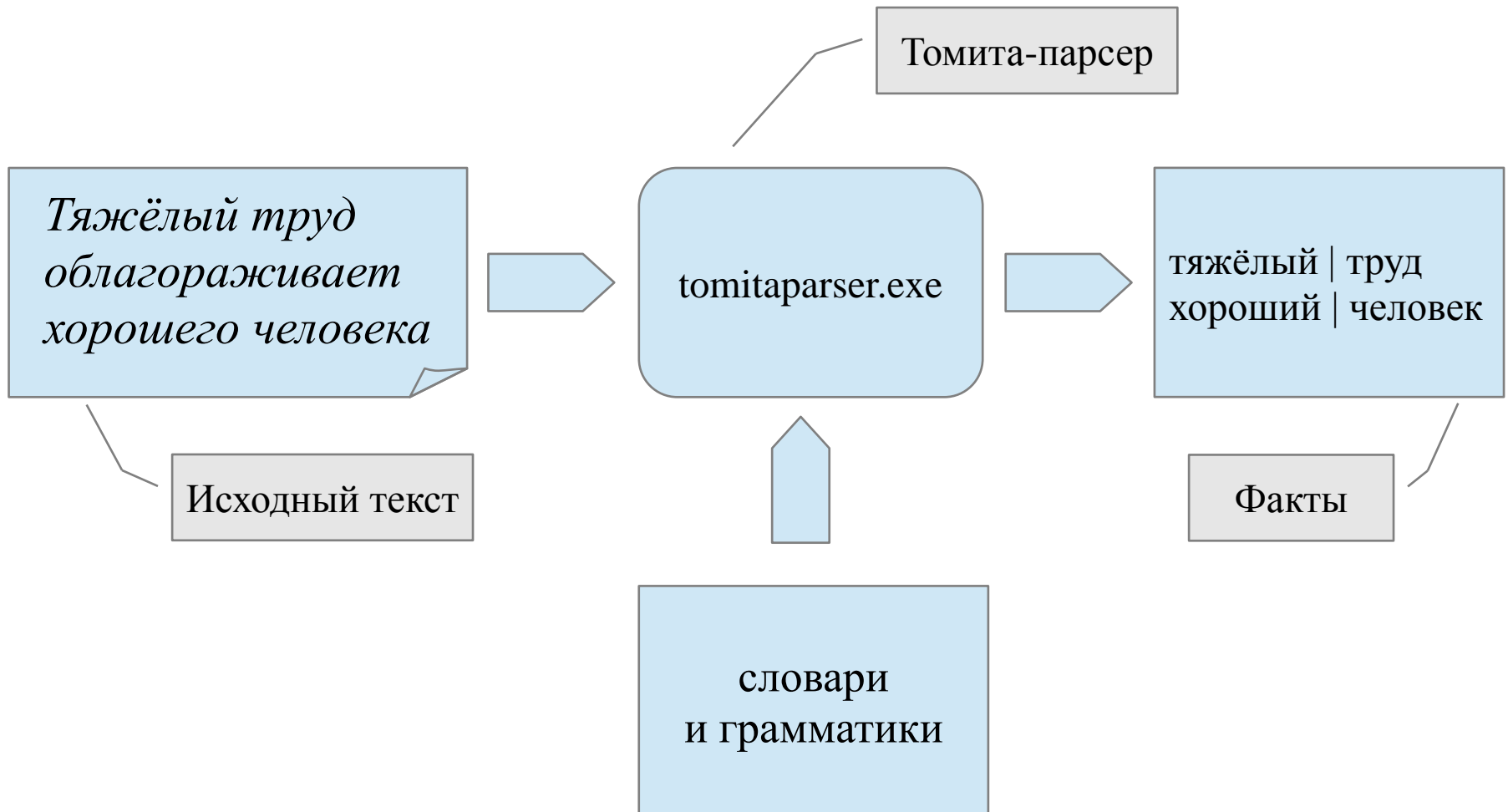
- язык, на котором можно записывать лексико-синтаксические шаблоны
- <http://www.lspl.ru/index.php>

AIRE (Artificial Intelligence Information Retrieval Engine)

- универсальный базовый компонент систем информационного поиска и автоматического перевода
- <http://clck.ru/4JKhe>

Как запустить Томита-парсер?

Что делает томита-парсер?



Откуда берется исходный текст?

- из одного текстового файла
 - один файл — один документ
 - одна строка — один документ (dpl)
- из нескольких текстовых файлов
 - папка
 - .tar архив
- из STDIN

Кодировка символов: UTF-8, Windows-1251

Вывод фактов

Куда сохраняются факты?

- в файл
- в STDOUT

Форматы:

- для автоматической обработки
 - Facts XML
 - Google Protobuf
- чтобы «смотреть глазами»
 - текстовый формат

Кодировка символов: UTF-8, Windows-1251

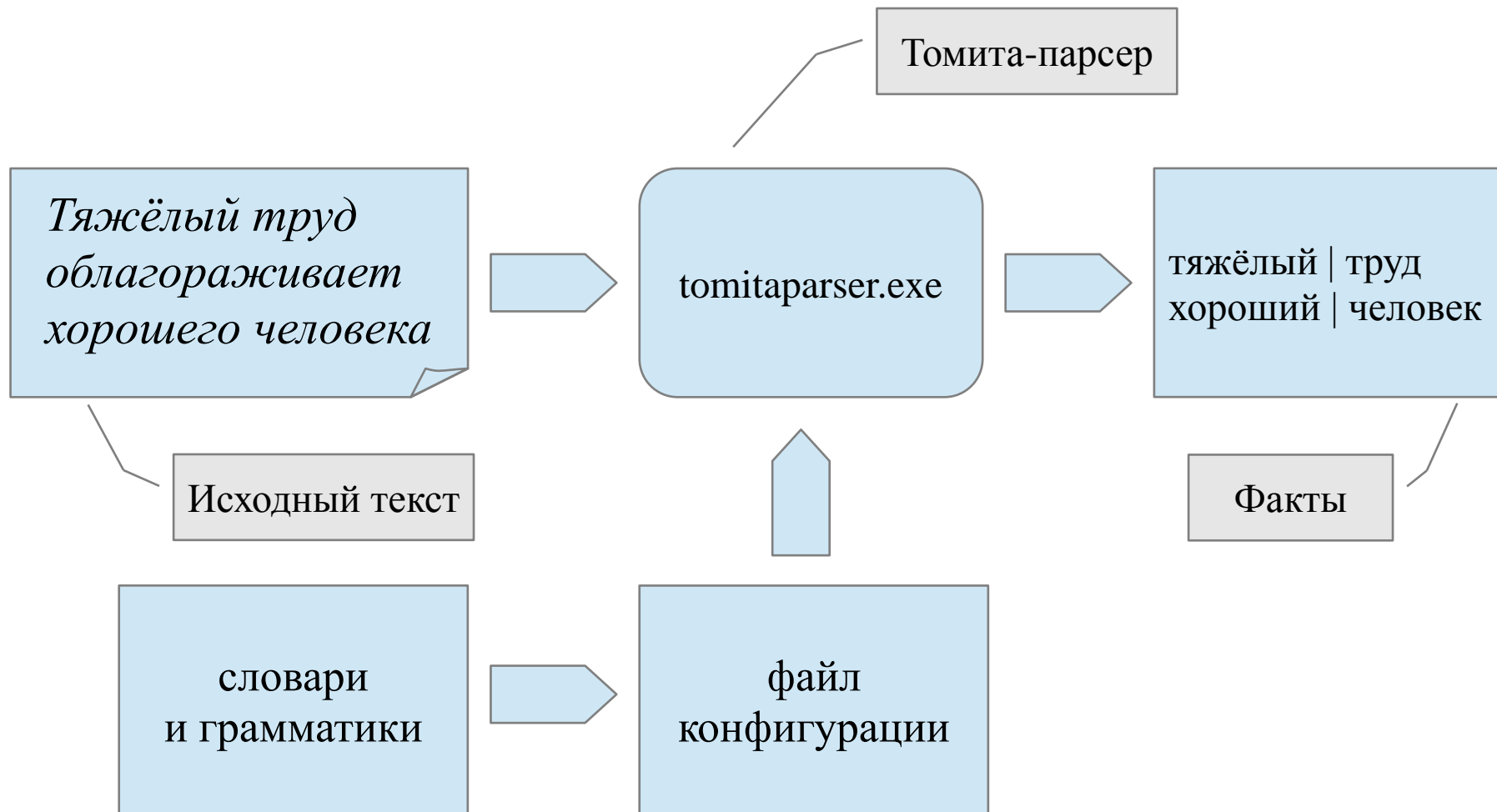
Конфигурация парсера

- Откуда читать текст?
- В какой кодировке?
- Куда записывать факты?
- В каком формате и кодировке?

+

- Какие грамматики запускать?
- Какие факты записывать?

Что делает Томита-парсер?



Запуск парсера

Windows:

```
tomitaparser.exe  
config.proto
```

Linux / *BSD (bash)

```
./tomitaparser config.proto
```

Файл конфигурации

- Формат — Google Protobuf
- Кодировка — UTF-8
- Обязательные параметры:
 - корневой словарь (Dictionary = ...)
- Передаётся в качестве аргумента при запуске парсера

Простой файл конфигурации

```
encoding "utf8";  
  
TTextMinerConfig {  
    Dictionary = "mydic.gz";  
}
```

Простой файл конфигурации

Действия парсера:

- скомпилирует и загрузит словарь `mydic.gzt`
- прочитает текст из `STDIN`

Для полноценной работы нужно указать:

- статьи корневого словаря, которые нужно запустить
- факты, которые нужно записать

**Какие грамматики и статьи
словаря запускать?**

Словари

Словари состоят из статей

- Формат — Google Protobuf
- Кодировка — UTF-8

Корневой словарь

- всегда один
- содержит ссылки на остальные словари и грамматики

Статьи корневого словаря

```
Articles = [  
    { Name = "статья1" }  
    { Name = "статья2" }  
    // можно указать  
    // несколько статей  
]
```

Корневой словарь

```
encoding "utf8";

import "base.proto";
import "articles_base.proto";

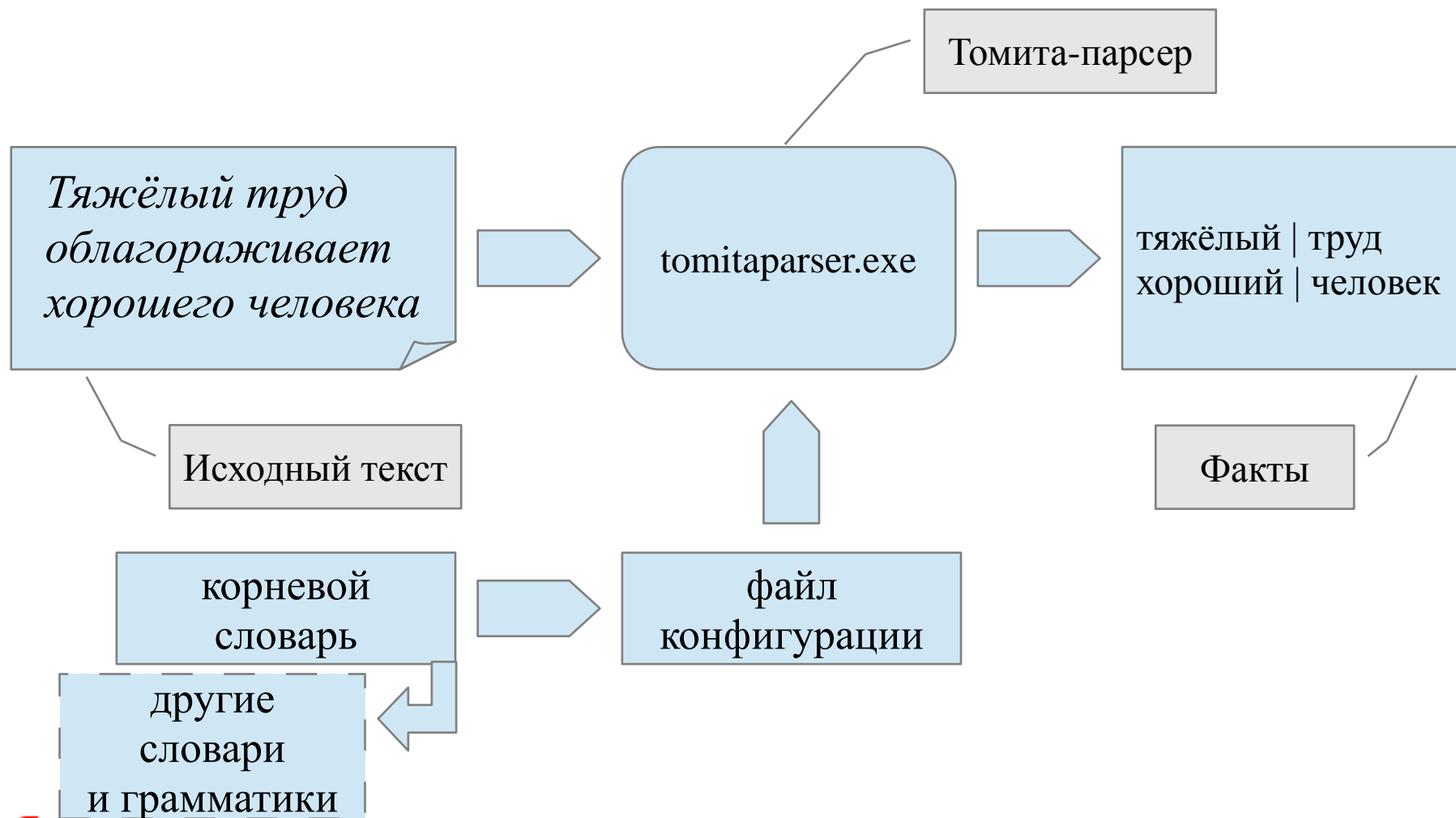
TAuxDicArticle "грамматика1"
{
    key = { "tomita:first.cxx"
           type=CUSTOM }
}
```

Файл конфигурации

```
encoding "utf8";

TTextMinerConfig {
    Dictionary = "mydic.gzt";
    Articles = [
        { Name = "грамматика1"
        }
    ]
}
```

Что делает Томита-парсер?



Какие факты записывать?

ФАКТЫ

```
Facts = [  
    { Name = "факт1" }  
    { Name = "факт2" }  
    // можно указать  
    // несколько фактов  
]
```

```
encoding "utf8";
```

```
TTextMinerConfig {  
    Dictionary = "mydic.gzt";  
    Articles = [  
        { Name = "дата" } ]  
    Facts = [  
        { Name = "Date" } ]  
}
```

**Куда и в каком формате
записывать факты?**

Секция Output

- File — в какой файл сохранять
- Format — в каком формате
 - xml, protobuf или text
- Mode: append/overwrite
- Encoding

По умолчанию:

STDOUT, xml, append, utf-8

...

```
Facts = [  
    { Name = "Date" } ]
```

```
Output = {  
    File = facts.txt;  
    Format = "text";  
}  
}
```

Откуда читать текст?

Секция Input

- File / Dir
- Format (plain или html)
- Type (no, dpl, tar, ...)
- Encoding

По умолчанию:

STDIN, plain, no, utf-8

...

```
Input = {  
    File = test.txt;  
}  
}
```

Конфигурация парсера

- Откуда читать текст? - секция Input (File, ...)
 - В какой кодировке? - Input.Encoding
 - Куда записывать факты? - секция Output
 - В каком формате и кодировке? - там же
- +
- Какие грамматики запускать? - Articles
 - Какие факты записывать? - Facts

Что ещё?

Хозяйке на заметку:

Ускорение работы парсера

- NumThreads — количество потоков

Отладка словарей и грамматик

- PrettyOutput — подробности разбора
- PrintTree — деревья разбора
- PrintRules — сработавшие правила

Разное

- ForceRecompile — перекомпиляция грамматик

Как писать грамматики?

Наша задача

Заполнение «карточки фильма» информацией, извлеченной из текста на естественном языке

название	
жанр	
год	
оригинальное название	
режиссер	

Исходный текст input.txt

Фильм Оливера Стоуна "Александр" основан на реальной жизни одного из самых выдающихся людей в истории.

«Титаник» (Titanic) — фильм-катастрофа 1997 года, снятый Джеймсом Кэмероном, в котором показана гибель легендарного лайнера «Титаник». Главные роли в фильме исполнили Кейт Уинслет (Роза Дьюитт Бьюкейтер) и Леонардо Ди Каприо (Джек Доусон).

«Неприкасаемые» (Intouchables) — трагикомедийный фильм 2011 года, основанный на реальных событиях. Главные роли исполняют Франсуа Клузе и Омар Си, удостоенный за эту актёрскую работу национальной премии «Сезар».

Правила в Томите

Грамматика состоит из правил, которые описывают цепочки

В правиле есть левая и правая часть, разделенные символом \rightarrow

В левой части стоит один нетерминал, правая состоит из терминалов и нетерминалов.

$S \rightarrow Noun;$

Превращаем правила в грамматику

Грамматика film.cxx

```
#encoding "utf-8"
```

```
#GRAMMAR_ROOT S
```

```
S -> Noun;
```

Корневой словарь mydic.gzt

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

} *всегда копируем*

```
TAuxDicArticle "фильм"  
{  
  key = { "tomita:film.cxx" type=CUSTOM }  
}
```

эту статью будем запускать

Я

Файл конфигурации config.proto

```
encoding "utf8";
```

```
TTextMinerConfig {
```

```
    Dictionary = "mydic.gzt";
```

корневой словарь

```
    Input = {File = "input.txt";}
```

входной файл

```
    Output = {File = "output.txt";
```

сюда записываем

```
        Format = text;}
```

результат

```
    Articles = [
```

ссылка на статью

```
        { Name = "фильм" }
```

из словаря

```
    ]
```

```
}
```

Я

Запускаем!

`tomitaparser.exe config.proto`

Результат output.txt

Фильм Оливера Стоуна "Александр" основан на реальной жизни одного из самых выдающихся людей в истории .

"Титаник" (Titanic) — фильм-катастрофа 1997 года , снятый Джеймсом Кэмероном , в котором показана гибель легендарного лайнера "Титаник" .

Главные роли в фильме исполнили Кейт Уинслет (Роза Дьюитт Бьюкейтер) и Леонардо Ди Каприо (Джек Доусон) .

«Неприкасаемые» (Intouchables) — трагикомедийный фильм 2011 года , основанный на реальных событиях .

Главные роли исполняют Франсуа Клузе и Омар Си , удостоенный за эту актёрскую работу национальной премии "Сезар" .

**Но если мы добавим
отладочный вывод...**

Файл конфигурации config.proto

```
encoding "utf8";
```

```
TTextMinerConfig {
```

```
    Dictionary = "mydic.gz";
```

```
    PrettyOutput = "pretty.html";
```

```
    Input = {File = "input.txt";}
```

```
    Output = {File = "output.txt";
```

```
        Format = text}
```

```
    Articles = [ { Name = "фильм" } ]
```

```
Я }
```


Подробный результат pretty.html

Фильм Оливера Стоуна "Александр" основан на реальной жизни одного из самых выдающихся людей в истории . EOS

"Титаник" (Titanic) — фильм-катастрофа 1997 года , снятый Джеймсом Кэмероном , в котором показана гибель легендарного лайнера "Титаник" . EOS

Главные роли в фильме исполнили Кейт Уинслет (Роза Дьюитт Бьюкейтер) и Леонардо Ди Каприо (Джек Доусон) . EOS

"Неприкасаемые" (Intouchables) — трагикомедийный фильм 2011 года , основанный на реальных событиях . EOS

Главные роли исполняют Франсуа Клузе и Омар Си , удостоенный за эту актёрскую работу национальной премии "Сезар" . EOS

Text	Type
фильм	TAuxDicArticle [фильм]
Стоун	TAuxDicArticle [фильм]
"Александра"	TAuxDicArticle [фильм]
жизнь	TAuxDicArticle [фильм]
человек	TAuxDicArticle [фильм]
история	TAuxDicArticle [фильм]
"титаник"	TAuxDicArticle [фильм]
фильм-катастрофа	TAuxDicArticle [фильм]
год	TAuxDicArticle [фильм]
гибель	TAuxDicArticle [фильм]
лайнер	TAuxDicArticle [фильм]
"Титаник"	TAuxDicArticle [фильм]
главное	TAuxDicArticle [фильм]
роль	TAuxDicArticle [фильм]
фильм	TAuxDicArticle [фильм]
Роза	TAuxDicArticle [фильм]
Ди	TAuxDicArticle [фильм]
фильм	TAuxDicArticle [фильм]
год	TAuxDicArticle [фильм]
событие	TAuxDicArticle [фильм]
главное	TAuxDicArticle [фильм]
роль	TAuxDicArticle [фильм]
Клуз	TAuxDicArticle [фильм]
Омар	TAuxDicArticle [фильм]

Пометы-ограничения

Регистр: *h-reg1, h-reg2, h-reg3, l-reg*

Многословная сущность: *mw*

Первое слово в предложении: *fw*

Вершина синтаксической группы: *rt*

И другие

**Теперь мы можем выделять
имена собственные**

Грамматика film.cxx

```
#encoding "utf-8"
```

```
#GRAMMAR_ROOT S
```

```
Name -> Word<h-reg1, ~fw>;
```

```
Name -> Word<h-reg1, ~fw> Word<h-reg1>;
```

```
Name -> Word<h-reg1, ~fw> Word<h-reg1>
```

```
Word<h-reg1>;
```

```
S -> Name;
```

**Не обязательно писать одно и
то же несколько раз**

Операторы

Позволяют получить более удобную сокращенную запись правил грамматики

* — символ повторяется 0 или более раз

$S \rightarrow Adj^* Noun;$

=

$S \rightarrow Noun;$

$S \rightarrow Adj Noun;$

$S \rightarrow Adj Adj Noun;$

...

Операторы

+ — символ повторяется 1 или более раз

$S \rightarrow Adj^+ Noun;$

=

$S \rightarrow Adj Noun;$

$S \rightarrow Adj Adj Noun;$

$S \rightarrow Adj Adj Adj Noun;$

...

Операторы

() — символ входит в правило 0 или 1 раз

$S \rightarrow (Adj) Noun;$

=

$S \rightarrow Noun;$

$S \rightarrow Adj Noun;$

Грамматика film.cxx

```
#encoding "utf-8"
```

```
#GRAMMAR_ROOT S
```

```
Name -> Word<h-reg1, ~fw> Word<h-reg1>*;
```

```
S -> Name;
```

**Имена собственные уже не
стыдно интерпретировать в
факты**

Описание фактов facttypes.proto

```
import "base.proto";  
import "facttypes_base.proto";  
  
message Film: NFactType.TFact  
{  
    required string Name = 1;  
}
```

всегда копируем

имя факта

поля факта

Корневой словарь mydic.gzt

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

```
import "facttypes.proto";
```

```
TAuxDicArticle "фильм"  
{  
  key = { "tomita:film.cxx" type=CUSTOM }  
}
```

Файл конфигурации config.proto

```
encoding "utf8";  
TTextMinerConfig {  
    Dictionary = "mydic.gz";  
    PrettyOutput = "pretty.html";  
  
    Input = {File = "input.txt";}   
    Output = {File = "output.txt"; Format = text;}  
  
    Articles = [ { Name = "фильм" } ]  
    Facts = [ { Name = "Film" } ]  
}
```

Грамматика film.cxx

```
#encoding "utf-8"
```

```
#GRAMMAR_ROOT S
```

```
Name -> Word<h-reg1, ~fw> Word<h-reg1>*;
```

```
S -> Name interp (Film.Name);
```

Результат output.txt

Фильм Оливера Стоуна "Александр" основан на реальной жизни одного из самых выдающихся людей в истории .

```
Film
{
    Name = Оливер Стоуна "Александр"
}
```

"Титаник" (Titanic) — фильм-катастрофа 1997 года , снятый Джеймсом Кэмероном , в котором показана гибель легендарного лайнера "Титаник" .

```
Film
{
    Name = Titanic
}
```

```
Film
{
    Name = 1997
}
```

```
Film
{
    Name = Джеймс Кэмероном
}
```

Результат output.txt

Главные роли в фильме исполнили Кейт Уинслет (Роза Дьюитт Бьюкейтер) и Леонардо Ди Каприо (Джек Доусон) .

Film { Name = Кейт Уинслет }

Film { Name = Роза Дьюитт Бьюкейтер }

Film { Name = Леонардо Ди Каприо }

Film { Name = Джек Доусон }

Результат pretty.html

Главные роли исполняют Франсуа Клузе и Омар Си , удостоенный за эту актёрскую работу национальной премии "Сезар" . EOS

Film
Name
Оливер Стоуна Александр
Titanic
1997
Джеймс Кэмероном
Титаник
Кейт Уинслет
Роза Дьюитт Бьюкейтер
Леонардо Ди Каприо
Джек Доусон
Intouchables
2011
Франсуа Клузе
Омар Си
Сезар

Text	Type
Оливер Стоуна "Александр"	TAuxDicArticle [фильм]
Titanic	TAuxDicArticle [фильм]

Что это?

Film
Name
<u>Оливер Стоуна Александр</u>
<u>Titanic</u>
<u>1997</u>
<u>Джеймс Кэмероном</u>

?

Я

Согласование

Согласование

По роду, числу и падежу: *gnc-agr*

По числу и падежу: *nc-agr*

По роду и числу: *gc-agr*

По падежу: *n-agr*

И другие

Грамматика film.cxx

```
#encoding "utf-8"
```

```
#GRAMMAR_ROOT S
```

```
Name -> Word<h-reg1, ~fw, nc-agr[1]> Word<h-reg1, nc-agr[1]>*;
```

```
S -> Name interp (Film.Name);
```

Результат pretty.html

Film
Name
<u>Оливер Стоун</u>
<u>Александр</u>
<u>Titanic</u>
<u>1997</u>
<u>Джеймс Кэмерон</u>

Переходим к делу

Выделяем название фильма

Файл film.cxx

```
#encoding "utf-8"
```

```
#GRAMMAR_ROOT S
```

```
//Name -> Word<h-reg1, ~fw, nc-agr[1]> Word<h-reg1, nc-agr[1]>*;
```

```
FilmName -> AnyWord<h-reg1, l-quoted> Word*  
(Word<r-quoted>);
```

```
S -> FilmName interp (Film.Name);
```


Результат pretty.html

Фильм Оливера Стоуна "Александр" основан на реальной жизни одного из самых выдающихся людей в истории . EOS

"Титаник" (Titanic) — фильм-катастрофа 1997 года , снятый Джеймсом Кэмероном , в котором показана гибель легендарного лайнера "Титаник"

Главные роли в фильме исполнили Кейт Уинслет (Роза Дьюитт Бьюкейтер) и Леонардо Ди Каприо (Джек Доусон) . EOS

"Неприкасаемые" (Intouchables) — трагикомедийный фильм 2011 года , основанный на реальных событиях . EOS

Главные роли исполняют Франсуа Клузе и Омар Си , удостоенный за эту актёрскую работу национальной премии "Сезар" . EOS

Film
Name
Александр
титаник
Титаник
неприкасаемый
Сезар

Что нам не нравится?

1. «Неприкасаемые» нормализовались
2. В качестве названия фильма выделилось название судна и название премии

Решаем проблему нормализации

Файл film.cxx

```
#encoding "utf-8"
```

```
#GRAMMAR_ROOT S
```

```
//Name -> Word<h-reg1, ~fw, nc-agr[1]> Word<h-reg1,  
nc-agr[1]>*;
```

```
FilmName -> AnyWord<h-reg1, l-quoted> Word*<r-  
quoted>;
```

```
S -> FilmName interp (Film.Name::not_norm);
```

Результаты pretty.html

Film
Name
<u>Александр</u>
<u>Титаник</u>
<u>Титаник</u>
<u>Неприкасаемые</u>
<u>Сезар</u>

Решаем проблему лишних срабатываний

Нужно, чтобы перед или после названия фильма стоял дескриптор.

Для дескрипторов удобнее всего создать словарь.

Словари

Словарь genre.gzt

encoding "utf8";

TAuxDicArticle "жанр"

{

key = "комедия"

key = "комедийный фильм"

key = "трагикомедийный фильм"

key = "фильм ужасов"

key = "фильм-катастрофа"

key = "триллер"

}

Корневой словарь mydic.gzt

```
encoding "utf8";
```

```
import "base.proto";  
import "articles_base.proto";
```

```
import "facttypes.proto";
```

```
import "genre.gzt";
```

```
TAuxDicArticle "фильм"  
{  
  key = { "tomita:film.cxx" type=CUSTOM }  
}
```


Посмотрим еще раз на входной файл

Фильм Оливера Стоуна "Александр" основан на реальной жизни одного из самых выдающихся людей в истории.

"Титаник" (Titanic) — фильм-катастрофа 1997 года, снятый Джеймсом Кэмероном, в котором показана гибель легендарного лайнера «Титаник». Главные роли в фильме исполнили Кейт Уинслет (Роза Дьюитт Бьюкейтер) и Леонардо Ди Каприо (Джек Доусон).

«Неприкасаемые» (Intouchables) — трагикомедийный фильм 2011 года, основанный на реальных событиях. Главные роли исполняют Франсуа Клузе и Омар Си, удостоенный за эту актёрскую работу национальной премии «Сезар».

Надо учесть:

1. После названия фильма на русском может идти оригинальное название в скобках. А может и не идти
2. Между названием и дескриптором может стоять тире
3. Дескриптор может быть как после, так и перед названием

Грамматика film.cxx

```
#encoding "utf-8"
```

```
#GRAMMAR_ROOT S
```

```
//Name -> Word<h-reg1, ~fw, nc-agr[1]> Word<h-reg1, nc-agr[1]>*;
```

```
FilmName -> AnyWord<h-reg1, l-quoted> Word* (Word<r-quoted>);
```

```
OriginalName -> (LBracket) Word<h-reg1, lat> Word<lat>* (RBracket);
```

```
Genre -> Word<kwtype="жанр">;
```

```
Film -> 'фильм';
```

```
Descr -> Genre | Film;
```

```
S -> Descr FilmName interp (Film.Name::not_norm) (OriginalName);
```

```
S -> FilmName interp (Film.Name::not_norm) (OriginalName) (Hyphen)
```

```
Descr;
```

Результат pretty.html

Фильм Оливера Стоуна "Александр" основан на реальной жизни одного из самых выдающихся людей в истории . EOS

"Титаник" (Titanic) — фильм-катастрофа 1997 года , снятый Джеймсом Кэмероном , в котором показана гибель легендарного лайнера "Титаник" . EOS

Главные роли в фильме исполнили Кейт Уинслет (Роза Дьюитт Бьюкейтер) и Леонардо Ди Каприо (Джек Доусон) . EOS

"Неприкасаемые" (Intouchables) — трагикомедийный фильм 2011 года , основанный на реальных событиях . EOS

Главные роли исполняют Франсуа Клузе и Омар Си , удостоенный за эту актёрскую работу национальной премии "Сезар" . EOS

Film
Name
Титаник
Неприкасаемые

Text	Type
"титаник" (Titanic) — фильм-катастрофа	TAuxDicArticle [фильм]
"неприкасаемый" (Intouchables) — трагикомедийный фильм	TAuxDicArticle [фильм]

**Кстати, оригинальное
название и жанр — тоже
полезная информация!**

Описание фактов facttypes.proto

```
import "base.proto";  
import "facttypes_base.proto";  
  
message Film: NFactType.TFact  
{  
    required string Name = 1;  
    optional string Genre = 2;  
    optional string OriginalName = 3;  
}
```

Грамматика film.cxx

```
#encoding "utf-8"
```

```
#GRAMMAR_ROOT S
```

```
//Name -> Word<h-reg1, ~fw, nc-agr[1]> Word<h-reg1, nc-agr[1]>*;
```

```
FilmName -> AnyWord<h-reg1, l-quoted> Word<r-quoted>*;
```

```
OriginalName -> Word<h-reg1, lat> Word<lat>*;
```

```
Genre -> Word<kwtype="жанр"> interp (Film.Genre);
```

```
Film -> 'фильм';
```

```
Descr -> Genre | Film;
```

```
S -> Descr FilmName interp (Film.Name::not_norm) (LBracket)  
(OriginalName interp (Film.OriginalName)) (RBracket);
```

```
S -> FilmName interp (Film.Name::not_norm) (LBracket)  
(OriginalName interp (Film.OriginalName)) (RBracket) (Hyphen)
```

```
Descr;
```

Результат pretty.output

Film		
Name	Genre	OriginalName
<u>Титаник</u>	фильм-катастрофа	Titanic
<u>Неприкасаемые</u>	трагикомедийный фильм	Intouchables

Извлекаем режиссера

Конструкции:

1. Родительный падеж после дескриптора
2. Дескриптор + снятый + ФИО в творительном падеже



Нужна морфология

Пометы gram

Проверяет значения грамматических характеристик отдельно для каждого

<gram> = "им, муж, ед"

<gram> = "прич"

<gram> = "brev"

Описание фактов

facttypes.proto

```
import "base.proto";
import "facttypes_base.proto";

message Film: NFactType.TFact
{
    required string Name = 1;
    optional string Genre = 2;
    optional string OriginalName = 3;
    optional string Director = 4;
}
```

Грамматика film.cxx

Name -> Word<h-reg1, ~fw, nc-agr[1]> Word<h-reg1, nc-agr[1]>*;

....

Director -> Name<gram="род"> interp (Film.Director);

Director -> Комма 'снять'<gram="прич">

Name<gram="твор"> interp (Film.Director);

DescrDirector -> Descr (Director);

S -> DescrDirector FilmName interp (Film.Name::not_norm)
(LBracket) (OriginalName interp (Film.OriginalName))
(RBracket);

S -> FilmName interp (Film.Name::not_norm) (LBracket)
(OriginalName interp (Film.OriginalName)) (RBracket)
(Hyphen) DescrDirector;

Результат pretty.html

Film			
Name	Genre	OriginalName	Director
Александр			Оливер Стоун
Титаник	фильм-катастрофа	Titanic	
Неприкасаемые	трагикомедийный фильм	Intouchables	

Даты

Включение грамматик

Правила, которые могут использоваться во многих грамматиках, целесообразно выделять в отдельную грамматику.

Например, грамматика дат.

Грамматика date.cxx

```
#encoding "utf-8"
```

```
#GRAMMAR_ROOT S
```

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

```
Descr -> 'год';
```

```
S -> Date Descr<gram='род'>;
```

Корневой словарь mydic.gzt

```
TAuxDicArticle "фильм"
```

```
{
```

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

```
}
```

```
TAuxDicArticle "даты"
```

```
{
```

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

```
}
```

facttypes.proto

optional string Date = 5;

Грамматика film.cxx

Date -> AnyWord<kwtype="даты">;

Director -> Name<gram="род"> interp (Film.Director);

Director -> Comma 'снять'<gram="прич">

Name<gram="твор"> interp (Film.Director);

DescrDirector -> Descr (Date interp (Film.Date)) (Director);

Результат pretty.output

Film				
Name	Genre	OriginalName	Director	Date
Александр			Оливер Стоун	
Титаник	фильм-катастрофа	Titanic	Джеймс Кэмерон	1997 года
Неприкасаемые	трагикомедийный фильм	Intouchables		2011 года

Ура, получилось!

Подробнее тут:

<http://api.yandex.ru/tomita/>

Вопросы?



Дмитрий Панкратов, **Н**аталья Остапук,
Виктор Бочаров

Группа извлечения фактов

Отдел лингвистических технологий

tomita@yandex-team.ru